Journal of Biomedical Informatics 66 (2017) 42–51

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Towards a privacy preserving cohort discovery framework for clinical research networks

Jiawei Yuan[a], Bradley Malin[b,c], François Modave[d], Yi Guo[d], William R. Hogan[d], Elizabeth Shenkman[d], Jiang Bian[d,*]

[a] Department of Electrical, Computer, Software, & Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL, United States
[b] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[c] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, United States
[d] Health Outcomes & Policy, University of Florida, Gainesville, FL, United States

ARTICLE INFO

Article history:
Received 1 September 2016
Revised 6 December 2016
Accepted 14 December 2016
Available online 19 December 2016

Keywords:
Privacy-preserving cohort discovery
Clinical research network (CRN)
Data privacy
Searchable encryption
Patient-Centered Clinical Research Network (PCORnet)
OneFlorida Clinical Data Research Network (CDRN)

ABSTRACT

Background: The last few years have witnessed an increasing number of clinical research networks (CRNs) focused on building large collections of data from electronic health records (EHRs), claims, and patient-reported outcomes (PROs). Many of these CRNs provide a service for the discovery of research cohorts with various health conditions, which is especially useful for rare diseases.

Background: Supporting patient privacy can enhance the scalability and efficiency of such processes; however, current practice mainly relies on policy, such as guidelines defined in the Health Insurance Portability and Accountability Act (HIPAA), which are insufficient for CRNs (e.g., HIPAA does not require encryption of data – which can mitigate insider threats). By combining policy with privacy enhancing technologies we can enhance the trustworthiness of CRNs. The goal of this research is to determine if searchable encryption can instill privacy in CRNs without sacrificing their usability.

Methods: We developed a technique, implemented in working software to enable privacy-preserving cohort discovery (PPCD) services in large distributed CRNs based on elliptic curve cryptography (ECC). This technique also incorporates a block indexing strategy to improve the performance (in terms of computational running time) of PPCD. We evaluated the PPCD service with three real cohort definitions: (1) elderly cervical cancer patients who underwent radical hysterectomy, (2) oropharyngeal and tongue cancer patients who underwent robotic transoral surgery, and (3) female breast cancer patients who underwent mastectomy) with varied query complexity. These definitions were tested in an encrypted database of 7.1 million records derived from the publically available Healthcare Cost and Utilization Project (HCUP) Nationwide Inpatient Sample (NIS). We assessed the performance of the PPCD service in terms of (1) accuracy in cohort discovery, (2) computational running time, and (3) privacy afforded to the underlying records during PPCD.

Results: The empirical results indicate that the proposed PPCD can execute cohort discovery queries in a reasonable amount of time, with query runtime in the range of 165–262 s for the 3 use cases, with zero compromise in accuracy. We further show that the search performance is practical because it supports a highly parallelized design for secure evaluation over encrypted records. Additionally, our security analysis shows that the proposed construction is resilient to standard adversaries.

Conclusions: PPCD services can be designed for clinical research networks. The security construction presented in this work specifically achieves high privacy guarantees by preventing both threats originating from within and beyond the network.

1. Introduction

Clinical research networks (CRNs) are receiving an increasing amount of attention due, in part, to their ability to offer a collaborative environment for researchers across disparate organizations [31]. Moreover, CRNs are designed to leverage various types of data

* Corresponding author.
E-mail address: bianjiang@ufl.edu (J. Bian).

http://dx.doi.org/10.1016/j.jbi.2016.12.008
1532-0464/© 2016 Elsevier Inc. All rights reserved.

collected by both the healthcare systems (e.g., electronic health records, or EHRs, and claims) and directly from patients themselves (e.g., patient-reported outcomes, or PROs). It is anticipated that the analysis of such data will lead to advances in medical knowledge, progress in healthcare delivery, and improvements in population health. For example, the national Patient-Centered Clinical Research Network (PCORnet) [37], funded by the Patient-Centered Outcomes Research Institute (PCORI) [30,36], is an expansive network of networks of organizations who are partnered to conduct research. These PCORnet sites collect data from multiple sources and make them available for research. In particular, they provide an invaluable cohort discovery service that proves particularly useful for identifying cohorts of a variety of health conditions, and especially for rare diseases.

However, there is a potential for significant data privacy threats to be realized in CRNs. Data are shared by multiple participating health care organizations (HCOs), across different technical infrastructures, and thus are more susceptible to breaches. Existing CRNs have invested substantial effort towards protecting the patients' privacy as well as other sensitive information (e.g., organizations' billing information) involved in their networks. However, existing effort does not sufficiently address all important adversary models. In particular, we think a CRN should be protected against not only outside attackers but also malicious insiders (e.g., employees of the participating HCOs). Further, the current practice of privacy protection in health care often relies heavily on policies and guidelines such as the de-identification process defined by the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), which are inadequate to cover all scenarios and use cases in these emerging research networks. For example, the HIPAA Security Rule only considers the use of encryption for *data at rest* and *in transit*, but not for *data in use* that has the potential to significantly reduce the risk of insider attacks. The Health Information Technology for Economic and Clinical Health Act (HITECH) of 2009 imposed additional security measures, such as the data breach notification requirements for unsecured Protected Health Information (PHI), but such measures are also insufficient in terms of privacy protection through technology.

Furthermore, from a technological perspective, the majority of efforts dedicated to protecting health data in the wild (beyond automating the de-identification process) are focused on authentication, authorization, and data encryption in transit and at rest. Health IT professionals and health practitioners often assume the data are sufficiently secure when they live in a data center that meets compliance responsibilities (e.g., the HIPAA Security Rule [22] and FISMA [32]). Many of these compliance requirements are based on security standards and cybersecurity frameworks established by the National Institute of Standard and Technology (NIST), such as NIST 800-53 for Security and Privacy Controls for Federal Information Systems and Organizations [33]. The security controls in the NIST frameworks are important to deploy, but they are insufficient to ensure all possible security and privacy guarantees, especially in the CRN environment. Consider several pressing concerns:

- How can we protect health data against insider attacks due to malicious system administrators or negligence?
- How can we resolve trust issues where data contributors want to have control over their own datasets?
- What security controls should be instituted to limit the damage of large-scale data breaches, such as the recent Anthem breach where the hackers gained access to data on over 80 million health care consumers [19]?

Due in part to such privacy and trust concerns, CRNs often restrict data sharing by providing access to de-identified data only. These practices limit the utility of the data, especially for cohort discovery queries. For example, a resistant hypertension phenotype specification is dependent on the dates of patients' medication refills [25]. However, while there is a potential for retaining dates under the Expert Determination implementation of HIPAA de-identification, it is explicitly forbidden as one of eighteen types of identifiers under the Safe Harbor implementation, which is often invoked in practice.

1.1. Background

To address the aforementioned data privacy and trust issues in CRNs, one solution is to permit participating organizations to share data in an encrypted format and, at a later point in time, directly perform processing tasks for cohort discovery without decrypting patient-specific records. In doing so, the participating organizations can reap the benefits of a CRN while ensuring that both patient, as well as proprietary business, data are obscured from other organizations in the CRN. To achieve such a solution, privacy-preserving techniques including symmetric searchable encryption (SSE) [13] and public searchable encryption (PSE) [1–5] schemes are promising candidates. Both SSE and PSE aim to query data directly over encrypted data without decryption.

In particular, SSE enables data owners to outsource their data to an untrusted server (e.g., a public cloud) in an encrypted format and, subsequently, search the server using predefined keywords without decryption [34,35]. Since SSE schemes require that the data be preprocessed and encrypted under the same key, these methods are only appropriate for scenarios that either (1) involve a single data source or (2) there exists a centralized fully trusted entity to collect and encrypt all data from different data sources. However, the former requirement contradicts the use cases of a CRN, which involves multiple data sources (organizations) contributing data to form a collaboration. Furthermore, different organizations in a CRN do not necessarily trust each other (or their ability to maintain data securely), such that they may only want to share their patients' sensitive data in a privacy-preserving manner. Thus, SSE techniques cannot be directly applied in a CRN.

To overcome the aforementioned limitations of SSE, the notion of PSE [1–5] was introduced. PSE methods enable multiple data sources to encrypt their data with a shared public key, and send their encrypted data to a third party, where the encrypted data can be searched by the organization holding the private key at a later point in time. However, these PSE methods are vulnerable with regard to both inside [6] and outside [7] keyword guessing attacks, because their encryptions for search requests are deterministic (i.e., encryption results of the same request are always the same). Thus, existing PSE methods should not be directly applied in a CRN.

1.2. Contribution

In this paper, we introduce a novel privacy-preserving cohort discovery (PPCD) technique for CRNs. Specifically, the contributions of this work are:

- The technique supports flexible privacy-preserving frequency counts over encrypted patient data shared by multiple health data sources.[1]
- By designing our algorithmic construction with an underlying elliptic curve cryptography (ECC) system [29], we provide strong data privacy guarantees for our construction.

[1] We use the concepts of cohort discovery, search, and count interchangeably.