Innovative Paradigms for Supporting Privacy-Preserving Multidimensional Big Healthcare Data Management and Analytics: The Case of the EU H2020 QUALITOP Research Project

Alfredo Cuzzocrea[1,2][0000−0002−7104−6415]⋆

[1] iDEA Lab, University of Calabria, Rende, Italy
[2] LORIA, Nancy, France
alfredo.cuzzocrea@unical.it

Abstract. QUALITOP is an authoritative EU H2020 research project whose main goal consists in supporting big data management, analytics and predictive analytics over cancer patients treated by the innovative immunotherapy therapeutic approach, in order to study and support decision making processes about their *Quality of Life* (QoL). Within the QUALITOP big data lake, a critical requirement consists in supporting *privacy-preserving multidimensional big healthcare data management and analytics*, which is addressed via innovative paradigms. This paper presents some relevant innovations developed in the context of the latter research area, as contextualized to the QUALITOP project, with also an overview of possible research challenges for future efforts in the investigated field.

Keywords: Big Data Management · Big Data Analytics · Privacy-Preserving Big Data Management · Privacy-Preserving Big Data Analytics · Privacy-Preserving Multidimensional Big Data Management and Analytics.

1 Introduction

Nowadays, *big healthcare data management and analytics* [22, 3] are playing the leading role within the broad context of big data applications and systems (e.g., [5, 21, 4, 2]). A clear example is represented by the modern COVID-19 outbreak that is spreading world-wide at an impressive rate. Here, healthcare analytics is critical for taking medical decision making at Country- and Regional-level,

⋆ This research has been made in the context of the Excellence Chair in Computer Engineering — Big Data Management and Analytics at LORIA, Nancy, France

thus determining the desired healthcare policy. *Monitoring multidimensional aspects of* **QUA***lity of* **L***ife after cancer* **I***mmuno***T***herapy - an* **O***pen smart digital* **P***latform for personalized prevention and patient management* (QUALITOP)[1] is an authoritative EU H2020 research project in this context. Within the QUALITOP project, a critical challenge is represented by the issue of *providing support for privacy-preserving multidimensional big healthcare data management and analytics*, where the main emphasis is on the two following keywords: *multidimensionality* [16] and *privacy preservation* [7]. In this paper, we present some relevant innovations developed in the context of the latter research area, as contextualized to the QUALITOP project, with also an overview of possible research challenges for future efforts in the investigated field.

2 The EU H2020 QUALITOP Research Project

QUALITOP aims at developing a European immunotherapy-specific open Smart Digital Platform and using big data analysis, artificial intelligence, and simulation modelling approaches. This approach enables collecting and aggregating efficiently real-world data to monitor health status and *Quality of Life* (QoL) of cancer patients given immunotherapy. Through causal inference analyses, QUALITOP identifies the determinants of health status regarding *Immunotherapy-Related Adverse Events* (IR-AEs) and defines patient profiles in a real-world context. For this, heterogeneous data sources (big data), both retrospective and prospective –collected for QUALITOP from clinical centres in four EU countries— integrate lifestyle, genetic, and psycho-social determinants of QoL. Using machine learning approaches, QUALITOP provides "real-time" recommendations stemming from patient profiles and feedbacks via the Smart Digital Platform. Furthermore, an increased visibility on patients' behaviour, a better IR-AEs prediction, and an improvement of care coordination help analysing through simulation modelling approaches, can be gained in cost-effectiveness. Guidelines are issued over the short and long-term.

The QUALITOP consortium is the following:

1. Hospices Civils de Lyon (HCL), France;
2. Université "Claude Bernard" Lyon 1 (UCBL), France;
3. Université "Lumière" Lyon 2 (Lyon2), France;
4. Centre Hospitalier Universitaire de Besançon (CHUB), France;
5. Instituto Português de Oncologia, Lisboa (IPOL), Portugal;
6. Consorci Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), Spain;
7. Academic Medical Center, University of Amsterdam (AMC), The Netherlands;
8. London School of Hygiene and Tropical Medicine (LSTHM), United Kingdom;
9. Scientific Academy for Service Technology (ServTech), Germany;
10. European Cancer Patient Coalition (ECPC), Belgium;

11. University Medical Center Groningen (UMCG), The Netherlands;
12. Massachusetts General Hospital (MGH), United States;
13. University of Calabria (UNICAL), Italy.

3 Privacy-Preserving Multidimensional Big Data Management and Analytics within the QUALITOP Big Data Lake

One essential component of the QUALITOP platform is represented by its big data lake. The QUALITOP big data lake is populated by big healthcare data coming from the reference data sources, and it supports both big data management and analytics procedures over them.

One critical aspect of the QUALITOP big data lake consists in ensuring the privacy of big healthcare data during the management and analytics tasks, as highlighted by several studies (e.g., [6, 17]). Consider Figure 1. Here, the QUALITOP big data lake is shown, with the detail on the *Privacy-Preserving Data Publishing* (PPDP) module.

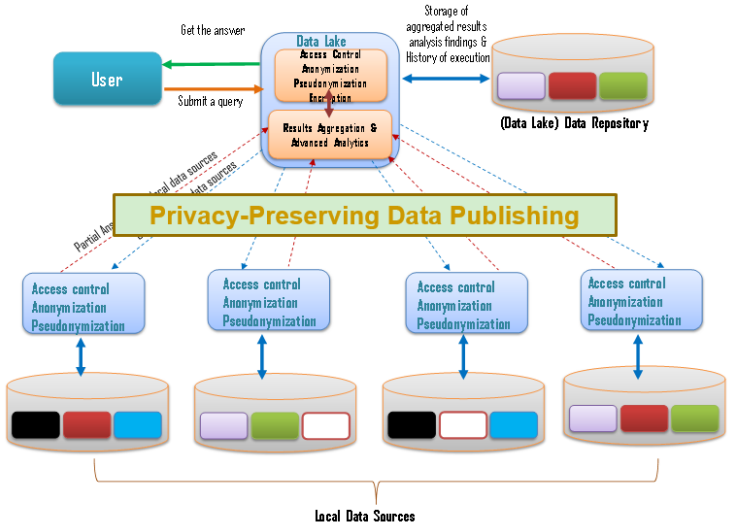

Fig. 1. The QUALITOP Big Data Lake

Within the PPDP module, several techniques for ensuring the privacy of big healthcare data are implemented. Among all alternatives, the choice is set on the so-called *privacy-preserving multidimensional big data management and analytics techniques* (e.g., [9, 8, 15]). Basically, these techniques extend the well-understood *privacy-preserving OLAP* paradigm (e.g., [14, 13, 11, 10]) to the innovative context of big data management and analytics.

Privacy-preserving multidimensional big data management and analytics predicates a collection of models, techniques and algorithms for making privacy-preserving data cubes on top of which management and analytics tasks are executed. This can be obtained, for instance, according to different alternatives:

1. SPPOLAP algorithm [14, 13], which introduces the two following main innovations: (*i*) a *novel privacy OLAP notion*; (*ii*) flexible adoption of *sampling-based techniques* in order to achieve the final privacy-preserving data cube;
2. SDO algorithm [11, 10], which introduces the novel concept of the so-called *secure distributed OLAP aggregation task* – basically, this task purses the idea of performing OLAP across multiple distributed SUM-based two-dimensional OLAP views extracted from data cubes under the *Secure Multi-party Computation* (SMC) requirements, by relaying on powerful *CUR-based matrix decomposition methods* used as a fundamental privacy-preserving tool [12].

Thanks to these fundamental algorithms, effective and efficient privacy-preserving multidimensional big data management and analytics can be obtained within the QUALITOP big data lake, obviously by taking into consideration the special features of big data processing (e.g., [25]), being *high-performance computing* and *fast data availability* two of the main research challenges among them.

4 Privacy-Preserving Big Data Management and Analytics: Future Research Challenges

With the impressive growth of big data and Cloud-based technologies, the need for privacy-preserving methodologies will be more and more relevant in both academic and industrial research, specially driven by modern ICT challenges of next-generation societies (e.g., social media, health intelligence, smart cities, and so forth). Among these challenges, we retain the following as the most relevant ones:

1. ***Integration with Cryptographic Techniques*** One of the most important challenge today is represented by integration of privacy-preserving approaches with traditional *cryptographic techniques*, as emerging in the last research trends (e.g., [23, 18]).
2. ***Balancing Accuracy and Privacy*** When dealing with analytics tools over big data, accuracy must be *balanced* with privacy (e.g., [13, 19]). Indeed, accuracy and privacy are two contrasting properties: when accuracy increases then privacy decreases, and viceversa. Therefore, in many contexts (e.g., healthcare analytics), intelligent tools must balance accuracy and privacy, in order to obtain good prediction performance at the cost of a safe privacy of data.

3. ***Inference Control Techniques*** *Inference methods* (e.g., [24, 20]) can still violate privacy of data when all (privacy-preserving) countermeasures are taken. For instance, *query inference* is a popular technique that allows us to derive (unknown) answers (or, approximate answers) to un-granted queries from (known) answers to granted queries. Next-generation privacy-preserving techniques must deal with this emerging issue, thus devising innovative *inference-control privacy preserving big data management and analytics approaches.*

5 Conclusions and Future Work

Starting from the context of the EU H2020 QUALITOP research project, this paper has presented some of the privacy-preserving multidimensional big data analytics techniques implemented within the big data lake of the reference platform. In addition to this, an overview of possible research challenges for future efforts in the investigated field has been proposed.

Future work is mostly oriented towards achieving the integration of privacy-preserving multidimensional big data analytics techniques with emerging machine-learning-based advanced analytical tools, such as *tensors*.

Acknowledgements

This research has been supported by the EU H2020 QUALITOP research project, H2020 Project ID: 875171.

References

1. QUALITOP research project. https://h2020qualitop.liris.cnrs.fr/wordpress/
2. Ahn, S., Couture, S.V., Cuzzocrea, A., Dam, K., Grasso, G.M., Leung, C.K., McCormick, K.L., Wodi, B.H.: A fuzzy logic based machine learning tool for supporting big data business analytics in complex artificial intelligence environments. In: 2019 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2019. pp. 1–6. IEEE (2019)
3. Batarseh, F.A., Ghassib, I., Chong, D.S., Su, P.: Preventive healthcare policies in the US: solutions for disease management using big data analytics. J. Big Data **7**(1), 38 (2020)
4. Bellatreche, L., Cuzzocrea, A., Benkrid, S.: *F&A*: A methodology for effectively and efficiently designing parallel relational data warehouses on heterogenous database clusters. In: Data Warehousing and Knowledge Discovery, 12th International Conference, DAWAK 2010. Lecture Notes in Computer Science, vol. 6263, pp. 89–104. Springer (2010)
5. Ceci, M., Cuzzocrea, A., Malerba, D.: Effectively and efficiently supporting roll-up and drill-down OLAP operations over continuous dimensions via hierarchical clustering. J. Intell. Inf. Syst. **44**(3), 309–333 (2015)

6. Chauhan, R., Kaur, H., Chang, V.: An optimized integrated framework of big data analytics managing security and privacy in healthcare data. Wirel. Pers. Commun. **117**(1), 87–108 (2021)
7. Cuzzocrea, A.: Privacy and security of big data: Current challenges and future research perspectives. In: Proceedings of the First International Workshop on Privacy and Secuirty of Big Data, PSBD@CIKM 2014. pp. 45–47. ACM (2014)
8. Cuzzocrea, A.: Scalable privacy-preserving big data management and analytics: Where we are and where we are going. In: Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing, ICCBDC 2018. pp. 52–56. ACM (2018)
9. Cuzzocrea, A.: Advanced, privacy-preserving and approximate big data management and analytics in distributed environments: What is now and what is next. In: 44th IEEE Annual Computers, Software, and Applications Conference, COMPSAC 2020. pp. 1564–1568. IEEE (2020)
10. Cuzzocrea, A., Bertino, E.: A secure multiparty computation privacy preserving OLAP framework over distributed XML data. In: Proceedings of the 2010 ACM Symposium on Applied Computing (SAC). pp. 1666–1673. ACM (2010)
11. Cuzzocrea, A., Bertino, E.: Privacy preserving OLAP over distributed XML data: A theoretically-sound secure-multiparty-computation approach. J. Comput. Syst. Sci. **77**(6), 965–987 (2011)
12. Cuzzocrea, A., Bertino, E.: Further theoretical contributions to a privacy preserving distributed OLAP framework. In: 37th Annual IEEE Computer Software and Applications Conference, COMPSAC 2013. pp. 234–239. IEEE Computer Society (2013)
13. Cuzzocrea, A., Saccà, D.: Balancing accuracy and privacy of OLAP aggregations on data cubes. In: DOLAP 2010, ACM 13th International Workshop on Data Warehousing and OLAP. pp. 93–98. ACM (2010)
14. Cuzzocrea, A., Saccà, D.: A theoretically-sound accuracy/privacy-constrained framework for computing privacy preserving data cubes in OLAP environments. In: On the Move to Meaningful Internet Systems: OTM 2012, Confederated International Conferences: CoopIS, DOA-SVI, and ODBASE 2012. Lecture Notes in Computer Science, vol. 7566, pp. 527–548. Springer (2012)
15. Cuzzocrea, A., Schuster, A., Vercelli, G.V.: PP-OMDS: an effective and efficient framework for supporting privacy-preserving olap-based monitoring of data streams. In: Proceedings of the 20th International Conference on Enterprise Information Systems, ICEIS 2018. pp. 282–292. SciTePress (2018)
16. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over large-scale multidimensional data: the big data revolution! In: DOLAP 2011, ACM 14th International Workshop on Data Warehousing and OLAP. pp. 101–104. ACM (2011)
17. Devi, R.R., Chamundeeswari, V.V.: Triple DES: privacy preserving in big data healthcare. Int. J. Parallel Program. **48**(3), 515–533 (2020)
18. Hamid, H.A.A., Rahman, S.M.M., Hossain, M.S., Almogren, A., Alamri, A.: A security model for preserving the privacy of medical big data in a healthcare cloud using a fog computing facility with pairing-based cryptography. IEEE Access **5**, 22313–22328 (2017)
19. Liu, Y., Sung, S.Y., Xiong, H.: A cubic-wise balance approach for privacy preservation in data cubes. Inf. Sci. **176**(9), 1215–1240 (2006)
20. Lu, H., Li, Y.: Practical inference control for data cubes. IEEE Trans. Dependable Secur. Comput. **5**(2), 87–98 (2008)

21. Morris, K.J., Egan, S.D., Linsangan, J.L., Leung, C.K., Cuzzocrea, A., Hoi, C.S.H.: Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: A machine learning approach for predictive analytics on big stock data. In: 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018. pp. 1486–1491. IEEE (2018)
22. Nazir, S., Khan, S., Khan, H.U., Ali, S., García-Magariño, I., Atan, R.B., Nawaz, M.: A comprehensive analysis of healthcare big data management, analytics and scientific programming. IEEE Access **8**, 95714–95733 (2020)
23. Ouazzani, Z.E., Bakkali, H.E.: A classification of non-cryptographic anonymization techniques ensuring privacy in big data. Int. J. Commun. Networks Inf. Secur. **12**(1) (2020)
24. Reza, K.J., Islam, M.Z., Estivill-Castro, V.: Privacy protection of online social network users, against attribute inference attacks, through the use of a set of exhaustive rules. Neural Comput. Appl. **33**(19), 12397–12427 (2021)
25. Russom, P., et al.: Big data analytics. TDWI best practices report, fourth quarter **19**(4), 1–34 (2011)