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

doi:10.3233/978-1-61499-852-5-76

Enabling Analytics on Sensitive Medical Data with Secure Multi-Party Computation

Meilof VEENINGEN[a,1], Supriyo CHATTERJEA[a], Anna Zsófia HORVÁTH[b], Gerald SPINDLER[b], Eric BOERSMA[c], Peter van der SPEK[c], Onno van der GALIËN[d], Job GUTTELING[e], Wessel KRAAIJ[f] and Thijs VEUGEN[g]
[a] *Philips Research, Eindhoven, The Netherlands*
[b] *University of Göttingen, Germany*
[c] *Erasmus Medical Center, Rotterdam, The Netherlands*
[d] *Achmea, Zeist, The Netherlands*
[e] *OLVG, Amsterdam, The Netherlands*
[f] *TNO, The Hague, The Netherlands and Leiden University, The Netherlands*
[g] *TNO, The Hague, The Netherlands and CWI, Amsterdam, The Netherlands*

Abstract. While there is a clear need to apply data analytics in the healthcare sector, this is often difficult because it requires combining sensitive data from multiple data sources. In this paper, we show how the cryptographic technique of secure multi-party computation can enable such data analytics by performing analytics without the need to share the underlying data. We discuss the issue of compliance to European privacy legislation; report on three pilots bringing these techniques closer to practice; and discuss the main challenges ahead to make fully privacy-preserving data analytics in the medical sector commonplace.

Keywords. Big data; privacy; data sharing; privacy-preserving data mining; secure multi-party computation; general data protection regulation

1. Introduction

The need to better use data analytics in the healthcare sector is nowadays well-understood. With healthcare costs already high and expected to rise even more (10% of the GDP of the European Union already and estimated to rise by 30% by 2060), it is important to control costs while not compromising on quality and access. The use of knowledge that is hidden in existing medical data (of which already zettabytes are available, soon rising to yottabytes) is seen as the "fastest, least costly, and most effective path to improving people's health" [3] and may lead to cost savings of over $250M in the US alone [4].

Gaining insights from medical data typically requires working with very sensitive data, often from multiple data sources. This can be within organisations (for instance, combining medical records in a hospital with real-time data from a tracking system); between organisations in the same vertical of the healthcare sector (for instance, using insights in the effectiveness of treatments for particular patient types across hospitals); or even between organisations in different verticals (for instance, combining data from hospitals and insurance companies to link medical data to data about treatment costs).

[1] M. Veeningen, Philips Research, High Tech Campus 34, Eindhoven, meilof.veeningen@philips.com

Unfortunately, while big data analytics in the healthcare sector is very relevant, it is also very challenging because it needs to be applied to very sensitive data. With recent stories about data breaches in mind (e.g., [8]), both healthcare organisations and patients are rightly reluctant to have their data being used for analytics. At the same time, the upcoming EU General Data Protection Regulation (GDPR) requires a high standard for data security, consent, and other measures to be in place before personal information is even allowed to be processed. As a result, for many healthcare organisations, "fears about data release outweigh their hope of using the information" [4].

In this paper, we show how the cryptographic technique of secure multi-party computation (MPC) can enable new data analytics applications by performing data analytics without the need to share the underlying data. As a consequence, data providers can contribute data to an analysis while being technically guaranteed that the data cannot be de-anonymised (e.g., [2]), decrypted, or used for any other purpose than the intended one. We demonstrate the potential of this technology by discussing three ongoing pilots: one inside a hospital, one between hospitals and one between a hospital and an insurance company. We also discuss the main technical and non-technical hurdles we see before these techniques can be applied in practice at a large scale.

This paper is structured as follows. We introduce privacy-preserving data mining based on MPC in Section 2, and discuss its relation to EU privacy legislation in Section 3. In Section 4, we discuss the three pilot projects. In Section 5, we discuss the outlook for MPC, including challenges to make privacy-preserving data analytics become reality.

2. Privacy-Preserving Data Mining based on Secure Multi-Party Computation

In the medical domain, there is a frequent need to perform data mining on sensitive data (e.g., medical/financial data), raising data sharing challenges that privacy-preserving data mining (PPDM) aims to address. PPDM traditionally assumes that there is one data owner, and aims at anonymizing data such that third parties can still mine patterns from it. Various anonymisation techniques are known, e.g., based on *k*-anonymity or differential privacy; but they generally reduce utility, and in many cases, partial de-anonymisation of supposedly anonymised datasets turned out to be possible [2].

PPDM using secure multi-party computation (MPC), introduced 17 years ago in a seminal paper by Lindell and Pinkas [6], goes beyond traditional techniques by assuming that multiple parties with confidential datasets want to mine their combined data. This data may be distributed horizontally (parties have the same kind of information on different data subjects) or vertically (parties have different kinds of information on the same data subjects). In the horizontal case, global analytics can often be approximated by locally computing aggregates and then combining them [5], but this does not work in the vertical case. In such cases, MPC allows analytics by distributing each sensitive data item between multiple processors who interact using a cryptographic protocol.

One main technique in such protocols is *secret sharing*, of which we give a simple example. Suppose three doctors want to know the total number of patients they treated, without sharing their individual subtotals. To do this, doctor 1 takes his subtotal x, adds a large random number, r, to it (chosen randomly, say, between 0 and 1000000), and sends $x+r$ to doctor 2. Note that r "statistically hides" x from doctor 2. Doctor 2 then adds his subtotal, y, to the running total and sends $x+r+y$ to doctor 3, who similarly adds his subtotal, z, and sends $x+r+y+z$ to Doctor 1. Doctor 1 subtracts r and reveals the end result, $x+y+z$. Note that secrets (x,y,z) are hidden by randomness (r) and a computation

is performed by exchanging randomised values between the participants. This basic principle can be used to perform any computation while only revealing its end result [6]. Other popular techniques are threshold and/or homomorphic encryption.

3. Legal framework for privacy-preserving data mining

Two aspects of the upcoming GDPR are particularly relevant for privacy-enhancing techniques (PETs) such as MPC. First, using state-of-the-art PETs, partially anonymised (pseudonymous or encrypted) personal data may qualify as anonymous data. According to the GDPR, data is personal if it contains any information relating to an identified or a directly or indirectly identifiable natural person. If a set of personal data has been fully anonymized, i.e. there is no mean reasonably likely to be used to identify any individual, the data will no longer be treated like personal data, and will not be subject to the GDPR. A limitation is that a potential possibility of an unauthorized access always has to be taken into account, but only if the means used are reasonably likely [1].

Second, the GDPR introduces a new definition for consent. Organizations that rely on the consent of data subjects as a lawful basis of processing are particularly affected, since the GDPR demands an explicit, unambiguous, specific informed consent given by a statement or clearly affirmative action in an opt-in mechanism. Controllers must in addition carefully evaluate the purposes of processing and are not allowed to collect data extensively. Data protection by design and by default set out the obligation for data processing systems to embed technical and organizational measures and integrate safeguards from the outset. So, while new PETs may introduce complex data protection risks, with minimizing the use of personal data they have a potential to reduce those risks.

4. Pilot projects

4.1. Privacy-preserving data collection: tracking staff and patients

Our first pilot, executed in a consortium including Philips and TNO, aims to apply MPC to put employees in charge of their location data in a hospital setting. Optimizing hospital workflows can help improve resource utilisation, reduce operational costs and, most importantly, improve quality of care. Traditionally, to find process improvements, consultants interview various stakeholders and patients, and spend days shadowing key staff members and patients in order to develop an accurate picture of how the hospital is functioning. However, individuals often tend to modify their natural behaviour the moment they are conscious about the fact that they are being observed (the Hawthorne effect). Also, as consultants are usually limited in numbers and do not stay at the hospital 24 hours a day, they are unable to get a global view of a department's operations.

One option to address this is to tag relevant staff members, patients and assets with a Real-Time Locating System (RTLS) which provides location information of all tagged entities every few seconds. However, while such systems can make a detailed and objective assessment of hospital operations, staff members are often reluctant to use RTLS as sharing their location data with hospital management is considered an invasion of privacy. Hence we propose for the hospital to only have access to location data of patients. Staff location data is made available only to staff members themselves, or to the worker's council that represents them. Using MPC, data analytics can be performed on

two (patient location data from the hospital and staff location data from the worker's council) or more data streams (patient location data from the hospital and staff location from each staff member). This way, staff members are put in control of their location data, individual data is never opened, and secondary use of their data is made impossible.

4.2. Intra-sector data analytics: analysing population health

The second pilot, part of the Horizon 2020 SODA project, is about performing joint analysis using data from different data providers in the same sector. The pilot focusses on prediction of population health, for instance to predict the risk of death within one year for chronic heart failure patients using logistic regression: such a model benefits from data from different regions or countries. Using MPC, the model (i.e., the logistic regression coefficients) can be built without hospitals needing to share any data about their patients. Other typical tasks include decision tree learning and computing statistics.

Compared to the single-organization case, this one is a lot more complex to deploy. A first challenge is to make organisations and data subjects willing to share their data. MPC should help with this by helping towards GDPR compliance and reducing the risk of unauthorized data re-use of the data. Other challenges include agreeing on common formats for data exchange and in interpreting concepts in a common way (in one case, conflicting definitions of "mortality rate" caused practical problems); and in formulating models, and dealing with outliers and missing data, without needing to inspect the data.

4.3. Inter-sector data analytics

The third pilot, by an academic hospital (Erasmus MC), a health-care insurance company (Achmea), and TNO in the Horizon 2020 BigMedilytics project, aims to shorten the lifecycle between research and clinical practice by combining data from multiple verticals, thus improving healthcare efficiency. In particular, the pilot combines population data, patient profiles, patient care, and financial claim data. One application is comorbidity-based risk stratification for heart failure patients. The expectation is that combining clinical data from the hospital with cardiovascular comorbidity information based on claims data from the insurance company leads to better treatment outcomes.

The main challenges for this pilot are, again, achieving GDPR compliance (toward which the use of MPC will contribute); but also ISO 27001 certification and linking the data from the various pockets, silos, and scientific software applications in the different institutions. In particular, for clinical data, this will be achieved through a link with the new electronic patient dossier, HiX, for which first steps have already been taken.

5. Conclusion, Challenges, and Outlook

As we aim to show in three pilots in the medical domain, secure multi-party computation enables data analytics on data from multiple providers, without requiring them to share data with anybody else. However, to make this vision a commonplace reality, several technical and non-technical challenges need to be addressed, which we now discuss.

A first challenge is integration with other systems. While building isolated MPC prototypes is possible, collecting data and inserting it into a MPC system today is largely a manual process. To simplify the use of MPC, it needs to be integrated with existing systems. This includes adding MPC to healthcare information standards like FHIR, and

working together with initiatives such as the Dutch Personal Health Train (PHT) initiative based on FAIR [7] data. A second step is to develop new techniques for checking consistency between datasets without manual inspection of the underlying data.

A second challenge is scalability. In theory, MPC can be applied to any kind of computation; in practice, its overhead compared to in-the-plain processing is large. For instance, performing regression on hundreds of millions of records using state-of-the-art techniques requires around 10 hours to complete. Solution directions include more scalable primitives, parallelisation, streaming, and mixing plain and hidden data.

A third challenge is GDPR compliance. The GDPR undoubtedly gives rise to several issues related to data processing with privacy-preserving techniques. Anonymisation may be a good strategy to benefit from big data, and to mitigate the data protection risks. It is highly probable that data which are de-identified with application of state-of-the-art PETs will fall outside the scope of the GDPR. However, large-scale processing of sensitive personal data is generally of greater risks, since sensitive personal data are subject to additional protections. Hence, a risk-based, multi-factor compliance analyses is advised in any case to ensure that data processing routines are compliant with the new requirements of the GDPR. Especially, because the legal evaluation of data processing activities will in any event be subject to a case-by-case assessment.

A final challenge is gaining trust from key stakeholders. Having technically sound solutions is just a first step towards convincing stakeholders (including patients, hospital staff, and information officers) to enable privacy-preserving data analytics on their data. Easy-to-understand ways of explaining cryptographic techniques in data analytics are needed. Moreover, rigorous qualitative and quantitative user studies are needed to better comprehend the concerns of stakeholders and their understanding of the issues involved.

Acknowledgements. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731583 and 780495 and is co-financed from a PPS allowance for research and innovation from the Dutch Ministry of Economic affairs.

References

[1] Article 29 Data Protection Working Party, Opinion 05/2014 on anonymisation techniques, WP 216, 2014.
[2] M. Barbaro, T. Zeller, S. Hansell, A face is exposed for AOL searcher no. 4417749, *New York Times*, 2006.
[3] B. Goldman, King of the mountain: Digging data for a healthier world, Stanford Medicine Summer, 2012.
[4] P. Groves *et al.*, The big-data revolution in us health care: Accelerating value and innovation, McKinsey Center for US Health System Reform, 2013.
[5] A. Jochems *et al.*, Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital: a real life proof of concept, *Radiotherapy and Oncology* **121**(3) (2016), 459–467.
[6] Y. Lindell, B. Pinkas, Privacy preserving data mining *Proceedings of CRYPTO*, 2000.
[7] M. Wilkinson et al., The fair guiding principles for scientific data management and stewardship, Scientific Data **3**:160018 (2016).
[8] D. Munro, Data breaches in healthcare totaled over 112 million records in 2015, Forbes, 2015.