Decision Support Systems and Education
J. Mantas et al. (Eds.)

doi:10.3233/978-1-61499-921-8-35

Development of a Secure Cross-Institutional Data Collection System Based on Distributed Standardized EMR Storage

Katsuya TANAKA[a,1], Ryuichi YAMAMOTO[b], Kazuhisa NAKASHO[c] and Atsuko MIYAJI[d,e]
[a] *The University of Tokyo Hospital, Tokyo, Japan*
[b] *Medical Information System Development Center, Tokyo, Japan*
[c] *Grad. Sch. of Sci. and Technol. for Innov., Yamaguchi University, Yamaguchi, Japan*
[d] *Grad. Sch. of Eng., Osaka University, Osaka, Japan*
[e] *Japan Advanced Institute of Science and Technology, Ishikawa, Japan*

Abstract. This paper describes a secure data collection infrastructure involving standardized electronic medical record (EMR) storage and Private Set Intersection, a secure data collection technology based on Bloom filter. The objective of this infrastructure is to facilitate rapid secondary use of exported EMR data in cross-patient or cross-institutional analyses based on the Standardized Structured Medical Information eXchange (SS-MIX), Japan's domestic standard for EMR exporting. Design of the infrastructure and its underlying concepts are described herein. In an experimental test, an intersection operation involving approximately 1 million records was completed within a minute; this result is expected to be representative of the system in actual use. In forthcoming work, we plan to verify the system performance using larger data sets.

Keywords. standardization, data collection method, information protection

1. Introduction

In Japan, the domestic standard for exporting whole electronic medical record (EMR) data to external storage is the Standardized Structured Medical Information eXchange (SS-MIX2) [1], which is based on Health Level-7 (HL7) v2 message files and is widely used for backing up data, regional collaboration, various disease registries, and other purposes. The directory structure for stored records is based on patient ID, clinical date, and clinical event type, making the exported stored data difficult to use for cross-patient analyses, such as epidemiological studies.

In May 2017, an amendment to Japan's Act on the Protection of Personal Information [2] designated medical information as important confidential information, requiring strict consent for its provision to a third party. Secondary use of the information for research was permissible if utilized in accordance with ethical guidelines; however, large-scale data collection and analyses that included secondary use of the information by third parties, including commercial companies, became technically difficult because

[1] Corresponding Author, Katsuya Tanaka, Department of Healthcare Information Management, The University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo, Tokyo, Japan; E-mail: katsuya@hcc.h.u-tokyo.ac.jp.

of the requirement to obtain the consent of each patient for information disclosure. Therefore, a new law was brought into force in May 2018. Under this new law, a certified business operator can collect clinical information, anonymize the information, and respond to a third party's request for an analysis. The new law allows the secondary usage of patient information on the supposition of consent unless a patient specifically opts out. Patients are provided with the opportunity to opt out.

Several forms of data collection are assumed, with large-scale collection using the SS-MIX2 being one of the leading candidates. With the SS-MIX2, personal information is collected and transferred to one data center without anonymization. The collection of personal information as raw data poses a risk of information disclosure that cannot completely be excluded, and there are inevitable maintenance costs according to the scale of collection and storage [3].

In this paper, we propose an alternative method of collecting and storing EMR data, wherein only necessary items are included in collected data, eliminating the need for individual identifiable information to spread outside the medical institution. The system facilitates EMR data distribution within each medical institution, enabling cross-patient or cross-facility data collection and analysis. Data integration and encryption of the extracted EMR data is achieved using the Private Set Intersection (PSI) library developed by Miyaji [4]. The purpose of this paper is to provide an overview of the system and its major technical elements and to evaluate the transaction performance of data extraction and collection from the distributed SS-MIX2 storage.

2. Methods

2.1. Overview and Concepts of the developed system

Figure 1 presents an overview of the system. The key concepts are as follows:

a) Each medical institution has EMR data in the SS-MIX2 storage, including billions of HL7 v2 messages.
b) HL7 v2 messages are periodically parsed and stored to relational database management system (RDBMS) tables, maintaining synchronization with the billions of message files in the SS-MIX2 system.
c) Analysis requests of researchers and data collection are managed by the PSI service on the cloud, which communicates with the client agent located at the client terminal and PSI agents located at each medical institution.
d) Target data criteria, such as diseases, age, and gender, must be defined before the PSI executes data collection. The PSI Party agent deploys the target data set in advance from the local RDBMS to memory.
e) Data collection is achieved using the PSI software, which is based on Bloom filter technology for record verification across institutions.
f) The collected data set can be verified considering the possibility of patient identification using the extracted attributes.
g) Patients can trace the use of their medical records during data collection.
h) Patients can withdraw from the secondary use of their data if they wish.

2.2 Experimental environment

Transaction performance of data extraction and collection from the distributed SS-MIX2 storage was evaluated using an experimental environment comprising a server (PSI Server), three data stores (PSI Party), and a client (PSI Client). The Server and Party machines were deployed as VMware ESXi virtual machines. The PSI Client can be deployed on any machine that can run Java.

Experimental data were virtually produced by anonymizing laboratory test result data in the SS-MIX2 storage exported from the EMR system of The University of Tokyo Hospital (Tokyo, Japan). Storage assumed to have 10% overlap between each node was arranged and used for the evaluation tests. The hash value of the character string combining the patient's name, date of birth, and sex was used as the key attribute of each record for the Bloom filter.

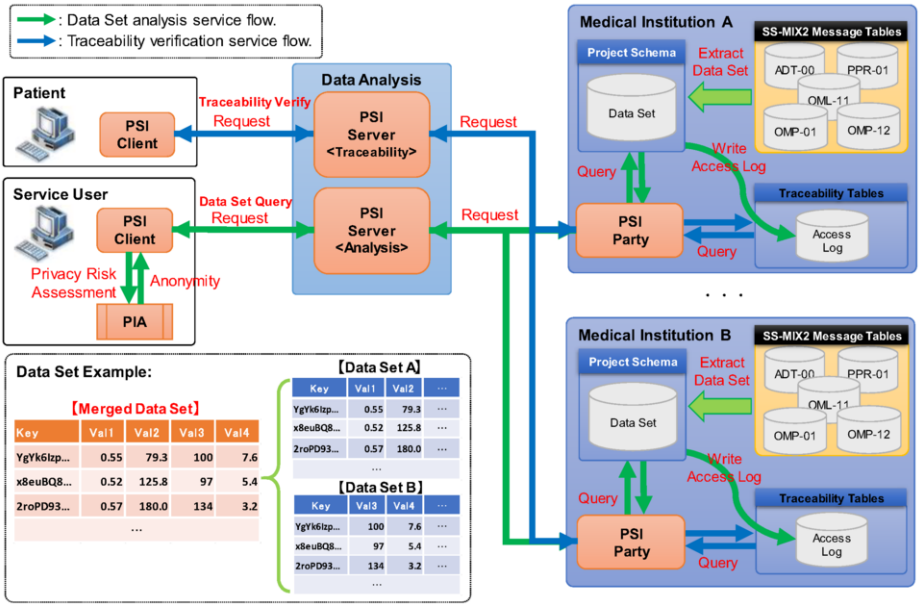

Figure 1. Overview of the System developed for Secure Data Collection and Analysis

2.3 Data collection with PSI

Figure 2 presents an overview of transaction flow during a secure data collection using the system. The entire system was designed as a Web service so that in the future we could make the service available via a commercial cloud. The PSI application programming interface was developed in Java using SOAP Web services and was deployed on an Apache Tomcat. All Web communications were implemented with client authentication under TLS 1.2. Extracted EMR data are encrypted by Cryptographic Message Syntax and can be decrypted only by the user requesting the collection.

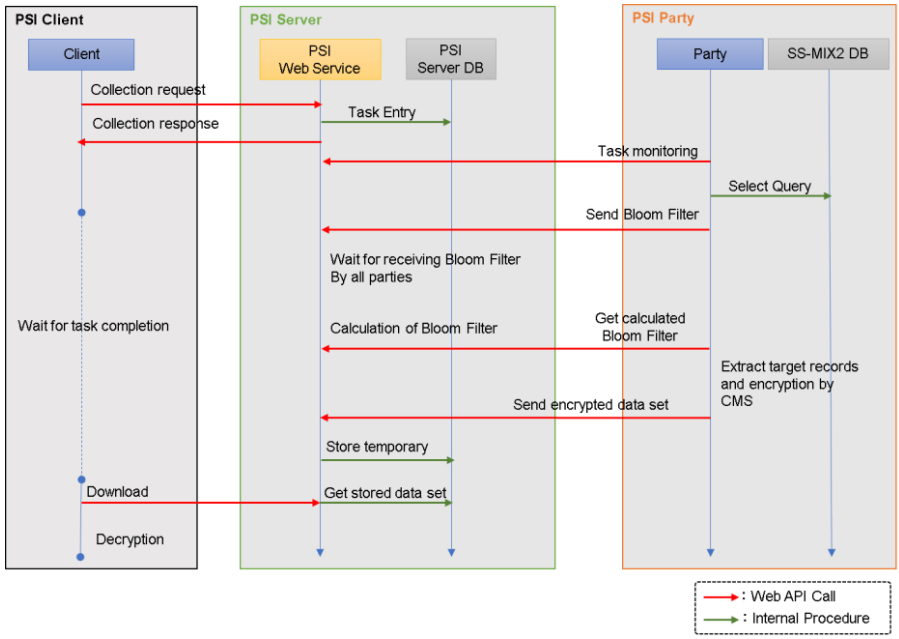

Figure 2. Overview of Transaction Flows during Data Collection

3. Results

Table 1 summarizes the evaluation test results for data queries for calculations, Bloom filter calculations, and result data extraction for increasing numbers of EMRs. The processing time linearly increased with the number of records.

Table 1. Evaluation Results (s) for Various Numbers of Medical Records

Records	20,000	40,000	80,000	160,000	320,000	640,000	960,000	1,280,000
Query Data	0.1	0.2	0.2	0.3	0.6	1.0	1.5	2.0
Bloom Filter Processing	0.5	0.5	1.0	2.0	2.9	7.4	13.5	20.7
Data Extraction	3.3	3.0	3.1	3.4	4.0	10.1	15.4	23.5
Total	3.8	3.7	4.3	5.7	7.4	18.5	30.4	46.2

4. Discussion

4.1. Significance of the system

The system was completely achieved using Web service architecture with encryption of the extracted EMR data, indicating that medical institutions participating in research would not need to maintain a secure connection to the specific service provider if the developed PSI services are operated on the commercial cloud. Encryption of EMR data avoids any disclosure of the extracted information to the cloud service providers. Furthermore, because the infrastructure makes it unnecessary to connect an EMR storage to the Internet, this eliminates the possibility of experiencing network attacks to the data

storage. To meet the requirements of a given analysis, the PSI can execute not only intersection operations but also union operations on distributed data sets.

4.2. Performance

The experimental results showed that an intersection operation involving approximately 1 million records was completed within a minute. With this level of processing performance, there should not be any problems with actual operations. We now intend to verify this with larger data sets.

4.3. Future work

The remaining issues for development include 1) the management of consent information, 2) risk assessment for the extracted data set, and 3) traceability management against data collection. The first issue can be addressed by scanning paper-based consent information related to patients opting out of secondary usage of their data and storing the scanned data files to the SS-MIX2 storage. We intend to represent consent information as XML files, such as HL7 CDA Privacy Consent Directives, Release 1 [5]. The other two issues are under discussion.

5. Conclusion

This paper describes the underlying concepts and implementation of a secure data collection infrastructure with distributed standardized EMR storage. Experimental results using the PSI data collection technology demonstrated high performance. A few issues remain for future implementation.

Acknowledgements

This study has been approved by the Research Ethics Committee of Graduate School of Medicine and Faculty of Medicine, The University of Tokyo (Permission number: 11787) and was supported by JST CREST Grant Number JPMJCR1404, Japan.

References

[1] M. Kimura, K. Nakayasu, Y. Ohshima, N. Fujita, N. Nakashima, H. Jozaki, T. Numano, T. Shimizu, M. S himomura, F. Sasaki, T. Fujiki, T. Nakashima, K. Toyoda, H. Hoshi, T. Sakusabe, Y. Naito, K. Kawagu chi, H. Watanabe, S. Tani, SS-MIX: a ministry project to promote standardized healthcare information e xchange. *Methods Inf Med*. **50** (2) **(**2011), 131-139.

[2] Amended Act on the Protection of Personal Information, 2016. Available from: https://www.ppc.go.jp/fil es/pdf/Act_on_the_Protection_of_Personal_Information.pdf.

[3] P. Pazos Gutierrez, Towards the Implementation of an openEHR-based Open Source EHR Platform, *Stud Health Technol Inform.* **216** (2015), 45-49.

[4] A. Miyaji, K. Nakasho, S. Nishida, Privacy-Preserving Integration of Medical Data, *Journal of Medical Systems* **41**(3) (2017), 37.

[5] HL7 Standards Product Brief - HL7 CDA R2 Implementation Guide: Privacy Consent Directives, Releas e 1 2017. Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=280.