Privacy-Preserving Data Integration for Health

Lisa Trigiante[1]

[1]First Year Ph.D. Student, ICT Doctorate at DBGroup, University of Modena and Reggio Emilia, Modena, Italy

Abstract

The digital transformation of health processes has resulted in the collection of vast amounts of health-related data that presents significant potential to support medical research projects and improve the healthcare system. Many of these possibilities arise as a consequence of integrating data from different sources to create an accurate and unified representation of the underlying data and enable detailed data analysis that is not possible through any individual source. Achieving this vision requires the collection and processing of sensitive health-related data about individuals, thus privacy and confidentiality implications have to be considered. In this paper, I describe my doctoral research topic: the design and development of a novel Privacy-Preserving Data Integration (PPDI) framework which aims to effectively address the challenges and opportunities of integrating Big Health Data (BHD) while ensuring compliance with the General Data Protection Regulation (GDPR). The paper describes the planned methodology for implementing the PPDI process through the usage of data pseudonymization techniques and Privacy-Preserving Record Linkage (PPRL) methods and provides an overview of the new framework, which is based on the re-implementation of MOMIS towards a microservices architecture with added PPDI functionalities.

Keywords

Privacy-Preserving, Data Integration, Healthcare, Pseudonymization, PPDI

1. Introduction and Preliminary Concepts

The digitization of health and administrative processes, among others, has led to vast amounts of data describing people and their behaviour being collected. These data are of great value to feed multiple research areas and enhance public and private sectors. In particular within the Healthcare[1] domain, the emergence of *Big Health Data* (BHD) in conjunction with advanced Data Analysis techniques, presents the opportunity to pave the way for a Data-Driven Healthcare System that supports the emerging P4M (Predictive, Preventive, Personalized, and Participatory Medicine) paradigm. To achieve this vision, an efficient *Data Integration* (DI) process is essential to create a unified and accurate view of BHD and enable in-depth analysis. A crucial step of DI is the *Record Linkage* (RL) [1, 2], which consists in identifying and linking records from various sources that belong to the same individual. However, the collection of sensitive personal medical data imposes the need to consider privacy legislation. Data protection in Europe is set off by the *General Data Protection Regulation* (GDPR), which is a comprehensive legal framework that

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy

lisa.trigiante@unimore.it (L. Trigiante)
0000-0002-2021-9259 (L. Trigiante)

[1] Healthcare serves as an application example; I referred to this domain because my PhD research involves a collaborative partnership with the Emilia Romagna Region's Department of Health.

sets guidelines for the collection and processing of personal information from individuals who live in the European Union. In a privacy-aware setting, the categorization of data that sources may contain is based on the twin principles of identifiability and privacy. *Personally Identifiable Information* (PII) refers to attributes that may be used alone or in combination to identify a specific individual e.g. national identification number or name, respectively called direct PII, and indirect PII or *Quasi-IDentifiers* (QID). *Sensitive Data* denotes attributes that contain confidential personal information that must be protected from privacy disclosure, e.g. medical history, diagnosis, and treatment outcomes. The process resulting from the addition of ethical privacy considerations is called *Privacy Preserving Data Integration* (PPDI) and aims to provide unified access to data residing in multiple autonomous data sources, using *pseudonymization* as an appropriate measure to implement data protection principles under the GDPR. The resulting process is called *Privacy Preserving Data Integration* (PPDI) and aims to provide unified access to data residing in multiple autonomous data sources, using *pseudonymization* as an appropriate measure to implement data protection principles under the GDPR. *Pseudonymization* refers to the process of replacing the identifiable information with a *pseudonym* or encrypted code, in such a manner that re-identification is not feasible, without the use of additional information. *Re-identification* refers to the process of identifying a specific data subject whose personal data has been pseudonymized.

This paper outlines the expected contributions of my PhD program, which will mainly focus on the design and development of a new framework to perform PPDI in compliance with GDPR. In particular, Section 2 describes different challenges and opportunities arising from the intersection of BHD Integration and Analysis, and privacy requirements. Section 3 presents the methodology designed to support the PPDI process and provides an overview of the novel framework, based on the re-implementation of *MOMIS* toward a microservices architecture with added PPDI functionalities. Finally, in Section 4, I conclude with some ideas on the possible future directions of my research.

2. Challenges and Opportunities

As Big Data Integration and Analysis meet healthcare applications in a privacy context, domain-specific challenges and opportunities materialize in various aspects of Data Science. One of these is the use of Data Mining and Artificial Intelligence (AI) techniques to advance medical knowledge discovery and clinical decision support. However, as per the Data-Centric principle, the quality and quantity of data used to train AI models are critical factors in determining their analysis capacity and accuracy, and raw health datasets often present intrinsic issues and sparse, scarce, and unbalanced nature. This imposes strict demands on the data resulting from the PPDI process in terms of volume, completeness, balance among target classes, consistency and regularity over time. Many solutions exist to handle various aspects of data quality within the Data Integration process [3], but the privacy requirements of PPDI pose additional challenges that require the adaptation of existing approaches or the development of new techniques.

To ensure privacy preservation, it is necessary to prevent an individual's sensitive personal information from being disclosed to internal parties involved in the process or to external adversaries and accordingly, prohibit plain data from leaving the local storage. To achieve this, encryption techniques must be used to create pseudonyms that allow the subsequent processing for efficient integration and accurate Record Linkage. In addition, while pseudonym re-identification must be prevented by adversaries, the ability to return results to a specific individual in situations where a preventable health-risk factor is discovered may be mandatory.

In real-world scenarios, with any information disclosure there is always some privacy loss, and with any pseudonymization technique there is always some information loss. The trade-off between privacy and usability of data is a complex issue that requires careful consideration of various factors, including the nature of the information involved, the performance of the integration process and the privacy risk. While the performance, especially in terms of scalability and linkage quality, can be assessed using measures also employed in a non-privacy-preserving setting, privacy assessment is the biggest impediment. Assessing privacy can be seen as the resistance to re-identification attacks and depends on aspects difficult to quantify, such as the different behaviors and background knowledge of the adversaries [4].

3. Privacy-Preserving Data Integration Framework Overview

The design of the PPDI process is the result of a Proof-of-Concept commissioned to the DBGroup by the Ministry of Justice. The implementation of the novel PPDI framework is the subject of an ongoing project with the region Emilia Romagna, which grants for my PhD project.

This Section presents an outline of the methodology designed to support the process without delving into the discussion of technical specifics. The idea behind the PPDI framework is an incremental extension of the *MOMIS (Mediator envirOnment for Multiple Information Sources)* [5] Data Integration system toward a microservice architecture, including specific software modules to realize PPDI in compliance with the GDPR and general IT security practices. As we can see in Figure 1, the framework will serve as Privacy-Preserving Integration Domain (or Trusted Third Party, TTP) to produce an integrated and consistent representation of different distributed and heterogeneous data sources (Source Domain) to provide researchers (Consumer Domain) with a unified view of the underlying data.

The PPDI process involves three steps: Schema Alignment, Privacy-Preserving Record Linkage (PPRL) and Data Fusion. The framework will provide one or more microservices for each step of the process, designed to fulfill specific functionalities and enable secure information exchange among internal parties via an encrypted communication channel. Moreover, the framework will incorporate a microservice that enables the possibility of the required re-identification of patients, through the management of additional information that must be stored separately to meet the requirements of GDPR.

3.1. Schema Alignment

Schema Alignment resolves inconsistencies at schema level by finding the semantic correspondences among the local schema and producing an integrated Global Schema. MOMIS Schema Integration [6] is performed in a semi-automatic way, exploiting the semantic relationships existing in the local data sources, the knowledge in thesauri and description logics. To address the privacy requirements MOMIS will be extended with health-domain specific thesauri and specialized rules. These enhancements will enable semantic annotations and the classification

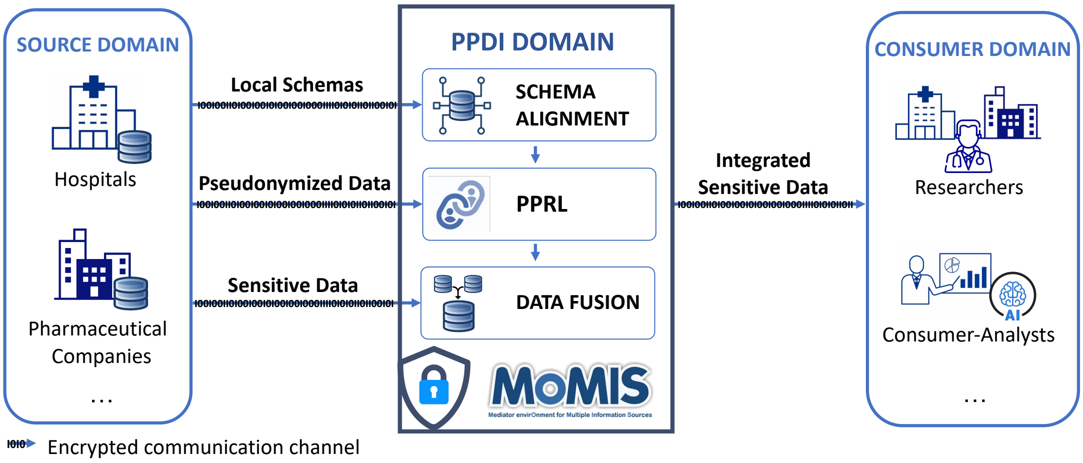

Figure 1: Schema of the PPDI Framework.

of schema elements based on identifiability (QID) and privacy (Sensitive Data). Classifying data in real-world scenarios is one of the challenges deriving from the trade-off between privacy and usability because different kinds of data can overlap and the potential privacy risks must be carefully considered. Biomedical research projects, for instance, require data on person-specific genomic sequences, but a subject's DNA is also a unique identifier and can be used for re-identification. Another challenge is handling the impact of data instability caused by the constant evolution of medical processes and data collection methods, which will be addressed by designing the Global Schema to be modifiable and expandable over time.

3.2. Privacy-Preserving Record Linkage (PPRL)

Privacy-Preserving Record Linkage (PPRL) resolves inconsistencies at tuple level by identifying and linking records about the same real-world entities from different sources while avoiding any privacy disclosure. Linkage of data about individuals is commonly based on QID since direct PII is more vulnerable to reidentification attacks. However, QID is neither unique nor stable over time and may be subject to recording errors and missing values. The process to achieve high linkage quality results in a privacy-preserving setting comprises different steps:

1. **Pre-processing** of raw data with attribute-specific functions to transform error-prone QID into a unique and comparable format, and sensitive data into a suitable format for the analysis.

2. **Pseudonymization** of QID into a unique pseudonym for each record, in such a manner that the pseudonyms will enable record linkage while preventing re-identification.

In order to avoid the transmission of non-pseudonymized data from local storage, the aforementioned steps are performed at the source site. Consequently, the actual linkage is established based on pseudonyms (or encrypted QID). The usage of pseudonyms in PPRL poses a significant challenge to the trade-off between privacy and performance, as the linkage of pseudonym pairs must be carried out with minimal bias in comparison to plain-text pairs while ensuring privacy. This necessitates the implementation of a microservice to standardize the pseudonym generation and similarity score estimation. To achieve this objective, the semantic knowledge and mapping rules extracted from Schema Alignment will be leveraged to select specific functions for each application context and provide them with the corresponding application rules to the local sources. Subsequently, the local sources transmit for each record the pair <pseudonym, record-ID> to the TTP, which performs the next steps in line with the Record Linkage process employed in non-privacy contexts:

3. **Blocking** of the pseudonyms that are likely to match into blocks, producing candidate pseudonym pairs. The blocking technique is crucial to face scalability issues, as it reduces the number of comparisons that need to be conducted.

4. **Comparison** of candidate pseudonym pairs in detail using approximate comparison (or similarity) functions.

5. **Classification** of candidate pseudonym pairs into a match or not match, using a decision model based on the result of the comparison.

The result of the process is a group of record-ID, in which each group represents a real-world entity. A lot of techniques have been studied in the literature to cover the various steps of the PPRL process [7]. The challenge lies in the choice of the best techniques because each step is dependent on and connected to the others and must take into account different aspects, such as the nature of data, the computational requirements and the performance and protection achieved. For example, a privacy technique for the encryption of a subject's DNA sequence that provides good protection and utility is Fully Homomorphic Encryption (FHE) [8]. However, FHE has poor performance and a massive overhead in computational and memory costs.

3.3. Data Fusion

Data Fusion resolves inconsistencies at value level by fusing the data of duplicate entries and creating a unique record for each distinct real-world entity [9]. In this step, the local sources transmit the pair <Sensitive Data, record-ID> for each record to the TTP. Afterward, the TTP performs the fusion of Sensitive Data belonging to the same real-world entity, using the previously computed record-ID groups. The objective of this process is to increase data conciseness and consistency by providing a unified and searchable privacy-preserving representation of the underlying data to the Consumer Domain. With regard to privacy requirements, the issue is that datasets containing more information on the same entity are more exposed to re-identification attacks. This problem is the central topic of statistical disclosure control [10], which aims to reduce the risk of information disclosure by restricting or modifying the amount of data released. *K-anonymity* is a widely used generalization technique that avoids the possibility of re-identification by ensuring that each record in the perturbed dataset is indistinguishable from at least $k-1$ other records in the same dataset. However, generalization techniques reduce the usability of the data and eliminate hidden patterns that are important for subsequent analyses. To address this limitation, a proposal is made for the exploration and application of data augmentation and imputation techniques [11] to satisfy k-anonymity and increase the fairness of the data for underrepresented subgroups in specific research studies.

4. Future Directions and Developments

In conclusion, the early stages of my PhD research encompassed the study of the research field of Privacy-Preserving Data Integration (PPDI) for Healthcare, along with the design concept for a novel PPDI framework and the envisaged methodology to address a range of related issues. Although much work remains to be done to fully develop the proposed framework, the open challenges in this field offer a vast panorama of potential future research directions. These may include, but are not limited to, the investigation of new *Privacy-Preserving Temporal Record Linkage (PPTRL)* [12] techniques that incorporate the temporal information available in dynamic datasets, as well as the optimization of PPRL by adapting advanced Blocking techniques designed by the DBGroup [13, 14] to the privacy context. Pursuing this line of thought, it is easy to notice that the PPDI process entails numerous tasks, for which different solutions exist in the literature and many others are awaiting exploration. A significant future enhancement of the proposed framework is therefore the creation and incorporation of microservices capable of automatically selecting the most suitable methods based on the specific application scenarios. However, the lack of metrics to accurately quantify the trade-off between privacy and utility presents a significant challenge. To this end, a future direction of my research is the study of specific privacy-loss metrics and approaches that hopefully will add new research contributions not only to PPDI but to the broader field of Data Management and Analysis. Expanding the view to the whole architecture, one promising opportunity for advancement involves leveraging and adapting Virtualized Data Integration facilitated by MOMIS to mitigate privacy risk and avoid moving data from local sources. This approach can serve as a starting point for future exploration into the area of Federated Learning [15].

Acknowledgments

I wish to thank my tutor, Prof. Domenico Beneventano, my co-tutor, Prof. Sonia Bergamaschi, and all the members of DBGroup. I would also like to acknowledge that this work is funded by the Emilia Romagna Region through the MIUR DM351 for doctoral research.

References

[1] G. Simonini, L. Gagliardelli, S. Bergamaschi, H. V. Jagadish, Scaling entity resolution: A loosely schema-aware approach, Inf. Syst. 83 (2019) 145–165.

[2] G. Simonini, L. Zecchini, S. Bergamaschi, F. Naumann, Entity resolution on-demand, Proc. VLDB Endow. 15 (2022) 1506–1518. URL: https://www.vldb.org/pvldb/vol15/p1506-simonini.pdf.

[3] S. Bergamaschi, D. Beneventano, F. Mandreoli, R. Martoglia, F. Guerra, M. Orsini, L. Po, M. Vincini, G. Simonini, S. Zhu, L. Gagliardelli, L. Magnotta, From data inte-

gration to big data integration, volume 31 of *Studies in Big Data*, Springer International Publishing, 2018, pp. 43–59. URL: https://doi.org/10.1007/978-3-319-61893-7_3. doi:10.1007/978-3-319-61893-7_3.

[4] A. Vidanage, T. Ranbaduge, P. Christen, R. Schnell, Taxonomy of attacks on privacy-preserving record linkage, J. Priv. Confidentiality 12 (2022). URL: https://doi.org/10.29012/jpc.764. doi:10.29012/jpc.764.

[5] S. Bergamaschi, D. Beneventano, F. Guerra, M. Orsini, Data integration, in: D. W. Embley, B. Thalheim (Eds.), Handbook of Conceptual Modeling - Theory, Practice, and Research Challenges, Springer, 2011, pp. 441–476. URL: https://doi.org/10.1007/978-3-642-15865-0_14. doi:10.1007/978-3-642-15865-0_14.

[6] M. Vincini, D. Beneventano, S. Bergamaschi, Semantic integration of heterogeneous data sources in the MOMIS data transformation system, J. Univers. Comput. Sci. 19 (2013) 1986–2012. URL: https://doi.org/10.3217/jucs-019-13-1986. doi:10.3217/jucs-019-13-1986.

[7] D. Vatsalan, P. Christen, V. S. Verykios, A taxonomy of privacy-preserving record linkage techniques, Inf. Syst. 38 (2013) 946–969. URL: https://doi.org/10.1016/j.is.2012.11.005. doi:10.1016/j.is.2012.11.005.

[8] M. Kantarcioglu, W. Jiang, Y. Liu, B. A. Malin, A cryptographic approach to securely share and query genomic sequences, IEEE Trans. Inf. Technol. Biomed. 12 (2008) 606–617.

[9] D. Beneventano, S. Bergamaschi, L. Gagliardelli, G. Simonini, Entity resolution and data fusion: An integrated approach, in: M. Mecella, G. Amato, C. Gennaro (Eds.), SEBD 2019, volume 2400 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2400/paper-17.pdf.

[10] M. Comerford, Statistical disclosure control : an interdisciplinary approach to the problem of balancing privacy risk and data utility, Ph.D. thesis, University of Glasgow, UK, 2014.

[11] F. M. Martins, V. M. G. Suárez, J. R. V. Flecha, B. García-López, Data augmentation effects on highly imbalanced EEG datasets for automatic detection of photoparoxysmal responses, Sensors 23 (2023) 2312. URL: https://doi.org/10.3390/s23042312. doi:10.3390/s23042312.

[12] T. Ranbaduge, P. Christen, Privacy-preserving temporal record linkage, in: IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018, IEEE Computer Society, 2018, pp. 377–386. URL: https://doi.org/10.1109/ICDM.2018.00053. doi:10.1109/ICDM.2018.00053.

[13] D. Beneventano, S. Bergamaschi, L. Gagliardelli, G. Simonini, *BLAST2*: An efficient technique for loose schema information extraction from heterogeneous big data sources, ACM J. Data Inf. Qual. 12 (2020) 18:1–18:22. URL: https://doi.org/10.1145/3394957. doi:10.1145/3394957.

[14] G. Simonini, L. Gagliardelli, M. Rinaldi, L. Zecchini, G. D. Sabbata, A. Aslam, D. Beneventano, S. Bergamaschi, Progressive entity resolution with node embeddings, in: SEBD 2022, volume 3194 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 52–60. URL: https://ceur-ws.org/Vol-3194/paper6.pdf.

[15] M. Ali, F. Naeem, M. Tariq, G. Kaddoum, Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey, IEEE J. Biomed. Health Informatics 27 (2023) 778–789. URL: https://doi.org/10.1109/JBHI.2022.3181823. doi:10.1109/JBHI.2022.3181823.