Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-581

Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data

Johan van SOEST[a,1], Chang SUN[b], Ole MUSSMANN[c], Marco PUTS[c], Bob van den BERG[c], Alexander MALIC[b], Claudia van OPPEN[b], David TOWEND[d], Andre DEKKER[a] and Michel DUMONTIER[b]

[a] *Department of Radiation Oncology (MAASTRO), GROW school for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands*

[b] *Institute of Data Science, Maastricht University, Maastricht, The Netherlands*

[c] *Centraal Bureau voor de Statistiek (CBS), Heerlen, The Netherlands*

[d] *Department of Health, Ethics and Society, CAPHRI Research School, Maastricht University, Maastricht, the Netherlands*

Abstract. Conventional data mining algorithms are unable to satisfy the current requirements on analyzing big data in some fields such as medicine, policy making, judicial, and tax records. However, applying diverse datasets from different institutes (both healthcare and non-healthcare related) can enrich information and insights. So far, analyzing this data in an automated, privacy-preserving manner does not exist to our knowledge. In this work, we propose an infrastructure, and proof-of-concept for privacy-preserving analytics on vertically partitioned data.

Keywords. Infrastructure, machine learning, data mining, statistics, privacy-preserving, secondary use of data

1. Introduction

Information exchange in the healthcare domain is becoming increasingly important. In first place for clinical purposes such as transfer of care documents among healthcare providers, however also increasingly for secondary use such as development of value-based healthcare and healthcare learning systems. For clinical purposes, information exchange systems are mostly targeted on the exchange of data, e.g. using syntactical standards (e.g. HL7) which facilitate semantic standards (e.g. terminological systems) [1]. For secondary use, purposes translate into e.g. business analytics, obligations to (governmental) registries, and scientific research. Although the use is different, the same clinical standards can be used for secondary purposes.

Although these standards provide transfer of information, they also raise questions about maintainability and ownership, and subsequently security and privacy. By transferring information between multiple health care providers, provenance and authorization become more complex. Furthermore, propagating provenance and

[1] Corresponding Author

authorization changes (e.g. changes in patient consent) becomes a complex task, as all health care providers who received the information need to re-validate their provenance and authorization, or even remove the data from their systems. Furthermore, public confidence regarding data security by large companies has been impaired by recent high-profile breaches (e.g. Equifax breach). Subsequently, policy makers and EU General Data Protection Regulations attempt to increase the requirements for data collection and use, however revise less what alternative options are [2].

One of the alternatives to data transfer is to investigate sending applications containing questions and algorithms to the data source. The goal of this paper is twofold: a) to develop an infrastructure to facilitate transfer and execution of algorithms, b) to apply this infrastructure in a proof-of-concept setup. This proof-of-concept (PoC) will focus on analyzing vertically partitioned data from two institutes. Beyond the scope of this paper, the PoC will be used as a baseline to investigate the causes of onset and progression of Diabetes Mellitus in a population cohort study; including socioeconomic and environmental factors. This paper is further organized into methods, results and conclusion/discussion. The methods section describes the development process of an infrastructure for communicating algorithms and results, called the Personal Health Train (PHT) infrastructure. Furthermore, this paragraph will explain the PoC setup. The results section briefly explains the developed infrastructure (open-source available), and a reference PoC implementation. Finally, the conclusion & discussion will explain our main findings, strengths and weaknesses, and future work.

2. Methods

The methods below are guided by the scientific question to perform analyses on a population cohort study, enriched with complementary information. Hence, we will first discuss the identification and development of required concepts, and afterwards define the PoC methods used for privacy-preserving processing using complementary data analytics.

2.1. Identification of complementary data analytics methods

We started this project by identifying options for complementary data analytics: performing analyses on datasets which have common patients, however have different data elements per patient. The main questions in this identification process were whether data should be transferred, and if yes, with or without patient identifiers. Afterwards, we identified current (commonly used) approaches to complementary data analytics. Finally, we chose the most appropriate starting point for implementation.

2.2. Development of the PHT infrastructure

The main goal of the PHT infrastructure is to provide a general-purpose infrastructure, where many different questions can be asked at multiple data owners (e.g. hospitals or even patients themselves). Using such an infrastructure, data owners should have more control over which questions and/or analytics are performed on their data. Furthermore, it should reduce data duplication, and its involved administrative issues [3]. Previously, we have successfully co-developed an infrastructure, which has been adopted by a

commercial entity (Varian Learning Portal, Varian Medical Systems, Palo Alto, CA, USA). This system has been successful for distributed machine learning on horizontally partitioned datasets, however is not flexible in terms of analysis tools used, or configuration within hospital infrastructures. In this work, we will continue on previous experience and developed several criteria for the newly developed infrastructure:

- Executing questions at a local institute should be operating system agnostic
- It should facilitate use of different (versions of) libraries
- Communication and computation should be separated
- The communication network should be as light-weight as possible
- IT administration and requirements on the client side should be as limited

One of the consequences of sending algorithms to the data, is that we cannot actually access (and "see") the data, and have to rely on information given regarding used data structures and systems where they are stored. Previously, we have developed an open-source infrastructure to extract data from clinical systems into standardized formats [4,5].

2.3. Proof of Concept (PoC) setup

The developed infrastructure was tested as a PoC in a collaborative information exchange project between a university and the national statistics agency. Specifically, this collaboration targets the vertically partitioned data problem, analyzing data from both participating. In the current PoC, we simulated two datasets:

- At the university: personal identifier and age
- At the statistics agency: personal identifier and income

The datasets were unbalanced in terms of number of patients, where the statistics agency has a large dataset, and the university dataset contains a small subset of patients. This resembles the actual situation, where the statistics agency has more data in comparison to the university. The goal of this PoC was to develop an automated system to plot the relationship between age and income.

3. Results

3.1. Identification of complementary data analytics methods

The final tree of complementary data analytics approaches, as a result from the brainstorm sessions, is shown in Figure 1. The tree in this figure also resulted in a development and validation flow; by starting with data transport and patient identifiers, we will have a validation method for more challenging approaches. In the current PoC, we will use the setup using a trusted third party (TTP) for linking datasets and performing the actual analysis. Using this TTP

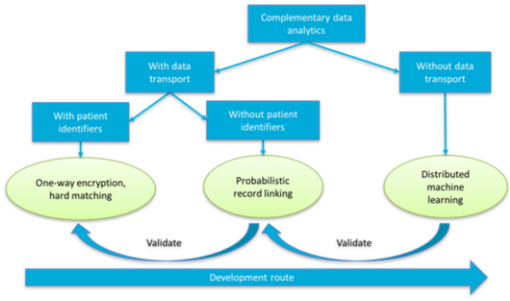

Figure 1. Approaches for complementary data analytics and the chosen development and validation workflows

and appropriate encryption methods, the chances of one party pertaining all datasets and being able to decrypt them are limited (as the TTP cannot retrieve the original patient IDs).

3.2. Development of the Personal Health Train infrastructure

In our current PoC, we split the Personal Health Train infrastructure into a client-server architecture, connected using the internet (HTTPS). These two developed applications are: a) central message dispatcher and b) client execution application. In this infrastructure, the client execution application (CEA) registers itself at the central message dispatcher (CMD). When successfully registered, the CEA will ask the CMD whether it needs to execute new tasks at regular time intervals. These tasks identify the execution of specific Docker images; hence a task only specifies the docker image identifier and optional (additional) input parameters, stored in a text-based format. The CEA will retrieve the Docker image form the central repository, will append several properties regarding access to the local data source (e.g. intranet URL of the data source), and executes the Docker container. The output of this container should be a text-based result, and is sent back to the CMD. This infrastructure is publically available at https://bitbucket.org/jvsoest/pytaskmanager.

3.3. Implementation of proof-of-concept

Based on the result of section 3.3, we simulated all involved parties: both institutes, and a TTP. Both institutes installed a CMD, to receive algorithms and work with the simulated data available. At the TTP, a modified version of the CEA was installed to fulfill the role of data receiver.

Three Docker containers were developed: a) for data extraction and encryption of the data at both institutes and b) for decryption at the TTP. The containers sent to the institutes (a) contained queries on the FAIR data sources. The personal identifiers would then be hashed with an agreed-upon salt at both sites. Afterwards the complete dataset would be signed, encrypted and signed, before sending the data to the TTP CEA. Encryption of the dataset was performed using symmetric encryption for performance reasons (symmetric keys are faster for large datasets, in comparison to public key encryption). The symmetric keys were exchanged separately using public key encryption. Finally, when encrypted data was sent, the container produced a positive (message: “OK”) result to the CMD that it performed the given task at hand.

After both centers had given a positive result, the final task (container b) was sent to the TTP CEA. First, this container would retrieve the signed and encrypted data from the CEA, and verify-decrypt-verify the data using the securely provided verification and symmetrical decryption key. The first verification was to verify the encryption of the data, and the second signature was to verify the actual dataset. Afterwards, it would perform the actual analysis (merging both datasets), resulting in a scatterplot of age and income of the matched patients in both datasets. Plots can be retrieved manually from the TTP, to ensure a manual validation of anonymity of the results. Docker containers including data were removed by the CEA when execution finished for all locations (data providers & TTP).

4. Conclusion and Discussion

We have successfully shown that the developed infrastructure and proof-of-concept worked with simulation data. Although the proof-of-concept implementation only shows one case example, the infrastructure can be reused for different questions; both for horizontally and vertically partitioned data.

Limitations of the current work pertain the limited scope of variables in the PoC, and the use of simulation data. Furthermore, the infrastructure will need more security enhancements before actually being implemented in practice. Ethical, legal and societal issues (ELSI) are also of importance in such an infrastructure, however were not scoped in the current prototype. The discussion between ELSI, technical and scientific challenges is a continuous debate among different stakeholders, and evolves over time. Hence, we developed this PoC as input for ELSI discussions, and to show the technical possibilities to the scientific field. Furthermore, our example relies on researchers/analysts which can develop algorithms without actually accessing the data directly. This was not an issue with simulated data in our PoC, however will be addressed using the FAIR principles [6,7] in future work.

Future work will include the discussion and development of an ELSI framework, where the different approaches for complementary data analytics will be discussed, from multiple stakeholder perspectives. In example, some scientific questions can only be answered with specific technical methods, which have certain ELSI requirements in terms of consent and privacy/security aspects. Likewise, ELSI insights may result in different technical opportunities or scientific directions.

From a technical perspective, future work will pertain further development of this reference infrastructure (e.g. security measures on executing applications), and case examples to use this network. FAIR descriptions of datasets will be part of these case examples, as well as measures to define of FAIR principles.

References

[1] G.W. Beeler, HL7 version 3--an object-oriented methodology for collaborative standards development, Int J Med Inform. 48 (1998) 151–161. doi:10.1007/978-0-387-34910-7_25.
[2] B.J. Koops, The trouble with European data protection law, International Data Privacy Law. 4 (2014) 250–261. doi:10.1093/idpl/ipu023.
[3] J.P.A. van Soest, A.L.A.J. Dekker, E. Roelofs, G. Nalbantov, Application of Machine Learning for Multicenter Learning, 2015. doi:10.1007/978-3-319-18305-3_6.
[4] J. van Soest, Data Integration Tutorial, Amsterdam, n.d. https://github.com/jvsoest/Data-Integration-Tutorial/wiki.
[5] J. van Soest, T. Lustberg, D. Grittner, M.S. Marshall, L. Persoon, B. Nijsten, et al., Towards a semantic PACS: Using Semantic Web technology to represent imaging data, Stud Health Technol Inform. 205 (2014) 166–170. doi:10.3233/978-1-61499-432-9-166.
[6] T.M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, et al., Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT, Clinical and Translational Radiation Oncology. 4 (2017) 24–31. doi:10.1016/j.ctro.2016.12.004.
[7] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, et al., The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data. 3 (2016) 160018–9. doi:10.1038/sdata.2016.18.