Normative and Empirical Evaluation of Privacy Utility Trade-off in Healthcare

Syeda Amna Sohail[0000−0001−8078−0411]

Department of Data Management and Biometrics (DMB), Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, The Netherlands
s.a.sohail@utwente.nl
http://www.utwente.nl

Abstract. Post-GDPR, the public/private (healthcare) enterprises, while performing (sensitive) Big Data Analytics (BDA), encounter the dilemma of abiding by the privacy regulations on one hand and extracting maximum value from (healthcare) metadata on the other. Concerning this, one of the major issues is the Privacy Utility trade-off (PUT). The PUT affects each phase including (healthcare) metadata collection, formulation, storage, and resharing amongst (healthcare) enterprises. So far in healthcare, PUT concerning issues are identified and resolved in a remote, disintegrated manner. It's high time to resolve the issue by taking a holistic approach. This Ph.D. research work strives to achieve the same with normative (should be) and empirical (as-is) evaluation of PUT in Dutch care metadata share landscape. For clarity, the problem area is segregated into four fundamental dimensions. For each dimension, empirical evaluation is performed using Process Mining (discovery/conformance checking) techniques on real-world healthcare event-log(s). Based on data analytics, the conceptual modeling frameworks are formulated using e^3 value modeling or/and REA ontologies. For normative evaluation, two alternative approaches; the 'Content Analysis', to formulate the conceptual modeling framework(s) and 'BPMN text extraction', for documents 'Rule Mining' for drawing the respective business model(s), are used. Later, the (in-field) IT expert(s) further evaluates the proposed conceptual model(s). The aim is to evaluate the technical (IS-based privacy-preserving tools and techniques) and respective organizational (access governance, data ownership) measures of Dutch healthcare providers. The research work will (ultimately) contribute standardized conceptual modeling framework(s) with technical and respective organizational measures to efficiently cope with the PUT in handling sensitive (healthcare) metadata.

Keywords: privacy utility tradeoff · conceptual modeling · process mining · healthcare.

1 Introduction

Information Technology (IT) is quintessential in how contemporary research and industrial undertakings proceed and execute [16]. IT (computers, software apps, and telecommunication) together with the business process modeling (and evaluation) created Information Technology Engineering (ITE) [10]. ITE is an amalgamation of business processes, techniques, and systems that improve business proceedings in better achieving business goals. Because of ITE, public/private enterprises formulate, store and share valuable metadata (i.e. data/information about other (big) data) for Big Data Analytics (BDA). Big Data Analytics (BDA) is (meta) data evaluation for valuable information gain. The BDA of metadata facilitates the extraction of insights and actionable decisions in achieving business goals in a cost and time-efficient manner [11]. In healthcare, the BDA improves the care business models, foresees the long/short term treatment outcomes, and does patient and disease centric stratification [11]. Post Covid-19, the BDA is essential in aiding infectious control measures and respective policies. However, the use of BDA predominantly relies upon metadata sharing and posits some serious ethical concerns including privacy preservation of patient's personally identifiable information [5, 6, 8] within and across healthcare sub-domains.

Privacy comprises the autonomous decision-making and direct/indirect control over personal information [13]. The privacy concerns do not imply the lack of trust in BDA, rather it demands responsible and fair metadata sharing [4, 6]. Unfortunately, the pace of privacy-preserving tools' (and techniques) formulation (and implementation) lag far behind in comparison to the use of BDA across domains especially in healthcare [8, 13]. Caregivers, for patients' effective/efficient clinical care, are bound to share the un-anonymized/pseudonymized patients' metadata with other internally and externally located counterparts such as labs and pharmacies [23]. Simultaneously, the care providers are obliged to fulfill the privacy legislature/regulations [2, 3] to avoid paying hundreds of thousands of Euros as compensation [23]. For example, Haga hospital and Menzis insurance company had to compensate for privacy lapses with hefty amounts [23]. **Privacy-Utility-Tradeoff (PUT)** is the performance impairment of data analytics in ascertaining data privacy [21]. The issue demands quick, comprehensive, and efficient technical/organizational solutions, to identify and evaluate the data utility-prone privacy-preserving measures for the care provider's Information Systems (ISs). Such measures will facilitate the healthcare providers to avoid paying hefty compensations and in turn infamous reputation.

Section 1 (introduction) comprises introduction of the problem domain and two subsections namely the related work and research objectives and research questions where we highlight the current state of the art and this research work's objectives and questions, Section 2 comprises the standardized data analytics approach and research methodology for all four dimensions. In Section 3, we present the current results employing the **two currently published papers and an under-review paper** in the first year of Ph.D. research work. Section 4 highlights the threats to the validity of this research work, Section 5 gives a detailed description of the dimension-wise contribution and the uniqueness of this

research work. Section 6 includes a conclusion, acknowledgments and is followed by references.

1.1 Related Work

In the contemporary world of the information economy, the BDA is essential for (private/public) enterprises to shape their business goals and information system engineering with privacy by design measures [6, 25]. To ensure the responsible Business Information System Engineering (BISE) the main concern, amongst others, is the Privacy Utility Tradeoff (PUT) [6, 22]. In healthcare, the PUT raises graver repercussions because of the involvement of (highly) sensitive personally identifiable data on one hand and the efficiency and effectiveness of healthcare performance on the other [23]. In this regard, research studies (in healthcare) remotely focus on privacy and IoT [12,16], privacy in AI and machine learning [13,19], privacy in Process Mining (focussing on third party process analytics) [17, 21, 27] by not paying much attention to the overall context of the issue. Similarly, the data pipeline (from data collection to valuable insights) and provenance records (pipeline description) is often ignored while sharing of the data is emphasized concerning PUT [20].

It is important here to realize that PUT concerning issues of (healthcare) BDA is both technical and organization-based. Thus, the issue requires an evaluation of (care provider's) integrated techniques, processes, and systems (i.e. ITE) in (care) metadata share landscape. Privacy by design only provides for the technical solutions of real-world problems [25] but the people performing the tasks behind their computer screens are the quintessential source of issue resolving. People who are interwoven in the organization's architectural setup and are accountable to the organization. Besides, to identify, 'what (i.e. techniques) is happening in the BISE', it is integral to identify and evaluate that how (i.e. business processes) it is happening in that fashion? Process Mining with the event logs gives us a glimpse of the same using the datasets extracted directly from an organization's IS. Moreover, to check the business system's and processes' compliance on both technical and organizational grounds, the normative (should be) evaluation is done. Normative evaluation is done using two alternative methodologies namely, 'Content Analysis' and organization's (official) documents' rule mining using 'BPMN text extraction'. For simplification, the findings are represented with conceptual modeling frameworks using REA and e^3 value modeling ontologies. The ontologies are further evaluated by (in-field: working in the same domain) IT experts.

1.2 Research Objectives and Research Questions

The **method** of the research work aims to answer dichotomous knowledge questions (Design Science Methodology [28]), that include *Analytical* research work, using Process Mining tools/algorithms on real-world healthcare event log(s) (i.e. data set(s) that are extracted directly from healthcare provider's Information Systems) for the empirical (as-is state of affairs) evaluation of PUT in Dutch care

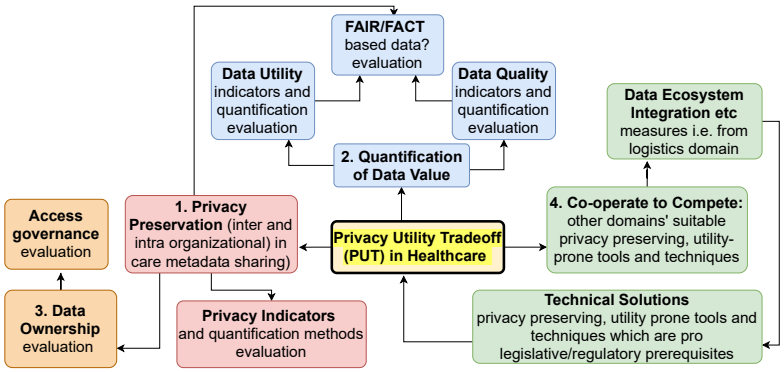

Fig. 1: Four fundamental dimensions to evaluate the Privacy Utility Trade-off (PUT) in (care) metadata share landscape.

metadata share landscape. And *Exploratory* research work, using either 'Content Analysis' or the 'BPMN text extraction' for normative (should-be state of affairs) evaluation. The **goal** is the normative and empirical evaluation of PUT concerning tools/techniques in the care provider's technical and organizational metadata sharing set-up.

For clarity the research goal is subdivided into *Four dimensions (see Fig.1) based on* ***four sub-research objectives***: *Objective 1*: Evaluate privacy measures in care metadata share landscape within Dutch care providers' inter/intra organizational setup. *Objective 2*: Identify the data utility and data quality indicators in healthcare and assess whether they are pro/against privacy indicators (PUT) on one hand and FAIR and FACT-based data indicators on the other. *Objective 3*: Evaluate both normatively and empirically the data ownership and access governance on both technical and organizational grounds in the care metadata share landscape. *Objective 4*: identify privacy-preserving, higher data utility prone measures from other domains (i.e. logistics, etc) that are effectively applicable to healthcare and are per regulatory and legislative requirements of the EU.

The respective **four *dimensions* (each with an assigned color see Fig.1) and their Research Questions (RQs)** are as follow: *Dimension 1*: Privacy evaluation in care metadata share landscape at the backdrop of care providers inter and intra-organizational setup. *RQ1*: What are the privacy-preserving indicators in the care metadata share the landscape, how are they implemented and assessed in Dutch inter and intra-organizational setup? *Dimension 2*: Quantification of the data value in healthcare and its relevance to privacy (PUT) and FAIR and FACT-based data. *RQ2.1*: What are the respective indicators for data utility and data quality while sharing healthcare metadata in inter/intra organizational setup. *RQ2.2*: How respective indicators support/ discourage the privacy indicators on one hand and FAIR (uninterrupted data) and FACT (responsible data) based data indicators on the other? *Dimen-*

sion 3: Privacy evaluation of data ownership in healthcare metadata within and amongst care providers. *RQ3*: What is the normative (should be) and empirical (as is) state of affairs of data ownership and access governance in care metadata share at inter/intra organizational levels? *Dimension 4*: Identify privacy-preserving, higher data utility prone measures from other domains i.e. logistics, etc, that are effectively applicable to healthcare and are per regulatory and legislative requirements of the EU. *RQ4.1*: What are useful privacy-preserving tools/techniques in safeguarding an organization's integrity in addition to the simultaneous sharing of valued metadata with other counterparts for the collective benefit? *RQ4.2*: How are those privacy-preserving, data utility-prone measures applicable to healthcare, and are they effective? Give normative and empirical evaluation.

2 Data Analytics Approach and Research Methodology

For each dimension (see Fig. 1), the following standardized data analytics approach (and methodology) is applied (see Fig. 2 for an overview). The approach combines the empirical and normative evaluation and aims to validate care providers' *integrated techniques, processes, and systems* concerning PUT in the Dutch (care) metadata share landscape. Empirical evaluation is done using Process Mining (PM) discovery and conformance checking techniques (to validate the empirical evaluation) on healthcare event logs i.e. datasets that are extracted directly from Hospital Information System (HIS) [26]. The objective is to evaluate data utility in comparison to privacy preservation of sensitive data. Based on the empirical analytics, conceptual modeling frameworks are drawn (using REA or e^3 value modeling ontologies) and are evaluated by in-field (i.e. from within the organization) IT expert(s). The aforementioned conceptual modeling frameworks are selected for clarity and convenient understanding of the key actors, their interactions, and mutual value gain for technical (and organizational) proceedings. [23]. For normative evaluation, two alternative approaches are followed as per the respective research objective to ascertain the provenance record (briefly explained in the next paragraph) making. One approach follows the 'Content Analysis' (using literature review, official websites, and (online) content) methodology to formulate conceptual modeling framework(s). We also plan to include the 'BPMN text extraction' methodology for document rule mining [1]. The deduced BPMN model will be evaluated by the IT expert(s). So far, publicly available event logs (comprising metadata share within and across a local hospital) are used for research findings, we are in process of a prospective collaboration with an EU project for some interesting datasets and respective project collaboration concerning inter-organization exchange of health data with a special focus on privacy preservation.

The PUT concerning issues either depend upon or directly influence the prescribed first three dimensions. Which in turn incorporate the data pipeline and data provenance [18]. Data pipeline allows automatic data gathering from diverse sources and its integration into a data warehouse. In healthcare, espe-

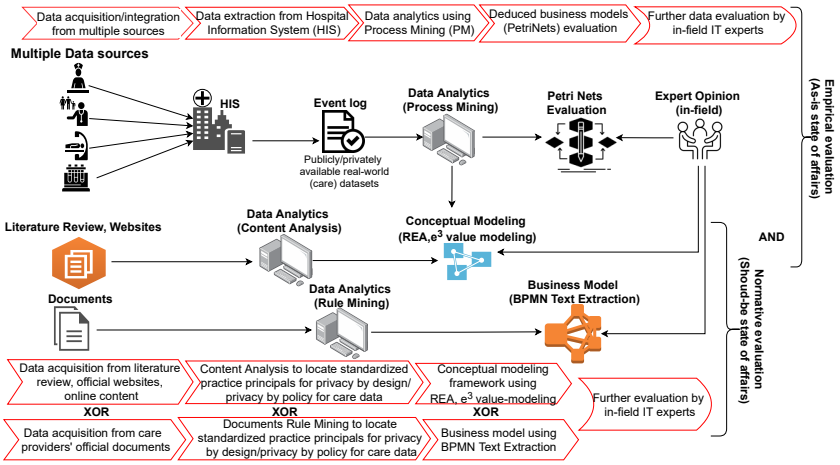

Fig. 2: Data Analytics Approach (with Methodology)

cially in HIS, the data is assimilated from diverse sources that include, amongst others, administrative, medical, and physiological (including sensor-based) data resources. Provenance is the documentation of the "data entities, systems and processes" to avoid data manipulation and system misuse [18]. The potential usefulness of provenance records depends directly upon the veracity (genuine description) of the data pipeline. Provenance records allow: easier assessment in complying with the regulatory prerequisites, identification, and recovery of bottlenecks concerning systems and processes, and data security and privacy preservation [18]. The aforementioned methods and their combination were intentionally selected for not only analytical but also normative evaluation of PUT to validate the integrity of the systems, tools, and processes in the contemporary Dutch healthcare landscape. The prescribed fourth dimension, however, largely relies upon the availability of the time.

3 Current Results

By following the above-mentioned approach and methodology, two papers are published and one paper is under review [23, 24]. Normative evaluation is done using the Content Analysis (explained above) methodology [24]. A conceptual modeling framework is designed on the fundamentals of the Padlock Chain Model of e^3 value modeling. The model is evaluated by the (in-field) IT expert (An IT head in a local hospital who is also affiliated with a local diagnostic lab). The model identified that the privacy is implemented with 'privacy by design, 'privacy by policy, and patients 'informed consent for care metadata sharing. The indicators for privacy by design vary as per each healthcare provider's business goals, and lack consistency across healthcare providers [24]. Privacy by pol-

icy (Information Security Management System) indicators are regulated by ISO (and NEN) and local regulatory authorities but these are only qualitative evaluations. The lack of; consistent/above board privacy by design indicators and quantitative evaluation of ISMS indicators, leaves ample room for technical and organizational ambiguity for the care providers and in turn, gives vent to the privacy lapses [24].

Another research work identifies that the un-anonymized/pseudonymized care metadata sharing within/across a local hospital poses serious (patients') privacy concerns [23]. The empirical evaluation was done using Process Mining on event logs from a local HIS. The evaluation identified that for the sake of efficient/effective performance-oriented care, the patients' un-anonymized metadata is shared within horizontally oriented intra-organizational (i.e. between multiple departments within the hospital, etc) caregivers [23]. The in-field IT expert evaluated that the findings are generalizable to the vertically located (i.e. outpatient caregivers such as lab, general practitioner, pharmacy, hospital) caregivers as well. Based on the aforementioned empirical evaluation and (normative) content analysis, the conceptual modeling framework using REA's Insurance Model [14] is extended [23]. The framework stresses recent advancements where 'Materialized Privacy Claims' are launched either by the patient or by any other potent authority, such as the Dutch Data Protection Officer (DPO) and costs hundreds of thousands of euros to the care providers [23].

Another under review research work evaluates the PUT on care event log from HIS. The PUT is evaluated using the ProM tool for identifying noise-adding plugins which are data utility efficient as well. The plugins are evaluated on three different datasets and two different versions of ProM and gave similar results. The research work will assist the ProM tool's end-users to make use of those plugins for privacy-preserving, utility-prone data analytics using Process Mining. So far, the first dimension and marginally the second dimension are covered. The rest of the dimensions will be covered in the remaining years to come.

4 Threats to Validity

To increase the scope and to ascertain the functionality of the data analytical approach, we are not confining ourselves to one disease-specific event log(s), rather an analysis of more diversified datasets (with a focus on inter/intra organizational data exchange) will allow us to conduct more realistic empirical evaluation (in avoiding the selection bias). But on the other hand, this approach can lead us to certain unforeseen pitfalls while aggregating the data information as the results will be diversified yet non-conforming to one another. To avoid this loophole, in the future, we aim to gather at least two or more event logs from a similar sub-domain. Additionally, to gather the diversified datasets, we are in discussion with the personnel who are directly involved with an EU project and are interesting in a prospective collaboration with us concerning inter/intra organizational care data exchange with a special focus on privacy preservation.

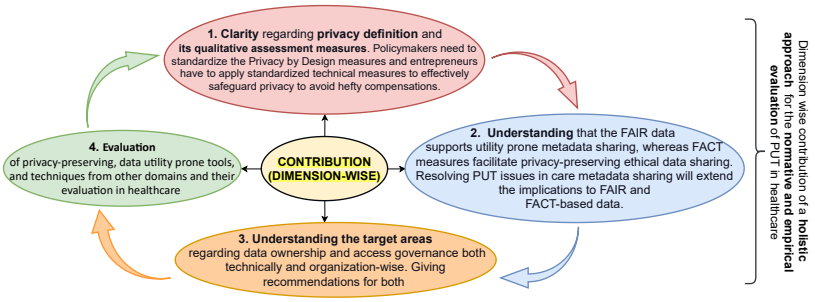

Fig. 3: Dimension-wise contribution to the PUT evaluation

So far the discursive technical solutions hamper the holistic understanding of the PUT in the care data share landscape. This research work aims to provide the footing for the same. Even if all four dimensions are not completed by the end of this Ph.D. work, at least it will provide the infrastructure with the first three dimensions which will (ideally) serve the basic purpose/goal.

5 Contribution

So far, various divergent, remotely conducted analytical research works in healthcare such as for Internet of Medical Things (IoMT) [12], Artificial Intelligence (AI) in healthcare [8,13], generic IT-related challenges [16], Process Mining [21] exist. Similarly, normative evaluation with the standardized privacy-aware conceptual modeling frameworks for the scalable, privacy-preserving systems for IoT exist [7,9,15]. The connection between the empirical and normative evaluations of PUT in care metadata sharing is still lacking. The proposed research work proposes a unique holistic approach in resolving PUT concerning issues in healthcare. Fig. 3 (with a recurring color scheme for each dimension as that of Fig. 1) gives a detailed description of the dimension-wise contribution of the proposed research work (see Fig. 3).

6 Conclusion

ITE is an amalgamation of business processes, techniques, and systems that improve business proceedings in better achieving business goals. Currently, enterprises encounter privacy concerning issues in performing performance-oriented data analytics. Privacy-Utility-Tradeoff (PUT) is the performance impairment of Big Data Analytics (BDA) in ascertaining data privacy. Normative (should be) and empirical (as-is) evaluation of PUT is essential to better find suitable techniques and processes for (care providers) Information Systems against PUT in healthcare.

The empirical evaluation is performed on real-world care event logs. The findings are drawn using conceptual modeling frameworks using REA and e^3 value

modeling ontologies. For normative evaluation, two alternatives approaches are followed. One content analysis technique provides a basis for the 'conceptual modeling' frameworks and the other 'BPMN text extraction' allows documents rule mining and formulation of BPM. (In-field) IT experts further evaluated the conceptual models.

By following the afore-mentioned data analytics approach and methodology, two papers are published and one paper is under review in the first year of this Ph.D. research work. The first paper identified the loopholes in privacy by design and privacy by policy measures. The second paper identified the unanonymized/pseudonymized care metadata share amongst horizontally and vertically located caregivers in the Dutch metadata share landscape. Another under-review paper locates the privacy-preserving data utility-prone noise-adding plugins in publicly available PM tool i.e. ProM.

The future work comprises the quantification of the data value in healthcare and its relevance to privacy (PUT) and FAIR and FACT-based data, privacy evaluation of data ownership/stewardship in healthcare metadata within and amongst care providers, identification of privacy-preserving, higher data utility prone measures from other domains (i.e. logistics, etc) and their evaluation in healthcare. PUT concerning issues of (healthcare) BDA is both technical and organization-based. Thus, the issue requires evaluation of (care provider's) integrated techniques, processes, and systems to find respective solutions in (care) metadata share landscape.

Acknowledgments. I am grateful to my mentor, dr. Maurice van Keulen, and my first supervisor, dr. Faiza Allah Bukhsh for their exceptionally discerning reviews, their undaunting support, and guidance, throughout.

References

1. Bpmntextextraction, `https://sudonull.com/post/299-Business-processes-Extract-BPMN-model-from-document-Part-1`, 20 March, 2020
2. Dutch-dpa, `https://autoriteitpersoonsgegevens.nl/en/about-dutch-dpa/board-dutch-dpa`, 11 Nov, 2020
3. Gdpr, `https://gdpr-info.eu/`, 11 Nov, 2020
4. go.fair, `https://www.go-fair.org/fair-principles/`, 20 March, 2020
5. thedigitalsociety, `https://www.thedigitalsociety.info/themes/responsible-data-science/`, 20 March, 2020
6. van der Aalst, W.M., Bichler, M., Heinzl, A.: Responsible data science (2017)
7. Arruda, M.F., Bulcão-Neto, R.F.: Toward a lightweight ontology for privacy protection in iot. In: Proceedings of the 34th ACM/SIGAPP symposium on applied computing. pp. 880–888 (2019)
8. Bohr, A., Memarzadeh, K.: Artificial intelligence in healthcare. Elsevier Science & Technology (2020)
9. Can, O., Yilmazer, D.: Improving privacy in health care with an ontology-based provenance management system. Expert Systems **37**(1), e12427 (2020)
10. Davenport, T.H., Short, J.E.: The new industrial engineering: information technology and business process redesign (1990)

11. Garattini, C., Raffle, J., Aisyah, D.N., Sartain, F., Kozlakidis, Z.: Big data analytics, infectious diseases and associated ethical impacts. Philosophy & technology **32**(1), 69–85 (2019)
12. Guan, Z., Lv, Z., Du, X., Wu, L., Guizani, M.: Achieving data utility-privacy tradeoff in internet of medical things: A machine learning approach. Future Generation Computer Systems **98**, 60–68 (2019)
13. Hlávka, J.P.: Security, privacy, and information-sharing aspects of healthcare artificial intelligence. In: Artificial Intelligence in Healthcare, pp. 235–270. Elsevier (2020)
14. Hruby, P.: Model-driven design using business patterns. Springer Science & Business Media (2006)
15. Iwaya, L.H., Giunchiglia, F., Martucci, L.A., Hume, A., Fischer-Hübner, S., Chenu-Abente, R.: Ontology-based obfuscation and anonymisation for privacy. In: IFIP International Summer School on Privacy and Identity Management. pp. 343–358. Springer (2015)
16. Kim, K.J., Joukov, N.: Information Science and Applications (ICISA) 2016, vol. 376. Springer (2016)
17. Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M., Michael, J.: Privacy-preserving process mining. Business & Information Systems Engineering **61**(5), 595–614 (2019)
18. McDaniel, P.: Data provenance and security. IEEE Security & Privacy **9**(2), 83–85 (2011)
19. McSherry, F.D.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. pp. 19–30 (2009)
20. Mohammed, N., Jiang, X., Chen, R., Fung, B.C., Ohno-Machado, L.: Privacy-preserving heterogeneous health data sharing. Journal of the American Medical Informatics Association **20**(3), 462–469 (2013)
21. Pika, A., Wynn, M.T., Budiono, S., Ter Hofstede, A.H., van der Aalst, W.M., Reijers, H.A.: Privacy-preserving process mining in healthcare. International journal of environmental research and public health **17**(5), 1612 (2020)
22. Pramanik, M.I., Lau, R.Y., Hossain, M.S., Rahoman, M.M., Debnath, S.K., Rashed, M.G., Uddin, M.Z.: Privacy preserving big data analytics: A critical analysis of state-of-the-art. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery p. e1387 (2020)
23. Sohail, S.A., Allah, F., Krabbe, J.G.: Identifying materialized privacy claims of clinical-care metadata share using process-mining and rea ontology. In: 15th International Workshop on Value Modelling and Business Ontologies, VMBO 2021 (2021)
24. Sohail, S.A., Krabbe, J., de Alencar Silva, P., Bukhsh, F.A.: Privacy value modeling: A gateway to ethical big data handling. In: 14th International Workshop on Value Modelling and Business Ontologies, VMBO 2020. pp. 5–15. CEUR (2020)
25. Srinivasan, S.: Guide to Big Data Applications, vol. 26. Springer (2017)
26. Van Der Aalst, W.: Data science in action. In: Process mining, pp. 3–23. Springer (2016)
27. von Voigt, S.N., Fahrenkrog-Petersen, S.A., Janssen, D., Koschmider, A., Tschorsch, F., Mannhardt, F., Landsiedel, O., Weidlich, M.: Quantifying the re-identification risk of event logs for process mining. In: International Conference on Advanced Information Systems Engineering. pp. 252–267. Springer (2020)
28. Wieringa, R.J.: Design science methodology for information systems and software engineering. Springer (2014)