Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-766

Privacy-preserving Statistical Query and Processing on Distributed OpenEHR Data

Meskerem Asfaw HAILEMICHAEL[a,1], Luis MARCO-RUIZ[b] and Johan Gustav BELLIKA[b,c]
[a] *Department of Computer Science, Faculty of Science and Technology, UiT The Artic University of Norway*
[b] *Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway*
[c] *Department of Clinical Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway*

Abstract. Reuse of data from EHRs is essential for many purposes. The objective of the study was to explore how distributed electronic health record (EHR) data can be reused for privacy-preserving statistical query and processing. Method: We have designed and created a proof of concept prototype solution based on the OpenEHR specification to ensure interoperability and to query the EHRs. XMPP was used for communication between the distributed processing components. Results: We have created a two-phased process where a distributed virtual dataset is first created and thereafter processed using distributed privacy-preserving statistical queries. Conclusion: Health authorities in Norway are currently defining the set of archetypes for the national interoperability program. This will create a common information schema enabling reuse of EHR data for statistical query and processing in a privacy-preserving manner. One benefit of the approach is that information transformation between information models for clinical use and statistical processing can be avoided.

Keywords. Privacy, Medical Records Systems, Information Storage and Retrieval

Introduction

The increasing uses of electronic health record (EHR) systems by different health institutions have led to huge collection of sensitive health data. Apart from patient treatments, patients' data can be used for various purposes that help to improve health outcomes and costs. These include clinical research [1], clinical audit [2], public health [3], and healthcare quality measurement and improvement.

Two main approaches for reuse exist when EHR data is distributed across multiple health institutions: centralized [4] and distributed [5]–[8]. The centralized approach involves collecting individual patient records into a large centralized database. This approach is easy to use, as it needs no complex mechanisms to explore the data, since all needed information is available in one place. On the other hand, the distributed approach employs different mechanisms to use data without the need to move individual patient records from its original place. Most importantly, the

[1] Corresponding Author.

distributed approach gives autonomy to the data owners regarding who uses what, and for what purpose, which could encourage them to participate. Thus, the focus of this study is the distributed approach.

There are several factors that should be considered for reuse of distributed health data. These include ethical issues, privacy concerns, cross-institutional contracting policies and regulations [5], the heterogeneity of healthcare information system [9], quality of data, and the complexity (lack of structure) and incompleteness of EHR data for reuse [10]. Among all the challenges in data reuse, the major portion is held by privacy [11] and interoperability issues [9].

The issue of privacy is a threat to patients to an extent that they self-medicate their illnesses, ask their information not to be registered in EHR, be unwilling to participate in clinical trials and be reluctant to give consent to any public health research [12], [13]. On the other hand, using distributed and heterogeneous EHRs data for secondary purpose poses the question of interoperability [9].

Ever since the concept of health data reuse was introduced, many efforts have been made to implement EHR query tools on both centralized and distributed approaches [6]. However, few of them have been deployed beyond the pilot stage. This is because they are yet to handle most important features such as transfer of "raw patient data" which violates privacy policies, user friendly interface to be used by non-technical users (e.g. researchers) [14]. Some of the available query tools to search patient databases are presented in [6], [15]–[17]. Generally, in order to get the full benefit of health data reuse, a reliable query tool that addresses both privacy and interoperability is needed. The objective of this paper is to present 1) a technique showing how distributed EHRs data can be reused for privacy-preserving statistical query and processing; 2) results from development of a proof of concept solution. The current strategic plan of Norwegian Health authorities is currently driving EHR vendors to adopt OpenEHR to enable health information interoperability. Therefore, we assume that all the EHRs are OpenEHR compliant.

1. Materials and methods

To be compatible with the future OpenEHR based hospital EHRs, we used the Think!EHR platform [18] for storing the data collections used in the development and testing of the query tool. Think!EHR enable persisting EHR extracts compliant with OpenEHR and query them using the Archetype Query Language (AQL). We used the XMPP protocol for communication and interaction between the distributed processing components. Strophe library was used to establish connections between a web client and OpenFire XMPP server. We used the Smack library to enable XMPP communications between the computation components. Iterative engineering approach was used to develop a proof of concept solution that achieved our objective.

2. Results

We developed a proof of concept solution that satisfies the requirements (listed in Table 1) specified based on literature study and interviews with users. The solution has a web client application where a user inputs a query and gets the result. As shown in Figure 1, the solution has, 1) virtual dataset creation and, 2) statistical computation

phases. Users specify the required datasets as AQL queries, which are executed against the Think!EHR at each participating site. To avoid transfer of a single patients data outside the site and possible changes to the original data during statistical processing, the query results (individual level data) are stored locally in a separate database. The combination of datasets created at all sites gives the overall query result; however, datasets are physically distributed, what we call virtual dataset. As datasets do not abandon the source organization, no de-identification of data is needed. The datasets is stored locally for the approved research duration. The dataset creation query is based on the universal information model shared by all OpenEHR based EHR installations. This specification will be limited by the archetypes listed and approved for usage in the Norwegian archetype registry available in the Clinical Knowledge Manager registry at http://arketyper.no. As extract, transform and load (ETL) process is not needed to transform the data, the AQL queries creating the datasets can be executed against the production systems at any time. The second phase performs statistical computations against the virtual dataset as illustrated in Figure 1. First, statistical computation query sent from the user is executed against the virtual dataset. Then, the coordination server uses secure summation [19] techniques to securely collect the intermediate results generated at the virtual dataset and perform the statistical computations upon them. Finally, the final result of the computation is displayed to the end user.

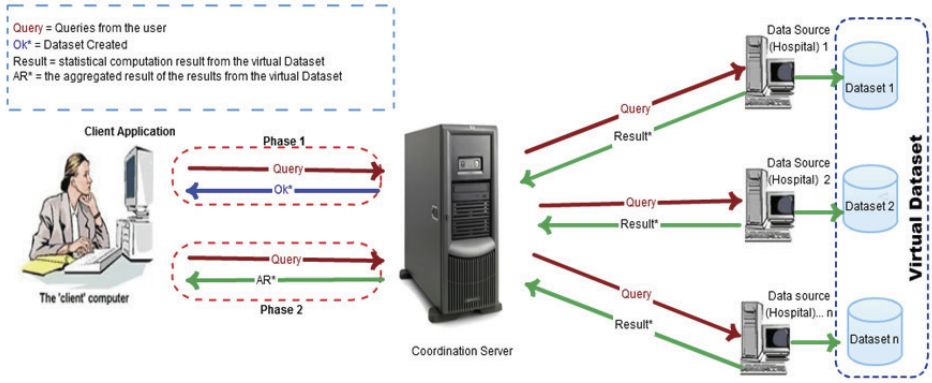

Figure 1. A two-phase process for privacy-preserving statistical query on distributed EHRs.

3. Discussion and conclusions

The existence of a national OpenEHR archetype repository for Norwegian EHRs has the potential to solve interoperability concerns which is the main data reuse challenge in addition to privacy. In this paper, we presented a proof of concept solution to overcome privacy issues in data reuse. Our solution creates a virtual dataset and then processes statistical queries on the virtual dataset. Having separate datasets, outside the EHR system, enables to improve the quality of data for reuse and to perform further queries in the future. Hence, this avoids data owners' fear of data loss or data modification, which gives them more confidence in contributing to data reuse.

SPIN [16], SHRINE [6] and EHR4CR [19, p. 4] are examples of query tools that use the distributed approach. In SPIN, each site executes a user query and de-identifies the result.

Table 1. System requirements for privacy-preserving statistical query on distributed EHRs.

#	Requirement	Rationale
1	The system must not allow any transfer of identifiable sensitive personal data.	To preserve patients' privacy and satisfy the Norwegian Personal Data Act §2 [20].
2	The system shall allow the user to create a query that extracts dataset from distributed EHRs.	To enable the user to create a query of interest.
3	The system shall broadcast queries to all sites	To enable execution of the queries against the EHRs.
4	The system shall execute the query against the EHRs.	To create a virtual dataset that matches the Dataset Creation query.
5	The system shall store the result of the dataset creation query locally at each EHR site.	To prepare the appropriate dataset for statistical computation.
6	The system shall allow the user to create queries for statistical computation.	To enable the user to perform the statistical computation of their interest.
7	The system shall execute the statistical computation against the virtual dataset.	To create a result that matches the statistical computation query.
8	The system shall compute the aggregate result without revealing combined statistics of <k number of sites' private data.	To avoid sending individual institution's result to the user and consequently preserve privacy.
9	The system shall display the aggregate result to the user.	To make the final result available to the user.

The de-identified data is extracted from all sites and stored on a central server for further processing. However, in our system, person level data records cannot be transported out of the original sources, only aggregated anonymous data. SHRINE uses different privacy-preserving techniques such as adding random number to a result, and using anonymous names for the hospitals. However, the result may still be exposed to privacy issues, since individual hospital level result is displayed. We overcome this problem by making a coordination server aggregate results from every institution before it is displayed to the user. Furthermore, both SHRINE and EHR4CR need to map heterogeneous EHRs into a common format to run the query tool.

Our approach is based on the OpenEHR specification, which implements multi-level modeling framework to facilitate interoperability. As a common data schema is used, no adaption of data access queries needs to be performed. Therefore the same queries can be applied to all systems. The drawback of having one data access mechanism is that we only support sources that provide an AQL interface. Non-OpenEHR based legacy systems still need to transform their data using an ETL mechanism, before data can be reused using the proposed approach. However, tools like LinkEHR [21] can be used to enable such transformations.

This work is a part of a project called SNOW [22] which is an agent based distributed EHR data processing system. The role of the SNOW system in the proposed approach is to perform initialization and coordination of the distributed computation components among the sites participating in the computations. We are expanding the solution using secure multi-party computation techniques to strengthen privacy, improve data quality and enhance the statistical computation functionalities [23]. The secure multi-party computation techniques will be used for privacy-preserving joint computation between the data sources to aggregate their local computation results. Thus, it avoids the need for revealing local computation results to a coordination server and only combined result of all the data sources will be revealed. The current solution assumes that a patient record only exist at one hospital. As future work we are developing solutions for patient records distributed across multiple hospitals where the patient received care.

References

[1] M. Bloomrosen and D. Detmer, "Advancing the Framework: Use of Health Data--A Report of a Working Conference of the American Medical Informatics Association," *J Am Med Inform Assoc*, vol. 15, no. 6, pp. 715–722, 2008.

[2] K. Dentler, A. ten Teije, N. de Keizer, and R. Cornet, "Barriers to the reuse of routinely recorded clinical data: a field report," *Stud Health Technol Inform*, vol. 192, pp. 313–317, 2013.

[3] D. P. Jutte, L. L. Roos, and M. D. Brownell, "Administrative record linkage as a tool for public health research," *Annu Rev Public Health*, vol. 32, pp. 91–108, 2011.

[4] "HHS awards $1M contract for effectiveness database | Government Health IT." [Online]. Available: http://www.govhealthit.com/news/hhs-awards-1m-contract-effectiveness-database. [Accessed: 10-Feb-2014].

[5] J. S. Brown, J. H. Holmes, K. Shah, K. Hall, R. Lazarus, and R. Platt, "Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care," *Med Care*, vol. 48, no. 6 Suppl, pp. S45–51, Jun. 2010.

[6] G. M. Weber, S. N. Murphy, A. J. McMurry, D. MacFadden, D. J. Nigrin, S. Churchill, and I. S. Kohane, "The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories," *J Am Med Inform Assoc*, vol. 16, no. 5, pp. 624–630, Sep. 2009.

[7] R. Lazarus, K. Yih, and R. Platt, "Distributed data processing for public health surveillance," *BMC Public Health*, vol. 6, no. 1, p. 235, Sep. 2006.

[8] J. C. Maro, R. Platt, J. H. Holmes, B. L. Strom, S. Hennessy, R. Lazarus, and J. S. Brown, "Design of a National Distributed Health Data Network," *Ann Intern Med*, vol. 151, no. 5, pp. 341–344, Sep. 2009.

[9] D. Ouagne, S. Hussain, E. Sadou, M.-C. Jaulent, and C. Daniel, "The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology," *Stud Health Technol Inform*, vol. 180, pp. 534–538, 2012.

[10] J. A. Berlin and P. E. Stang, "CLINICAL DATA SETS THAT NEED TO BE MINED," in *Learning What Works: Infrastructure Required for Comparative Effectiveness Research: Workshop Summary*, vol. 1, National Academies Press (US), 2011.

[11] A. Geissbuhler, C. Safran, I. Buchan, R. Bellazzi, S. Labkoff, K. Eilenberg, A. Leese, C. Richardson, J. Mantas, P. Murray, and G. De Moor, "Trustworthy reuse of health data: A transnational perspective," *International Journal of Medical Informatics*, vol. 82, no. 1, pp. 1–9, Jan. 2013.

[12] B. A. Malin, K. E. Emam, and C. M. O'Keefe, "Biomedical data privacy: problems, perspectives, and recent advances," *J Am Med Inform Assoc*, vol. 20, no. 1, pp. 2–6, Jan. 2013.

[13] M. A. Rothstein, "Is Deidentification Sufficient to Protect Health Privacy in Research?," *Am J Bioeth*, vol. 10, no. 9, pp. 3–11, Sep. 2010.

[14] S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. C. Chueh, S. Churchill, and I. Kohane, "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)," *J Am Med Inform Assoc*, vol. 17, no. 2, pp. 124–130, Mar. 2010.

[15] D. F. Sittig, B. L. Hazlehurst, J. Brown, S. Murphy, M. Rosenman, P. Tarczy-Hornoch, and A. B. Wilcox, "A Survey of Informatics Platforms That Enable Distributed Comparative Effectiveness Research Using Multi-institutional Heterogenous Clinical Data," *Medical Care July 2012*, 2012.

[16] C. J. McDonald, P. Dexter, G. Schadow, H. C. Chueh, G. Abernathy, J. Hook, L. Blevins, J. M. Overhage, and J. J. Berman, "SPIN Query Tools for De-identified Research on a Humongous Database," *AMIA Annu Symp Proc*, vol. 2005, pp. 515–519, 2005.

[17] "EHR4CR: Electronic Health Records for Clinical Research." [Online]. Available: http://www.ehr4cr.eu/index.cfm. [Accessed: 25-Mar-2014].

[18] "Think!EHR Platform." [Online]. Available: http://www.marand-thinkmed.com/thinkehr. [Accessed: 10-Mar-2014].

[19] S. Wang, X. Jiang, Y. Wu, L. Cui, S. Cheng, and L. Ohno-Machado, "EXpectation Propagation LOgistic REgRession (EXPLORER): Distributed Privacy-Preserving Online Model Learning," J Biomed Inform, vol. 46, no. 3, pp. 480–496, Jun. 2013.

[20] Norwegian Data Protection Authority, "Act of 14 April 2000 No. 31 relating to the processing of personal data (Personal Data Act)," 2000.

[21] J. A. Maldonado, D. Moner, D. Boscá, J. T. Fernández-Breis, C. Angulo, and M. Robles, "LinkEHR-Ed: a multi-reference model archetype editor based on formal semantics," *Int J Med Inform*, vol. 78, no. 8, pp. 559–570, Aug. 2009.

[22] J. G. Bellika, T. Henriksen, and K. Y. Yigzaw, "The Snow System – A Decentralized Medical Data Processing System," in *Data Mining in Clinical Medicine*, vol. 1246, Spinger, 2014.

[23] K. Y. Yigzaw, J. G. Bellika, A. Andersen, G. Hartvigsen, and C. Fernandez-Llatas, "Towards Privacy-preserving Computing on Distributed Electronic Health Record Data," in *Proceedings of the 2013 Middleware Doctoral Symposium*, New York, NY, USA, 2013, pp. 4:1–4:6.