

Privacy-preserving personal data analysis

Citation for published version (APA):

Sun, C. (2022). *Privacy-preserving personal data analysis*. [Doctoral Thesis, Maastricht University]. Maastricht University. <https://doi.org/10.26481/dis.20221114cs>

Document status and date:

Published: 01/01/2022

DOI:

[10.26481/dis.20221114cs](https://doi.org/10.26481/dis.20221114cs)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Doctoral thesis

**PRIVACY-PRESERVING PERSONAL
DATA ANALYSIS**

Chang Sun

2022

PRIVACY-PRESERVING PERSONAL DATA ANALYSIS

Dissertation

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. P. Habibović,
in accordance with the decision of the Board of Deans,
to be defended in public
on Monday 14 November 2022, at 10.00 AM (GMT+2)

by

Chang Sun

Promotor

Prof. dr. M. Dumontier

Copromotor

Dr. J. van Soest

Assessment Committee

Prof. dr. I. Arts (Chair)

Prof. dr. O. Beyan, University of Cologne

Prof. dr. G. Weiss

Prof. dr. I. Lagendijk, TU Delft

© Chang Sun, Maastricht 2022.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the author.

Cover	Chang Sun, 2022
Production	Proefschriftenprinten.nl
ISBN	978-90-832727-5-7

To my parents and my grandparents

Contents

1	Introduction	1
2	A Systematic Review on Privacy-Preserving Distributed Data Mining	13
3	A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario	59
4	Studying the Association of Diabetes and Healthcare Cost on Distributed Data from The Maastricht Study and Statistics Netherlands using a Privacy-Preserving Federated Learning infrastructure	73
5	Privacy-Preserving Generalized Linear Models on Vertically Partitioned Data using Distributed Block Coordinate Descent	103
6	Generating Synthetic Tabular Data using Conditional GANs combining with Differential Privacy	135
7	ciTizen-centric Data Platform (TIDAL): Using Distributed Personal Data in a Privacy-Preserving Manner for Health Research	167
8	General Discussion	197
	Summary	211
	Samenvatting	215
	Impact Paragraph	219
	Acknowledgments	223
	List of manuscripts	225
	About the author	227

1

Introduction

1.1 Uncover potential high value from data sharing

The amount of data is growing every single day. The data produced globally is expected to be 175 zettabytes in 2025, growing dramatically from 33 zettabytes in 2018 [1]. These massive amounts of data are used to improve digital technologies and develop data-driven innovations that can impact every aspect of people's lives [2]. For example, data can help improve health outcomes, optimize health services, and reduce energy consumption at home. The data generated in the past can influence the way we live, produce, and consume goods and services in the future.

The value of data lies in its use and re-use. At present, data is scattered and maintained by many different organizations of the public sector and private companies such as hospitals, governments, banks, and insurance companies. Based on statistics reported by the European Commission in 2020, 80% of the processing and analysis of data is conducted at the data sources in a centralized manner [3]. Lack of sharing of, access to, and reuse of data from multiple organizations hinders the analysis possibilities and hence potential insights from the data.

The significance and benefits of data sharing and data (re-)use are increasingly valued by the public and private sectors [2, 4]. From a societal perspective, sharing data across organizations can contribute to delivering timely healthcare services, combating climate change, and improving public services and policymaking. Financially, the economic value of data sharing is expected to increase to approximately 533 billion euros in Europe by having higher quality products, increasing productivity, and delivering better and on-time public services [1].

From a scientific perspective, sharing and mining distributed data can incentivize new scientific insights and a wide variety of applications. For example, learning from sufficient healthcare data from patients helps health professionals in the decision-making process, which potentially tailors the timely prevention and treatment strategies to the patient's particular needs and preferences [5, 6, 7]. However, human health and well-being are not only determined by genetic and biological factors and received medical care, but also by individual behavior, social circumstances, and physical environment [8]. These data are collected by different, independent organizations. Analyzing patient data from one single organization might lead to incomplete knowledge and unreliable decision-making. Combining personal health data (e.g., health status, current and prior medications) with other information (e.g., socio-economic factors and lifestyle data) offers new opportunities to improve our understanding of human health and to develop more accurate and reliable analytical models for health prognosis and predictions [9, 10].

1.2 Challenges in personal data sharing

Although the potential value of sharing data is high, sharing personal data across multiple organizations has a number of challenges. These challenges include (among others): technical barriers, security, data protection compliance to one or more legal jurisdictions, privacy concerns, and trust issues. Technical challenges, for example, include insufficiency of data interoperability and quality, which causes difficulties when combining and analyzing data that are dispersed in their terminologies and representation (structured, semi-structured, unstructured). Second, security is a major concern for sharing personal data for all data organizations (organizations that collect, maintain, and provide data for primary and secondary use). Innovative technologies to foster data sharing such as Federated Learning, Blockchain or Secure Multiparty Computation bring new challenges for preserving a high level of security [2]. Third, the European Union has established a legal framework for protecting personal data, including the General Data Protection Regulation (GDPR) and ePrivacy Directive. However, a legal framework such as GDPR leaves room for interpretation at the discretion of the member countries, which causes inconsistent actions between member countries [11]. While data sharing involves a third country¹ legislation such as the U.S. Privacy Act or China's Personal Information Protection Law, the contradictions need to be addressed between different jurisdictions and legal frameworks [3]. Furthermore, individuals highly value their data rights and privacy. The public consultation on the European strategy for data reports a majority of respondents would be willing to share their data, if sufficient mechanisms were in place to protect privacy [13]. Nonetheless, individuals whose data have been collected currently have very limited control over their data. Technical tools and standards to empower individuals to exercise their data rights are lacking.

1.3 Analyzing distributed data from multiple sources

Addressing the above challenges effectively requires a great level of technical sophistication to simultaneously address legal and/or privacy constraints. Instead of centralizing all personal data from various data organizations, data-processing algorithms can be sent to each site, and only return the results of analysis rather than the source data. One such initiative for a data-sharing platform is the Personal Health Train (PHT) [14, 15, 16], where applications containing data queries and algorithms are sent to the data organizations. The data organizations can inspect whether the application is

¹A third country, defined by the European Commission, is a country that is not a member of the European Union as well as a country or territory whose citizens do not enjoy the European Union's right to free movement [12].

allowed to be executed on (a subset of) the available data. The PHT initiative facilitates authorized algorithmic processing securely at multiple data sites without requiring the transfer of the source data to a centralized location.

Infrastructures such as Vantage6 have been developed implementing the PHT initiative and applied to horizontally partitioned data where multiple organizations hold the same information elements (i.e., variables, attributes, or features) from different data providers (such as study participants) [17, 18, 19]. Correspondingly, vertically partitioned data represents multiple organizations that have different information elements from the same data providers. For example, a hospital has information elements on the same individuals as the tax office while the type of information elements collected differs per data organization. When analyzing vertically partitioned data, data records must be linked across multiple datasets and some intermediate results and/or encrypted information are exchanged between data organizations. In such a situation, the data needs to be safeguarded not only by the infrastructure but also by the analysis algorithms that are sent over the infrastructure. In Chapter 3, we investigate the technical and ethical-legal challenges of analyzing vertically partitioned data in practice. Following the PHT initiative, we develop an infrastructure that transfers and executes machine learning models on vertically partitioned data in a trust secure environment. Chapter 4 presents an application that uses the developed infrastructure and analysis algorithms on real-life personal data. To enhance the privacy preservation and tackle the practical challenge of setting up the trust secure environment, Chapter 5 further extends to a generalized linear model without requiring a trusted secure environment or any third party.

1.4 Generating synthetic data for data sharing

Like other infrastructures such as Privacy-Preserving Federated Neural Network Learning (POSEIDON) [20] and Swarm Learning [21] which analyze distributed data from multiple sources, the infrastructure we developed and practically implemented (in chapter 3-5) shares two common concepts: 1) keeping the source data locally with the data owner, 2) transferring the machine learning models to the data rather than outsourcing the data. However, the practical implementation of these infrastructures remains challenging such as low data interoperability, inconsistent data standards, and uneven data quality from different data parties. These challenges hinder researchers to build up accurate and reliable machine learning models using these infrastructures [22, 23, 24]. Researchers need samples of data early on to determine the usability of the data elements and the feasibility of answering their research questions. Especially, when the researcher is conducting exploratory

research studies. The utility of the requested data is not known to the researcher until they conduct the preliminary examination. This uncertainty may cause a severe delay and unnecessary costs for research projects.

To address this challenge, Chapter 6 proposes to generate synthetic data that is structurally and statistically similar to the source data. The synthetic data derives meaningful insights from the source data so that researchers can select the relevant data elements and achieve an optimal performance in machine learning models without accessing the source data or before starting the data requesting process. The synthetic data in this study represents data that is similar to the source data at the population level (i.e., distributions of single variables, correlations between variables), and at the machine learning utility level (i.e., the analysis results on synthetic data are close to the results on the source data). Meanwhile, the synthetic data should be realistic at the individual level. For example, a record whose sex is male and has a positive Pregnancy test is not realistic and does not appear in the source data. Hence, the synthetic data should not have such a record. To protect the privacy of the source data, the synthetic data should offer strong privacy guarantees to prevent adversaries from extracting any sensitive information about the source data. Chapter 6 discusses this trade-off between data utility and data privacy in the generation of synthetic data, which leads to the potential and challenges of future use of synthetic data.

1.5 Empowering individuals to control over their data

After addressing challenges from data organizations and the research community, every citizen in our society should be empowered to take control over data describing or being related to themselves and to make better decisions based on insights learned from their data. Citizens highly value their data rights and privacy [13]. However, surveys from the European Commission show our citizens have limited access to their personal data [3]. They suffer from a lack of technical tools and standards to exercise their data rights. For the time being, researchers and organizations must act on their behalf, but in the future technological innovation will give them greater control over how their data are used. With this goal in mind, Chapter 7 extends the developed infrastructure with a citizen-centric data platform (TIDAL) that gives individuals better access to more of their data and ensures citizen-controlled data are processed in a predefined manner. Chapter 7 shifts data access control from data organizations to citizens, and give them the means to decide at a granular level how their data is shared and used. Citizens, as being a custodian of their own data, can be connected with researchers, and data organizations to increase the trust placed by citizens in the processing of their personal data.

1.6 Objectives and outline of chapters

The overall aim of this thesis is to develop new privacy-preserving data sharing and analysis techniques that creates new possibilities for (re-)use of personal data while maximally protecting individual privacy. In this research, preserving privacy in data sharing means:

- restricting the access to source data that are available to share or be analyzed;
- restricting the results of the analysis to only processed data, rather than source data;
- preventing individuals/organizations from seeing the data of other individuals/organizations in the network of data sharing without appropriate permissions;
- being able to learn new insights by advanced analysis techniques where the above-mentioned points are carefully considered;
- providing a realistic, but synthetic alternative to source data.

To achieve this overall research aim, this thesis gradually addresses the research challenges of analyzing distributed personal data from three aspects - data organizations, researchers and scientists, and individuals (such as data providers, study participants). We defined the key research objectives as follows:

- For the data organizations: to develop a secure infrastructure that can combine and analyze individual data from multiple sources without revealing sensitive personal information;
- For the scientific community: 1) to develop and apply privacy-preserving distributed data mining methods to analyze vertically partitioned data with and without a third party; 2) to build a synthetic data generator to simulate the personal data so that the researchers can have an insight about data before the lengthy data request process or build-up analysis model without accessing the source data;
- For the individual: to design a novel citizen-controlled technology that enables individuals to access and control their personal data and monitor the (re-)use of their data.

This thesis is structured into four sections (see Table 1.1) and the contribution statement of each chapter is presented in Table 1.2. A systematic literature review is presented in **Chapter 2** to analyze and define the current status of existing privacy-preserving distributed data mining (PPDDM) techniques. This

Table 1.1: Outline of each chapter

Topic	Chapter	Title
Introduction	Chapter 1	General introduction and outline of the thesis
Background	Chapter 2	A systematic review of privacy-preserving distributed data mining
Privacy-preserving federated learning	Chapter 3	A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario
	Chapter 4	Studying the association of diabetes and health-care cost on distributed data from the Maastricht Study and Statistics Netherlands using a privacy-preserving federated learning infrastructure
	Chapter 5	Generalized Linear Models on Vertically Partitioned Data using Distributed Block Coordinate Descent
Synthetic data generation	Chapter 6	DP-CGANS: Generating synthetic tabular data using conditional GANs combining with differential privacy
Citizen-controlled data platform	Chapter 7	ciTizen-centric Data pLatform (TIDAL): Using Distributed Personal Data in a Privacy-Preserving Manner for Health Research
Discussion	Chapter 8	Discoveries and general summary

chapter elaborates on the issue of executing data mining algorithms on sensitive, and/or confidential data from multiple data organizations while maintaining privacy and presents the developments in the past 20 years. **Chapter 3** applies a federated learning infrastructure, which supports secure and privacy-preserving analysis of personal health data from multiple independent organizations with different governance policies. To prove the feasibility of the infrastructure, **Chapter 4** presents a real-life application using the proposed privacy-preserving infrastructure on personal health data which are vertically partitioned at The Maastricht Study (at Maastricht University) and Statistics Netherlands. After implementation and utilization of the proposed infrastructure, **Chapter 5** proposes a new privacy-preserving generalized linear model to address the observed limitations in the previous chapters by removing a third party in the infrastructure. **Chapter 6** develops a conditional Generative Adversarial Network framework addressing differential privacy to generate synthetic tabular data that is structurally and statistically similar to the source data. This model enables researchers to get insights into the source data and provide a simulation dataset to develop the actual analyses, potentially in a privacy-preserving infrastructure. **Chapter 7** focuses on the future perspective of transferring data control from data organizations to individuals. This chapter presents a citizen-centric data platform (TIDAL) which includes mechanisms to provide fine-grained access to external parties. Finally, **Chapter 8** presents the general discussion of this thesis including the overall conclusion, highlights, lessons learned, remaining challenges and

future perspectives.

1.6.1 Contribution statements

Table 1.2: Author statement of each chapter of this thesis.

	Conceptualization	Methodology	Formal Analysis	Experimenting	Software	Writing
Ch 1	X					X
Ch 2	X	X	X	X		X
Ch 3		X	X	X	X	X
Ch 4	X	X	X	X	X	X
Ch 5		X	X	X		X
Ch 6	X	X	X	X	X	X
Ch 7	X	X	X	X	X	X
Ch 8	X					X

Table 1.2 presents the contributions the Ph.D. candidate made for each chapter. The contribution categories in the table are based on Contributor Roles Taxonomy [25]. **Chapter 1 (Introduction), Chapter 8 (Discussion), Summary, and Impact paragraph** were fully structured and drafted by the Ph.D. candidate. The evolution of overarching research goals and the conclusion of research discoveries were refined based on the outcome of several discussion rounds with my supervisors. In **Chapter 2**, the Ph.D. candidate designed the systematic review study including defining the research questions, setting up the search strategy, the analysis plan, and evaluation criteria for the existing methods. The Ph.D. candidate searched and read all included literature, performed the analysis, and wrote and revised the manuscript based on the input and feedback from the co-authors. In **Chapter 3**, the Ph.D. candidate developed and programmed the infrastructure (in a collaboration with other co-authors), designed simulation experiments, and conducted the analyses. The code of the software and the draft of the manuscript were fully written by the Ph.D. candidate. In **Chapter 4**, the Ph.D. candidate designed and implemented the whole study including requesting data and computational resources, designing and programming the analysis models, setting up and conducting the experiments, writing and finalizing the manuscript. Based on Chapter 3 and 4, the Ph.D. candidate collaborated with another Ph.D. student to advance the theoretical work in **Chapter 5**. The Ph.D. candidate contributed to designing the method and experiments, conducted the analysis (with the first author), and drafted and finalized the manuscript (with the first author). The idea generation and research questions in **Chapter 6 and chapter 7** were based on the discussions between the Ph.D. candidate and

supervisors. The Ph.D. candidate designed and set up the study, studied literature, developed and programmed the frameworks and models, designed and conducted the experiments, discuss the results with supervisors and improved the work. The Ph.D. candidate drafted, revised, and finalized the manuscripts for Chapter 6 and 7 based on the feedback from the supervisors. The Ph.D. candidate was responsible for the manuscript submission of **Chapter 2-7** and source code creation and maintenance for the work from **Chapter 2,3,4,6,7**.

References

- [1] European Commission, Directorate-General for Communications Networks, Content, and Technology. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance (Data Governance Act)*. 2020. URL: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020PC0767>.
- [2] European Commission. Directorate General for Communications Networks, Content and Technology., CEPS., ICF., and Wavestone. *Study to support an impact assessment of regulatory requirements for Artificial Intelligence in Europe: final report*. LU: Publications Office, 2021. URL: <https://data.europa.eu/doi/10.2759/523404>.
- [3] European Commission, Directorate-General for Communications Networks, Content, and Technology. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European strategy for data*. 2020. URL: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020DC0066>.
- [4] Research Roundtable on Environmental Health Sciences, Board on Population Health Practice, Public Health, Health Division, Medicine, and Engineering National Academies of Sciences. *The Benefits of Data Sharing*. Publication Title: Principles and Obstacles for Sharing Data from Environmental Health Research: Workshop Summary. National Academies Press (US), 2016. URL: <https://www.ncbi.nlm.nih.gov/books/NBK362433/> (visited on 12/31/2021).
- [5] Isaac S. Chan and Geoffrey S. Ginsburg. "Personalized Medicine: Progress and Promise". In: *Annual Review of Genomics and Human Genetics* 12.1 (2011). eprint: <https://doi.org/10.1146/annurev-genom-082410-101446>, pp. 217–244. DOI: 10.1146/annurev-genom-082410-101446.

- [6] Davide Cirillo and Alfonso Valencia. “Big data analytics for personalized medicine”. In: *Current Opinion in Biotechnology*. Systems Biology • Nanobiotechnology 58 (Aug. 1, 2019), pp. 161–167. DOI: 10.1016/j.copbio.2019.03.004.
- [7] Holger Fröhlich et al. “From hype to reality: data science enabling personalized medicine”. In: *BMC Medicine* 16.1 (Aug. 27, 2018), p. 150. DOI: 10.1186/s12916-018-1122-7.
- [8] Commission on Social Determinants of Health. *Closing the gap in a generation : health equity through action on the social determinants of health : final report : executive summary*. World Health Organization, 2008, p. 33. URL: https://www.who.int/social.determinants/final_report/csdh-finalreport.2008.pdf (visited on 01/11/2022).
- [9] Suranga N Kasthurirathne, Joshua R Vest, Nir Menachemi, Paul K Halverson, and Shaun J Grannis. “Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services”. In: *Journal of the American Medical Informatics Association* 25.1 (2017), pp. 47–53. DOI: 10.1093/jamia/ocx130.
- [10] Jessica S Ancker, Min-Hyung Kim, Yiye Zhang, Yongkang Zhang, and Jyotishman Pathak. “The potential value of social determinants of health in predicting health outcomes”. In: *Journal of the American Medical Informatics Association* 25.8 (2018), pp. 1109–1110. DOI: 10.1093/jamia/ocy061.
- [11] Birgit Wouters et al. “Putting the GDPR into Practice: Difficulties and Uncertainties Experienced in the Conduct of Big Data Health Research”. In: *European Data Protection Law Review* 7.2 (2021). DOI: 10.21552/edpl/2021/2/9.
- [12] European Foundation for the Improvement of Living and Working Conditions. *Free movement of citizens*. Eurofound. 2011. URL: <https://www.eurofound.europa.eu/observatories/eurwork/industrial-relations-dictionary/free-movement-of-citizens> (visited on 02/12/2022).
- [13] European Commission. *Summary report of the public consultation on the European strategy for data*. Tech. rep. COM(2018) 233 final. European Commission, 2020. URL: <https://ec.europa.eu/digital-single-market/en/news/summary-report-public-consultation-european-strategy-data>.
- [14] Dutch Techcentre for Life Sciences (DTL). *Personal Health Train*, <https://www.dtls.nl/fair-data/personal-health-train/>. Access on 12-8-2021.

-
- [15] Johan van Soest, Chang Sun, Ole Mussmann, Marco Puts, Bob van den Berg, Alexander Malic, Claudia van Oppen, David Townend, Andre Dekker, and Michel Dumontier. "Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data." In: *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*. Vol. 247. IOS Press, 2018, pp. 581–585. DOI: 10.3233/978-1-61499-852-5-581.
- [16] Oya Beyan et al. "Distributed Analytics on Sensitive Medical Data: The Personal Health Train". In: *Data Intelligence 2.1-2* (2020), pp. 96–107. DOI: 10.1162/dint.a.00032.
- [17] Timo M. Deist et al. "Distributed learning on 20 000+ lung cancer patients – The Personal Health Train". In: *Radiotherapy and Oncology* 144 (2020), pp. 189–200. DOI: 10.1016/j.radonc.2019.11.019.
- [18] Timo M. Deist et al. "Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT". In: *Clinical and Translational Radiation Oncology* 4 (2017), pp. 24–31. DOI: 10.1016/j.ctro.2016.12.004.
- [19] Arthur Jochems, Timo M Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. "Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital—a real life proof of concept". In: *Radiotherapy and Oncology* 121.3 (2016), pp. 459–467. DOI: 10.1016/j.radonc.2016.10.002.
- [20] Sinem Sav, Apostolos Pyrgelis, J. Troncoso-Pastoriza, David Froelicher, Jean-Philippe Bossuat, João Sá Sousa, and J. Hubaux. "POSEIDON: Privacy-Preserving Federated Neural Network Learning". In: *NDSS* (2021). DOI: 10.14722/NDSS.2021.24119.
- [21] Stefanie Warnat-Herresthal et al. "Swarm Learning for decentralized and confidential clinical machine learning". In: *Nature* 594.7862 (2021), pp. 265–270. DOI: 10.1038/s41586-021-03583-3.
- [22] David Roschewitz, Mary-Anne Hartley, Luca Corinzia, and Martin Jaggi. "IFedAvg: Interpretable Data-Interoperability for Federated Learning". In: *arXiv:2107.06580 [cs]* (2021). URL: <http://arxiv.org/abs/2107.06580> (visited on 02/12/2022).
- [23] Juan González-García et al. "Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations from the JA-InfAct". In: *Archives of Public Health* 79.1 (2021), p. 221. DOI: 10.1186/s13690-021-00731-z.

- [24] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. “Federated Learning for Healthcare Informatics”. In: *Journal of Healthcare Informatics Research* 5.1 (2021), pp. 1–19. DOI: 10.1007/s41666-020-00082-4.
- [25] Liz Allen, Alison O’Connell, and Veronique Kiermer. “How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship”. In: *Learned Publishing* 32.1 (2019), pp. 71–74. DOI: 10.1002/leap.1210.

2

A Systematic Review on Privacy-Preserving Distributed Data Mining

Adapted from: Chang Sun, Lianne Ippel, Andre Dekker, Michel Dumontier, and Johan van Soest. “A systematic review on privacy-preserving distributed data mining”. English. In: *Data Science* 4.2 (Oct. 2021), pp. 121–150. DOI: 10.3233/DS-210036.

Abstract

Combining and analysing sensitive data from multiple sources offers considerable potential for knowledge discovery. However, there are a number of issues that pose problems for such analyses, including technical barriers, privacy restrictions, security concerns, and trust issues. Privacy-preserving distributed data mining techniques (PPDDM) aim to overcome these challenges by extracting knowledge from partitioned data while minimizing the release of sensitive information. This paper reports the results and findings of a systematic review of PPDDM techniques from 231 scientific articles published in the past 20 years. We summarize the state of the art, compare the problems they address, and identify the outstanding challenges in the field. This review identifies the consequence of the lack of standard criteria to evaluate new PPDDM methods and proposes comprehensive evaluation criteria with 10 key factors. We discuss the ambiguous definitions of privacy and confusion between privacy and security in the field, and provide suggestions of how to make a clear and applicable privacy description for new PPDDM techniques. The findings from our review enhance the understanding of the challenges of applying theoretical PPDDM methods to real-life use cases, and the importance of involving legal-ethical and social experts in implementing PPDDM methods. This comprehensive review will serve as a helpful guide to past research and future opportunities in the area of PPDDM.

2.1 Introduction

Mining distributed, sensitive data offers tantalising potential for new insights and a wide variety of applications, but is generally fraught with concerns of model accuracy and data privacy. Consider the case of analyzing patient data in the healthcare domain: hospitals have used patient data to improve diagnostic accuracy and efficiency [1, 2] and to fuel the transition to preventive [3] and precision medicine [4, 5, 6]. However, learning patient data from a single hospital might cause limited model performance and incomplete knowledge discovery [7]. Patients' health are not only affected by genetic and biological factors, but also by individual behaviour and social circumstances [8]. Combining various patient data from multiple sources offers one pathway to obtain more accurate and reliable analytical models for health outcomes [9, 10]. However, combining distributed sensitive data faces a number of challenges including: data protection compliance to one or more legal jurisdictions, privacy concerns, security, and trust issues. Beyond the healthcare domain, this also applies to applications in many other fields, such as finance and law [11, 12]. Conventional centralised data mining techniques are challenged in this environment and require viable alternatives.

Privacy-preserving distributed data mining (PPDDM), which focuses on the analysis of decentralised data without leaking sensitive information from any party to the other parties, offers one way forward for multiple data parties to overcome the challenges posed by centralising the data for analysis [13]. PPDDM techniques, whether data mining or machine learning, aim to make it technically or mathematically infeasible to deduce the original data from a communication message, and certainly from the final analysis result. To make use of PPDDM in practical applications, we should consider the data problems (e.g., classification, regression), the adversarial concerns the involving data parties have (e.g., malicious, honest), and the balance between data privacy and model performance. PPDDM is sometimes referred to privacy-preserving federated learning after Google first proposed the concept in 2016 [14, 15]. However, privacy-preserving federated learning can be regarded as a specific category of PPDDM, in which there is a federation of autonomous organisations that express an interest to contribute to a joint analysis [16].

A number of PPDDM methods have been reported in the last 20 years. The existing survey papers have compared the theoretical backgrounds, strengths, and limitations. However, the analysis of distributed data has been poorly addressed as only one special case of privacy-preserving data mining [17, 18, 19, 20]. The distributed data problem has been addressed to a limited extent in the survey of Hina Vaghashia [21] and Suchitra Shelke [22]. Vassilios S. et al [20] presented five dimensions of state-of-the-art privacy-preserving data mining algorithms where the problem of analysing distributed data was merely considered to be addressed by cryptography-based techniques and only the association rule mining problem and decision tree induction were presented in this survey. Several surveys summarized the evaluation parameters to assess privacy-preserving techniques including privacy level, hiding failure, data quality, complexity, efficiency, and resistance of different data mining algorithms [18, 20, 23, 24]. Others have a major focus on the definition and construction of Secure Multiparty Computation (SMC) and how SMC can be combined with data mining algorithms [13, 25, 26]. In a recent survey [27], privacy-preserving approaches were summarized for data collection, data publishing, data mining output, and distributed learning. The majority of the published surveys have typically treated PPDDM as a specialised subtopic of either distributed data mining or privacy-preserving data mining. As an emerging field, PPDDM is under-reported in the existing surveys and now requires a more comprehensive and complete analysis.

This study aims to provide an overview of existing approaches and identify outstanding challenges in the field of PPDDM. We report the results and findings of a comprehensive review of PPDDM techniques from 231 scientific

articles published in the past 20 years. We present the characteristics of the 18 most cited studies and analyze their influence on other studies in the field. The results show a wide range of privacy-preserving methods and data mining algorithms have been well-studied. We highlight the findings showing a lack of standard evaluation criteria in the field, the ambiguous definition of privacy, and insufficient experimental information in some studies. These findings enhance the understanding of the challenges of applying the theoretical PPDDM methods to real-life use cases, and the importance of involving legal-ethical experts in implementing PPDDM methods.

The main contributions of this work to the literature in PPDDM field are:

1. to propose comprehensive criteria with 10 key factors to evaluate the new PPDDM techniques. The evaluation criteria include adversarial behaviour of data parties, data partitioning, experiment datasets, privacy/security analysis, privacy-preserving methods, data mining problems, analysis algorithms, complexity and cost, performance measures, and scalability.
2. to present different definitions of privacy, distinguish information privacy from information security in the PPDDM field, and provide suggestions of how to make clear and applicable privacy descriptions to propose new PPDDM techniques.
3. to identify the most cited PPDDM articles, analyze their characteristics and how these articles influence other studies in the field, and
4. to provide a guideline based on the proposed evaluation criteria for researchers to conduct future research and publications in PPDDM field.

This systematic review offers new insights into the important factors that should be considered to propose and evaluate new PPDDM techniques and how to bridge the gap between theoretical methods and practical applications in the field. We present this review paper as a helpful guide to past research and future opportunities in the area of PPDDM.

The outline of this paper is as follows. In the next section, we present existing privacy-preserving methods and define terms related to PPDDM. In Section 2.3, we describe the approach in conducting this systematic review. In Section 2.4, we provide the results of our review, including evaluation criteria. In Section 2.5, we compare the key influential papers. In the last section, we summarize our main findings, present a list of recommendations, and discuss future directions.

2.2 Privacy-preserving methods

Privacy-preserving methods, as the major component of PPDDM techniques, are used to minimize the release of information during data mining model training and communication among multiple parties. Various privacy-preserving methods have been proposed from different communities such as statistics, cryptography, data mining, and secure data transfer. In this section, we summarize the most commonly-used privacy-preserving methods in PPDDM.

2.2.1 Secure Multiparty Computation (SMC)

Secure multiparty computation protocols are designed for multiple parties to jointly compute some function over their own data without revealing the original data to any other parties [13]. The foundation for SMC started from cryptography. In addition to protect the participants from being attacked by external parties (who are outside of the system or protocol), SMC also protects the participants from each other. For example, some SMC protocols are implemented to prevent participants from learning private information from other parties or deliberately sending incorrect computation results to other parties. The following sub-sections describe some protocols in SMC.

Building Blocks (primitives) SMC of Protocols.

Secure protocols that are deployed as building blocks of secure computation are used to prevent data being revealed or deduced from the communication and/or computation between data parties [13]. Commonly used encryption protocols include oblivious transfer and homomorphic encryption. Oblivious transfer, first developed by Even et al. [28], considers two data parties, a requester and a sender, where the requester obtains exactly one instance without the sender knowing which element was queried, and without the requester knowing about the other instances that were not retrieved. Oblivious transfer protocols iteratively pass over the data many times during training, and as a result are computationally expensive. Another technique, homomorphic encryption, was introduced by Rivest [29]. This technique supports certain algebraic operations such as additions and multiplications on encrypted text (i.e., ciphertext). The decrypted result from the operations on ciphertext matches the result of the operations performed on the plain text. Homomorphic encryption systems are grouped into fully homomorphic encryption (FHE) or partial homomorphic encryption (PHE) [30]. As the initial scheme of a homomorphic cryptosystem, PHE can only perform a specific algebra operation such as addition or multiplication in each iteration. This limits the

usability for data mining algorithms, as the algorithms consist of several complex operations. On the contrary, FHE supports any desirable operation and functionality that can run on the ciphertext. Since the ciphertext is never decrypted, the input from each data party is not revealed. The first generation of FHE system was proposed by Gentry in 2009 [31]. However, FHE systems are not sufficiently efficient due to the high computational cost of performing iterative operations over encrypted data during the training epochs.

Generic SMC Protocols.

Generic SMC protocols were implemented for any probabilistic polynomial-time function [13]. Unlike homomorphic encryption systems, these generic protocols are sensitive to the number of data parties. The commonly-used protocol of secure two-party computation is Yao's garbled circuit protocol [32]. The protocol is based on evaluating the function that needs to be computed by two data parties as a combinatorial circuit with a collection of gates (e.g., AND, XOR gate). These gates connect with circuit-input wires, circuit-output wires and intermediate wires. Each gate has two input wires and one single output wire. The required communication of the protocol depends on the size of the circuit, while the computation cost depends on the number of input wires. Extensions to more than two data parties, i.e. the cases of multiparty computation, have been developed by Micali et al. [33], Beaver et al. [34], and Ben-Or et al. [35]. Following Yao's theory, these protocols are based on designing the function as a circuit and applying a secure computation protocol to the circuit [13]. Beside computational complexity, communication cost is a considerable factor in these protocols. All protocols need a one-to-one communication channel between every pair of parties. Some require a broadcast channel for all parties.

Specialized SMC Protocols.

Specialized SMC protocols are commonly used as primitives to the data mining algorithms including secure sum, secure set union, secure size of intersection, and secure scalar product protocols. These protocols allow certain operations without revealing any inputs from any of the participating data parties.

Secure sum as a basic and simple example of secure multiparty computation was introduced by Clifton et al. to obtain the sum of the inputs [25]. The protocol is as follows: data party A has V_1 local value. Party A generates a

random number R and calculates $(R + V1)$ and sends this result to data party B (PB). Then, Party B adds their local value to the received value and sends it $(R + V1 + V2)$ to the next party. In the end, to obtain the final result, the last sum value will be sent back to party A to subtract R . The protocol ends with sending this final result to all participating parties. An example of securely computing a sum among 4 four parties is shown in Fig.2.1a.

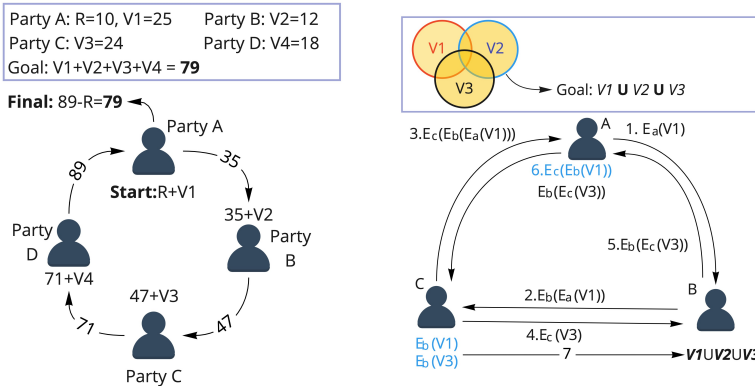
Secure set union has been applied to the case where data parties want to jointly create unions of sets from rules and itemsets shared by multiple parties but not leaking the owner of each set. To guarantee a secure computation, one approach is to apply a commutative encryption system in computing the set union [25, 36]. A commutative encryption system can encrypt original data multiple times using different users' public keys. The final encrypted data can be decrypted without considering the order of the public keys in the encryption process [37]. In the secure set union protocol, one data party encrypts its own itemsets using commutative encryption and transfers them to other parties. The receiver party encrypts both its own sets and the received encrypted sets and passes it to the next party. Once the data is encrypted by all parties, decryption can start at each party in any order. The permutation of the encryption order prevents the participating parties from tracking the ownership of itemsets. However, if one item is present at multiple data parties, then the number of the item will be exposed because of duplication. Fig.2.1b presents an example of securely computing a set union among three data parties.

Secure size of set intersection is solving the problem that multiple data parties want to obtain the size of set intersection of their local datasets without revealing the ownership. Similar to secure set union, each data party encrypts its own item sets by using commutative encryption and sends it to another data party. The receiver encrypts these items, arbitrarily permutes the order, and sends it to the next data party. This process ends when all item sets are encrypted by all data parties. Due to the commutative encryption, if and only if the original inputs are the same, then the final outcomes from two different item sets can be equal. Therefore, the number of values that occur in all encrypted item sets is the size of the set intersection. No input will get exposed since only encryption (no decryption) is required. Fig.2.1c demonstrates the protocol of securely computing the size of set intersection.

Secure scalar product protocols are essential and powerful. It has been widely applied in many data mining algorithms which can be decomposed to the calculation of scalar products. As a notable example, Vaidya and Clifton extended a secure scalar products protocol to solve association rule mining problems between two parties [38]. The general idea is as follows:

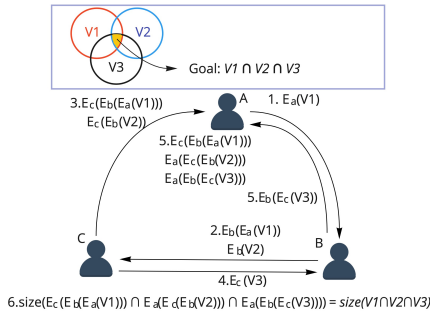
1. Data party A has $X = \{x_1, \dots, x_n\}$, while data party B has $Y = \{y_1, \dots, y_n\}$. The goal is to calculate $X * Y = \sum_{i=1}^n (x_i * y_i)$ without

Chapter 2. A Systematic Review on Privacy-Preserving Distributed Data Mining



(a) An example of secure computation of a sum among four parties. R is a random number generated by Party A. $V1, V2, V3$, and $V4$ presents private data from party A to party D.

(b) An example of secure computation of a set union among three parties. Party A, B, C encrypts their private data using a commutative encryption scheme respectively (E_a, E_b, E_c). Text in blue is decrypted text.



(c) An example of secure computation of a size of set intersection among three parties. Party A, B, C encrypts their private data using a commutative encryption scheme respectively (E_a, E_b, E_c).

Figure 2.1: Examples of three secure multiparty computation protocols - Fig.2.1a. Secure sum protocol, Fig.2.1b. Secure set union protocol, Fig.2.1c. Secure size of set intersection protocol.

revealing inputs to the other party. Both parties share a matrix C which is generated by random numbers.

2. The protocol starts at Party A who generates n random numbers $Ra = \{r_1, \dots, r_n\}$. Then, party A calculates $X' = X + C * Ra$ and send to party B.
3. Party B generates $m(< n)$ random numbers Rb and calculate $Y' = C_1 * Y + Rb_1, \dots, C_{n/m} * Y + Rb_1, \dots, C_{2n/m} * Y + Rb_2, \dots, C_n * Y + Rb_n$ and $S' = \sum_{i=1}^n (x'_i * y_i)$. Y' and S are sent to party A.
4. Party A calculates $S'' = S' - \sum_{i=1}^n (Ra * Y')$ and m sets of sum of Ra which is $Ra' = Ra_1 + Ra_2 + \dots + Ra_{n/m} + Ra_{n/m+1} + \dots + Ra_{2n/m}, \dots, Ra_{((m-1)n/m)+1} + Ra_{((m-1)n/m)+2} + \dots + Ra_n$. Party A sends S'' and Ra' for final result calculation.
5. Party B computes the final scalar product as $S = S'' + Ra' * Rb$.

The security of this secure scalar product protocol is guaranteed by the inability of either side to deduce k equations with more than k unknowns. As with many other existing scalar product protocols [39, 40], it is limited to the collaboration between only two parties because of the lack of efficiency in practice [25].

2.2.2 Data Perturbation

Data Perturbation preserves data privacy by adding ‘noise’ to the individual records but still keeps the key summary information about the data [41]. One major approach of data perturbation is to use statistical techniques to replace the original data with synthetic values which have the same or comparable statistical information (e.g., distributions) as the original values. The synthetic data can be generated by a statistical model which learns from the original data. The other main approach is to distort the values by applying additive noise, multiplicative noise, or other randomization procedures [42]. Data swapping, another method of data perturbation, switches a set of (sensitive) attributes between different data entities to prevent the linkage of records to identities [43, 44]. The major drawback of these methods is the decrease of data quality and accuracy of the learning model. Data perturbation techniques are more commonly used to protect privacy in data publishing problems [27].

2.2.3 Local Learning and Global Integration

The method that integrates local models to one global model uses the foundation of ensemble learning that trains a set of models in order to enhance the performance of one single model [45, 46]. Each data party can train their own local data miners independently. Then, these local data miners are sequentially or parallelly integrated to compose a center or global data miner which can generate the final results. Consequently, the original data of each party is never transferred to other data parties. A majority of data mining algorithms have been theoretically developed to this approach including Support Vector Machine [47, 48, 49, 50], Decision Tree [51, 52, 53], Neural Networks [54, 55, 50, 56] and so forth. A few of them have been successfully implemented, applied and evaluated in practical use cases [7, 57].

2.3 Methodology

This paper follows the systematic review procedures described by Kitchenham [58]. In this section, we discuss the inclusion and exclusion criteria of study selection, followed by the search strategies, and evaluation criteria for reviewing selected studies.

2.3.1 Eligibility Criteria

We selected papers that are peer-reviewed publications in English between 2000 and 2020 working on data mining and machine learning techniques that solve problems of classification, regression, clustering, or association rule mining. The eligible papers must take privacy preservation into account when data mining and machine learning models are executed on partitioned data. Partitioned data includes horizontally partitioned/homogeneous data, vertically partitioned/heterogeneous data, and arbitrarily partitioned data (The definitions are presented in section 3.3). Furthermore, included papers must 1) propose and/or implement a new approach and/or; 2) apply existing approaches to a practical case and/or; 3) improve the performance of existing approaches.

To narrow down the number of publications, we excluded poster and workshop abstracts, survey papers, and articles that only contain discussions on current concerns and future research. To set the scope of this survey, the authors screened titles, keywords, and abstracts to exclude the papers that 1) only focus on privacy-preserving data mining/machine learning on centralised data, 2) solve problems of parallel computing, cloud computing, grid computing, edge computing, and fog computing to

improve computational performance rather than the complexity of the data analysis problem, 3) solve privacy issues in data collecting, data publishing, data storage, and data querying, and 4) focus on Blockchain, web attacks detection, intrusion detection, data privacy focusing on mobile devices, geographic data privacy, and differential privacy. If the papers could not be identified based on its title, keywords, and abstract, the authors reviewed the full paper.

2.3.2 Search Strategy

We used the following search engines and digital libraries: IEEE Xplore Digital Library ¹, ACM Digital Library ², Science Direct ³, ISI Web of Science ⁴, Springer Link ⁵, PubMed ⁶. Based on the inclusion criteria, we formulated the following terms to search in the title, abstract, and keywords of papers. The entire workflow for selecting relevant studies is presented with search results in Figure 2.3 in Section 2.4.1.

1. *privacy* and (*distributed* or *de-centralized* or *de-centralised* or *partitioned*) and *machine learning* (PPDML)
2. *privacy* and (*distributed* or *de-centralized* or *de-centralised* or *partitioned*) and *data mining* (PPDDM)
3. *privacy* and (*vertically* or *heterogeneous*) and *machine learning* (PPVML)
4. *privacy* and (*vertically* or *heterogeneous*) and *data mining* (PPVDM)
5. *privacy* and (*horizontally* or *homogeneous*) and *machine learning* (PPHML)
6. *privacy* and (*horizontally* or *homogeneous*) and *data mining* (PPHDM)

2.3.3 Evaluation Criteria for Reviewing Papers

To evaluate the paper on PPDDM techniques, conventional data mining evaluation criteria are not adequate [45]. Beside conventional evaluation methods, additional factors such as communication costs, data partitioning, adversary behavior, privacy measures should be considered. To the best of our knowledge, there are no standard criteria for evaluating new PPDDM approaches. Consequently, studies selected a various set of evaluation methods

¹IEEE Xplore: <https://ieeexplore.ieee.org/Xplore/home.jsp/>

²ACM Digital Library: <https://dl.acm.org/>

³ScienceDirect: <https://www.sciencedirect.com/>

⁴Web of Science - Clarivate: <https://clarivate.com/products/web-of-science/>

⁵Springer Link: <https://link.springer.com/>


⁶PubMed: <https://www.ncbi.nlm.nih.gov/pubmed/>


which they think are necessary for their approaches. In this review, we assessed selected papers considering the following 10 factors including adversarial behavior of data party, data partitioning, experimented datasets, privacy/security analysis, privacy-preserving methods, data mining problems, analysis algorithms, complexity and cost, performance measures, and scalability. The authors initially generated and modified these evaluation criteria by reviewing 10% of the included articles. Then, the evaluation criteria have been discussed by the co-authors in several iterations of reviewing until an agreement has been made on these 10-factor evaluation criteria. All selected papers have been reviewed and assessed again using the criteria.

1) Adversarial behavior of data parties covers the assumed adversarial behavior that involved data parties have. In this review, we consider two types of adversarial behavior of involved parties - semi-honest and malicious. A semi-honest (passive or honest-but-curious) party follows the protocol properly, however is also curious about other parties' data [13]. The semi-honest party will attempt to learn or deduce data from other parties. A malicious (or active) party will arbitrarily deviate from the protocol and will make deliberate attacks to obtain access to data from other parties [59]. For example, possible malicious behavior might be not starting the execution of protocols at all or suspending (or aborting) the execution at any desired point in time. Papers that use ambiguous expressions such as "untrusted" or "non-trusting" or "non-collaborative" are not classified into any category, because they did not clearly indicate the adversarial property of data parties, nor did they provide any privacy or security proof of their methods. In addition, we include the situation where a third party was involved. A third party, as another independent entity, can combine data from multiple parties, execute analysis on the joint datasets, or do the final computation based on information from data parties. A third party can be fully-honest, semi-honest, and malicious.

2) Data partitioning Figure 2.2 shows three scenarios of data partitioning which are considered in this review: 1) Horizontally partitioned data which contains the same attributes from different data instances (see Figure 2.2a). For example, different hospitals see different patients, though they collect the same patient attributes; 2) Vertically partitioned data which contains the same data instances but with different attributes (see Figure 2.2b). For example, a hospital has data on the same individuals as the tax office, while the attributes collected differs per data party; 3) Arbitrarily partitioned data, the hybrid situation of horizontally and vertically partitioned data. In this scenario, the data providing institutes hold different attributes for different data instances (see Figure 2.2c).


ID	Age	Gender	Education	Wellbeing	Diabetes
1	56	Male	University	Good	YES
2	25	Female	University	Medium	NO
3	31	Female	High School	Good	NO
4	45	Male	Primary School	Poor	YES
5	32	Male	Primary School	Good	No
6	60	Female	High School	Poor	YES
7	55	Male	University	Medium	NO


Party A (Rows 1-4) → 

Party B (Rows 5-7) → 

(a) An example of horizontally partitioned data.


ID	Age	Gender	Education	Wellbeing	Diabetes
1	56	Male	University	Good	YES
2	25	Female	University	Medium	NO
3	31	Female	High School	Good	NO
4	45	Male	Primary School	Poor	YES
5	32	Male	Primary School	Good	No
6	60	Female	High School	Poor	YES
7	55	Male	University	Medium	NO


Party A (Columns 1-4) → 


Party B (Columns 5-6) → 

(b) An example of vertically partitioned data.

ID	Age	Gender	Education	Wellbeing	Diabetes
1	56	Male	University	Good	YES
2	25	Female	University	Medium	NO
3	31	Female	High School	Good	NO
4	45	Male	Primary School	Poor	YES
5	32	Male	Primary School	Good	No
6	60	Female	High School	Poor	YES
7	55	Male	University	Medium	NO

Party A (Columns 1-3) → 

Party B (Columns 4-6, Row 1) → 

Party C (Columns 4-6, Rows 2-7) → 

(c) An example of arbitrarily partitioned data.

Figure 2.2: Examples of three different partitioned data. Fig.2.2a shows horizontally partitioned data which contains the same attributes/features from different data instances. Fig.2.2b shows vertically partitioned data which contains the same data instances but with different attributes/features. Fig.2.2c shows arbitrarily partitioned data which is a hybrid situation of horizontally and vertically partitioned data.

3) Dataset information factor indicates whether the study provides adequate information about the applied datasets in their experiments. Basic information of datasets including sources, names, numbers of features and instances, categorical or numeric type (if available) were recorded. Considering the readability, collected information is composed into 5 categories:

1. Datasets that are publicly available (e.g., UCI repository) [60]
2. Datasets from practical cases (e.g., real patients data from a clinic)
3. Synthetic datasets and datasets which were generated by authors
4. Experiments are presented but dataset information is missing
5. No experiments are presented in the paper

4) Privacy definition or measurement describes whether the study gave an explicit privacy definition, analyses, or measurements. Due to a lack of a universally accepted standard definition, there are many different definitions of privacy from various aspects such as law and philosophical point of view covering personal information, body, communications, and territory [61, 62]. This review only focuses on information privacy which concerns the control of collection, use, retention, and distribution of personal information. During reviewing, we do not assess if the privacy definitions are correct and the levels of privacy these studies can preserve though whether they gave a sufficient description, measurement, or analysis of privacy.

5) Privacy-preserving methods are classified into 5 categories: 1) secure multiparty computation - building blocks, 2) secure multiparty computation - generic and specialized construction protocols, 3) data modification, 4) local learning and global integration, and 5) others. First 4 categories have been explained in detail in the Privacy-Preserving Method Section. The papers which did not use any method from above are categorized to others.

6) Types of problems covers four main data mining areas: i.e., classification, regression, clustering, and association rule mining. Classification predicts a class with categorical labels. These categorical labels can be represented by discrete values, where the ordering among values has no meaning. In contrast, regression is to predict continuous-valued function or ordered value. Clustering is to group a set of data objects into multiple groups so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Association rule mining is to discover interesting associations and correlations between itemsets in transactional and relational databases [63]. Additionally, we labeled the studies as general that solved some mathematical or statistical problems which are applied to classification, regression, and clustering. The studies which worked on outlier detection, record linkage,

recommendation system, attribute/dimension reduction, feature selection, and probabilistic graph are categorized into others.

7) Data mining algorithms present the algorithms which have been developed in a privacy-preserving manner and which ones lack attention. There are plenty of algorithms across the data mining and statistics domain [64, 18]. In this review, the top eight algorithms are listed including decision tree, K-nearest neighbor, bayesian networks, support vector machine, neural networks, K-means, linear/logistic regressions, and A-priori algorithms.

8) Complexity and cost indicates whether the study explicitly measures computational complexity, time, and communication cost. The papers which did not present experiments but only briefly discussed computation, time, and communication costs are counted as 'No Measurement'.

9) Performance measures covers whether the study compared the performance of their approaches with 1) other published PPDDM methods, 2) centralised data mining methods, and 3) distributed without preserving privacy methods. The performance measures include accuracy, precision, recall, F1 score, AUC (Area Under the Curve), mean squared error, mean absolute error, and other standard evaluation criteria in the data mining domain [63, 65, 66, 67, 68]. Owing to the high degree of heterogeneity in the reporting of performance measures across the reviewed papers, we determine whether any performance measure was applied to evaluate the methods rather than comparing different performance measures. The papers which contained experiments but did not compare their results with other methods are categorized into 'No comparison (with experiment)'. The studies which did not provide any experiments are classified to 'No experiments'.

10) Scalability covers whether the study presented a scalability analysis or the experiments prove the scalability of their approach. The scalability in this review means if the approach can tackle large-size datasets which contain a large number of either features or instances. It is noteworthy that only discussing scalability or mentioning their approaches are scalable were not included.

2.4 Results

In this section, we first describe the number and distribution of search results retrieved from the six search engines in the last 20 years. Detailed reviews of selected papers based on the evaluation criteria are elaborated in section 2.4.2. The analysis of the relations among selected papers is described in section 2.4.3.

2.4.1 Search Results

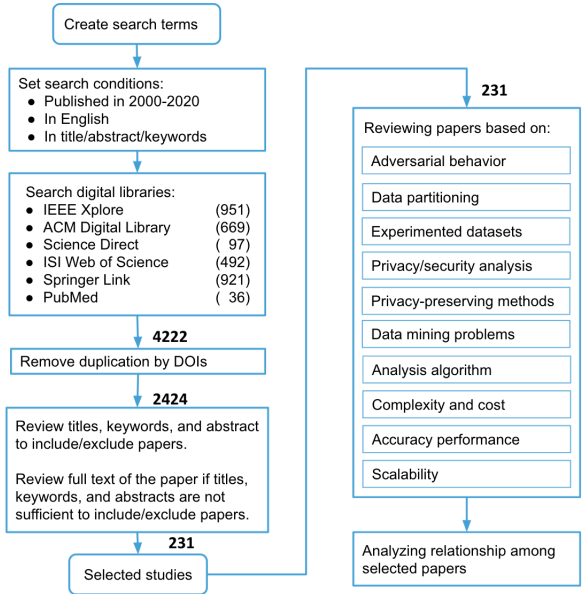


Figure 2.3: Workflow of conducting this systematic review

Figure 2.3 presents the workflow of this systematic review with the number of papers included in each step. Following the inclusion criteria, 4222 publications including duplicates were retrieved from six search engines. Most papers were from IEEE and Springer Link followed by ACM Digital Library. To remove the duplicates, we used Digital Object Identifiers (DOI) to keep the unique papers. The number of publications was reduced from 4222 to 2424. Furthermore, we filtered out irrelevant papers by screening the titles and abstracts of the retrieved papers. Papers that focused on parallel computing, cloud computing, edge computing, network security, intrusion detection, web attack detection, privacy in mobile data and geographic data, differential privacy, privacy in data collecting, data publishing, data storing, data querying were excluded. In the end, 231 papers were selected to be preliminarily reviewed.

To improve the insight of the search result, we map the selected papers into graphs by using the Gephi visualization tool [69]. In Figure 2.4, the distribution of 231 selected papers using different search terms is presented. Papers are presented as nodes and clustered by the search terms. For instance, 182 selected papers were found by using the search term - PPDDM, while 38 of them were findable in PPHDM category and 50 of them were findable in

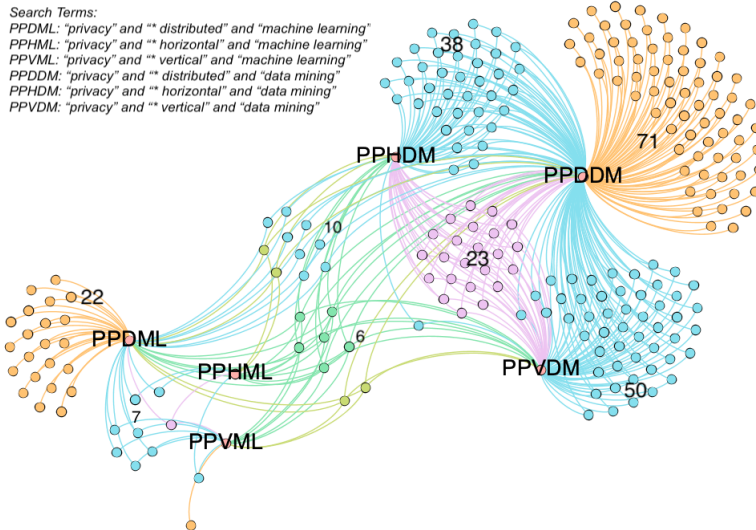


Figure 2.4: Numbers and clusters of papers from different search terms. Papers are presented as nodes and clustered by the search terms. The number of papers is labeled in the figure. The edges show which terms were used to find the papers.

PPVDM. It is obvious that data mining papers are the majority of the search outcomes. It is reasonable as data mining covers a larger scope than machine learning. Privacy issues should be considered in the entire data processing procedure instead of only the part of analysis and building machine learning models. Moreover, a large number of papers (71 papers from PPDDM, 22 papers from PPDML) did not indicated what exact data partitioning problems their method can solve in their titles, abstracts, and keywords. This increases difficulties for other researchers and practitioners to find the correct papers based on their needs.

2.4.2 Review Results

Fig. 2.5 we summarizes the review results of 231 papers using the 10 evaluation factors. The full review results of 231 papers are publicly available in the data repository: <https://doi.org/10.6084/m9.figshare.14239937.v4>.

Adversarial behavior of data parties. About half of the reviewed studies assuming their approaches are applicable for the data parties with semi-honest adversary behavior. In contrast, only 17 reviewed studies developed their methods against malicious parties. Third party constructions were applied in

the method of 47 studies. More than half of them handled semi-honest behavior data parties together with employing the third party. However, it is worth noting that over 30% of selected papers did not state a clear assumption that which adversarial behavior their approach can deal with.

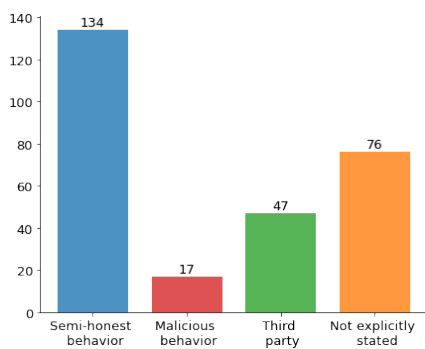
Data partitioning. Horizontally partitioned data (105 reviewed papers) and vertically partitioned data (112 reviewed papers) seem to be represented equally in the selected literature. There are 35 papers handling both horizontally partitioned data and vertically partitioned data. However, only 9 reviewed studies developed PPDDM methods on arbitrarily partitioned data which can work with semi-honest data parties. Additionally, 20% of selected studies did not indicate in which data partitioning situation their methods can be applied.

Privacy is one of the most important evaluation parameters for PPDDM techniques. However, only one fifth of selected studies describe an explicit definition of privacy and mathematical analysis of how much information is leaked by the proposed method. There are 81 papers proving the security of their approaches rather than a privacy analysis. The difference between security and privacy will be discussed in the next section. The majority of studies describe privacy preservation briefly in their own understanding. These descriptions are heterogeneous: e.g., not revealing privacy of any database, not compromising the privacy of the data owners, preserving the confidentiality of datasets, and no important information leakage. The remaining 30 papers proposed new PPDDM methods without indicating any definition or description about privacy.

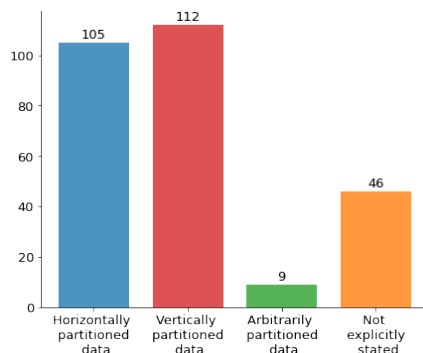
Privacy-preserving methods. Secure multiparty computation techniques are the most encountered solutions in the PPDDM domain. The generic and specialized protocols were applied in 101 papers, while 89 studies employed homomorphic encryption or oblivious transfer protocols. A minority of reviewed studies used data modification, or methodologies to train local models and combine these local models into a global model. A combination of techniques such as combining data modification and homomorphic encryption protocols has been applied by 41 studies.

Types of data problems and data mining algorithms. Classification problems attracted the most attention from researchers in the PPDDM domain, followed by association rule mining and clustering. By contrast, a minority of studies deal with regression modeling. The most implemented data mining algorithms tackling these data problems are: Tree-based algorithms such as decision tree, random forest (35 papers), A-priori-based algorithms (34), Neural Networks (21), Bayesian Networks (18), Support Vector Machine (17), K-Nearest Neighbor (16), Linear/Logistic/Ridge Regression (16), and K-means (9). There are over 10% of reviewed papers studied on generic algorithms

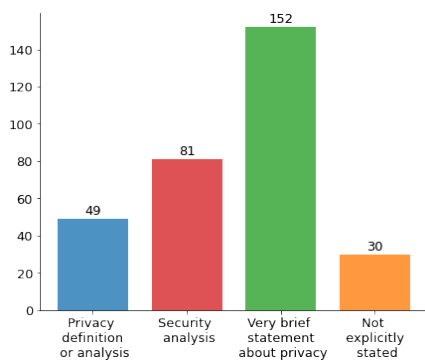
that can be applied to multiple data mining techniques such as gradient descent. About 12% of reviewed papers worked on solving privacy problems in outlier detection, record linkage, recommendation system approaches, attribute/dimension reduction, feature selection, and probabilistic graphs.



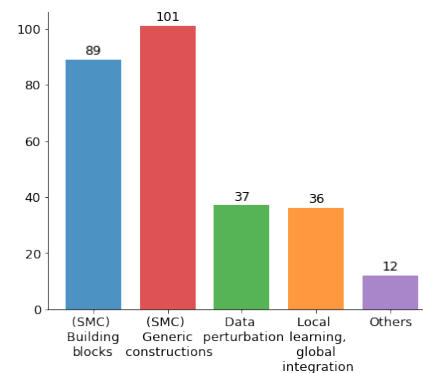
(a) Adversarial behavior of data parties



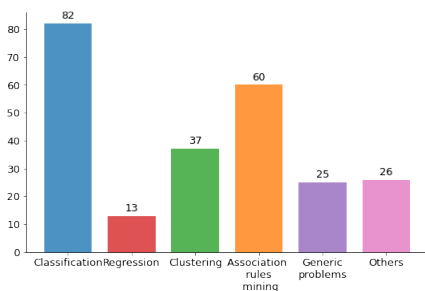
(b) Data partitioning



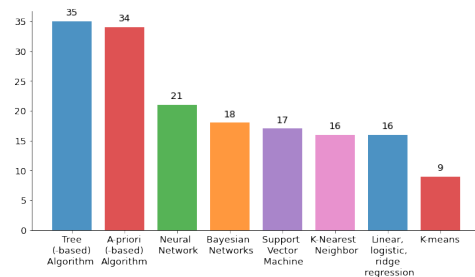
(c) Privacy definition or analysis



(d) Privacy-preserving methods



(e) Types of data problems



(f) Data mining algorithm

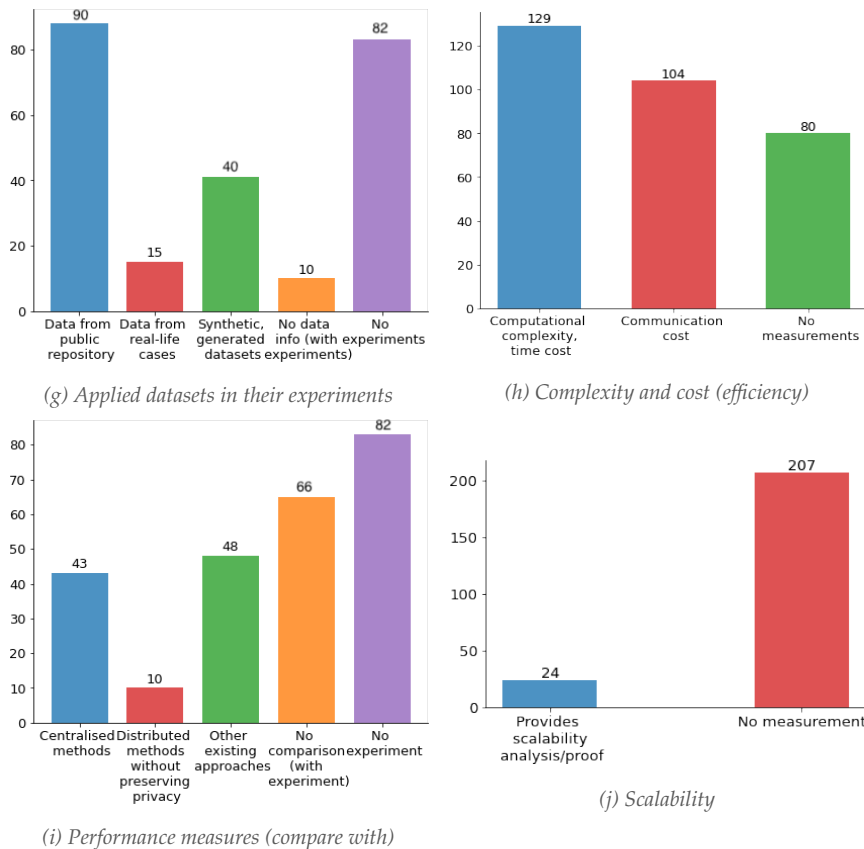


Figure 2.5: Bar charts of presenting review results using 10-factor evaluation criteria. Papers can cover one or more items in the factors except Privacy Definition/Analysis and Scalability.

Applied datasets in their experiments. From the selected studies, we identified the datasets that were applied in their experiments, measurement of complexity and cost, and performance on accuracy and scalability. We found 90 studies used datasets from public repositories, while 40 studies generated synthetic datasets to conduct their experiments. It is noteworthy that only 15 papers applied real-world datasets in practical use cases. Furthermore, it is remarkable to find that 82 papers proposed new methods by only presenting mathematical theories without any experiments, while 10 papers conducted experiments but did not provide any information about the datasets.

Complexity and cost. To prove the efficiency of proposed methods, 129 papers calculated computational complexity and/or time cost, while 104 papers reported communication cost of their approaches. Among them, 85

papers measured both computational complexity/time cost and communication cost. However, one third of (80) reviewed papers did not have any measurement of computation, running time, or communication cost.

Accuracy performance. We found 82 reviewed papers were lacking in evaluating accuracy performance of their methods because no experiments were conducted in these studies. In the rest of the papers, 43 papers proved their PPDDM methods can achieve comparable accuracy as the centralised data mining methods, while 48 studies proved their methods exceeded other existing PPDDM methods or achieved the same accuracy with higher efficiency. A small proportion of (10) studies proved their privacy-preserving models have comparable performance on learning partitioned data as the non-privacy-preserving models. Lastly, 66 papers conducted experiments but did not compare with any other methods or situations.

Scalability. The last factor - scalability - shows 10% papers proved or analyzed the scalability of their proposed methods. The majority of papers either only provided very brief statements in the discussion and future work section of the paper, or did not consider the scalability challenge.

2.4.3 Referencing Relationship among Selected Papers

We investigated how selected papers influence each other based on their references and citations. We extracted text from reference sections of all selected studies and recognized titles and authors from the text. As DOIs are not available in the reference section of all papers, only titles and authors were used to recognize different studies. Figure 2.6 illustrates the citation network, where papers are represented as nodes, and citing relations are represented as edges. The size of nodes are proportional to the number of citations among the 231 papers. Papers [38, 70, 71] are most cited, with 1354, 1320, and 875 citations respectively (until 2021 Feb).

Table 2.1 lists the attributes of the most cited articles. Semi-honest behavior is the most common assumption, while none of these influential papers addressed malicious adversarial behavior. 3 out of 18 studies considered a third party. Two papers [72, 47] took all possible data distribution situations (horizontally, vertically, and arbitrarily partitioned data) into account. Horizontally and vertically partitioned data problems have been covered with a good balance. Although the vertically partitioned data problem is more complicated than the horizontally one [73, 74], our review indicates that they have been developed at the same pace.

A similar balance is apparent in the types of problems as well. Seven papers focused on solving a classification problem by using SVM, decision tree,

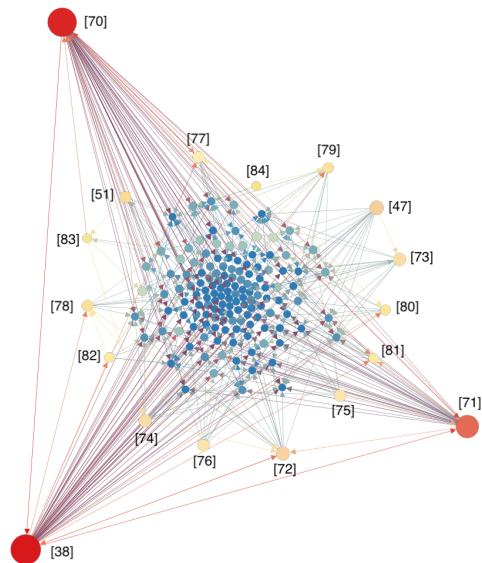


Figure 2.6: Citation network among the selected papers. Papers are presented as nodes, while the citing relations are edges. The size of nodes are proportional to the number of citations.

bayesian networks, while 8 papers looked at clustering problems particularly at K-means, Expectation Maximization algorithms (EM), Local Outlier Factor (LOF) algorithm. Association rule mining problem has fewer influential papers, but the top 2 influential papers [38, 70] both focused on this problem. In contrast to the balance in the types of problems, privacy-preserving solutions from the influential papers are completely dominated by SMC. 16 out of 18 influential papers covered SMC [75, 76] combined SMC with homomorphic encryption, while [47, 77, 78] combined it with structuring local and global data miners. More than half of existing studies in our review applied SMC as the major privacy-preserving method.

It is notable that 12 out of 18 studies did not conduct experiments, but they provided explicit privacy/security analyses and costs measurements instead. These privacy/security analyses have been presented in different ways, but the main objectives were similar. All influential papers described what information their approaches can protect, what information have to be disclosed, and what potential risks, problems or troubles might exist. Moreover, their computational complexity and communication costs of their approaches were clearly presented as one of the evaluation parameters. Hence, the described performance evaluation on privacy and efficiency may be the reasons why these papers are often cited.

Table 2.1: Review results for the 18 most cited papers in this review. (Hor: Horizontally partitioned data; Ver: Vertically partitioned data; Arb: Arbitrarily partitioned data. PP method: Privacy-preserving methods; LG: Local learning and global integration. CLF: Classification; CLS: Clustering; ARM: Association Rule Mining. CPT: Computational cost; COM: Communication cost.)

Ref	User scenario		Data distribution			Privacy/ security analysis	PP method		Type of problems			Cost	
	Semi honest	Third party	Hor	Ver	Arb		SMC	LG	CLF	CLS	ARM*	CPT	COM
[38]	✓			✓		✓				✓			✓
[70]	✓		✓			✓	✓			✓			✓
[71]	✓			✓		✓	✓		✓				✓
[47]	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓
[79]	✓			✓		✓	✓	✓	✓			✓	✓
[77]				✓		✓	✓	✓					
[72]	✓	✓	✓	✓	✓	✓	✓		✓			✓	✓
[80]	✓		✓			✓	✓		✓				
[81]			✓			✓		✓					✓
[51]	✓			✓		✓		✓				✓	✓
[76]				✓		✓		✓			✓		✓
[82]	✓		✓	✓		✓		Perturbation	✓	✓		✓	✓
[78]	✓		✓			✓	✓	✓	✓			✓	✓
[83]	✓		✓			✓	✓	✓		✓			
[84]	✓	✓	✓			✓	✓	✓		✓		✓	✓
[85]				✓		✓	✓	✓		Probabilistic graph		✓	✓
[75]	✓		✓			✓	✓	✓	✓			✓	✓
[86]	✓			✓		✓	✓	✓	✓	✓		✓	✓

2.5 Discussion

PPDDM has been rapidly developing through active research programs across different scientific communities including data mining and machine learning, mathematics and statistics, cryptography, and data management. The total number of publications in this domain has dramatically increased in the last 20 years. Many of the studies included promising results in the efficiency and accuracy of their models in an experimental environment. These promising experimental results helped move the field forward towards practical applications. In the past five years, use cases have been developed in healthcare [7, 87, 88, 89], finance [90], and technology companies [14, 91, 92] to examine different PPDDM methods. Participation of industry partners accelerates the transformation of PPDDM theoretical methods to practical applications. The existing PPDDM methods have been well-developed to solve a wide range of data problems (e.g., classification, clustering, association rule mining) using various data mining algorithms. To achieve the goal of PPDDM methods in practical studies, methods that will preserve privacy require legal, ethical, and social scholars in addition to scientific and technical experts. Successful implementation of PPDDM needs a joint effort from researchers with diverse backgrounds.

2.5.1 Inadequate definition and measurement of privacy

There are some challenges hindering PPDDM methods to be further developed and widely applied in practice. One of the key issues is the lack of the definition and measurement of (information) privacy. The meaning and operational definition of privacy is commonly ambiguous and subjective in the selected papers. It is not sufficiently expressed by the papers what privacy means to them, and what their proposed approaches can preserve. The three most common definitions of privacy preservation in the selected papers are 1) not revealing sensitive information; 2) not revealing private information; 3) not revealing raw data. However, it is unclear if sensitive information or private information or raw data is equal to personal information privacy. To understand personal information privacy from a legal and ethical perspective, it is the right of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively [93, 94, 95]. Similarly, privacy is seen as the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others [96]. In relation to controlling and protecting privacy, two definitions from legal literature state Privacy, as a whole or in part, represents the control of transactions between person(s) and other(s), the ultimate aim of which is to enhance autonomy and/or to

minimize vulnerability [97] and Privacy is to protect personal data and information related to a communication entity to be collected from other entities that are not authorized [98].

According to privacy definitions above, any information about a person can be considered as privacy regardless of its sensitivity, originality, and transformation. It is the data subject that determines what data is private. For instance, a data subject might consider their state of mental health more private than their date of birth. However, existing PPDDM methods have not yet addressed different privacy requirements from each data subject. All data elements have equal treatment for all data subjects. This might cause insufficient privacy preservation for some data elements and data subjects, while over-protection for the others. To personalize the privacy preservation, Xiao and Tao [99] proposed a new generalization framework using personalized anonymity that data subjects can specify the degree of privacy protection for her/his data elements. In the study, Xiao and Tao [99] assume: 1) data subjects can easily set/change their privacy requirements with data parties, 2) data subjects are knowledgeable about the benefits and consequences of setting different degrees of privacy. This method is only applicable when the data is centralized. In the partitioned data scenario, there is no platform yet facilitating data subjects to customize privacy requirements for each data element across multiple parties. Second, privacy requirements can be satisfied when using one single data source. However, analyzing an amount of partitioned data from multiple sources increases risk of privacy violation. As indicated by the 2020 European Commission White Paper on Artificial Intelligence [100], data about persons can be re-identified through the analysis of large amounts of other non-private data.

2.5.2 Ambiguity between privacy and security

Another ambiguity lies in the difference between (information) privacy and (information) security. Different from privacy, security has an explicit definition and measurement from the cryptography domain, separating the problem into semantic security and technical security [59]. Semantic security is a computational-complexity analogue of Shannon's definition of perfect privacy (which requires that the ciphertext yield no information regarding the plaintext). Technical security is the infeasibility of distinguishing between encryptions of a given pair of messages. Generally speaking, security focuses on maximally protecting information/data from malicious attacks and stealing data. Satisfying security requirements is not always sufficient for addressing privacy issues [101]. However, in the majority of the reviewed papers, the difference between security and privacy is not clearly stated. For example,

some studies defined the data privacy but evaluated the methods by conducting security analysis [102, 103, 104]. Certain approaches guarantee that the data used for the analyses remain unknown to other parties through secure computation. However, this does not mean that the resulting output from the analyses is equally privacy-preserving [13, 101, 105]. The output can reveal information about the person so that the privacy is still not preserved according to the privacy definition we discussed above. For instance, the outcome of the analysis might portray a harmful profile for individuals sharing certain characteristics. Some essential problems are not taken into consideration, such as how much data or information will be revealed by the output although the output is computed securely [89], whether the models and algorithms are harmless to the data party or individuals, does the purpose of formula or function satisfy the legal and ethical concerns [106, 107]. A typical example is building a decision tree on vertically partitioned data in a privacy-preserving way. The decision tree model can be securely and correctly built up. However, to some extent, the decision tree, as an output, leaks information about the input data [108]. Decision tree algorithm splits nodes based on attributes or features, while the splitting decision is dictated by the data. When the final decision tree is completed, the leaf nodes in the tree might reveal some information about the input data such as class counts. Therefore, releasing the final decision tree to all participating parties could potentially breach privacy.

Providing an applicable privacy description is significant to any PPDDM studies. What data or information should be preserved from mining can be influenced by different legal restrictions, ethical concerns, organizational regulations, personal preference, and application domains. Instead of generalizing the solution of a specific scheme to all situations, it is more reasonable to make a precise statement on the specific scenario to address. Therefore, the authors could provide a clear description to readers about what privacy means to them, and in which situation the proposed approach is privacy preserving by answering the following questions:

1. *What is the operational definition of privacy-preservation for the work?*
2. *Which data are deemed sensitive or require protection, and why?*
3. *What computational operation is intended to preserve privacy, and where does it fail?*
4. *What is the role or responsibility of each actor (e.g., data collector, data holder, data publisher, data analyst) in the scenario?*

2.5.3 Inadequate experiments and practical use cases

Our review result shows half of the reviewed papers did not provide any experiments to evaluate their methods, and as such there were no reports of accuracy, efficiency, and scalability in these papers. This is probably one of the gaps between the theoretical research and practical use cases in this domain. Solutions based on theory might not solve real world problems. In our review, only a few papers applied real-world use cases to evaluate their methods. It reflects a fact in this domain that many solutions have been proposed by researchers, but only a few of them were implemented in practice. Without experimenting on real data, the proposed approaches might neglect essential problems such as sparse or biased datasets [54, 109], or record linkage problems in vertically partitioned data [110, 111, 112]. Future research in PPDDM should consider conducting experiments using real-world datasets and provide adequate information about the experiments. Meanwhile, we observed most real-life use cases to examine existing PPDDM approaches from the healthcare domain [87, 106, 89]. We suggest researchers apply the PPDDM methods to practical cases also in other research domains such as social sciences. In addition to developing new theories, implementing and improving existing approaches in practice can also make a meaningful contribution to the PPDDM domain.

Nevertheless, these findings were observed in the light of limitations in our search strategy, which are elaborated in section 2.5.6. This review did not specifically search for follow-up studies of reviewed papers. A possible effect is that papers which lack experiments might present their experiments in the follow-up studies, and might introduce selection bias towards the low number of practical experiments. However, we would argue that our search strategy would have found these papers if proper terminology was used.

2.5.4 Challenge of linking data in vertically partitioned data scenario

The accurate linking of entities across distributed datasets is of crucial importance in vertically partitioned data mining. Data parties must link their data and/or order them in an identical manner prior to data analysis. However, most papers assume this correspondence between data entities (records) exist by default. Matching data entities from multiple datasets can be error-prone particularly where the use of direct identifiers - even encrypted - are prevented by law, as is the case in the use of the national Citizen Service Number (Burgerservicenummer) in the Netherlands [113]. Sharing such identifiers compromises privacy as the sole information that a data subject is known to another data entity might be sensitive. Furthermore, one often assumes

that records can be linked by doing exact matching on this unique identifier. However, exact matching can be very difficult due to the unstable and incorrect identifiers. Winkler and Schnell showed that 25% true matches would have been missed by exact matching in a census operation [114, 115]. In another case, two data parties do not share the unique identifiers but have some features in common. As an alternative solution, two parties can match the data entities based on their common features. The matching accuracy will be affected by the correctness, completeness, and updating promptness of these common features from both data parties. In addition, privacy needs to be preserved in the matching procedure. Some efficient and privacy-compliant algorithms for the field of privacy-preserving entity matching have been developed [116, 117, 118, 119] in the past 10 years.

2.5.5 A recommendation list of key parameters for PPDDM studies

It is challenging to compare similar PPDDM methods where there is a lack of key parameters presented. For instance, approaches which are designed for semi-honest parties might not be comparable with the approaches aiming to handle malicious behavior. The privacy-preserving methods for semi-honest parties will fail if involved parties show malicious behavior such as manipulating the input or output or completely aborting the protocol. Thus, the allowed adversarial behavior of participating parties is essential to be explicitly stated in the PPDDM papers. To consider all key parameters in PPDDM techniques, we provide a list of recommendations for the reporting of studies proposing new PPDDM methods or improving existing PPDDM methods as Table 2.2 shows. The recommendations detail the key parameters that should be described in each section of the paper of PPDDM. The factors in Table 2.2 refer to the 10 factors in the evaluation criteria which were discussed in the Methodology Section.

Section	Factor	Recommendations
Title and abstract		
Title and key-words	2,7	Identify the study as developing new or improving existing PPDDM algorithms to solve which data problem by using which type of partitioned data in a privacy-preserving manner
Abstract	1,2,4,6,7	Summarize the problems, objectives covering assumed adversarial behavior of data parties, data partitioning, brief description about privacy-preserving method, data mining algorithms, and applied dataset in the experiments.
Introduction		

Table 2.2 continued from previous page

Section	Factor	Recommendations
Problem statement and background	2,3,5,6	Describe how data partitioned in which domain are considered by this study, what privacy issues are involved in that domain, which data mining algorithm is studied to solve what problems. Additionally, the number of participating parties and if all parties or only some parties have the target class should be also covered by this section.
Objectives and study design	1,3,4,7	Specify the objectives and study design include what level of privacy (or information leakage) is preserved against what adversarial behavior, applied privacy-preserving methods, evaluation criteria (for accuracy, efficiency, and privacy level), applied datasets in the experiments.
Methods		
Method design	4,5,6	Clearly explain which privacy-preserving methods are applied including the specific protocols/structures, proofs of preserving information leakage. Then, describe how certain data mining algorithms are adapted to combine with privacy-preserving methods, what information is communicated among parties, and complexity in different scenarios such as using categorical or numerical data, or involving different numbers of data parties. Lastly, make the code publicly available so that other researchers can reproduce the work.
Data	7	Describe data sources (and where and how other researchers can request the same dataset), the type and size of the datasets, basic description about data, what the target features/attributes are, missing values, and other basic information about the datasets.
Data analysis design	5,6	If real-life datasets are applied in the study, this subsection should describe the pre-processing of features/attributes (such as normalization, re-sampling), data analysis algorithms, parameter setting, and so on with reference to other comparable studies.
Experiment design	7	Describe how the datasets are partitioned (both feature-wise and instances-wise), how data parties communicate/-transfer files, what validation is used, and what machine(such as CPU, memory) and software(versions) are used to do the experiments. In addition, experiments should be set up to compare with other existing PPDDM methods, or compare with privacy-preserving centralised data mining methods, or compare with distributed data mining methods without preserving privacy.
Evaluation design	8,9,10	Describe the evaluations of accuracy, efficiency (computational complexity, time cost on computation and communication among parties), privacy/security (such as information disclosure measurement)
Result		
Discovery from datasets	7	If real-life datasets are applied in the study, this subsection should describe what new knowledge was obtained from their analysis

Table 2.2 continued from previous page

Section	Factor	Recommendations
Model performance	8	Present the performance measures such as accuracy scores of the proposed models in comparison with other existing PPDDM methods, or privacy-preserving centralised data mining methods, or distributed data mining methods without preserving privacy. Performance will be presented based on the evaluation criteria which was described in the methods section.
Privacy and/or security analysis	9	Provide sufficient privacy/security analysis based on the assumed adversarial behavior (semi-honest or malicious). Describe what information is exchanged among parties, what can be learnt from the exchanged information, if the models as a final outcome can cause information leakage, what the potential risks exist during the training process or in the final model.
Scalability analysis	10	Present the computation complexity and time consumption of the methods and describe what the volume (number of instances) and variety (number of features/attributes) of data can be handled by the proposed methods
Discussion		
Limitations	/	Discuss any limitations of proposed methods such as special cases where the methods are not applicable or certain assumptions which are not common in practice.
Interpretation	/	If real-life datasets are applied in the study, this subsection should discuss the findings with reference to any other validation data from other studies. Then, interpret the model performance on accuracy, efficiency, feasibility in practice, strengths and weaknesses with reference to other existing PPDDM methods.
Implementation	/	Discuss what other resources, paperwork, or supports are needed to implement the proposed methods, what potential challenges or risks will appear if apply the methods on real-life data.

Table 2.2: A list of recommendations for reporting PPDDM studies

2.5.6 Potential limitations

The findings of this review have to be seen in light of some potential limitations. First, the 231 reviewed studies were searched from only 6 digital bibliographic databases (IEEE Xplore Digital Library, ACM Digital Library, Science Direct, ISI Web of Science, SpringerLink, and PubMed) and must be peer-reviewed publications. Some relevant studies may be missed in this review because they were not findable in these 6 bibliographic databases during searching. Studies that have not been peer-reviewed such as relevant articles published on arXiv.org⁷ were excluded.

⁷arXiv - a free distribution service and an open-access archive: <https://arxiv.org/>

Second, we did not apply an iterative ‘snowballing’ approach to further identify more relevant studies [120]. ‘Snowballing’ searching includes 1) reference tracking which identifies relevant studies from the reference lists of the primarily selected papers, 2) citation tracking which identifies relevant articles that cite primarily selected papers. We decided not to apply ‘snowballing’ approach is because it may introduce a bias in favour of what authors think is relevant to their narrative [121]. Contrary, omitting the ‘snowballing’ approach results in omitting follow-up studies of the reviewed papers. We decided to choose the latter approach, as we deemed our search criteria to be broad enough to cover follow-up studies. We have found several follow-up papers, where these papers present an extension of their existing methods to: 1) solve other data partitioning problems [77, 78]; 2) apply to more advanced data analysis algorithms [122, 123]; 3) to include more complicated user scenarios [124, 125]; 4) to conduct more experiments by using real-life datasets [7, 87, 112, 107].

Moreover, due to the scope of this review (providing a general overview of existing PPDDM methods and identifying outstanding challenges), more details of some privacy-preserving methods were not extensively discussed. For instance, in the category of ‘local learning and global integration’, multiple different methods can be applied to integrate the local miner (model) into a global miner (model) such as stacked generalization [126] and meta-learning[127]. In our belief this field warrants a separate in-depth review. Additionally, it has been well-recognized that there is an important trade-off between leakage of information and effectiveness or efficiency of learning in PPDDM technologies [14, 27, 128, 90]. In practice, it is crucial to balance this trade-off depending on the specific use cases, the purposes of the data analysis, and the urgency of the problems. Although we included the privacy and efficiency factors in our review, we did not further investigate how each method weights the trade-off between them. For example, we did not measure how much and in which way information loss was tolerated to increase efficiency. We believe this specific trade-off issue between privacy (information leakage) and learning performance (effectiveness or efficiency) deserves further investigation.

2.6 Conclusion

Privacy-preserving distributed data mining (PPDDM) techniques consider the issue of executing data mining algorithms on private, sensitive, and/or confidential data from multiple data parties while maintaining privacy. This review presented a comprehensive overview of current PPDDM methods to

help researchers better understand the development of this domain and assist practitioners to select the suitable solutions for their practical cases.

In this review, we discovered there is a lack of standard criteria for evaluating new PPDDM techniques. The previous studies applied a variety of different evaluation methods, which brings challenges to objectively comparing existing PPDDM techniques. Therefore, an comprehensive evaluation criteria was proposed in this review including 10 key factors - adversarial behavior of data parties, data partitioning, experiment datasets, privacy/security analysis, privacy-preserving methods, data mining problems, analysis algorithms, complexity and cost, performance measures, and scalability to assess 231 recent studies published between 2000 to 2020 (August). We highlighted the characteristics of the 18 most cited studies and analyzed their influence on other studies in the field. Furthermore, a variety of definitions of privacy and distinguishment between information privacy and information security in the PPDDM field were discussed in this review, followed by some suggestions of making applicable privacy descriptions for new PPDDM methods. Finally, we also provided a list of recommendations for future research such as explicitly describing the privacy aspect under consideration, and evaluating new approaches using real-life data to narrow the gap between theoretical solutions and practical applications.

References

- [1] Geoff Dougherty. *Digital image processing for medical applications*. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511609657.
- [2] Hanaa Elshazly, Ahmed Taher Azar, Abeer El-Korany, and Aboul Ella Hassanien. "Hybrid system for lymphatic diseases diagnosis". In: *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2013, pp. 343–347. DOI: 10.1109/ICACCI.2013.6637195.
- [3] EA Clarke. "What is preventive medicine?" In: *Canadian Family Physician* 20.11 (1974), p. 65. DOI: 10.1145/772862.772867.
- [4] Jacques S Beckmann and Daniel Lew. "Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities". In: *Genome medicine* 8.1 (2016), pp. 1–11. DOI: 10.1186/s13073-016-0388-7.
- [5] Shawn Dolley. "Big data's role in precision public health". In: *Frontiers in public health* 6 (2018), p. 68. DOI: 10.3389/fpubh.2018.00068.

-
- [6] Raphael B Stricker and Lorraine Johnson. “Lyme disease: the promise of Big Data, companion diagnostics and precision medicine”. In: *Infection and drug resistance* 9 (2016), p. 215. DOI: 10.2147/IDR.S114770.
- [7] Arthur Jochems, Timo M Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. “Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital—a real life proof of concept”. In: *Radiotherapy and Oncology* 121.3 (2016), pp. 459–467. DOI: 10.1016/j.radonc.2016.10.002.
- [8] Commission on Social Determinants of Health. *Closing the gap in a generation : health equity through action on the social determinants of health : final report : executive summary*. World Health Organization, 2008, p. 33. URL: https://www.who.int/social_determinants/final_report/csdh-finalreport_2008.pdf (visited on 01/11/2022).
- [9] Jessica S Ancker, Min-Hyung Kim, Yiye Zhang, Yongkang Zhang, and Jyotishman Pathak. “The potential value of social determinants of health in predicting health outcomes”. In: *Journal of the American Medical Informatics Association* 25.8 (2018), pp. 1109–1110. DOI: 10.1093/jamia/ocy061.
- [10] Suranga N Kasthurirathne, Joshua R Vest, Nir Menachemi, Paul K Halverson, and Shaun J Grannis. “Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services”. In: *Journal of the American Medical Informatics Association* 25.1 (2017), pp. 47–53. DOI: 10.1093/jamia/ocx130.
- [11] Olga Ohrimenko, Felix Schuster, Cédric Fournet, Aastha Mehta, Sebastian Nowozin, Kapil Vaswani, and Manuel Costa. “Oblivious multi-party machine learning on trusted processors”. In: *25th {USENIX} Security Symposium ({USENIX} Security 16)*. USENIX Association, 2016, pp. 619–636. URL: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/paper.pdf>.
- [12] Ruili Wang, Wanting Ji, Mingzhe Liu, Xun Wang, Jian Weng, Song Deng, Suying Gao, and Chang-an Yuan. “Review on mining data from multiple data sources”. In: *Pattern Recognition Letters* 109 (2018), pp. 120–128. DOI: 10.1016/j.patrec.2018.01.013.
- [13] Yehida Lindell. “Secure multiparty computation for privacy preserving data mining”. In: *Encyclopedia of Data Warehousing and Mining*. IGI global, 2005, pp. 1005–1009. DOI: 10.4018/978-1-59140-557-3.ch189.

- [14] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. “Federated learning: Strategies for improving communication efficiency”. In: *arXiv preprint arXiv:1610.05492* (2016).
- [15] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 1273–1282. URL: <http://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf>.
- [16] Sheng Shen, Tianqing Zhu, Di Wu, Wei Wang, and Wanlei Zhou. “From distributed machine learning to federated learning: In the view of data privacy and security”. In: *Concurrency and Computation: Practice and Experience* (2020). DOI: 10.1002/cpe.6002.
- [17] Yousra Abdul Alsaheb S Aldeen, Mazleena Salleh, and Mohammad Abdur Razzaque. “A comprehensive review on privacy preserving data mining”. In: *SpringerPlus* 4.1 (2015), pp. 1–36. DOI: 10.1186/2193-1801-4-1.
- [18] Elisa Bertino, Dan Lin, and Wei Jiang. “A survey of quantification of privacy preserving data mining algorithms”. In: *Privacy-preserving data mining*. Springer, 2008, pp. 183–205. DOI: 10.1007/978-0-387-70992-5_8.
- [19] Alpa Shah and Ravi Gulati. “Privacy preserving data mining: techniques, classification and implications-a survey”. In: *Int. J. Comput. Appl* 137.12 (2016), pp. 40–46. DOI: 10.5120/IJCA2016909006.
- [20] Vassilios S Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. “State-of-the-art in privacy preserving data mining”. In: *ACM Sigmod Record* 33.1 (2004), pp. 50–57. DOI: 10.1145/974121.974131.
- [21] Hina Vaghashia and Amit Ganatra. “A survey: privacy preservation techniques in data mining”. In: *International Journal of Computer Applications* 119.4 (2015). DOI: 10.5120/21056-3704.
- [22] Suchitra Shelke and Babita Bhagat. “Techniques for Privacy Preservation in Data Mining”. In: *International Journal of Engineering Research* 4.10 (2015). DOI: 10.17577/ijertv4is100473.
- [23] Elisa Bertino and Igor Nai Fovino. “Information driven evaluation of data hiding algorithms”. In: *International Conference on Data Warehousing and Knowledge Discovery*. Springer, 2005, pp. 418–427. DOI: 10.1007/11546849_41.

-
- [24] Sam Fletcher and Md Zahidul Islam. "Measuring information quality for privacy preserving data mining". In: *International Journal of Computer Theory and Engineering* 7.1 (2015), p. 21. DOI: 10.7763/IJCTE.2015.V7.924.
- [25] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y Zhu. "Tools for privacy preserving distributed data mining". In: *ACM Sigkdd Explorations Newsletter* 4.2 (2002), pp. 28–34.
- [26] Jaideep Vaidya. "A survey of privacy-preserving methods across vertically partitioned data". In: *Privacy-preserving data mining*. Springer, 2008, pp. 337–358. DOI: 10.1007/978-0-387-70992-5_14.
- [27] Ricardo Mendes and João P Vilela. "Privacy-preserving data mining: methods, metrics, and applications". In: *IEEE Access* 5 (2017), pp. 10562–10582. DOI: 10.1109/ACCESS.2017.2706947.
- [28] Shimon Even, Oded Goldreich, and Abraham Lempel. "A randomized protocol for signing contracts". In: *Communications of the ACM* 28.6 (1985), pp. 637–647. DOI: 10.1145/3812.3818.
- [29] Ronald L Rivest, Len Adleman, Michael L Dertouzos, et al. "On data banks and privacy homomorphisms". In: *Foundations of secure computation* 4.11 (1978), pp. 169–180. URL: <http://people.csail.mit.edu/rivest/RivestAdlemanDertouzos-OnDataBanksAndPrivacyHomomorphisms.pdf>.
- [30] Monique Ogburn, Claude Turner, and Pushkar Dahal. "Homomorphic encryption". In: *Procedia Computer Science* 20 (2013), pp. 502–509. DOI: 10.1016/j.procs.2013.09.310.
- [31] Craig Gentry et al. *A fully homomorphic encryption scheme*. Vol. 20. Stanford university Stanford, 2009. URL: <https://crypto.stanford.edu/craig/craig-thesis.pdf>.
- [32] Andrew Chi-Chih Yao. "How to generate and exchange secrets". In: *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*. IEEE, 1986, pp. 162–167. DOI: 10.1109/SFCS.1986.25.
- [33] Silvio Micali, Oded Goldreich, and Avi Wigderson. "How to play any mental game". In: *Proceedings of the Nineteenth ACM Symp. on Theory of Computing, STOC*. Association for Computing Machinery, 1987, pp. 218–229. DOI: 10.1145/28395.28420.
- [34] Donald Beaver, Silvio Micali, and Phillip Rogaway. "The round complexity of secure protocols". In: *Proceedings of the twenty-second annual ACM symposium on Theory of computing*. Association for Computing Machinery, 1990, pp. 503–513. DOI: 10.161145/100216.100287.

- [35] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. "Completeness theorems for non-cryptographic fault-tolerant distributed computation". In: *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*. Association for Computing Machinery, 2019, pp. 351–371. DOI: 10.1145/3335741.3335756.
- [36] Stephen Pohlig and Martin Hellman. "An improved algorithm for computing logarithms over GF (p) and its cryptographic significance (corresp.)" In: *IEEE Transactions on information Theory* 24.1 (1978), pp. 106–110. DOI: 10.1109/TIT.1978.1055817.
- [37] Kaibin Huang and Raylin Tso. "A commutative encryption scheme based on ElGamal encryption". In: *2012 International Conference on Information Security and Intelligent Control*. IEEE, 2012, pp. 156–159. DOI: 10.1109/ISIC.2012.6449730.
- [38] Jaideep Vaidya and Chris Clifton. "Privacy preserving association rule mining in vertically partitioned data". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. Association for Computing Machinery, 2002, pp. 639–644. DOI: 10.1145/775047.775142.
- [39] Mikhail J Atallah and Wenliang Du. "Secure multi-party computational geometry". In: *Workshop on Algorithms and Data Structures*. Springer, 2001, pp. 165–179. DOI: 10.1007/3-540-44634-6.16.
- [40] Ioannis Ioannidis, Ananth Grama, and Mikhail Atallah. "A secure protocol for computing dot-products in clustered and distributed environments". In: *Proceedings International Conference on Parallel Processing*. IEEE, 2002, pp. 379–384. DOI: 10.1109/ICPP.2002.1040894.
- [41] Rick L Wilson and Peter A Rosen. "Protecting data through perturbation techniques: The impact on knowledge discovery in databases". In: *Journal of Database Management (JDM)* 14.2 (2003), pp. 14–26. DOI: 10.4018/jdm.2003040102.
- [42] Nabil R Adam and John C Worthmann. "Security-control methods for statistical databases: a comparative study". In: *ACM Computing Surveys (CSUR)* 21.4 (1989), pp. 515–556. DOI: 10.1145/76894.76895.
- [43] Tore Dalenius and Steven P Reiss. "Data-swapping: A technique for disclosure control". In: *Journal of statistical planning and inference* 6.1 (1982), pp. 73–85. DOI: 10.1016/0378-3758(82)90058-1.
- [44] Stephen E Fienberg and Julie McIntyre. "Data swapping: Variations on a theme by dalenius and reiss". In: *International Workshop on Privacy in Statistical Databases*. Springer, 2004, pp. 14–29. DOI: 10.1007/978-3-540-25955-8.2.

-
- [45] Diego Peteiro-Barral and Bertha Guijarro-Berdiñas. “A survey of methods for distributed machine learning”. In: *Progress in Artificial Intelligence* 2.1 (2013), pp. 1–11. DOI: 10.1007/s13748-012-0035-5.
- [46] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. “A survey on distributed machine learning”. In: *ACM Computing Surveys (CSUR)* 53.2 (2020), pp. 1–33. DOI: DOI:10.1145/3377454.
- [47] Jaideep Vaidya, Hwanjo Yu, and Xiaoqian Jiang. “Privacy-preserving SVM classification”. In: *Knowledge and Information Systems* 14.2 (2008), pp. 161–178. DOI: 10.1007/s10115-007-0073-7.
- [48] Yunmei Lu, Piyaphol Phoungphol, and Yanqing Zhang. “Privacy aware non-linear support vector machine for multi-source big data”. In: *2014 IEEE 13th international conference on trust, security and privacy in computing and communications*. IEEE. 2014, pp. 783–789. DOI: 10.1109/TrustCom.2014.103.
- [49] Dashan Gao, Yang Liu, Anbu Huang, Ce Ju, Han Yu, and Qiang Yang. “Privacy-preserving heterogeneous federated transfer learning”. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 2552–2559. DOI: 10.1109/BigData47090.2019.9005992.
- [50] Anup Tuladhar, Sascha Gill, Zahinoor Ismail, Nils D Forkert, Alzheimer’s Disease Neuroimaging Initiative, et al. “Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling”. In: *Journal of biomedical informatics* 106 (2020), p. 103424. DOI: 10.1016/j.jbi.2020.103424.
- [51] Jaideep Vaidya, Chris Clifton, Murat Kantarcioglu, and A Scott Patterson. “Privacy-preserving decision trees over vertically partitioned data”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2.3 (2008), pp. 1–27. DOI: 10.1145/1409620.1409624.
- [52] Eakalak Suthampan and Songrit Maneewongvatana. “Privacy preserving decision tree in multi party environment”. In: *Asia Information Retrieval Symposium*. Springer. 2005, pp. 727–732. DOI: 10.1007/11562382_75.
- [53] Weiwei Fang and Bingru Yang. “Privacy preserving decision tree learning over vertically partitioned data”. In: *2008 International Conference on Computer Science and Software Engineering*. Vol. 3. IEEE. 2008, pp. 1049–1052. DOI: 10.1109/CSSE.2008.731.

- [54] Elena Czeizler, Wolfgang Wiessler, Thorben Koester, Mikko Hakala, Shahab Basiri, Petr Jordan, and Esa Kuusela. "Using federated data sources and Varian Learning Portal framework to train a neural network model for automatic organ segmentation". In: *Physica Medica* 72 (2020), pp. 39–45. DOI: 10.1016/j.ejmp.2020.03.011.
- [55] Ye Dong, Xiaojun Chen, Liyan Shen, and Dakui Wang. "EaSTFLy: Efficient and secure ternary federated learning". In: *Computers & Security* 94 (2020), p. 101824. DOI: 10.1016/j.cose.2020.101824.
- [56] Qi Zhao, Chuan Zhao, Shujie Cui, Shan Jing, and Zhenxiang Chen. "PrivateDL: Privacy-preserving collaborative deep learning against leakage from gradient sharing". In: *International Journal of Intelligent Systems* 35.8 (2020), pp. 1262–1279. DOI: 10.1002/int.22241.
- [57] Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, et al. "DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data". In: *International journal of epidemiology* 39.5 (2010), pp. 1372–1382. DOI: 10.1093/ije/dyq111.
- [58] Barbara Kitchenham. "Procedures for performing systematic reviews". In: *Keele, UK, Keele University* 33.2004 (2004), pp. 1–26. URL: http://artemisa.unicauca.edu.co/~ecaldon/docs/spi/kitchenham_2004.pdf.
- [59] Oded Goldreich. *Foundations of cryptography: volume 2, basic applications*. Cambridge university press, 2009. DOI: 10.1017/CBO9780511721656.
- [60] *UCI Machine Learning Repository*, url = <https://archive.ics.uci.edu/ml/index.php>.
- [61] Judith Wagner DeCew. *In pursuit of privacy: Law, ethics, and the rise of technology*. Cornell University Press, 1997. URL: <https://www.jstor.org/stable/10.7591/j.ctv75d3zc>.
- [62] H Jeff Smith, Tamara Dinev, and Heng Xu. "Information privacy research: an interdisciplinary review". In: *MIS quarterly* (2011), pp. 989–1015. DOI: 10.2307/41409970.
- [63] Jianwei Han, Micheline Kamber, and Jian Pei. *Data Mining Concepts and Techniques*. Third Edition. Elsevier, 2011. DOI: 10.1016/C2009-0-61819-5.
- [64] Alex A Freitas. "A survey of evolutionary algorithms for data mining and knowledge discovery". In: *Advances in evolutionary computing*. Springer, 2003, pp. 819–845. DOI: 10.1007/978-3-642-18965-4_33.

-
- [65] Johannes Fürnkranz and Peter A Flach. "An analysis of rule evaluation metrics". In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003, pp. 202–209. URL: <https://www.aaai.org/Papers/ICML/2003/ICML03-029.pdf>.
- [66] Mohammad Hossin and MN Sulaiman. "A review on evaluation metrics for data classification evaluations". In: *International Journal of Data Mining & Knowledge Management Process* 5.2 (2015), p. 1. DOI: 10.5121/ijdkp.2015.5201.
- [67] Julio-Omar Palacio-Niño and Fernando Berzal. "Evaluation metrics for unsupervised learning algorithms". In: *arXiv preprint arXiv:1905.05667* (2019).
- [68] Alexei Botchkarev. "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology". In: *arXiv preprint arXiv:1809.03006* (2018).
- [69] *Gephi - The Open Graph Viz Platform*, url = <https://gephi.org/>.
- [70] Murat Kantarcioglu and Chris Clifton. "Privacy-preserving distributed mining of association rules on horizontally partitioned data". In: *IEEE transactions on knowledge and data engineering* 16.9 (2004), pp. 1026–1037. DOI: 10.1109/TKDE.2004.45.
- [71] Jaideep Vaidya and Chris Clifton. "Privacy-preserving k-means clustering over vertically partitioned data". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. Association for Computing Machinery, 2003, pp. 206–215. DOI: 10.1145/956750.956776.
- [72] Geetha Jagannathan and Rebecca N Wright. "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data". In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. Association for Computing Machinery, 2005, pp. 593–599. DOI: 10.1145/1081870.1081942.
- [73] John Wang. *Encyclopedia of data warehousing and mining*. iGi Global, 2005. DOI: 10.1108/14684520610675852.
- [74] Jaideep Vaidya, Christopher W Clifton, and Yu Michael Zhu. *Privacy preserving data mining*. Vol. 19. Springer Science & Business Media, 2006. DOI: 10.1007/978-0-387-29489-6.
- [75] Ming-Jun Xiao, Liu-Sheng Huang, Yong-Long Luo, and Hong Shen. "Privacy preserving id3 algorithm over horizontally partitioned data". In: *Sixth international conference on parallel and distributed computing applications and technologies (PDCAT'05)*. IEEE, 2005, pp. 239–243. DOI: 10.1109/PDCAT.2005.191.

- [76] Justin Zhan, Stan Matwin, and LiWu Chang. "Privacy-preserving collaborative association rule mining". In: *Journal of Network and Computer Applications* 30.3 (2007), pp. 1216–1227. DOI: 10.1016/j.jnca.2006.04.010.
- [77] Hwanjo Yu, Jaideep Vaidya, and Xiaoqian Jiang. "Privacy-preserving svm classification on vertically partitioned data". In: *Pacific-asia conference on knowledge discovery and data mining*. Springer, 2006, pp. 647–656. DOI: 10.1007/11731139_74.
- [78] Hwanjo Yu, Xiaoqian Jiang, and Jaideep Vaidya. "Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data". In: *Proceedings of the 2006 ACM symposium on Applied computing*. Association for Computing Machinery, 2006, pp. 603–610. DOI: 10.1145/1141277.1141415.
- [79] Rebecca Wright and Zhiqiang Yang. "Privacy-preserving bayesian network structure computation on distributed heterogeneous data". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. Association for Computing Machinery, 2004, pp. 713–718. DOI: 10.1145/1014052.1014145.
- [80] Xiaodong Lin, Chris Clifton, and Michael Zhu. "Privacy-preserving clustering with distributed EM mixture modeling". In: *Knowledge and information systems* 8.1 (2005), pp. 68–81. DOI: 10.1007/s10115-004-0148-7.
- [81] Srujana Merugu and Joydeep Ghosh. "Privacy-preserving distributed clustering using generative models". In: *Third IEEE International Conference on Data Mining*. IEEE, 2003, pp. 211–218. DOI: 10.1109/ICDM.2003.1250922.
- [82] Kun Liu, Hillol Kargupta, and Jessica Ryan. "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining". In: *IEEE Transactions on knowledge and Data Engineering* 18.1 (2005), pp. 92–106. DOI: 10.1109/TKDE.2006.14.
- [83] Mark Shaneck, Yongdae Kim, and Vipin Kumar. "Privacy preserving nearest neighbor search". In: *Machine Learning in Cyber Trust*. Springer, 2009, pp. 247–276. DOI: 10.1007/978-0-387-88735-7_10.
- [84] Ali Inan, Selim V Kaya, Yücel Saygın, Erkay Savaş, Ayça A Hintoğlu, and Albert Levi. "Privacy preserving clustering on horizontally partitioned data". In: *Data & Knowledge Engineering* 63.3 (2007), pp. 646–666. DOI: 10.1016/j.datak.2007.03.015.
- [85] Da Meng, Krishnamoorthy Sivakumar, and Hillol Kargupta. "Privacy-sensitive Bayesian network parameter learning". In: *Fourth IEEE International Conference on Data Mining (ICDM'04)*. IEEE, 2004, pp. 487–490. DOI: 10.1109/ICDM.2004.10076.

-
- [86] Boris Rozenberg and Ehud Gudes. "Association rules mining in vertically partitioned databases". In: *Data & Knowledge Engineering* 59.2 (2006), pp. 378–396. DOI: 10.1016/j.datak.2005.09.001.
- [87] Timo M. Deist et al. "Distributed learning on 20 000+ lung cancer patients – The Personal Health Train". In: *Radiotherapy and Oncology* 144 (2020), pp. 189–200. DOI: 10.1016/j.radonc.2019.11.019.
- [88] Hiroaki Kikuchi, Chika Hamanaga, Hideo Yasunaga, Hiroki Matsui, Hideki Hashimoto, and Chun-I Fan. "Privacy-preserving multiple linear regression of vertically partitioned real medical datasets". In: *Journal of Information Processing* 26 (2018), pp. 638–647. DOI: 10.2197/ipsjip.26.638.
- [89] Jin Li, Yu Tian, Yan Zhu, Tianshu Zhou, Jun Li, Kefeng Ding, and Jingsong Li. "A multicenter random forest model for effective prognosis prediction in collaborative clinical research network". In: *Artificial intelligence in medicine* 103 (2020), p. 101814. DOI: 10.1016/j.artmed.2020.101814.
- [90] Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. "Federated learning for privacy-preserving AI". In: *Communications of the ACM* 63.12 (2020), pp. 33–36. DOI: 10.1145/3387107.
- [91] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. "Protection against reconstruction and its applications in private federated learning". In: *arXiv preprint arXiv:1812.00984* (2018).
- [92] Yaochen Hu, Di Niu, Jianming Yang, and Shengping Zhou. "Fdml: A collaborative machine learning framework for distributed features". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2019, pp. 2232–2240. DOI: 10.1145/3292500.3330765.
- [93] Charles A Shoniregun, Kudakwashe Dube, and Fredrick Mtenzi. *Electronic healthcare information security*. Vol. 53. Springer Science & Business Media, 2010. DOI: 10.1007/978-0-387-84919-5.
- [94] Maria Manuela Cruz-Cunha. *Handbook of research on digital crime, cyberspace security, and information assurance*. IGI Global, 2014. DOI: 10.4018/978-1-4666-6324-4.
- [95] Fatima-Zahra Benjelloun and Ayoub Ait Lahcen. "Big data security: challenges, recommendations and solutions". In: *Web Services: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2019, pp. 25–38. DOI: 10.4018/978-1-4666-8387-7.CH014.
- [96] Alan F Westin. "Privacy and freedom". In: *Washington and Lee Law Review* 25.1 (1968), p. 166. DOI: 10.2307/3102188.

- [97] Stephen T Margulis. "Conceptions of privacy: Current status and next steps". In: *Journal of Social Issues* 33.3 (1977), pp. 5–21. DOI: 10.1111/j.1540-4560.1977.tb01879.x.
- [98] Yacine Djemaiel, Slim Rekhis, and Noureddine Boudriga. "Trustworthy Networks, Authentication, Privacy, and Security Models". In: *Handbook of Research on Wireless Security*. IGI Global, 2008, pp. 189–209. DOI: 10.4018/978-1-59904-899-4.ch014.
- [99] Xiaokui Xiao and Yufei Tao. "Personalized privacy preservation". In: *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. Association for Computing Machinery, 2006, pp. 229–240. DOI: 10.1145/1142473.1142500.
- [100] European Commission. *White paper on artificial intelligence: a european approach to excellence and trust*. Tech. rep. European Commission, 2020. URL: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.
- [101] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. "Big data privacy: a technological perspective and review". In: *Journal of Big Data* 3.1 (2016), pp. 1–25. DOI: 10.1186/s40537-016-0059-y.
- [102] Wei Jiang and Maurizio Atzori. "Secure distributed k-anonymous pattern mining". In: *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, 2006, pp. 319–329. DOI: 10.1109/ICDM.2006.140.
- [103] Lu Li, Liusheng Huang, Wei Yang, Xiaohui Yao, and An Liu. "Privacy-preserving lof outlier detection". In: *Knowledge and Information Systems* 42.3 (2015), pp. 579–597. DOI: 10.1007/s10115-013-0692-0.
- [104] Bin Gu, Zhiyuan Dang, Xiang Li, and Heng Huang. "Federated doubly stochastic kernel learning for vertically partitioned data". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2020, pp. 2483–2493. DOI: 10.1145/3394486.3403298.
- [105] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. "Secure, privacy-preserving and federated machine learning in medical imaging". In: *Nature Machine Intelligence* 2.6 (2020), pp. 305–311. DOI: 10.1038/s42256-020-0186-1.
- [106] Timo M. Deist et al. "Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT". In: *Clinical and Translational Radiation Oncology* 4 (2017), pp. 24–31. DOI: 10.1016/j.ctro.2016.12.004.

-
- [107] Chang Sun, Lianne Ippel, Johan van Soest, Birgit Wouters, Alexander Malic, Onaopepo Adekunle, Bob van den Berg, Ole Mussmann, Annemarie Koster, Carla van der Kallen, et al. "A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario." In: *MEDINFO 2019: Health and Wellbeing E-Networks for All: Proceedings of the 17th World Congress on Medical and Health Informatics*. Vol. 264. IOS Press, 2019, pp. 373–377. DOI: 10.3233/SHTI190246.
- [108] Sam Fletcher and Md Zahidul Islam. "Decision tree classification with differential privacy: A survey". In: *ACM Computing Surveys (CSUR)* 52.4 (2019), pp. 1–33. DOI: 10.1145/3337064.
- [109] Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records". In: *Journal of biomedical informatics* 99 (2019), p. 103291. DOI: 10.1016/j.jbi.2019.103291.
- [110] Alexandros Karakasidis and Vassilios S Verykios. "A sorted neighborhood approach to multidimensional privacy preserving blocking". In: *2012 IEEE 12th International Conference on Data Mining Workshops*. IEEE, 2012, pp. 937–944. DOI: 10.1109/AINA.2017.52.
- [111] Aleksandra B. Slavkovic, Yuval Nardi, and Matthew M. Tibbits. "Secure Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases". In: *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 723–728. DOI: 10.1109/ICDMW.2007.84.
- [112] Johan van Soest, Chang Sun, Ole Mussmann, Marco Puts, Bob van den Berg, Alexander Malic, Claudia van Oppen, David Townend, Andre Dekker, and Michel Dumontier. "Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data." In: *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*. Vol. 247. IOS Press, 2018, pp. 581–585. DOI: 10.3233/978-1-61499-852-5-581.
- [113] Binnenlandse Zaken en Koninkrijksrelaties. "Wet van 21 juli 2007, houdende algemene bepalingen betreffende de toekenning, het beheer en het gebruik van het burgerservicenummer (Wet algemene bepalingen burgerservicenummer)". In: (2018). URL: <https://wetten.overheid.nl/BWBR0022428/2018-07-28>.

- [114] R. Schnell. “Efficient private record linkage of very large datasets”. In: *59th World Statistics Congress of the International Statistical Institute*. International Statistical Institute, 2013. URL: <https://openaccess.city.ac.uk/id/eprint/14652/>.
- [115] William E Winkler. “Record linkage”. In: *Handbook of statistics*. Vol. 29. Elsevier, 2009, pp. 351–380. DOI: 10.1016/S0169-7161(08)00014-X.
- [116] Rob Hall and Stephen E Fienberg. “Privacy-preserving record linkage”. In: *International conference on privacy in statistical databases*. Springer, 2010, pp. 269–283. DOI: 10.1007/978-3-642-15838-4_24.
- [117] Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012. DOI: 10.1007/978-3-642-31164-2.
- [118] Dinusha Vatsalan, Peter Christen, and Vassilios S Verykios. “A taxonomy of privacy-preserving record linkage techniques”. In: *Information Systems* 38.6 (2013), pp. 946–969. DOI: 10.1016/j.is.2012.11.005.
- [119] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. “Privacy-preserving distributed linear regression on high-dimensional data”. In: *Proceedings on Privacy Enhancing Technologies* 2017.4 (2017), pp. 345–364. DOI: 10.1007/978-3-540-71701-0.
- [120] Trisha Greenhalgh and Richard Peacock. “Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources”. In: *Bmj* 331.7524 (2005), pp. 1064–1065. DOI: 10.1136/bmj.38636.593461.68.
- [121] Matthias Egger, George Davey-Smith, and Douglas Altman. *Systematic reviews in health care: meta-analysis in context*. John Wiley & Sons, 2008. DOI: 10.1002/9780470693926.
- [122] Hiroaki Kikuchi, Hideki Hashimoto, Hideo Yasunaga, and Takamichi Saito. “Scalability of Privacy-Preserving Linear Regression in Epidemiological Studies”. In: *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*. 2015, pp. 510–514. DOI: 10.1109/AINA.2015.229.
- [123] Hiroaki Kikuchi, Chika Hamanaga, Hideo Yasunaga, Hiroki Matsui, Hideki Hashimoto, and Chun-I Fan. “Privacy-preserving multiple linear regression of vertically partitioned real medical datasets”. In: vol. 26. Information Processing Society of Japan, 2018, pp. 638–647. DOI: 10.2197/ipsjip.26.638.

-
- [124] Kaleb L. Leemaqz, Sharon X. Lee, and Geoffrey J. McLachlan. "Corruption-Resistant Privacy Preserving Distributed EM Algorithm for Model-Based Clustering". In: *2017 IEEE Trustcom/BigDataSE/ICSS*. 2017, pp. 1082–1089. DOI: 10.1109/Trustcom/BigDataSE/ICSS.2017.356.
- [125] Sharon X. Lee, Kaleb L. Leemaqz, and Geoffrey J. McLachlan. "PPEM: Privacy-preserving EM learning for mixture models". In: *Concurrency and Computation: Practice and Experience* 31.24 (2019), e5208. DOI: 10.1002/cpe.5208.
- [126] David H Wolpert. "Stacked generalization". In: *Neural networks* 5.2 (1992), pp. 241–259. DOI: 10.1016/S0893-6080(05)80023-1.
- [127] Phillip K Chan, Salvatore J Stolfo, et al. "Toward parallel and distributed learning by meta-learning". In: *AAAI workshop in Knowledge Discovery in Databases*. 1993, pp. 227–240. DOI: <https://dl.acm.org/doi/10.5555/3000767.3000789#d49627527e1>.
- [128] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. "Federated learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 13.3 (2019), pp. 1–207. DOI: 10.2200/S00960ED2V01Y201910AIM043.

3

A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario

Adapted from: Chang Sun, Lianne Ippel, Johan van Soest, Birgit Wouters, Alexander Malic, Onaopepo Adekunle, Bob van den Berg, Ole Mussmann, Annemarie Koster, Carla van der Kallen, Claudia van Oppen, David Townsend, Andre Dekker, and Michel Dumontier. "A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario". In: *Studies in Health Technology and Informatics* 264 (2019), pp. 373–377. DOI: 10.3233/SHTI190246.

Abstract

It is widely anticipated that the use and analysis of health-related big data will enable further understanding and improvements in human health and wellbeing. Here, we propose an innovative infrastructure, which supports secure and privacy-preserving analysis of personal health data from multiple providers with different governance policies. Our objective is to use this infrastructure to explore the relation between Type 2 Diabetes Mellitus status and healthcare costs. Our approach involves the use of distributed machine learning to analyze vertically partitioned data from the Maastricht Study, a prospective population-based cohort study, and data from the official statistics agency of the Netherlands, Statistics Netherlands (Centraal Bureau voor de Statistiek; CBS). This project seeks an optimal solution accounting for scientific, technical, and ethical/legal challenges. We describe these challenges, our progress towards addressing them in a practical use case, and a simulation experiment.

3.1 Introduction

A growing amount of personal health data are being collected by a variety of entities, such as healthcare providers, insurance companies, and wearable device manufacturers. Use of personal health data such as health status, current and prior medications, lifestyle and behavior offers unprecedented opportunities to augment our understanding of human health and disease. This contributes to improved diagnostic accuracy and efficiency [1, 2], and facilitates the transition to preventive [3, 4] and precision medicine [5, 6, 7]. Moreover, the analysis of health data can help governments pursue effective health policies while minimizing healthcare costs. Such innovation arises from the secondary use of health data for research.

However, a major barrier to research lies in the difficulty of accessing and analyzing health data that are dispersed in both their form (e.g. medical records, consumer activity, and social media), representation (structured, semi-structured, and/or unstructured), and stewardship (who is responsible for data collection and governance?). While many methods to represent and exchange healthcare data have been developed [8], there has been a lack of focus on legal-ethical concerns such as data ownership and data stewardship as well as issues relating to privacy, security, and confidentiality [9]. Such considerations are particularly crucial when use and analysis of health data involve multiple legal entities, different data standards, a lack of detailed provenance, and unclear access authorization procedures.

Another significant challenge lies in the analysis of personal health data from multiple sources. The simplest case is where data are horizontally partitioned, such that data about different sets of individuals are located in different sites. Analyzing these distributed data is relatively well understood and reduces to combining a set of models from each site. A more challenging case is where data are vertically partitioned: different attributes about a particular individual are distributed over a set of data sources. While in the case of horizontally partitioned data analytical results are combined afterwards, this is not possible in the vertically partitioned case since none of the data providers can execute the complete analysis independently of the other providers. This is particularly challenging either when there is a legal impediment to link records across data providers with a unique identifier or when this unique identifier is unavailable. Addressing this challenge effectively requires a great level of technical sophistication to simultaneously address legal and/or privacy constraints.

Instead of centralizing the data for the analysis, one could use distributed learning methods, which operate over vertically partitioned data. In such a scenario, data-processing algorithms are sent to each site, and can only return

the results of an analysis rather than any of the original data. One such infrastructure is the Personal Health Train (PHT) [10, 11], which sends applications (the trains) containing algorithms to the data sources (the stations). The station can inspect whether the train is allowed to execute the application on (a subset of) the available data. The PHT empowers data subjects with more control (who can access the data?) and transparency (what are the trains requesting?). Hence, the PHT facilitates authorized algorithmic processing in a secure manner at multiple data sites without requiring a transfer of (original) data to a centralized location. Moreover, the PHT implements privacy-by-design in the following ways: 1) it can restrict which data elements are available to an application, 2) it can restrict the results of the analysis to only processed data, rather than original data, and 3) no data party can see the data of other parties in the network.

Here, we describe an implementation of the PHT that uses a Trusted Secure Environment (TSE) to analyze vertically partitioned data that are prepared in line with the FAIR principles (Findable, Accessible, Interoperable, Reusable) [12]. By describing data using the FAIR principles, the infrastructure becomes ambivalent to certain syntactic data structures (e.g. OHDSI, CDISC-ODM or HL7 v2/v3/FHIR), as the applications, executed at the data source, should be able to interpret different types of data structures. To test the feasibility of this infrastructure, we combine data from two independent data providers to investigate how Type 2 Diabetes Mellitus (T2DM) status affects healthcare cost. The first dataset comes from the Maastricht Study¹, an observational prospective population-based cohort study focusing on the etiology of T2DM, and the second comes from the official statistics office in the Netherlands: Statistics Netherlands² (Centraal Bureau voor de Statistiek; CBS). We present preliminary results involving simulated data and discuss the challenges and feasibility of such an infrastructure to be scalable and secure.

3.2 Methods

In this section, we describe the development of our proposed infrastructure from a scientific, technical, and legal perspective to support the workflow. Following is the description of our simulation experiment to test the usability of our infrastructure.

¹The Maastricht Study is an observational prospective population-based cohort study focusing on the etiology, pathophysiology, complications and comorbidities of T2DM. <https://www.demaastrichtstudie.nl/>

²Statistics Netherlands is a Dutch governmental institution that gathers statistical information about the Netherlands: <https://www.cbs.nl/en-gb>

3.2.1 Development Workflow

The PHT architecture has been previously used to analyze horizontally partitioned datasets [13, 14, 15, 16]. Here, we extend this work to include vertically partitioned data. While several studies discuss exchanging and analyzing vertically partitioned data [17, 18], these are largely theoretical and overlook practical challenges, e.g. legal and ethical considerations, incompatible data management standards, scalability of the infrastructure, lack of financial support to sustain such efforts, and the technical requirements of learning from vertically partitioned data. To tackle these challenges, our team has established three interlocking work packages that target: i) the scientific questions in the medical domain; ii) the ethical, legal, and societal issues; and iii) the technical aspect. These packages are highly intertwined to ensure the development of practical solutions.

3.2.2 Scientific Perspective

To develop infrastructure that is useful to scientific researchers, we have identified key research questions that the infrastructure should help answer. Answering these research questions should require the combination of sensitive (non-public) data from multiple providers. To combine data from multiple providers, a substantive set of individuals should be shared by the providers and at least some attributes of these individuals are present in both datasets to enable linking of the data records (and not necessarily by some specific individual identifier).

3.2.3 ELSI Perspective

The Ethical, Legal, and Societal Issues (ELSI) team deals with two types of challenges: i) privacy concerns that arise from the special nature of personal health data³; and ii) the legal challenges that arise from working with multiple data providers with each a distinct governance framework. Combining data from multiple parties is a relatively new phenomenon, and often not foreseen when establishing the legal framework when the data are collected. Therefore, one of the major challenges has been to facilitate this study whilst adhering to the original legal framework and defined purpose. In doing so, the ELSI team has examined the reach of the original legal basis (i.e. informed consent) and purpose for which each data provider obtained the personal data, and is further analyzing the legal basis and purpose for which secondary processing can occur. Options that are being considered include but are not limited to the route of compatible processing and the route of scientific research in the public interest. Additionally, there are a number of

limitations from the data providers themselves regarding accessing, sharing, and linking data. In addition, for this challenge, a legal framework has to be formulated in order to establish collaborations between the data providers, among themselves and with the research team. Constructing this legal framework and finding the proper legal basis for the researchers team is a valuable contribution from the ELSI team.

3.2.4 Technical Perspective

Following the PHT architecture³, we use the concepts of (FAIR data) stations⁴, rails (infrastructure) and (applications) trains. The minimal requirement of a FAIR data station is to enable execution of applications, where data providers decide whether to execute the application. These FAIR data stations are based on Semantic Web technologies such as the Resource Description Framework (RDF) [19], to convert the source data⁵, and make the converted data FAIR.

Application (train) developers (i.e., researchers) can create the application trains using Docker containers [20], which are lightweight virtual machines. The Docker container carries all required software packages to execute the application on board. These applications can for instance query data available in the data station, perform data cleaning/formatting, and execute machine learning or statistical analysis [15]. Only the results of these (analytical) applications are sent back to the application developers.

To implement the proposed infrastructure, we created three stations. Two FAIR data stations are at the Maastricht Study and at CBS. A third station was configured as a Trusted Secure Environment (TSE), containing no data by itself, however, acting as a trusted and independent entity. Additionally, we created two application trains. The first application train extracts the data from two data stations, pseudonymizes the personal identifiers, encrypts the dataset, and sends the data to the TSE station. The second application train decrypts the data and analyzes the data at the TSE. For every execution, both application trains are configured for proper encryption and security measures.

3.3 Experiment design

Prior to feeding our infrastructure with real data, we conducted a simulation experiment with two scenarios where researchers combine data from two in-

³PHT architecture: <https://bitbucket.org/jvsoest/pytaskmanager.git>

⁴FAIR stations: <http://github.com/maastroclinic/DataFAIRifier>

⁵Convert CSV file to RDF file: <https://github.com/sunchang0124/FAIRHealth/>

dependent providers using a TSE station. We monitor time to obtain the analytical results for each scenario. Scenario 1 consists of two providers, A and B, each having the same (small) number of individuals; Scenario 2 consists of providers A and B, but provider B has a much larger set of individuals, including all Provider A's individuals. For these scenarios, we use data from a publicly available dataset which contains attributes that could be interpreted as sex, body mass index (BMI), number of children, smoking status, region, and health insurance reimbursement of participants [21]. Additionally, we generated artificial personal identifiers including date of birth, zip code, house number, and sex for linking purpose [22]. In practice, combining multiple datasets might be prone to record-linking errors. We will discuss this in more detail in the Discussion section. Please find this synthetic dataset in Figshare⁶. This dataset is vertically split over the two providers: both have artificial personal identifiers (date of birth, zip code, house number, and sex). Only Provider A has BMI, number of children, and smoking status, while only Provider B has living region and health insurance reimbursement. In scenario 1, both providers have 1338 patients. In scenario 2, Provider A still has 1338 patients while Provider B hosts 64,400 patients. Since, Provider A in the second scenario only hosts a small subset of Provider B, a single record of Provider A might match with several records from Provider B. Even though this scenario is often encountered in practice, few solutions are available to address this linking challenge for vertically partitioned data [23].

For our experiment, we developed application trains using Docker 18.03.1. Pseudonymization, encryption, verification, and record linkage were implemented in Python 2.7. The infrastructure was tested with a 2.5GHz PC with 16GB RAM and 500 GB hard disk.

3.4 Results

In this section, we detail the contributions of each of the three work packages. Next, we discuss the outcome of the experiment. Figure 3.1 and Figure 3.2 provide an illustration of the infrastructure. In Figure 3.1, an overview of the operational framework for two providers, A and B, and a trusted secure environment, TSE, is presented. In Figure 3.2, we present the technical and legal requirements of the FAIR data stations. Researchers request permission to access and process data from the data provider. Once permission is granted, application trains to pseudonymize and encrypt the data are sent and executed in the data stations. Next, the encrypted data are sent to the TSE, followed by the data analysis application (from the researchers).

⁶Find our synthetic datasets: <https://doi.org/10.6084/m9.figshare.7379810.v2>.

In the FAIR data stations (Figure 3.2), personal identifiers are pseudonymized by one-way hashing and salting techniques. One-way hashing turns any format of data into a fixed-length “fingerprint” that cannot be reversed. Salt, as a random string, is appended to data before hashing, to eliminate the risk of malicious decryption. We used Secure Hash Algorithm 2 (SHA- 512) as the one-way hashing function and random salts are shared by two data providers to make personal identifiers pseudonymized on both sites. This results in a unique code per record, allowing linking the same records from all data providers. Every time data providers grant researchers permission to process/analyze the data, the personal identifiers get pseudonymized using different salts. The salt needs to be created and agreed upon by all data providers. Additionally, to safeguard secure transfer, processed data are encrypted, prior to sending them to TSE. The same as with the salt, encryption keys are re-generated every time.

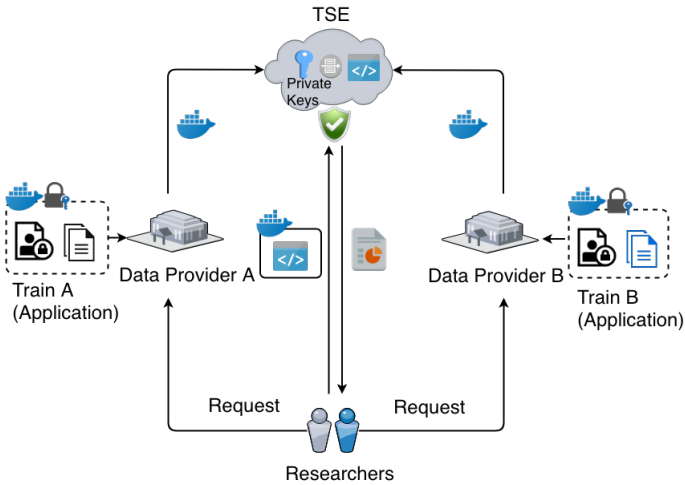


Figure 3.1: Conceptual overview of the proposed infrastructure. Data access is regulated by the data provider hosting the stations. If access is granted, the data providers encrypt the data and send these to the TSE. The TSE executes the researchers application and allows aggregated results to be returned to the researcher.

The procedure then continues as follows: when the encrypted data are sent to the TSE, a notification is generated by the data stations to confirm the successful execution and departure of the train. After all encrypted data arrive at the TSE station, the researchers trigger analysis at the TSE with a set of keys and an application that includes code for the analysis. There is one private key per data station to decrypt the dataset, and one verification key to test the dataset integrity. The data station can only encrypt using the public key but

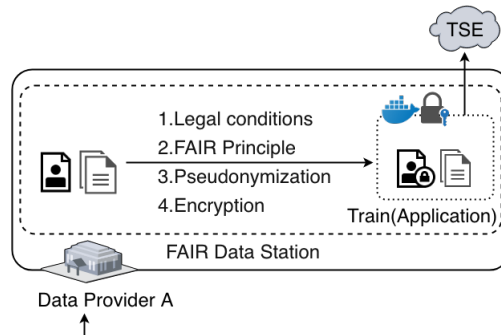


Figure 3.2: Overview of data stations and application trains. Within each station, data are prepared, i.e., legal conditions are checked, FAIR principles implemented, personal identifiers pseudonymized, and encrypted. The application train enters the data station with algorithms and leaves with results or processed data.

cannot decrypt. The TSE station maintains the private key to decrypt for this specific data provider. After getting verified and decrypted data from both providers, the data can be linked and merged by pseudonymized personal identifiers. As the salted hashes performed at the data station are unknown to the TSE, it is not able to reverse or decrypt sensitive data such as personal identifiers. Thus, in addition to pseudonymization and encryption, the privacy of information is further protected as no data provider has direct access to the TSE. After executing the analytical algorithms on the merged dataset, the TSE checks whether the results reveal any personal identifiable information. Only the validated results such as figures and/or tables that do not contain any personal identifiable information are returned to the researchers. Finally, all (received and created) data in the TSE are destroyed.

3.5 Simulation experiments

We used our proposed infrastructure to analyze synthetic data (discussed in the Methods section) that was vertically partitioned to form two datasets, each with a different data provider. Figure 3.3 shows one such result: a plot of BMI and health insurance reimbursement over one calendar year. While simple, the simulation experiment provides evidence for the feasibility of the infrastructure to execute an analysis, in this case, retrieval of a relation between two attributes in separate datasets in a secure and privacy-preserving manner.

We conducted an experiment with two scenarios. In the first scenario, where

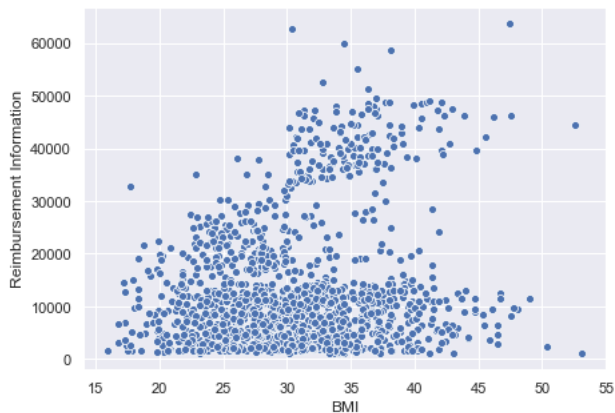


Figure 3.3: Plot of body mass index (BMI) versus health insurance reimbursement in the past year (dollars) from the analysis of a synthetic and vertically partitioned dataset using the proposed infrastructure.

both providers host 1338 individuals, pseudonymization took 0.4-0.5 seconds and encryption took 0.1-0.2 seconds for each data station. At the TSE, verification and decryption spent merely 0.1-0.2 seconds, while record linkage took around 7.2 seconds. For the second scenario, where provider A hosts 1338 individuals and Provider B 64,400 individuals, pseudonymization for Provider B took 7.3 seconds and 2.5 seconds to encrypt. The total time cost at the TSE increased to about 15 seconds. From our experiment, we found that pseudonymization (at data stations) and record linkage (at the TSE) consumed approximately 80% of the running time. Future work will focus on operational performance measures, and among others, the size of provider datasets and number of attributes considered in linking.

3.6 Discussion

We have described and demonstrated a distributed learning infrastructure using artificial and vertically partitioned data involving two providers and a trusted secure environment. This is a preliminary, but promising result.

Our long-term goal is to deploy the infrastructure to analyze actual data from two independent organizations - Statistics Netherlands (CBS) and the Maastricht Study. Thus far, we have requested data for 3451 consenting participants from the Maastricht Study, which is characterized by extensive phenotyping and provides information on the etiology, pathophysiology, com-

plications, and comorbidities of T2DM. All participants are aged between 40 and 75 years and live in the southern part of The Netherlands. We requested those attributes which were complete and consented. Attributes include socio-demographic factors, lifestyle factors, the status of T2DM, physical function, mental functions, BMI, and cardiovascular disease history. From CBS, we requested regional population data of health insurance reimbursement from 2010-2016. As of November 2018, all application trains have been developed. We are in the stage of approving and building data stations for the Maastricht Study and CBS. A joint controllership agreement between the two organizations is established to enable the TSE. We are preparing analytic algorithms that will 1) answer scientific questions regarding the associations between T2DM status and healthcare costs, and 2) to evaluate the performance and security of our infrastructure.

Applying the infrastructure to real-world situations will present several challenges. Although we have only explored a two-data-provider scenario, we anticipate that it can be extended to more than two providers. While the performance of this system will depend on the size of the data and the algorithms used for encryption, merging, and analysis, we believe that the biggest bottleneck is in creating consortium agreements and deploying the infrastructure in individual facilities. As such, there is a need to further develop a PHT deployment kit that enables stakeholders to consider all the issues and options and make informed decisions in the most efficient manner. A second challenge in our implementation lies in the possibility of errors caused by linking vertically partitioned datasets. The accuracy of matching across these will decrease owing to missing data, typographical errors, differences in pseudonymization procedures, and different formats of identifying information. In addition, to match a fraction of records from multiple large datasets, the data providers could limit the size of their data by sending only a selection to TSE. This selection can be discovered and defined by sending exploratory or individual selection algorithms first. For instance, in our case, instead of sending the information of the entire Dutch population to the TSE, only a subset of the Dutch population which meets the criteria of the Maastricht Study sample is sent to the TSE. However, note that this selection might also leak information about the individuals in the data of (one or more) data providers. We intend to explore the impact of such aspects in future studies. A third challenge is how to manage and transport the keys securely among different parties. The TSE requires decryption and verification keys to decrypt the data and run the analysis algorithms. This approach must be agreed on by all parties from both technical and ethical-legal perspectives.

3.7 Conclusions

To analyze vertically partitioned data, we extended a Personal Health Train (PHT) infrastructure to send data analysis algorithms to multiple data stations and return only the results instead of the original data to the researchers. This infrastructure was developed in a coordinated manner across multiple scientific, technical, ethical, legal, and societal aspects involving several units and organizations. This coordination across interests is essential to explore viable solutions for data sharing and reuse, as envisioned by the proponents of the FAIR principles. In particular, the idea of bringing the algorithm to the data, rather than obtaining consent to receive a copy of the data, offers an entirely new paradigm that has not been considered by most organizations. Having a new paradigm will require stakeholders to take the time and effort to thoroughly evaluate this in terms of their legal and technical requirements. However, as our experiment shows, it offers a more scalable and secure solution to analyze vertically partitioned data in a secure and privacy-preserving manner. Additional operational and security enhancements are still needed before the infrastructure is suited to deal with real (sensitive) data. Future work will explore the quality of scientific discovery (accuracy of outcome), the security, scalability, sustainability, and performance of computation. While no solution will be perfect for all situations, we believe that this adaptation of the PHT model will find utility in situations involving sensitive data with a multitude of stakeholders.

References

- [1] Geoff Dougherty. *Digital image processing for medical applications*. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511609657.
- [2] Hanaa Elshazly, Ahmed Taher Azar, Abeer El-Korany, and Aboul Ella Hassanien. "Hybrid system for lymphatic diseases diagnosis". In: *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2013, pp. 343–347. DOI: 10.1109/ICACCI.2013.6637195.
- [3] EA Clarke. "What is preventive medicine?" In: *Canadian Family Physician* 20.11 (1974), p. 65. DOI: 10.1145/772862.772867.
- [4] Muhammad Imran Razzak, Muhammad Imran, and Guandong Xu. "Big data analytics for preventive medicine". In: *Neural Computing and Applications* 32.9 (2020), pp. 4417–4451. DOI: 10.1007/s00521-019-04095-y.

-
- [5] Raphael B Stricker and Lorraine Johnson. "Lyme disease: the promise of Big Data, companion diagnostics and precision medicine". In: *Infection and drug resistance* 9 (2016), p. 215. DOI: 10.2147/IDR.S114770.
- [6] Jacques S Beckmann and Daniel Lew. "Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities". In: *Genome medicine* 8.1 (2016), pp. 1–11. DOI: 10.1186/s13073-016-0388-7.
- [7] Shawn Dolley. "Big data's role in precision public health". In: *Frontiers in public health* 6 (2018), p. 68. DOI: 10.3389/fpubh.2018.00068.
- [8] Edward H Shortliffe, Edward H Shortliffe, James J Cimino, and James J Cimino. *Biomedical informatics: computer applications in health care and biomedicine*. Springer, 2014. DOI: <https://doi.org/10.1007/978-1-4471-4474-8>.
- [9] Matthew J Bietz, Cinnamon S Bloss, Scout Calvert, Job G Godino, Judith Gregory, Michael P Claffey, Jerry Sheehan, and Kevin Patrick. "Opportunities and challenges in the use of personal health data for health research". In: *Journal of the American Medical Informatics Association* 23.e1 (2016), pp. 42–48. DOI: 10.1093/2Fjamia/2Focv118.
- [10] Johan van Soest, Chang Sun, Ole Mussmann, Marco Puts, Bob van den Berg, Alexander Malic, Claudia van Oppen, David Townend, Andre Dekker, and Michel Dumontier. "Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data." In: *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*. Vol. 247. IOS Press, 2018, pp. 581–585. DOI: 10.3233/978-1-61499-852-5-581.
- [11] Dutch Techcentre for Life Sciences (DTL). *Personal Health Train*, <https://www.dtls.nl/fair-data/personal-health-train/>. Access on 12-8-2021.
- [12] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1 (2016), pp. 1–9. DOI: 10.1038/sdata.2016.18.
- [13] Johan van Soest, Andre Dekker, Erik Roelofs, and Georgi Nalbantov. "Application of machine learning for multicenter learning". In: *Machine Learning in Radiation Oncology*. Springer, 2015, pp. 71–97. DOI: 10.1007/978-3-319-18305-3_6.

- [14] Andrea Damiani, Mauro Vallati, Roberto Gatta, Nicola Dinapoli, Arthur Jochems, Timo Deist, Johan van Soest, Andre Dekker, and Vincenzo Valentini. “Distributed learning to protect privacy in multi-centric clinical studies”. In: *Conference on artificial intelligence in medicine in europe*. Springer. 2015, pp. 65–75. DOI: 10.1007/978-3-319-19551-3_8.
- [15] Timo M. Deist et al. “Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT”. In: *Clinical and Translational Radiation Oncology* 4 (2017), pp. 24–31. DOI: 10.1016/j.ctro.2016.12.004.
- [16] Arthur Jochems, Timo M Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. “Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital—a real life proof of concept”. In: *Radiotherapy and Oncology* 121.3 (2016), pp. 459–467. DOI: 10.1016/j.radonc.2016.10.002.
- [17] Jaideep Vaidya. “A survey of privacy-preserving methods across vertically partitioned data”. In: *Privacy-preserving data mining*. Springer, 2008, pp. 337–358. DOI: 10.1007/978-0-387-70992-5_14.
- [18] Tamas Zoltan Gal, Gábor Kovács, and Zsolt T Kardkovács. “Survey on privacy preserving data mining techniques in health care databases”. In: *Acta Universitatis Sapientiae, Informatica* 6.1 (2014), pp. 33–55. DOI: 10.2478/ausi-2014-0017.
- [19] Frank Manola, Eric Miller, Brian McBride, et al. “RDF primer”. In: *W3C recommendation 10.1-107* (2004), p. 6.
- [20] *What is a Container?*, <https://www.docker.com/resources/what-container>. Access on 22-10-2018.
- [21] Brett Lantz. *Machine Learning with R*. Packt Publishing, 2013. DOI: 10.5555/2588158.
- [22] *Welcome to Faker’s documentation!*, <https://faker.readthedocs.io/en/master/>. Access on 22-03-2019.
- [23] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. “Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption”. In: *arXiv preprint arXiv:1711.10677* (2017).

4

Studying the Association of Diabetes and Healthcare Cost on Distributed Data from The Maastricht Study and Statistics Netherlands using a Privacy-Preserving Federated Learning infrastructure

Adapted from: Chang Sun, Johan van Soest, Annemarie Koster, Simone J.P.M. Eussen, Miranda T. Schram, Coen D.A. Stehouwer, Pieter C. Dagnelie, and Michel Dumontier. "Studying the association of diabetes and healthcare cost on distributed data from the Maastricht Study and Statistics Netherlands using a privacy-preserving federated learning infrastructure". In: *Journal of Biomedical Informatics* (2022). DOI: 10.1016/j.jbi.2022.104194.

Abstract

Mining personal data collected by multiple organizations remains challenging in the presence of technical barriers, privacy concerns, and legal and/or organizational restrictions. While a number of privacy-preserving and data mining frameworks have recently emerged, much remains to show their practical utility. In this study, we implement and utilize a secure infrastructure using data from the Statistics Netherlands and the Maastricht Study to learn the association between Type 2 Diabetes Mellitus (T2DM) and individuals' healthcare expenses considering the impact of lifestyle, physical activities, and T2D complications. Through experiments using real-world distributed personal data, we present the feasibility and effectiveness of the secure infrastructure for the practical use cases of linking and analyzing vertically partitioned data across multiple organizations. We discovered that individuals diagnosed with T2DM had significantly higher expenses than those with prediabetes, while participants with prediabetes spent more than those without T2DM in all the included healthcare categories to different degrees. We further discuss a joint effort from technical, ethical-legal, and domain-specific experts that is highly valued for applying such a secure infrastructure to real-life use cases to protect data privacy.

4.1 Introduction

The amount of personal data generated from individuals is dramatically growing. This massive data can be used by the research community to study unresolved research questions and gain new scientific insights. However, one of the major barriers that researchers are often faced with is the difficulty of accessing and jointly analyzing the personal data that are distributed at multiple data organizations such as healthcare providers, banks, retailers, insurance companies, and governmental organizations. Sharing and analyzing distributed personal data across multiple organizations remains challenging from technical, ethical-legal, administrative, and political aspects owing to such as inconsistent data standards, a lack of data provenance, and insufficient FAIR (Findable, Accessible, Interoperable, Reusable) data management [1]. This hinders discovering more potential knowledge from distributed personal health data, as well as the secondary use of health data for intra- and inter-disciplinary research.

To tackle this challenge, we have proposed a prototype secure infrastructure which supports analysis of personal health data that are vertically partitioned data at multiple organizations with different governance policies with preserving individual privacy [2]. Vertically partitioned data represents multiple organizations hosting different features from the same group of data subjects. Following the Personal Health Train (PHT) initiative, our approach sends data analysis algorithms to multiple data organizations and returns only the results instead of the original data to the researchers [2, 3, 4]. From our previous systematic literature study on privacy-preserving distributed data mining [5], several remaining challenges have been recognized in the existing methods such as the lack of data linkage process in the vertically partitioned data and the shortage of the practical implementation of the theoretical methods. In vertically partitioned data, participating organizations must accurately link their data or sort them in an identical order prior to data analysis. A majority of studies assume this correspondence between data records exists by default which is often not the case in reality. Second, a large number of privacy-preserving data analysis studies did not implement their theoretical solutions to practical applications. Without practical implementation, the approaches might neglect essential problems such as the impact of the ethical-legal or organizational regulations on the technical development and data quality issues (e.g., the sparse or biased datasets) [6, 7, 8].

Unlike the most existing studies, our proposed infrastructure covers accurate data linkage in vertically partitioned data using pseudomized identifiers and analyzes encrypted data in a trusted secure environment which is independent from the parties who provide the data and researchers who conduct the data analysis. We implemented and evaluated the proposed infrastructure

in a practical use case to study the association between the status of Type 2 Diabetes Mellitus (T2DM) (normal glucose metabolism, pre-diabetes, and T2DM) and individuals' healthcare expenses using vertically partitioned data from The Maastricht Study and Statistics Netherlands while meeting their legal and technical requirements. The Maastricht Study is an observational prospective population-based cohort study focusing on the etiology, pathophysiology, complications and comorbidities of T2DM and is characterized by an extensive phenotyping approach [9]. Statistics Netherlands (CBS: Centraal Bureau voor de Statistiek) is the Dutch national statistical agency that provides reliable statistical information and data about the Netherlands.

The motivation of this use case study is T2DM and diabetes complications creating a significant economic burden on patients and their families in terms of higher healthcare payments and loss of family income [10]. This economic burden is estimated to rise further due to changes in demographics and lifestyles [11]. Modern lifestyles such as increasing intake of processed food, longer sedentary time, and physical inactivity are the most influential factors of developing T2DM and its complications [12, 13]. Indirectly, lifestyle might have a considerable impact on healthcare costs of people with T2DM. Unfortunately, the health-related data, lifestyle data, and health care expense data from individuals are held by different organizations (e.g., healthcare providers, insurance companies, and statistics offices) and these data are restricted to access and be shared across organizations. Hence, only limited evidence has been obtained on the economic impact of T2DM status in different categories of healthcare providers observed from an individual level. The key drivers of the economic impact across different healthcare categories is poorly studied. Therefore, our infrastructure can be applied to jointly analyze the distributed personal data in this case and gain a better understanding of the economic impact of T2DM in order to help inform effective health policies while minimizing healthcare costs or optimizing cost allocation.

In this study, the development and practical implementation of proposed infrastructure are based on a collaboration between technical and ethical-legal experts from all participating parties. We summarize the key contributions:

1. Proving the feasibility of the proposed privacy-preserving infrastructure using distributed personal health data from two independent organizations as a practical use case,
2. Accurately linking real-life data of unequally sized datasets - 3283 records from the Maastricht Study with data of over 1 million records from the Statistics Netherlands using a pseudonymization method,
3. Jointly analyzing linked vertically partitioned data using machine learning models combined with encryption methods to examine the

association between T2DM status and annual healthcare costs in different healthcare provider categories.

This paper is structured as follows. Section 4.2 will describe the personal data from two data sources and our analysis methods. The settings of experiments and the secure infrastructure will be presented in Section 4.3 followed by the results and findings in Section 4.4. Section 4.5 will describe the strengths and limitations of the study. Lastly, we conclude the work in Section 4.6.

4.2 Methods

The proposed infrastructure designed for analyzing data from multiple sources without revealing any original data was preliminarily proposed in the previous study [2, 4]. It is an extension of Personal Health Train (PHT) architecture which facilitates researchers to send data-processing algorithms to each data source instead of centralizing the data required for the analysis. The PHT architecture has been implemented and evaluated by several practical use cases on horizontally partitioned data¹ [14, 15, 16]. In the case of horizontally partitioned data, a set of models, individually trained on each data source, are combined over multiple iterations, resulting in a global model. However, this is not possible for vertically partitioned data since not all of the input features are available in all data sources. Hence, data sources cannot execute the complete analysis independently of the other sources. Additional challenges need to be considered when there is a legal impediment to link data on each individual level across sources with a unique identifier or when this unique identifier is unavailable [17]. This study aims to tackle the challenges in vertically partitioned data.

To tackle the challenges, we proposed and practically implemented a secure infrastructure to analyze vertically partitioned data. Figure 4.1 illustrates the workflow of the proposed infrastructure. The infrastructure consists of application trains which contain analysis models (trains) designed by researchers, data stations where data organizations can execute the models, and a legal-ethical framework (railway tracks) supporting such data analyses. In the use case, we created data stations at the Maastricht Study and CBS, respectively, and a third station which was configured as a Trusted Secure Environment (TSE) supported by a legal agreement (joint controller agreement) between the two data organizations. The TSE station does not contain data by itself, however, acting as a trusted and independent entity. No data organization or

¹Horizontally partitioned data scenarios are when multiple organizations host the same features from different individuals, while vertically partitioned data scenarios are when these organizations host different features from the same group of individuals.

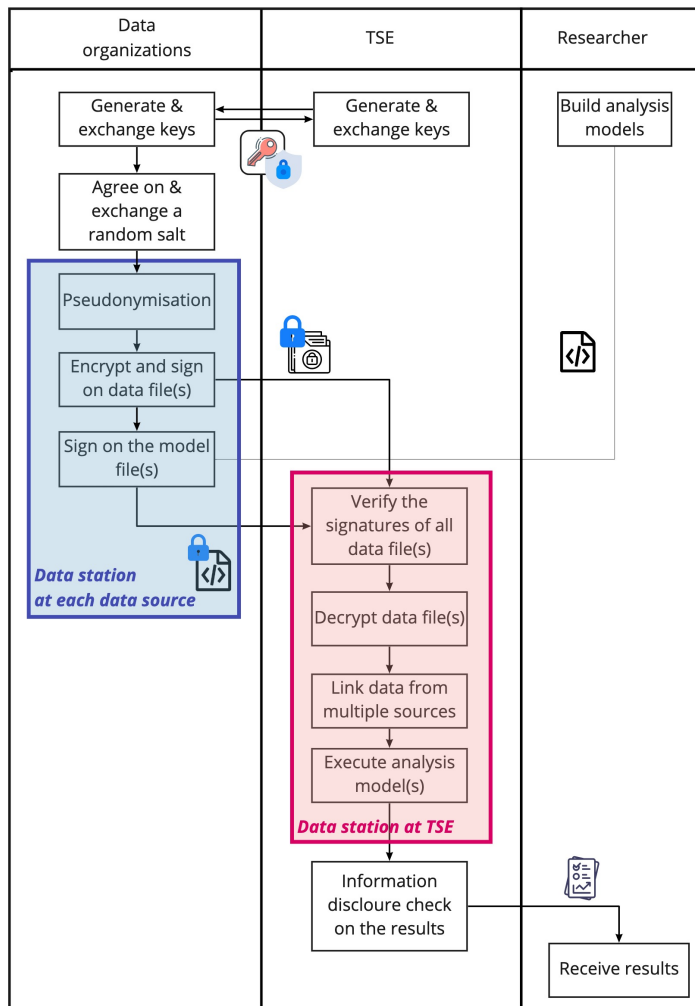


Figure 4.1: A simplified working process of using the privacy-preserving infrastructure. As the Statistics Netherlands and The Maastricht Study conducted the same operations, the data organizations in the figure present both of them.

researcher has direct access to the TSE. To be able to have secure communications between data stations and verify authentications of each other, all participating organizations are required to first generate encryption key pairs and securely exchange generated keys with the TSE station using Diffie-Hellman key exchange method and Secure File Transfer Protocol over public channels.

Other encryption schemes and details which are applied in the infrastructure are elaborated in the following sections.

The application trains are built as Docker containers, which carry all required software packages to execute the application on board. These applications can query data, perform data cleaning/formatting, and execute machine learning or statistical analysis. To run the application train, each source deploys a data station which contains the data required for the analysis. The data station only returns the results of the analysis rather than any of the original data. In the use case, we created two application trains. The first application train is executed at the data station of each organization respectively. The application extracts the data from two data stations, pseudonymises the personal identifiable features for linking purpose, encrypts the data files, sign on the encrypted data files and analysis model file, and sends them to the TSE data station. The data station generates a random encryption key using Salsa20 256-bit stream cipher [18] to encrypt the data files. The encrypted files are then digitally signed by each data station using Elliptic Curve Digital Signature Algorithm (EdDSA) [19]. Salsa20 and EdDSA are both widely-applied encryption schemes in the field and have the advantage of high-speed and high-security [20, 21]. However, they suffer from the data-dependent timing variation which means the execution time of the encryption algorithms is dependent on the size of the data [22]. By measuring the time for each operation in the encryption and analyzing the encryption time taken, an attacker could possibly trace back to the input data. Therefore, we applied the McBits algorithm, a constant-time fast implementation for the public-key encryption systems in the infrastructure to have a full protection against the timing attacks [22, 23]. The analysis model file is a configurable Python script where the researcher can choose different predefined models and modify the parameters based on their analysis plans and send them to every participating data organization. Only when the model file is evaluated and approved by all organizations, can it be signed and transferred to the TSE data station.

The second application train is executed at the TSE data station. The TSE station maintains the private keys to verify the signatures and decrypt the data files from each data organization. After getting verified and decrypted, the data can be linked and merged by pseudonymized personal identifiers. As the pseudonymization is one-way hashing and performed at the data stations of sources, it is not able to reverse or decrypt the hashed personal identifiable features. After executing the analytical algorithms on the merged dataset, the TSE checks whether the results reveal any personal identifiable information. Only the validated results that do not contain any personal identifiable information are returned to the researchers. Finally, all received and created data in the TSE are destroyed.

4.2.1 Data linkage

To link data from multiple sources on each individual record, unique identifiers, such as national identification numbers or social security numbers, must be available in every data source. The GDPR leaves it up to the national governments to determine the use of the national identification number. The Netherlands has adopted a very restrictive approach regarding the use of the national identification number (Burgerservicenummer - BSN) to prevent linking individuals over multiple sources. There are limited situations that allow for the use of the BSN, for example uses by governmental entities, matters related to tax, and health and educational institutes (for administrative / financial purposes) However, BSN as the most reliable identifier cannot be used for scientific purposes. To achieve compliance with Dutch law, we pseudonymised a combination of personal features including gender, date of birth, zip code, and house number. Two data parties formalized these four features into the same data formats and presentations before pseudonymisation. Since neither of the two parties is allowed to know the data subjects from the other party, CBS provided all residents who were between 40 and 75 years old and lived in the south of the Netherlands in 2010 to 2013. Considering people might move out/in this area during these years, the address histories of the residents were also included to increase the number of matches using our created identifier. The healthcare cost dataset from CBS contains more than 1 million records including duplicate citizens with different addresses.

To pseudonymise the combination of personal features at data sources, one-way hashing method (SHA-512) was applied for pseudonymisation. One-way hashing method turns any format of data into a fixed-length “fingerprint” that cannot be reversed. Salt, as a random string, is appended to data before hashing, to eliminate the risk of malicious decryption. This salt needs to be agreed and shared by two data organizations. The pseudonymized personal features are not related to a specific person anymore, but shared hashing function and salts make it possible to link two datasets.

4.2.2 Data Analysis and experiment settings

After data linkage, all analyses of this study were conducted on the merged dataset. All steps (data linkage, creating the merged subset, and the analysis) were composed into one application train and executed at the TSE without human interference. The descriptive characteristics of the study population were first generated and presented as mean with standard deviation for continuous features or as numbers and percentages for categorical features based on the three groups of participants without diabetes, with pre-diabetes, and

with T2DM. The pairwise correlations were calculated using Pearson's correlation method to observe the linear correlations between features [24]. The healthcare costs were calculated based on the average of the annual healthcare cost over all categories from 2010 to 2016. Regression models have been commonly employed in the previous studies to examine the relationship between diabetes and healthcare costs [25, 26, 27]. In our study, we applied ordinary least squares linear regression models to examine the associations of T2DM with every category of healthcare cost and with covariates of lifestyle and complications. The associations were presented using regression coefficients (coef) with 95% Confidence Intervals (CI) and P-value. The basic linear regression model (model 1) was adjusted for sex, age, and educational level. Model 2 was additionally adjusted for complications features including BMI, movement limitation, history of cardiovascular disease, hypertension, and depression. Model 3 was additionally adjusted for smoking status, alcohol consumption, energy intake, Mediterranean-diet score, physical activity.

All analyses were built up as an executable application in Docker (V19.03.8) using Python (V3.6.9), Scikit Learn (V0.21.3), and Statsmodel (V0.11.1) Python libraries. Encryption methods were applied using Pynacl (V1.3.0), Crypto (V1.4.1) and Pycrypto (V2.6.1). Two data stations at CBS and the Maastricht Study were built up in an Ubuntu 18.4 LTS machine with 2 CPU, 6GB RAM and 60GB storage. The TSE station was employed in an Ubuntu 18.04 LTS machine with 2 CPU, 4GB RAM and 40GB storage. The code of the infrastructure and data analysis is published on: <https://gitlab.com/CBDS/DataSharing>.

4.3 Datasets and materials

We requested health-related data including demographic data, life-style, T2D status and its complications from the Maastricht Study and data about individuals annual healthcare costs from the Statistics Netherlands (CBS). The annual healthcare costs data is provided by Vektis, a company in the Netherlands that streamlines healthcare processes and collects information about healthcare cost declarations. This section describes the datasets and the approaches of linking and analyzing two datasets without revealing original data using the secure infrastructure. At last, the analysis models are presented to examine the association between T2DM and healthcare costs.

4.3.1 Health-related data from The Maastricht Study

The health-related data from the study participants is provided by the Maastricht Study including individuals' demographic, socioeconomic, lifestyle, T2DM status, and its complication data. The rationale and methodology have

been described previously [9]. We requested cross-sectional data from participants who completed the baseline survey between 2010 and 2013. The eligibility of participation of the Maastricht Study were individuals aged between 40 and 75 years and living in the southern part of the Netherlands. Participants were recruited through mass media campaigns, from the national municipal registries, and the regional Diabetes Patient Registry via mailings [9]. Due to the purpose of data linking and joint analysis, this study requested additional permissions to link participants' data between the Maastricht Study and CBS. In the analysis, we included 13 features which were categorized into diabetes outcome, demographic characteristics, socioeconomic conditions, T2DM complications, and lifestyle. An overview of features is shown in Figure 4.2. The collection method and measure of variables is summarized in the following subsections.

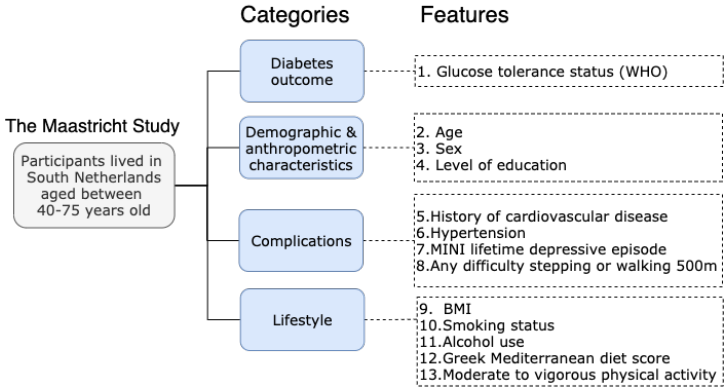


Figure 4.2: Overview of health-related data collected from the Maastricht Study.

Diabetes outcome. The definition of T2DM status was based on the World Health Organization's diagnostic criteria of glucose tolerance status [28]. Participants were categorized into no-diabetes, prediabetes, and T2DM. Participants with no diabetes have normal glucose metabolism (fasting plasma glucose < 6.1 mmol/l and 2 h plasma glucose (after glucose load) < 7.8 mmol/l), while participants with pre-diabetes have impaired fasting glucose (fasting plasma glucose 6.1-6.9 mmol/l and 2 h plasma glucose < 7.8 mmol/l) or impaired glucose tolerance (fasting plasma glucose < 7.0 mmol/l and 2 h plasma glucose ≥ 7.0 - 11.1 mmol/l). Pre-diabetes is an intermediate but high-risk state for diabetes. 5-10% of people per year with prediabetes will progress to diabetes, with the same proportion converting back to normoglycemia [29]. Participants with T2DM have fasting plasma glucose ≥ 7.0 mmol/l or 2 h plasma glucose ≥ 11.1 mmol/l.

Demographic characteristics include age, sex, and educational level. Educational level of the participant was collected by a questionnaire and was categorized into 3 classes: 1) no education or primary education not completed or primary education or lower vocational education; 2) intermediate vocational education or higher secondary education; 3) higher professional education or university education.

Complications category includes calculated BMI and self-reported information on mobility limitation, history of cardiovascular disease, hypertension, and lifetime depression. BMI was calculated as $\text{weight (kg)} / \text{height}^2 \text{ (m)}$. Mobility limitation (yes/no) was defined as whether the participant has difficulty walking 500 m or climbing the stairs. History of cardiovascular disease (yes/no) was defined as whether the participant has a history of disorders of the heart and blood vessels. Hypertension (yes/no) was based on the average blood pressure and antihypertensive medication use. Lifetime depression (yes/no) was defined as if participants experienced depressive disorder in their lifetime.

Lifestyle factors include the following self-reported health behavior: physical activity, dietary, smoking and alcohol consumption. Physical activity was measured and categorized into low, medium, and high physical activity. Participants' dietary patterns were scored by the 10-point Greek Mediterranean-diet scale as the main measurement of dietary intakes [30, 31]. Smoking status was categorized as: never-smokers, former-smokers, and current-smokers. Alcohol use was categorized as: non-consumers, low-consumers, and high-consumers. Extended explanation of each variable in Complications and Lifestyle categories can be found in the Appendix.

4.3.2 Study population healthcare costs data from Statistics Netherlands (CBS)

Due to privacy concerns, the two data organizations cannot share identifiable information or communicate which individuals are included in this study. To link two datasets, CBS prepares data of the annually declared healthcare cost data from all residents aged between 40 and 75 years living in the southern part of the Netherlands between 2010 and 2016. Over 1 million individuals are included in the dataset from CBS. The healthcare costs in this study comply with the Health Insurance Act [32] in the Dutch Basic Health Package that healthcare providers and insured persons have declared to health insurers and have got reimbursed between 2010 and 2016. The basic health package is mandatory in the Netherlands and contains the same healthcare products for everyone in the country. The coverage and availability of the

healthcare services from the Dutch Basic Health Package provides the most comprehensive and reliable healthcare costs data for this study [32, 33].

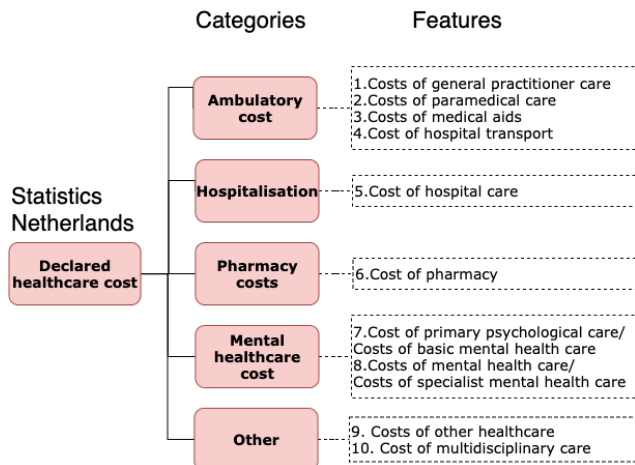


Figure 4.3: Overview of healthcare cost features collected from 2010 to 2016 by Vektis and accessed from the Statistics Netherlands (CBS)

This study includes 10 categories of healthcare costs (as Figure 4.3 shows): general practitioner (family doctor) care, paramedical care, medical aids, hospital transport, hospital care, pharmacy, primary (first-line) psychological care (2010-2013) / basic mental health care (2014-2016), (second-line) mental health care (2010-2013) / specialist mental health care (2014-2016), multidisciplinary care (2015-2016), and other health care. The multidisciplinary care is for patients with chronic conditions requiring the help of multiple doctors and health care providers. Type 2 diabetes, chronic obstructive pulmonary disease, and cardiovascular diseases were included in this healthcare category. Definitions and the coverage of each healthcare cost feature are presented in Table A.4 in the Appendix.

Figure 4.4 illustrates the flowchart of how the data was selected for the final analyses from two data sources. First, we excluded 168 participants who did not give consent to link their health data from the Maastricht Study with any data from CBS. After pseudonymizing and matching the two datasets, we excluded 845 participants who had missing values in their health-related features and 125 participants who had missings in their healthcare cost data. At the end, data from 2313 participants' are included in the final analysis.

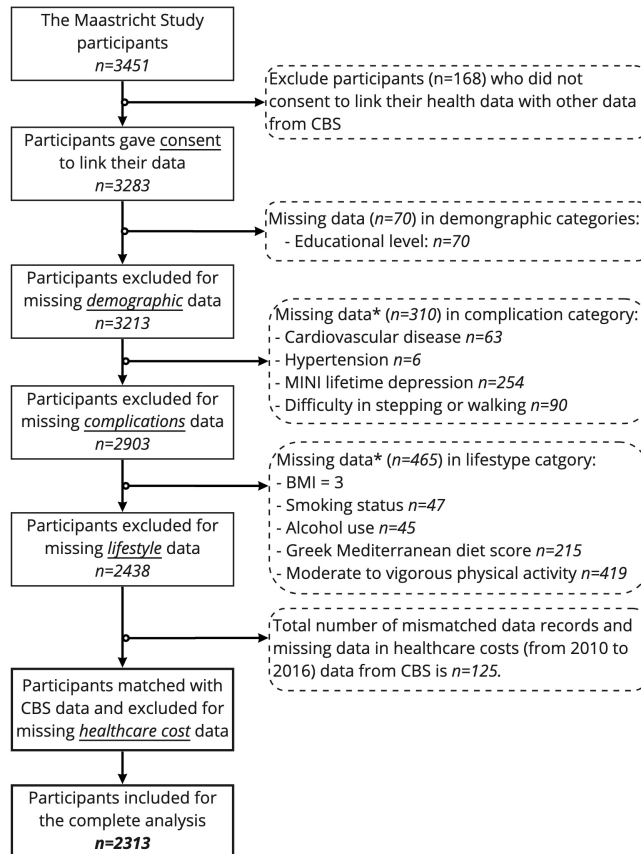


Figure 4.4: A flowchart of data selection by excluding participants with missing values in the features. Numbers of missing values do not add up to the number of excluded participants. Participants may have missing values in multiple features.

4.4 Results

4.4.1 Data linkage between two data sources

The data with 3283 participants from the Maastricht Study and data with 1009309 participants from CBS were matched and linked by the pseudonymized linking features in the TSE data station. Pseudonymizing linking features and encrypting the data files only happen at the data sources, while verification, decryption and matching datasets are only executed at the TSE station.

Table 4.1 shows the matching results and time costs using the real data in a practical setting. 96% of the participants from the Maastricht Study were successfully found and matched with the unique corresponding healthcare cost data records in the CBS dataset. There were 121 participants not being found in the CBS dataset, while 17 were found with multiple data records. The total time consumption of the whole process is approximately 37 mins. The most time-consuming step is the pseudonymization of linking features at the party with most data entries (1838 sec), followed by the matching process with 316 sec.

Table 4.1: Time performance for each step of method at different data stations.

Data Station		CBS	Maastricht Study	TSE
#Records		1 009 309	3 283	-
Time (sec)	Pseudonymize	1838.75	9.96	-
	Encrypt	47.46	0.18	-
	Verify & decrypt	-	-	32.10
	Match	-	-	316.67
Matches	Unique	-	-	3145
	Multi	-	-	17
	Non	-	-	121

4.4.2 Descriptive analysis

After linking two datasets and removing instances with missing values, the study population consisted of 2313 participants, of whom 580 (25.1%) were diagnosed with T2DM. This portion is higher than the prevalence of diabetes in the Netherlands which is 6.1% (male: 7.0%, female:5.3%) [34], because the recruitment of the Maastricht Study was stratified according to known T2DM status, with an oversampling of individuals with T2DM [9]. In addition, 15.7% of participants (n=364) were recognized with pre-diabetes, while 59.2% (n=1369) of participants had no diabetes. Table 4.2 presents the baseline characteristics of the study population stratified by diabetes status.

The entire study population had a gender ratio at almost 1:1 (50.2% were male, 49.8% were female) and a mean age of 59.7 (\pm 8.1) years. However, more females (58.4%) were in the group of participants without diabetes, while males became the majority in the groups of participants with prediabetes (53.3%) and T2DM (68.3%). The participants with prediabetes and T2DM had a slightly higher age at 61.7 (\pm 7.7) and 62.4 (\pm 7.6). The majority of participants without diabetes had a higher level of education and averagely lower BMI compared to people with prediabetes and T2DM. In the lifestyle category, participants with T2DM were more often current smokers, were less often consumed high levels of alcohol, and spent less hours doing moderate to vigorous physical activities on a weekly basis compared to people with prediabetes or no diabetes. Participants with and without T2DM

have the same dietary score, which was gently higher than that of participants with prediabetes. Regarding T2DM complications, participants with T2DM had significantly higher percentages of people suffering from limited mobilities, hypertension, and cardiovascular disease.

Table 4.2: Descriptive characteristics of the study population

Health-related features from the Maastricht Study					
Features	Unit	Total n= 2313	No Diabetes n=1369	Prediabetes n=364	T2DM n=580
Sex	% men	1160 (50.2%)	570 (41.6%)	194 (53.3%)	396 (68.3%)
Age	years	59.7±8.1	58.0±8.1	61.7±7.7	62.4±7.6
Education level (%)	Low	732 (31.6%)	348 (25.4%)	130 (35.7%)	254 (43.8%)
	Medium	664 (28.7%)	391 (28.6%)	103 (28.3%)	170 (29.3%)
	High	917 (39.6%)	630 (46.0%)	131 (36.0%)	156 (26.9%)
BMI	kg/m2	27.0±4.4	25.5±3.5	27.9±4.4	29.9±4.9
Smoking status	None	842 (36.4%)	568 (41.5%)	110 (30.2%)	164 (28.3%)
	Former	1209 (52.3%)	651 (47.6%)	217 (59.6%)	341 (58.8%)
	Current	262 (11.3%)	150 (11.0%)	37 (10.2%)	75 (12.9%)
Alcohol consumption	Never	384 (16.6%)	171 (12.5%)	55 (15.1%)	158 (27.2%)
	Low	1317 (56.9%)	809 (59.1%)	197 (54.1%)	311 (53.6%)
	High	612 (26.5%)	389 (28.4%)	112 (30.8%)	111 (19.1%)
Dietary score	1-9	4.7(1.7)	4.5(1.6)	4.2(1.5)	4.5(1.6)
Physical Activity	hr/week	5.6±4.3	6.2±4.4	5.2±4.1	4.4±4.1
Mobility limitation	% yes	455 (19.7%)	175 (12.8%)	84 (23.1%)	196 (33.8%)
MINI lifetime depression	% yes	690 (29.8%)	398 (29.1%)	115 (31.6%)	177 (30.5%)
Hypertension	% yes	1287 (55.6%)	566 (41.3%)	236 (64.8%)	485 (83.6%)
Cardiovascular disease	% yes	358 (61.7%)	160 (6.9%)	48 (3.5%)	150 (41.2%)
Average healthcare cost data from CBS (Currency: euros)					
Category	Years	Total	No Diabetes	Prediabetes	T2DM
GP care	2010-16	168.9±84.4	151.5±68.6	172.6±73.9	207.8±108.1
Pharmacy	2010-16	460.6±738.5	269.5±457.8	399.7±577.8	950.0±1070.0
Hospital care	2010-16	1550.5±2649.2	1205.5±2099.6	1517.9±2413.7	2385.2±3602.2
Paramedical	2010-16	64.4±290.5	52.4±281.6	69.5±254.7	89.4±328.6
Medical aids	2010-16	128.8±508.8	61.8±265.9	88.8±186.5	311.9±894.4
Hospital transport	2010-16	20.5±160.0	12.3±44.3	15.2±51.5	43.4±308.7
Specialist mental health	2010-13	10.0±87.1	9.7±93.2	16.5±97.7	6.4±61.2
	2014-16	125.3±731.4	118.1±658.8	159.0±1086.9	121.2±607.5
Multidisciplinary care	2015-16	124.8±160.6	49.4±94.7	117.4±140.2	307.6±150.1
Other care	2010-16	88.0±135.8	28.0±70.0	74.0±149.0	238.5±129.7

Chapter 4. Studying the Association of Diabetes and Healthcare Cost on Distributed Data from The Maastricht Study and Statistics Netherlands using a Privacy-Preserving Federated Learning infrastructure

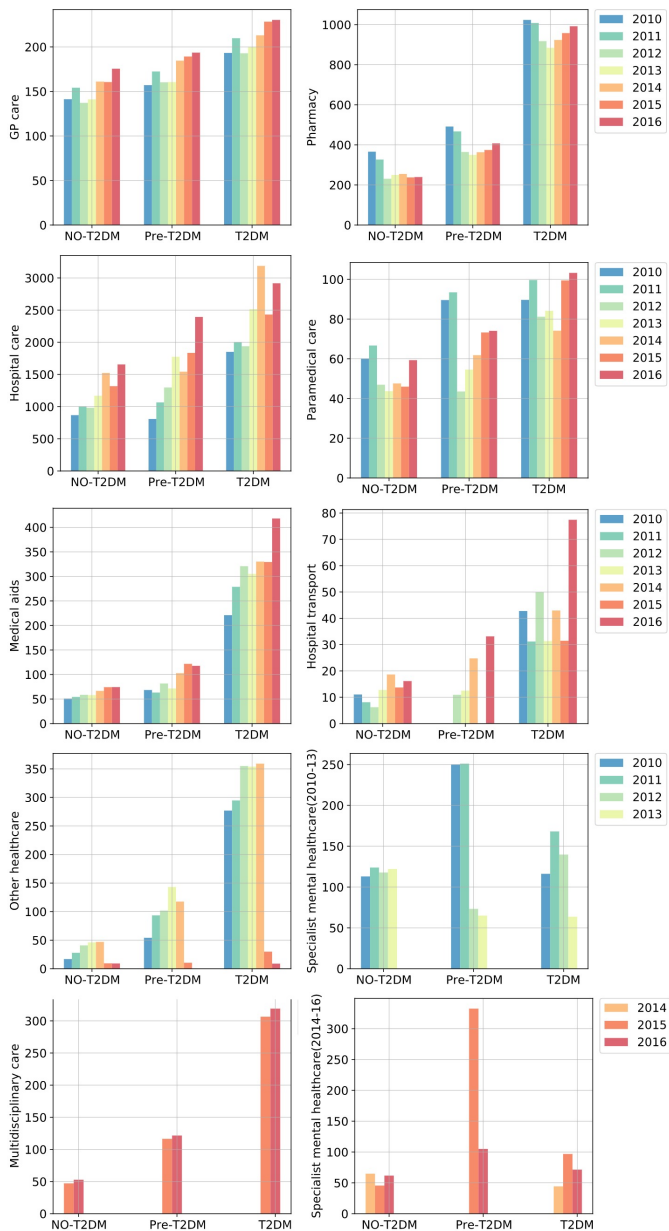


Figure 4.5: Annual healthcare costs on each healthcare category from 2010 to 2016.

The average annual healthcare cost from 2010 to 2016 are presented in Table 4.2. Participants with T2DM spent significantly more expenses than participants with prediabetes, notwithstanding specialist mental healthcare, multidisciplinary care, and other healthcare costs. Participants with prediabetes incurred additional costs than people without diabetes. Costs associated with multidisciplinary care and other healthcare were twice that for those with prediabetes compared to those without diabetes, while the costs of T2DM patients had doubled multidisciplinary care and tripled other healthcare costs than participants with prediabetes. Participants with prediabetes had the highest average cost in specialist mental health care.

Annual healthcare costs of each year from 2010 to 2016 are presented in Figure 4.5. For every year from 2010 to 2016, participants with T2DM spent more healthcare costs than people with prediabetes and no diabetes in all categories except mental health care. The excess amounts are most noticeable in the categories of pharmacy, medical aids, multidisciplinary care, and other healthcare. We observe the healthcare costs of participants had corresponding increases and decreases in all groups, but with different trends and degrees. Participants from all groups had the similar changing trend of GP care and pharmacy costs from 2010 to 2016. By contrast, in the paramedical care category, in 2012 participants with prediabetes had a significant decrease compared to participants with and without T2DM. In 2013, participants with T2DM and prediabetes had increased healthcare costs in paramedical care, while people without diabetes kept decreasing the costs.

4.4.3 Regression analysis

Table 4.3 describes the associations of the average healthcare costs from 2010 to 2016 from different healthcare categories with prediabetes and T2DM status. Model 1 is adjusted for sex, age, and level of education. Model 2 is additionally adjusted for the T2DM complication features including BMI, mobility limitation, history of cardiovascular disease, history of hypertension, and lifetime depression. Model 3 is additionally adjusted for the lifestyle features including smoking status, alcohol consumption, diet score, (moderate to vigorous) physical activities. Compared to those without diabetes, model 1 shows people with prediabetes and T2DM had significantly higher costs in GP care and pharmacy [Coef (Prediabetes - GP) = 12.10 euros (95% CI = 2.93, 21.28), [Coef (Prediabetes - pharmacy) = 96.01 euros (95% CI = 16.46, 175.55)] , [Coef (T2DM - GP) = 45.91 euros (95% CI = 37.85, 53.97), Coef (T2DM - pharmacy) = 629.81 euros (95% CI = 559.89, 699.73)]. After additional adjustment for complication and lifestyle features in model 3, T2DM was still associated with costs in GP care and pharmacy [Coef (T2DM - GP) = 26.39 euros (95% CI = 17.73, 35.05), Coef (T2DM - pharmacy) = 387.14 euros (95% CI = 313.67,

Chapter 4. Studying the Association of Diabetes and Healthcare Cost on Distributed Data from The Maastricht Study and Statistics Netherlands using a Privacy-Preserving Federated Learning infrastructure

Table 4.3: Associations of the average healthcare costs (2010-2016) in different categories with prediabetes and T2DM status. Results are presented as regression coefficients (B) with 95% confidence interval (CI) using Ordinary Least Square in linear regression models. Bold fonts indicate statistical significance ($p < 0.05$).

		Unadjusted model		Model 1		Model 2		Model 3	
		B	95% CI	B	95% CI	B	95% CI	B	95% CI
GP Care	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	21.17	11.79, 30.54	12.10	2.93, 21.28	5.56	-3.52, 14.64	4.20	-4.85, 13.24
	DM	56.33	48.46, 64.21	45.91	37.85, 53.97	30.45	21.86, 39.03	26.39	17.73, 35.05
Pharmacy	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	130.26	51.52, 208.99	96.01	16.46, 175.55	15.06	-61.86, 91.98	-2.78	-79.49, 73.94
	DM	680.54	614.40, 746.70	629.81	559.89, 699.73	430.55	357.83, 503.27	387.14	313.67, 460.62
Hospital care	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	312.48	11.39, 613.57	131.36	-173.01, 435.73	-11.05	-310.58, 288.47	-74.86	-374.78, 225.07
	DM	1179.76	926.81, 1432.72	921.00	653.48, 1188.53	476.22	193.06, 759.38	343.16	55.90, 630.43
Paramedical care	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	17.12	-16.44, 50.68	11.27	-22.80, 45.34	-0.54	-34.71, 33.62	-2.94	-37.32, 31.43
	DM	37.01	8.81, 65.20	31.02	1.07, 60.97	-5.94	-38.24, 26.36	-13.88	-46.80, 19.04
Medical aids	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	27.04	-30.53, 84.61	19.73	-38.82, 78.29	-9.66	-68.44, 49.12	-8.33	-67.24, 50.57
	DM	250.03	201.66, 298.39	236.87	185.4, 288.34	159.32	103.75, 214.89	143.70	87.29, 200.12
Hospital transport	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	2.94	-15.51, 21.39	0.49	-18.29, 19.28	-3.86	-22.94, 15.21	-4.44	-23.59, 14.71
	DM	31.15	15.65, 46.66	28.36	11.85, 44.88	14.39	-3.64, 32.42	10.52	-7.83, 28.86
Specialist mental health (2010-13)	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	40.90	-43.70, 125.49	70.99	-14.80, 156.77	47.08	-39.68, 133.85	48.59	-38.55, 135.72
	DM	3.03	-68.04, 74.10	44.45	-30.96, 119.85	-9.53	-91.55, 72.50	-14.22	-97.67, 69.23
Specialist mental health (2014-16)	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	124.28	21.60, 226.95	158.98	54.54, 263.42	152.63	45.88, 259.39	147.66	40.35, 254.97
	DM	13.19	-73.07, 99.44	67.73	-24.07, 159.53	57.50	-43.43, 158.42	40.00	-62.78, 142.78
Multidisciplinary care	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	68.02	54.31, 81.73	55.77	42.05, 69.49	44.39	30.63, 58.15	45.02	31.20, 58.84
	DM	258.25	246.74, 269.76	240.61	228.55, 252.67	217.00	204.99, 231.00	218.91	205.67, 232.15
Other care	No-DM	Ref.		Ref.		Ref.		Ref.	
	Pre-DM	46.04	34.13, 57.95	39.20	27.16, 51.25	31.94	19.78, 44.10	32.04	19.84, 44.25
	DM	210.54	200.54, 220.55	200.49	189.90, 211.08	183.49	171.99, 194.98	181.75	170.06, 193.44

460.62)]. But the association between prediabetes with these two costs is no longer observable in model 3. We also found only T2DM were statistically

significant associated with the costs of hospital care and medical aids independent of complication and lifestyle features in model 3 [Coef (T2DM - hospital) = 343.16 euros (95% CI = 55.90, 630.43), Coef (T2DM - medical aids) = 143.70 euros (95% CI = 87.29, 200.12)].

An association between T2DM and the costs of paramedical care was observed in model 1 [Coef (T2DM - paramedical)= 31.02 euros (95% CI = 1.07, 60.97)]. In the adjusted models (model 2 and 3), this association was no longer statistically significant in our study [Coef (T2DM - transport) = 14.39 euros (95% CI = -3.64, 32.42)]. T2DM was associated with the cost of hospital transport, independent of complication features in model 2. We observed that prediabetes was associated with specialist mental healthcare cost (2014-2016) when adjusting for complications and lifestyle features in model 3 [Coef (prediabetes-mental)= 147.66 euros (95% CI = 40.35, 254.97)]. However, the similar association was not found from people with T2DM. People with prediabetes and T2DM had significantly higher costs from multidisciplinary care and other healthcare services independent of all potential confounders from the complications and lifestyle categories [Coef (prediabetes - multidis) = 45.02 euros (95% CI = 31.20, 58.84), [Coef (prediabetes - others) = 32.04 euros (95% CI = 19.84, 44.25)], [Coef (T2DM - multidis) = 218.91 euros (95% CI = 205.67, 232.15), [Coef (T2DM - others) = 181.75 euros (95% CI = 170.06, 193.44)]. Both presented tight intervals of 95% CI and high statistical significance. From the estimated coefficients, participants with T2DM spent much more on multidisciplinary care and other healthcare than people with prediabetes.

4.5 Discussion

4.5.1 Technical implementation of the secure infrastructure

This study examined the associations between healthcare costs from ten different healthcare categories and T2DM status considering complications and lifestyle factors on vertically partitioned data using a new privacy preserving data analysis infrastructure. We proved the feasibility of using the infrastructure to securely analyze sensitive personal health data across multiple independent organizations in a real-life use case. The health-related data of 2313 participants from the Maastricht Study was matched and linked with healthcare cost data from 2010 to 2016 of over 1 million data records from Statistics Netherlands using pseudonymised personal identifiable features (gender, date of birth, zip code, house number) achieving the accuracy of data linkage at 96%. Instead of centralizing multiple datasets to researchers or any data sources, our approach established a Trusted Secure Environment (TSE) data

station in the infrastructure with an ethical-legal framework under GDPR to link multiple datasets and execute analyses on the linked data. Researchers are able to send their analysis models to the TSE data station and get the analysis results back after information disclosure checking. The researchers never received any original data or intermediate results which could potentially reveal the original data. In this study, we worked closely with our partners (The Maastricht Study and Statistics Netherlands) and collaborated across disciplines including data science, statistics, computer science, law, and health to install the infrastructure, and establish an ethical-legal framework between two organizations.

Every time the analysis algorithms need to be executed, the requested data is queried, prepared, and pseudonymized at its source. Any changes and updates in the source data can be synchronously reflected in the study. For instance, when participants inform the data organization to withdraw the permissions for using their data, the organization only needs to remove their data records in the data station. The future analysis will not contain any data from the participants who withdraw the permission. It is significantly helpful for data subjects to practice their right to withdraw their consent at any time and reflect on the data use immediately.

We did experiments between two independent data parties using linear regression models, but the number of participating parties and analysis algorithms are not limited. Technically, new parties are able to participate in the joint analysis by employing the infrastructure and generating public-private key pairs and exchanging the public key with TSE. From an ethical-legal perspective, the new parties can construct a joint controller agreement similar to the existing agreement with other participating parties and get informed consent from the participants to link their personal data across organizations.

In the future, some improvements can be made to address the vulnerabilities of the current infrastructure. Firstly, researchers design and send the analysis models to the original data without seeing the actual data. However, in order to maximize the models' performance, researchers usually need to include new or exclude existing features, give different weights to features based on their prior knowledge, or tune the parameters of the models by training the models multiple times. Every model training using the infrastructure requires new pseudonymization and encryption at all data sources and decryption, verification, linking and analysis at TSE which consumes time, computation, and human resources. One approach to solve this limitation can be providing researchers with synthetic data at data stations which is structurally and statistically similar to the real data. Researchers will be able to build the analysis models on the synthetic data, select the important

features, and tune the parameters of models. Once the models perform sufficiently accurate on the synthetic data, researchers can inform all data parties and send the models to the real data. Furthermore, to comply with the data use requirements from one of the participating parties, the final analysis results were checked by experts for the possibility of information disclosure before being exported from the current infrastructure. For example, if the dataset has less than 10 data points, the analysis cannot be conducted because of the high risk of re-identification. In the future work, the output checking can be automated and integrated into the infrastructure to some degree.

4.5.2 Data analysis in the use case

However, our findings have to be seen in light of some potential limitations. Firstly, the declared healthcare costs data in our study covers the major healthcare costs from the Dutch Basic Health Package, but it is possible that people have additional expenses in other national health packages such as Long-Term Care Package for whom require permanent or 24-hour home care or Supplementary Insurance Package which is fully private in nature. Some previous studies, which examined the relation between diabetes and healthcare cost using data from different sources found comparable discovery but using different data sources [35, 11]. So far, there is no single source that could provide the full healthcare costs data [11]. Secondly, the policies for the declaration and the coverage of healthcare services in the Dutch Basic Health Package may change according to the updated policies in the Dutch Health Insurance Act. For example, in 2012, the number of physiotherapy and remedial therapy sessions for chronic disorders that could be reimbursed was increased from 12 to 20. As participants got more treatments reimbursed, their costs in the category of paramedical care increased correspondingly. Another change happened in the mental health care category in 2014. The primary (first-line) psychological care and (second-line) mental health care were replaced by basic and specialist mental health care. Patients with non-complex mental disorders or requiring mild to moderate treatment are referred to the basic mental health care. Patients with complex disorders are categorized as specialist mental health care. This change could have a potential impact on the change of healthcare costs of the participants.

Furthermore, rather than only focusing on the direct costs for T2DM treatments or medications, this study includes the total healthcare costs which includes T2DM and other comorbidities. The previous studies showed 85% of T2DM patients suffering from at least one other chronic condition [11, 36]. However, due to the absence of the detailed information on the provided healthcare service itself, we could not distinguish the costs for T2DM from

other comorbidities. Finally, due to differences in healthcare systems, medical practices, costs and declaration policies, it is difficult to accurately compare our results with those based on other populations [37].

4.6 Conclusion

In this paper, we proved the feasibility of the proposed privacy-preserving infrastructure in a real-life use case using personal health data which are vertically partitioned at The Maastricht Study and Statistics Netherlands. As an extension of Personal Health Train architecture, the infrastructure sends the analysis models built by the researcher to the data, in an attempt to minimize centralizing data from multiple sources to the researcher's site. The data from different sources was linked and analyzed at a Trusted Secure Environment which is considered as an independent entity supported by an ethical-legal framework. In the use case, we examined the association between prediabetes and T2DM and annual healthcare costs in different healthcare categories considering the impact of complications and lifestyle factors. We discovered that individuals diagnosed with T2DM had significantly higher expenses than those with prediabetes, while participants with prediabetes spent more than those without T2DM in all the included healthcare categories.

References

- [1] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1 (2016), pp. 1–9. DOI: 10.1038/sdata.2016.18.
- [2] Chang Sun et al. "A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario". In: *Studies in Health Technology and Informatics* 264 (2019), pp. 373–377. DOI: 10.3233/SHTI190246.
- [3] Oya Beyan et al. "Distributed Analytics on Sensitive Medical Data: The Personal Health Train". In: *Data Intelligence* 2.1-2 (2020), pp. 96–107. DOI: 10.1162/dint.a.00032.

-
- [4] Johan van Soest, Chang Sun, Ole Mussmann, Marco Puts, Bob van den Berg, Alexander Malic, Claudia van Oppen, David Townend, Andre Dekker, and Michel Dumontier. "Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data." In: *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*. Vol. 247. IOS Press, 2018, pp. 581–585. DOI: 10.3233/978-1-61499-852-5-581.
- [5] Chang Sun, Lianne Ippel, Andre Dekker, Michel Dumontier, and Johan van Soest. "A systematic review on privacy-preserving distributed data mining". English. In: *Data Science* 4.2 (Oct. 2021), pp. 121–150. DOI: 10.3233/DS-210036.
- [6] R. Schnell. "Efficient private record linkage of very large datasets". In: *59th World Statistics Congress of the International Statistical Institute*. International Statistical Institute, 2013. URL: <https://openaccess.city.ac.uk/id/eprint/14652/>.
- [7] Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records". In: *Journal of biomedical informatics* 99 (2019), p. 103291. DOI: 10.1016/j.jbi.2019.103291.
- [8] Elena Czeizler, Wolfgang Wiessler, Thorben Koester, Mikko Hakala, Shahab Basiri, Petr Jordan, and Esa Kuusela. "Using federated data sources and Varian Learning Portal framework to train a neural network model for automatic organ segmentation". In: *Physica Medica* 72 (2020), pp. 39–45. DOI: 10.1016/j.ejmp.2020.03.011.
- [9] Miranda T Schram, Simone JS Sep, Carla J van der Kallen, Pieter C Dagnelie, Annemarie Koster, Nicolaas Schaper, Ronald Henry, and Coen DA Stehouwer. "The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities". In: *European journal of epidemiology* 29.6 (2014), pp. 439–451. DOI: 10.1007/s10654-014-9889-0.
- [10] Gojka Roglic and World Health Organization, eds. *Global report on diabetes*. Geneva, Switzerland: World Health Organization, 2016. 86 pp. URL: <https://www.who.int/publications/i/item/9789241565257>.
- [11] M L Peters, E L Huisman, M Schoonen, and B H R Wolffenbuttel. "The current total economic burden of diabetes mellitus in the Netherlands". In: *The Netherlands Journal of Medicine* 75.7 (2017), p. 17. URL: <https://pubmed.ncbi.nlm.nih.gov/28956787/>.

- [12] International Diabetes Federation. *IDF Diabetes Atlas 9th edition 2019*. Brussels, 2019. URL: <https://www.diabetesatlas.org/en/> (visited on 07/19/2021).
- [13] Matti Uusitupa. “Lifestyles Matter in the Prevention of Type 2 Diabetes”. In: *Diabetes Care* 25.9 (Sept. 2002), pp. 1650–1651. DOI: 10.2337/diacare.25.9.1650.
- [14] Timo M. Deist et al. “Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT”. In: *Clinical and Translational Radiation Oncology* 4 (2017), pp. 24–31. DOI: 10.1016/j.ctro.2016.12.004.
- [15] Arthur Jochems, Timo M Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. “Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital—a real life proof of concept”. In: *Radiotherapy and Oncology* 121.3 (2016), pp. 459–467. DOI: 10.1016/j.radonc.2016.10.002.
- [16] Timo M. Deist et al. “Distributed learning on 20 000+ lung cancer patients – The Personal Health Train”. In: *Radiotherapy and Oncology* 144 (2020), pp. 189–200. DOI: 10.1016/j.radonc.2019.11.019.
- [17] Birgit Wouters et al. “Putting the GDPR into Practice: Difficulties and Uncertainties Experienced in the Conduct of Big Data Health Research”. In: *European Data Protection Law Review* 7.2 (2021). DOI: 10.21552/edpl/2021/2/9.
- [18] Daniel J. Bernstein. “The Salsa20 Family of Stream Ciphers”. In: *New Stream Cipher Designs: The eSTREAM Finalists*. Ed. by Matthew Robshaw and Olivier Billet. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 84–97. DOI: 10.1007/978-3-540-68351-3_8.
- [19] Simon Josefsson and Ilari Liusvaara. “Edwards-curve digital signature algorithm (eddsa)”. In: *Internet Research Task Force, Crypto Forum Research Group, RFC*. Vol. 8032. 2017, pp. 257–260. URL: <https://www.rfc-editor.org/rfc/pdfrfc/rfc8032.txt.pdf>.
- [20] Alex Biryukov and Léo Perrin. “State of the Art in Lightweight Symmetric Cryptography”. In: *IACR Cryptol. ePrint Arch.* 2017 (2017), p. 511. URL: <https://eprint.iacr.org/2017/511.pdf>.
- [21] Jacqueline Brendel, Cas Cremers, Dennis Jackson, and Mang Zhao. “The Provable Security of Ed25519: Theory and Practice”. In: *2021 IEEE Symposium on Security and Privacy (SP)*. 2021, pp. 1659–1676. DOI: 10.1109/SP40001.2021.00042.

-
- [22] Daniel J Bernstein, Tung Chou, and Peter Schwabe. "McBits: fast constant-time code-based cryptography". In: *International Conference on Cryptographic Hardware and Embedded Systems*. Springer. 2013, pp. 250–272. DOI: 10.1007/978-3-642-40349-1_15.
- [23] Tung Chou. "McBits Revisited". In: *Cryptographic Hardware and Embedded Systems – CHES 2017*. Ed. by Wieland Fischer and Naofumi Homma. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 213–231. DOI: 10.1007/978-3-319-66787-4_11.
- [24] Karl Pearson. "Note on Regression and Inheritance in the Case of Two Parents". In: *Proceedings of the Royal Society of London Series I* 58 (1895), pp. 240–242. DOI: 10.1098/rspl.1895.0041.
- [25] Emelie Andersson, Sofie Persson, Nino Hall'en, Åsa Ericsson, Desir'ee Thielke, Peter Lindgren, Katarina Steen Carlsson, and Johan Jendle. "Costs of diabetes complications: hospital-based care and absence from work for 392,200 people with type 2 diabetes and matched control participants in Sweden". In: *Diabetologia* 63.12 (2020), pp. 2582–2594. DOI: 10.1007/s00125-020-05277-3.
- [26] Katharina Kahm, Michael Laxy, Udo Schneider, Wolf H Rogowski, Stefan K Lhachimi, and Rolf Holle. "Health care costs associated with incident complications in patients with type 2 diabetes in Germany". In: *Diabetes Care* 41.5 (2018), pp. 971–978. DOI: 10.2337/dc17-1763.
- [27] Ulf-G Gerdtham, Philip Clarke, Alison Hayes, and Soffia Gudbjornsdottir. "Estimating the cost of diabetes mellitus-related events from inpatient admissions in Sweden using administrative hospitalization data". In: *Pharmacoeconomics* 27.1 (2009), pp. 81–90. DOI: 10.1155/2019/2363292.
- [28] World Health Organization. "Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia : report of a WHO/IDF consultation". In: *Geneva: World Health Organization* 3 (2006). URL: <https://apps.who.int/iris/handle/10665/43588>.
- [29] Adam G Tabak, Christian Herder, Wolfgang Rathmann, Eric J Brunner, and Mika Kivimaki. "Prediabetes: a high-risk state for diabetes development". In: *The Lancet* 379.9833 (2012), pp. 2279–2290. DOI: 10.1016/S0140-6736(12)60283-9.
- [30] Antonia Trichopoulou, Tina Costacou, Christina Bamia, and Dimitrios Trichopoulos. "Adherence to a Mediterranean diet and survival in a Greek population". In: *New England Journal of Medicine* 348.26 (2003), pp. 2599–2608. DOI: 10.1056/NEJMoa025039.

- [31] Martien CJM van Dongen, Nicole EG Wijckmans-Duysens, Louise JCD den Biggelaar, Marga C Ocké, Saskia Meijboom, Henny AM Brants, Jeanne HM de Vries, Edith JM Feskens, H Bas Bueno-de-Mesquita, Anouk Geelen, et al. "The Maastricht FFQ: development and validation of a comprehensive food frequency questionnaire for the Maastricht study". In: *Nutrition* 62 (2019), pp. 39–46. DOI: 10.1016/j.nut.2018.10.015.
- [32] Welzijn en Sport Ministerie van Volksgezondheid. *Healthcare in the Netherlands - Leaflet - Government.nl*. Dec. 17, 2018. URL: <https://www.government.nl/documents/leaflets/2016/02/09/healthcare-in-the-netherlands> (visited on 07/19/2021).
- [33] Centraal Bureau voor de Statistiek. *Zvwzorgkostentab: Zorgkosten personen basisverzekering*. Centraal Bureau voor de Statistiek. URL: <https://www.cbs.nl/nl-nl/onzediensten/maatwerk-en-microdata/microdata-zelf-onderzoekdoen/microdatabestanden/zvwzorgkostentab-zorgkosten-personen-basisverzekering>.
- [34] World Health Organization. "Noncommunicable diseases country profiles 2018". In: (2018). URL: <https://apps.who.int/iris/handle/10665/274512>.
- [35] L. M. M. Janssen et al. "Burden of disease of type 2 diabetes mellitus: cost of illness and quality of life estimated using the Maastricht Study". In: *Diabetic Medicine* 37.10 (2020), pp. 1759–1765. DOI: <https://doi.org/10.1111/dme.14285>.
- [36] GEHM Rutten, WJC De Grauw, Giel Nijpels, ST Houweling, FA Van de Laar, HJ Bilo, Frits Holleman, JS Burgers, TJ Wiersma, and PGH Janssen. "NHG-Standaard Diabetes mellitus type 2 (derde herziening)". In: *Huisarts Wet* 56.10 (2013), pp. 512–525. URL: <https://richtlijnen.nhg.org/standaarden/diabetes-mellitus-type-2>.
- [37] Carola A Huber, Matthias Schwenkglens, Roland Rapold, and Oliver Reich. "Epidemiology and costs of diabetes mellitus in Switzerland: an analysis of health care claims data, 2006 and 2011". In: *BMC endocrine disorders* 14.1 (2014), pp. 1–9. DOI: 10.1186/1472-6823-14-44.
- [38] John E Ware Jr and Cathy Donald Sherbourne. "The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection". In: *Medical care* (1992), pp. 473–483. URL: <https://pubmed.ncbi.nlm.nih.gov/1593914/>.
- [39] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. "'Minimal state": a practical method for grading the cognitive state of patients for the clinician". In: *Journal of psychiatric research* 12.3 (1975), pp. 189–198. DOI: 10.1016/0022-3956(75)90026-6.

-
- [40] KEN Resnicow, Frances McCarty, Dhana Blissett, Terry Wang, Carrie Heitzler, and Rebecca E Lee. "Validity of a modified CHAMPS physical activity questionnaire among African-Americans." In: *Medicine & Science in Sports & Exercise* (2003). DOI: 10.1249/01.MSS.0000084419.64044.2B.
- [41] Welzijn en Sport Ministerie van Volksgezondheid. *Basispakket Zvw - Verzekerde zorg - Zorginstituut Nederland*. July 24, 2017. URL: <https://www.zorginstituutnederland.nl/Verzekerde+zorg/basispakket-zorgverzekeringwet-zvw>.

Supplementary

Extended details in collection method and measurement of health-related data

Diabetes outcome. The definition of T2DM status was based on the World Health Organization's diagnostic criteria of glucose tolerance status [28]. All participants underwent a standardized 7-point OGTT after overnight fasting except the participants who were insulin-dependent and participants with a fasting glucose level higher than 11.0 mmol/l (as determined by finger prick). Participants were categorized into no-diabetes, prediabetes, and T2DM. Participants with no diabetes have normal glucose metabolism (fasting plasma glucose < 6.1 mmol/l and 2 h plasma glucose (after glucose load) < 7.8 mmol/l). Participants with pre-diabetes have impaired fasting glucose (fasting plasma glucose ≥ 6.1 mmol/l and 2 h plasma glucose (after glucose load) < 7.8 mmol/l) or impaired glucose tolerance (fasting plasma glucose < 7.0 mmol/l and 2 h plasma glucose (after glucose load) $\geq 7.0 - 11.1$ mmol/l). Pre-diabetes is an intermediate but high-risk state for diabetes. People with pre-diabetes have glycaemic variables that are higher than normal, but lower than diabetes thresholds. 5-10% of people per year with prediabetes will progress to diabetes, with the same proportion converting back to normoglycemia [29]. Participants with T2DM have fasting plasma glucose ≥ 7.0 mmol/l or 2 h plasma glucose (after glucose load) ≥ 11.1 mmol/l. Participants on diabetes medication and without type 1 diabetes were also categorized into T2DM class.

Complications category includes calculated BMI and self-reported information on mobility limitation, history of cardiovascular disease, hypertension, and lifetime depression. BMI was calculated as weight (kg) / height² (m) based on the measurement of weights and heights of the participant. Mobility limitation (yes/no) was defined as whether the participant has difficulty walking 500 m or climbing the stairs in the 36-Item Short Form Health Survey questionnaire [38]. History of cardiovascular disease (yes/no) was defined as a history of myocardial infarction, cerebrovascular infarction or hemorrhage, or percutaneous artery angioplasty of, or vascular surgery on, the coronary, abdominal, peripheral, or carotid arteries. Hypertension (yes/no) was based on the average blood pressure (systolic blood pressure ≥ 140 mmHg) and antihypertensive medication use (diastolic blood pressure ≥ 90). Lifetime depression (yes/no) was assessed by the Mini International Neuropsychiatric Interview (MINI) which is a short diagnostic structured interview to assess the presence of minor or major depressive disorder in the participant's lifetime [39].

Lifestyle factors include the following self-reported health behavior: physical activity, dietary, smoking and alcohol consumption. Physical activity was measured by modified CHAMPS questionnaire [40] and categorized into 1) low physical activity (0 h to 9.75 h per week); 2) medium physical activity (9.76 h to 16.25 h per week); and 3) high physical activity (more than 16.25 hours per week). Participants' dietary patterns were collected by a Dutch national tailor-made Food Frequency Questionnaire (FFQ) which assesses dietary intake frequency, amount of foods, and nutrients [31]. The outcome of FFQ was scored (0-9 points) by the Greek Mediterranean-diet scale as the main measurement of dietary intakes [30, 31]. The diet score ranged from 0 to 9 indicating minimal to maximal adherence to the Greek Mediterranean diet. The Greek Mediterranean diet is low in saturated fat, high in monounsaturated fat, high in complex carbohydrates (from grains and legumes), and high in fibre (vegetables and fruits) [30]. Smoking (cigarettes, cigars and/or pipe tobacco) status was categorized as: never smokers, former smokers, and current smokers. Alcohol use was categorized as: non-consumers (never consumed alcohol), low-consumers (for women consuming ≤ 7 glasses of alcohol per week and for men consuming ≤ 14 glasses of alcohol per week), for high consumers (for women consuming > 7 glasses per week and for men >14 glasses a week).

The coverage of each Dutch healthcare category

Table .4: Cost coverage of each healthcare cost feature of healthcare providers [41]

Cost	Period	Coverage
General practitioner (family doctor)	2010-2016	Costs of registration fees, consultation fees, GP practice module (POH), GP practice support (POH) GGZ, arrears fund, modernization and innovation, evening, night and weekend services (ANW), and other costs for general practitioner care (including passer-by rates).
Pharmacy	2010-2016	Costs of pharmaceutical assistance, including pharmacists' fees and practical costs, module rate pharmaceutical help by dispensing general practitioners and module rate separate from care and trade.
Hospital care	2010-2016	Costs for regulated and free segment DBC care products, add-ons, specialists for oral diseases and oral surgery, extramural specialists, and other hospital care and curative care costs.
Paramedical care	2010-2016	Costs for physiotherapy, remedial therapy, speech therapy, occupational therapy and dietary advice.
Medical aids	2010-2016	Costs for the medical devices that people use at home. The content of the cover varies from personal care items (e.g. incontinence materials and diabetes test strips) to equipment (e.g. hearing aids and orthopaedic footwear).
Hospital transport	2010-2016	Cost of transport by ambulance, helicopter, taxi, public transport, and own car.
Primary psychological / basic mental healthcare	2010-2013	Cost of the treatment of mild to moderate psychological problems
GGZ / specialist GGZ care	2014-2016	Cost of the treatment of more serious psychological problems
Special cost of general practitioner care	2015-2016	Costs of multidisciplinary care, remuneration of results and care innovation for GPs, and reward for results and innovation in multidisciplinary care.

5

Privacy-Preserving Generalized Linear Models on Vertically Partitioned Data using Distributed Block Coordinate Descent

Adapted from: Erik-Jan van Kesteren, Chang Sun, Daniel L Oberski, Michel Dumontier, and Lianne Ippel. “Privacy-Preserving Generalized Linear Models using Distributed Block Coordinate Descent”. In Revision: *International Journal of Data Science and Analytics* (2022) DOI: 10.48550/arXiv.1911.03183

Abstract

Combining data from varied sources has considerable potential for knowledge discovery: collaborating data parties can mine data in an expanded feature space, allowing them to explore a larger range of scientific questions. However, data sharing among different parties is highly restricted by legal conditions, ethical concerns, and/or data volume. Fueled by these concerns, the fields of cryptography and distributed learning have made great progress towards privacy-preserving distributed data mining. However, practical implementations have been hampered by the limited scope or computational complexity of these methods. In this paper, we greatly extend the range of analyses available for vertically partitioned data, i.e., data collected by separate parties with different features on the same subjects. To this end, we present a novel approach for privacy-preserving generalized linear models, a fundamental and powerful framework underlying many prediction and classification procedures. We base our method on a distributed *block coordinate descent* algorithm to obtain parameter estimates, and we develop an extension to compute accurate standard errors without additional communication cost. We critically evaluate the information transfer for semi-honest collaborators and show that our protocol is secure against data reconstruction. Through both simulated and real-world examples we illustrate the functionality of our proposed algorithm. Without leaking information, our method performs as well on vertically partitioned data as existing methods on combined data – all within mere minutes of computation time. We conclude that our method is a viable approach for vertically partitioned data analysis with a range of real-world applications.

5.1 Introduction

With technological developments in computational power and storage capacity, an increasing amount of data is collected and stored by a variety of data parties [1]. Over the past decades, data mining has been successful in extracting information from such datasets, but it is especially powerful when various data sources are combined: collaborating data parties can mine data in a larger feature space, allowing them to discover knowledge beyond their individual potential. For example, in the medical domain, personal health conditions are significantly affected not only by genetic and biological factors, but also by individual behaviour and social circumstances [2]. Combining those sources has the potential to improve analytical models for health outcomes [3, 4].

However, there is a pertinent obstacle to unlocking the potential of combining datasets: integrating various sources may reveal private information about individual data subjects to the collaborating parties. Hence, data sharing is highly restricted by legal and ethical concerns. This highlights the need for privacy-preserving techniques which perform data mining tasks on multiple sources without explicitly sharing their full data [5, 6, 7, 8]. In this paper, we extend the range of analyses available in such a vertically partitioned data situation. Specifically, the contributions of this paper are as follows:

1. we develop a distributed block coordinate descent (BCD) algorithm for performing generalized linear modeling (GLM) in vertically partitioned data across two or more parties.
2. we create a completely novel extension to BCD for computing standard errors without additional communication cost.
3. we analyze the privacy-preserving properties and information transfer associated with this algorithm for semi-honest parties.
4. we provide an open-source implementation of this algorithm and show experimentally that it performs as well as existing GLM methods in real-world datasets in acceptable time.

These contributions push the boundary of data analysis with vertically partitioned data, as GLM is a powerful framework for prediction and classification at the basis of a wide range of analysis applications, including linear models, count and survival models, and logistic regression [9, 10, 11]. All of these methods are implemented in `privreg`, an open-source software package for the R programming language [12]. This implementation includes encryption for all communication across parties based on a pre-shared key, and includes a user-friendly interface based around an object-oriented architecture. The package is available for installation from the supplementary materials.

This paper is organized as follows. In Section 5.2, related work is discussed to contextualize our contribution. In Section 5.3, we introduce our proposed method for GLM on vertically partitioned data. Next, we describe the information sharing characteristics of this protocol in Section 5.4, and we analyze how the information transfer affects the ability of the partner organisation to recover the collaborator's data. In Section 5.5, we benchmark our implementation of the protocol against full-data analysis using three different real-life data sets from the UCI Machine Learning repository [13]. Finally, we discuss the strengths and limitations of our approach in Section 5.6 and provide suggestions for future research.

5.2 Related work

In practice, there are two main types of data partitioning [14]. Different data sources might collect the same features of different data subjects, e.g., different hospitals collect the same type of data from their own set of patients. This situation is referred to as *horizontally partitioned* data. Alternatively, separate sources might collect different information from the same data subjects, e.g., medical data collected by a hospital may be combined with socioeconomic data collected by a governmental department from the same group of people. This situation is referred to as *vertically partitioned* data, which is the focus of the current paper. A third scenario, where data are both vertically and horizontally partitioned, is referred to as *hybrid partitioning*.

This study aims to perform supervised learning on vertically partitioned data without raw data leaks between the collaborating parties (*Alice* and *Bob*). In order to analyze such data, either the dataset may be combined but hidden from the collaborating parties, or the analytical procedure should prevent leaking of information. The former relies on the inclusion of a trusted third party (TTP): Each party sends their encrypted data to the TTP, who then performs the required analyses on the combined data sets. Afterwards, the TTP returns the results to all data parties and the raw data of *Alice* stays hidden to *Bob*. However, this solution requires all parties to fully trust the TTP, which might not be possible in the face of restrictive legislation or sensitive data.

There is another class of methods which do not rely on a TTP, instead using cryptography to perform data mining tasks on vertically partitioned data. These methods focus on preventing information leakage by creating protocols which hide the raw data from the collaborators (e.g., for the construction of decision trees [15]). Secret sharing protocols, encryption schemes, or a combination of them are commonly applied in this class of methods. Du et al. [5, 16] investigated combining an oblivious transfer protocol and homomorphic encryption schemes to perform secure matrix computation for linear

least squares regression and classification problems. However, the methods are limited to only two participating parties. Several studies [8, 17, 18, 19] applied and extended more general secure multiparty computation protocols (e.g., the garbled circuit protocol [20]) to perform linear regression on vertically partitioned data. [8, 17] and [18] assume the model training is delegated to a small group of computing servers that do not collude with each other. However, due to the expensive communication cost, the computation protocols are challenging to be ported cross multiple servers in practice [21]. If any collusion exists among the parties, the methods are not sufficiently secure anymore [22, 23]. This class of methods also have the potential risk of information leakage due to intermediate data shared by all parties [19]. It is possible for some parties to deduce the original data from other parties. Fang et al. [19] applied fully homomorphic encryption in their computation protocols and required no interaction among parties to avoid information leakage from intermediate data. However, their method cannot be generalized to non-linear models.

Yet another approach leverages algorithms from *federated* or *distributed learning*, a field researching data mining on separated datasets [24, 25, 26]. One of the early implementations and canonical examples is proposed by [27], who developed a method to compute global linear regression coefficients iteratively based on an algorithm by [28]. However, the security during computation is not equally guaranteed for different participating parties. The party who has smaller number of variables or who starts the training iteration is less secure compared to others. Other authors leverage specific distributed learning algorithms to implement regression models for vertically partitioned data such as ridge regression [29], logistic regression [30, 31], and a three-server regression model [32]. Our method is closely related to this branch of research. Unlike existing regression methods from the TTP or cryptography fields, our method does not make use of a trust assumption or complex cryptographic protocols, but it relies on a federated learning algorithm which never moves the data from its original location. In the next section, we explain the concept and implementation behind our proposed privacy-preserving GLM technique.

5.3 Materials and methods

We propose using a distributed form of *block coordinate descent* (BCD) to estimate generalized linear models in a situation where data is vertically partitioned across two or more parties. In BCD, parameters are iteratively updated for each block of features, cycling over the blocks until an optimum is found [33]. This optimization algorithm can be seen as a form of distributed

learning [34, 35] where the features remain in different locations. Only linear predictions need to be transferred across the feature blocks – the full data is never shared.

For the remainder of the paper, we assume that the records of the data subjects are in the same order across databases, in line with [8]. This can be done using pre-existing identifiers or via probabilistic record linkage [36]. Furthermore, we only consider the situation where the target attribute is available to both parties, following [27]. This restriction can be relaxed in case correlated targets are available at different locations – a situation akin to distributed multi-task learning [37, 38, 39] – but for simplicity we leave this extension to future work.

In this section, we build up the BCD algorithm from the simpler case of linear regression before extending it to full GLM. Therefore, we first explain the necessary background on linear regression, as well as the notation used throughout this paper. Then, coordinate descent estimation is introduced as a means to estimate its maximum likelihood coefficients. In Section 5.3.3, this algorithm is then extended to accommodate a vertically partitioned data structure, and in Section 5.3.4 we generalize it to different outcome families in order to estimate GLMs. Finally, we develop a novel method to obtain standard errors within this framework.

5.3.1 Background

We consider the centered design matrix with features $\mathbf{X} \in \mathbb{R}^{N \times P}$ and the centered target variable $\mathbf{y} \in \mathbb{R}^{N \times 1}$, where N is the sample size, or number of observations, and P is the number of features. The p^{th} column in \mathbf{X} is represented as \mathbf{x}_p . The columns in \mathbf{X} excluding the p^{th} are denoted as \mathbf{X}_{-p} .

The basic regression model is then as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^P$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, and $\boldsymbol{\epsilon} \perp \mathbf{X}$. The well-known closed-form maximum likelihood estimator of the P regression coefficients $\boldsymbol{\beta}$ in this model is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \tag{5.2}$$

We further define the vector of predicted values as $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and the vector of residuals as $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}$.

5.3.2 Cyclic coordinate descent estimation

When instead of the full design matrix \mathbf{X} we consider only the p^{th} variable, the estimator in Equation 5.1 yields the *marginal* regression coefficient. Thus, by simplifying Equation 5.1 to the univariate case, the marginal coefficient for the p^{th} variable β_p^* is estimated as

$$\hat{\beta}_p^* = \frac{\langle \mathbf{x}_p, \mathbf{y} \rangle}{\langle \mathbf{x}_p, \mathbf{x}_p \rangle} = \frac{\text{cov}(\mathbf{x}_p, \mathbf{y})}{\text{var}(\mathbf{x}_p)} \quad (5.3)$$

where $\langle \cdot, \cdot \rangle$ indicates the inner product of two vectors. The covariance/variance notation holds because we assume a centered design matrix \mathbf{X} and outcome variable \mathbf{y} .

If \mathbf{x}_p covaries with any of the predictors in \mathbf{X}_{-p} , the marginal coefficient β_p^* is different from the *conditional* coefficient β_p . The estimate of this coefficient is an element of $\hat{\beta}$ in Equation 5.1, but it can equivalently be estimated in a coordinate-wise, univariate manner [33] as follows:

$$\begin{aligned} \hat{\beta}_p &= \frac{\langle \mathbf{x}_p, \hat{\boldsymbol{\epsilon}}_{-p} \rangle}{\langle \mathbf{x}_p, \mathbf{x}_p \rangle} = \frac{\langle \mathbf{x}_p, \mathbf{y} - \mathbf{X}_{-p} \hat{\boldsymbol{\beta}}_{-p} \rangle}{\langle \mathbf{x}_p, \mathbf{x}_p \rangle} \\ &= \frac{\langle \mathbf{x}_p, \mathbf{y} \rangle}{\langle \mathbf{x}_p, \mathbf{x}_p \rangle} - \frac{\langle \mathbf{x}_p, \mathbf{X}_{-p} \hat{\boldsymbol{\beta}}_{-p} \rangle}{\langle \mathbf{x}_p, \mathbf{x}_p \rangle} \end{aligned} \quad (5.4)$$

The residual $\hat{\boldsymbol{\epsilon}}_{-p} = \mathbf{y} - \mathbf{X}_{-p} \hat{\boldsymbol{\beta}}_{-p}$ is the residual with respect to the variables excluding \mathbf{x}_p , evaluated at the maximum likelihood (ML) estimates of $\boldsymbol{\beta}$. Equation 5.4 states that the conditional regression coefficient can be obtained by computing the marginal regression coefficient of $\hat{\boldsymbol{\epsilon}}_{-p}$ on \mathbf{x}_p . This relation holds because $\hat{\boldsymbol{\epsilon}}_{-p}$ represents the part of the outcome variable unrelated to \mathbf{X}_{-p} – by definition, $\hat{\boldsymbol{\epsilon}}_{-p} \perp \mathbf{X}_{-p}$. In addition, the last part of Equation 5.4 shows that the marginal and conditional estimate of the p^{th} regression coefficient are equal if \mathbf{x}_p and \mathbf{X}_{-p} do not covary, because the last term drops out.

The coordinate-wise estimation of $\hat{\beta}_p$ (Equation 5.4) requires the maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{-p}$ of the remaining variables to be known. However, when estimation of $\hat{\beta}$ is the goal, these estimates are not available. This can be solved by an iterative updating procedure of the $\hat{\beta}$ estimates:

Algorithm 1: Cyclic coordinate descent [33]

1. Initialize $\hat{\beta} \leftarrow \hat{\beta}^*$ (marginal coefficients)
2. For each $p \in P$:
 - a) $\hat{\epsilon}_{-p} \leftarrow \mathbf{y} - \mathbf{X}_{-p}\hat{\beta}_{-p}$
 - b) $\hat{\beta}_p \leftarrow \langle \mathbf{x}_p, \hat{\epsilon}_{-p} \rangle / \langle \mathbf{x}_p, \mathbf{x}_p \rangle$
3. Repeat step (2.) for R iterations until convergence (i.e., the change in parameter estimates over iterations becomes negligible)

An advantage of this method is that it does not require storing the full $P \times P$ covariance matrix in memory, and this matrix does not need to be inverted – an $\mathcal{O}(P^3)$ operation. This advantage becomes especially relevant as P grows [33]. Another advantage is that this estimation method allows for regularization to be implemented naturally. For example, the ℓ_1 penalized parameters can be computed by soft-thresholding $\langle \mathbf{x}_p, \hat{\epsilon}_{-p} \rangle$ in each iteration. This is the approach taken by the popular regularized regression package `glmnet` [40].

A graphical display of the behaviour of the estimated parameters during the cyclical coordinate descent procedure is shown in panel A of Figure 5.1. Here, 9 covarying features \mathbf{X} were generated from a multivariate normal distribution. Then random parameter values β and random normal errors ϵ were created and used to generate the target variable $\mathbf{y} = \mathbf{X}\beta + \epsilon$.

Next, we show how coordinate descent generalizes to blocks of variables, and how it may be used to estimate linear regression coefficients in the vertically partitioned data scenario described above.

5.3.3 Securely estimating coefficients for linear regression

In this section, we develop the framework for analysing vertically partitioned data. Our key contribution is the combination of two observations:

1. Coordinate descent estimation works the same for single features as well as for blocks of features – resulting in a variant called block coordinate descent (BCD) [33].
2. Vertically partitioned data is blocked data – the features held by *Alice* can be considered the 1st block, and those held by *Bob* the 2nd block.

Following these two observations, Algorithm 2 below thus provides an iterative estimator for the parameters of *Alice* (β_a) and those of *Bob* (β_b) through

sharing of predictions. Predictions from *Alice* are written as $\hat{\mathbf{y}}_a = \mathbf{X}_a \hat{\boldsymbol{\beta}}_a$, and the working residual with respect to *Alice*, i.e., the part of \mathbf{y} not related to the features in \mathbf{X}_a is then $\hat{\boldsymbol{\epsilon}}_a = \mathbf{y} - \hat{\mathbf{y}}_a$.

Algorithm 2: Secure block coordinate descent

1. Initialize $\hat{\mathbf{y}}_b \leftarrow \mathbf{0}$
2. *Alice*:
 - a) $\hat{\boldsymbol{\epsilon}}_b \leftarrow \mathbf{y} - \hat{\mathbf{y}}_b$
 - b) $\hat{\boldsymbol{\beta}}_a \leftarrow (\mathbf{X}_a^T \mathbf{X}_a)^{-1} \mathbf{X}_a^T \hat{\boldsymbol{\epsilon}}_b$
 - c) $\hat{\mathbf{y}}_a \leftarrow \mathbf{X}_a \hat{\boldsymbol{\beta}}_a$
 - d) Send $\hat{\mathbf{y}}_a$ to *Bob*
3. *Bob*:
 - a) $\hat{\boldsymbol{\epsilon}}_a \leftarrow \mathbf{y} - \hat{\mathbf{y}}_a$
 - b) $\hat{\boldsymbol{\beta}}_b \leftarrow (\mathbf{X}_b^T \mathbf{X}_b)^{-1} \mathbf{X}_b^T \hat{\boldsymbol{\epsilon}}_a$
 - c) $\hat{\mathbf{y}}_b \leftarrow \mathbf{X}_b \hat{\boldsymbol{\beta}}_b$
 - d) Send $\hat{\mathbf{y}}_b$ to *Alice*
4. Repeat step (2.) and (3.) for R iterations until convergence.

Since the least-squares objective is strictly convex, the above algorithm is guaranteed to converge to the global minimum [41]. Upon convergence, the concatenated parameter estimates vector $(\hat{\boldsymbol{\beta}}_a, \hat{\boldsymbol{\beta}}_b)$ is equal (up to a small pre-determined tolerance value) to the parameter estimates vector that would be obtained using the standard maximum likelihood estimator in the combined data set [42]. It follows that the element-wise summed prediction $\hat{\mathbf{y}}_a + \hat{\mathbf{y}}_b$ is equal to the prediction $\hat{\mathbf{y}}$ that would be obtained from the combined dataset. Thus, prediction can be done without sharing the parameter estimates. Further analysis of the privacy-preserving properties of this procedure is discussed in Section 5.4.

In panel B of Figure 5.1 we illustrate BCD, applied to the same data set as in panel A. However, instead of P blocks of 1 feature each, now there are two blocks with 5 and 4 features. BCD reaches convergence with fewer iterations than the cyclic version, because it uses more information about the covariance between the features. In general, convergence is obtained faster with fewer blocks, and with less covariance between blocks [35]. In the case of orthogonal blocks, only a single iteration is needed for convergence as the marginal estimates equal the conditional estimates. Li et al. [43, Theorem 8] derived a general result about the iteration complexity of BCD, showing that

for smooth convex losses such as the GLM log-likelihood, the number of iterations required for convergence is linear in the number of features P .

In the next section, we show how our BCD approach may be modified to estimate generalized linear model coefficients for a wide range of applications. Then, we provide a way to estimate standard errors within this framework.

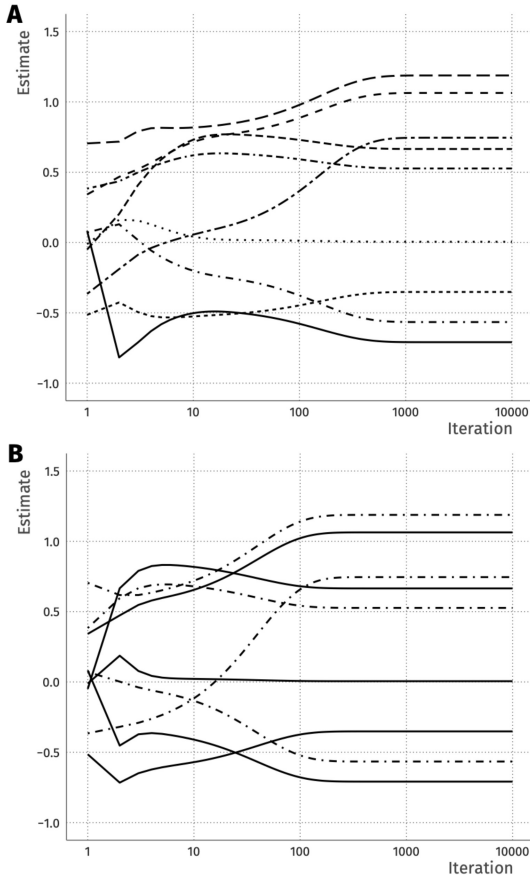


Figure 5.1: Panel A: Coordinate descent paths for linear regression with 9 covarying features, simulated from a multivariate normal distribution. The parameter lines converge from the marginal ML estimates (iteration 1) to the conditional ML estimates (iteration 10000). Note that the x-axis is on a logarithmic scale and convergence happens around iteration 1000. Panel B: Block coordinate descent path for regression with 9 covarying predictors, applied to the same simulated dataset. There are two blocks, indicated by the line types. Note that convergence happens before iteration 500, faster than the cyclic coordinate descent algorithm.

5.3.4 Extension to generalized linear models

Extending this procedure to generalized linear models (GLM) requires a slightly different estimation approach: whereas the parameter estimates of full-data linear regression can be found analytically (Equation 5.2), GLM requires an iteratively reweighted least squares (IRLS) procedure [44, 45]. In each iteration i in full-data GLM, the estimates are computed as follows:

$$\hat{\boldsymbol{\beta}}^{(i+1)} = (\mathbf{X}^T \mathbf{W}^{(i)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(i)} \mathbf{z}^{(i)} \quad (5.5)$$

Here, \mathbf{W} is a diagonal weights matrix and \mathbf{z} is a transformation of the target variable called the *working response*, computed as

$$\mathbf{z}^{(i)} = \boldsymbol{\eta}^{(i)} + (\mathbf{y} - \boldsymbol{\mu}^{(i)}) \left(\frac{d\boldsymbol{\mu}^{(i)}}{d\boldsymbol{\eta}^{(i)}} \right) \quad (5.6)$$

where $\boldsymbol{\eta}^{(i)} = \mathbf{X} \hat{\boldsymbol{\beta}}^{(i)}$ and $\boldsymbol{\mu}^{(i)}$ is a function of $\boldsymbol{\eta}^{(i)}$ as predefined in the link function [e.g., logit link for logistic regression; 9]. From this working response, a *working residual* needs to be obtained which acts like $\hat{\boldsymbol{\epsilon}}_{-p}$ in Equation 5.4: a response vector orthogonal to the predictors excluding feature p . We define this working residual as follows [40]:

$$\hat{\boldsymbol{\epsilon}}_{-p} = \mathbf{z} - \mathbf{X}_{-p} \hat{\boldsymbol{\beta}}_{-p} \quad (5.7)$$

Using this working residual and the usual weights matrix from GLM, the coordinate descent algorithm proceeds in a similar fashion to that of linear regression (Algorithm 1). Just as with coordinate descent for linear regression, this algorithm readily extends to a blockwise procedure, meaning it can be adapted for the private regression method as discussed in Section 5.3.3. In many practical cases (e.g., no degeneracies in data matrix), the negative log-likelihood used as the objective in GLM is strictly convex [46], hence the block coordinate descent algorithm is guaranteed to converge to a global minimum [41].

5.3.5 Computing standard errors

A key component of inference in regression models is obtaining a measure of sampling uncertainty about the obtained estimates, usually standard errors. Under the assumptions of maximum likelihood theory, the limiting distribution of the deviation of the parameter estimates is the following:

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\beta) \quad (5.8)$$

where Σ_β is the asymptotic variance-covariance matrix of $\hat{\beta}$:

$$\Sigma_\beta = \text{var}(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (5.9)$$

In linear regression, $\hat{\sigma}^2 = \langle \hat{\epsilon}, \hat{\epsilon} \rangle / (N - P)$ and the standard errors of $\hat{\beta}$ can be computed as

$$\hat{\text{se}}_{\hat{\beta}} = \sqrt{\text{diag}(\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1})} \quad (5.10)$$

Thus, to compute an estimate of the variance-covariance matrix of the sampling distribution of the $\hat{\beta}$ parameters, the inverse covariance matrix of the features is needed. However, when the data is vertically partitioned, part of this covariance matrix is missing for each party. As a result, computing standard errors using the above information matrix approach is impossible for vertically partitioned data without sharing the features.

We present a novel approach to compute standard errors of the regression coefficient through creating a substitute \mathbf{V}_b of the partner's data matrix \mathbf{X}_b . This substitute is then used as the partner's data in the computation of the asymptotic variance-covariance matrix as in Equation 5.9.

The substitute \mathbf{V}_b needs to contain the same information for the parameters of Alice as the real data. This information is in the predictions received from Bob – the parameter estimates of Alice depend only on Bob's linear predictions. Consider the inputs and outputs of Bob, as seen by Alice: as the coordinate descent algorithm progresses along the R iterations, Alice can create two $N \times R$ matrices, $\hat{\mathbf{E}}_a$ and $\hat{\mathbf{Y}}_b$

$$\begin{aligned} \hat{\mathbf{E}}_a &= [\hat{\epsilon}_a^{(1)}, \dots, \hat{\epsilon}_a^{(R)}] \\ \hat{\mathbf{Y}}_b &= [\hat{\mathbf{y}}_b^{(1)}, \dots, \hat{\mathbf{y}}_b^{(R)}] \end{aligned} \quad (5.11)$$

These are the input and output matrices, respectively, from the projection that Bob applies in each iteration. This projection is commonly known as the *hat matrix* $\mathbf{H}_b \in \mathbb{R}^{N \times N}$. The hat matrix relates to Bob's data matrix \mathbf{X}_b as follows:

$$\begin{aligned} \hat{\mathbf{Y}}_b &= \mathbf{H}_b \hat{\mathbf{E}}_a \\ \hat{\mathbf{Y}}_b &= \mathbf{X}_b (\mathbf{X}_b^T \mathbf{X}_b)^{-1} \mathbf{X}_b^T \hat{\mathbf{E}}_a \\ \hat{\mathbf{Y}}_b &= \mathbf{X}_b \mathbf{X}_b^+ \hat{\mathbf{E}}_a \end{aligned} \quad (5.12)$$

where \mathbf{X}_b^+ indicates the Moore-Penrose generalized inverse of \mathbf{X}_b [47].

Alice can compute the projection that Bob applies in each iteration \mathbf{H}_b , as follows:

$$\hat{\mathbf{H}}_b = \hat{\mathbf{Y}}_b \hat{\mathbf{E}}_a^+ \quad (5.13)$$

Across iterations, this minimum-norm solution $\hat{\mathbf{H}}_b$ performs the same projection as the true hat matrix of Bob. Using this projection, Alice can then create the data substitute $\mathbf{V}_b \in \mathbb{R}^{N \times P_b}$. For this, \mathbf{V}_b should have the property $\hat{\mathbf{H}}_b = \mathbf{V}_b \mathbf{V}_b^+$. Such a \mathbf{V}_b has the same effect on the coefficient estimates of Alice that \mathbf{X}_b does, because it generates the same predictions that Bob does:

$$\begin{aligned} \hat{\mathbf{Y}}_b &= \hat{\mathbf{H}}_b \hat{\mathbf{E}}_a \\ \hat{\mathbf{Y}}_b &= \mathbf{V}_b \mathbf{V}_b^+ \hat{\mathbf{E}}_a \end{aligned} \quad (5.14)$$

There is no unique solution to decomposing $\hat{\mathbf{H}}_b$ into an $N \times P$ matrix \mathbf{V}_b and its pseudoinverse. However, a numerically convenient \mathbf{V}_b solution can be found as the first P_b eigenvectors of $\hat{\mathbf{H}}_b$. This is a convenient choice, because the columns of \mathbf{V}_b are then orthogonal, meaning they also have the following property: $\mathbf{V}_b^+ = (\mathbf{V}_b^T \mathbf{V}_b)^{-1} \mathbf{V}_b^T = \mathbf{I}^{-1} \mathbf{V}_b^T = \mathbf{V}_b^T$. As follows from Equations 5.12 and 5.14, the \mathbf{V}_b matrix relates to \mathbf{X}_b by means of an unknown P_b -dimensional positive definite rotation matrix $\mathbf{V}_b = \mathbf{X}_b \mathbf{R}$ [48].

By leveraging this similarity of \mathbf{V}_b to \mathbf{X}_b , Alice can create an augmented data matrix of the following form: $\mathbf{Z}_a = [\mathbf{X}_a, \mathbf{V}_b]$. The augmented data matrix replaces the full data matrix in the computation of the asymptotic covariance matrix: $\Sigma_\beta^{(a)} = \sigma^2 (\mathbf{Z}_a^T \mathbf{Z}_a)^{-1}$. The partition of $\Sigma_\beta^{(a)}$ belonging to β_a is then identical to its counterpart from the full data asymptotic covariance matrix Σ_β (for proof see Appendix 5.6). The square root of its diagonal elements are thus the correct standard errors that would be obtained had the full data been available.

Alternative standard error procedures are available, e.g., profile likelihood methods or bootstrapping, but those require additional iterations of the main block coordinate descent algorithm. This yields additional information leakage and dramatically increases time requirements. Conversely, in the novel procedure we suggest here, both parties efficiently leverage the information in the existing iterations to compute standard errors without additional communication.

5.4 Privacy considerations for block coordinate descent

In this section, we analyze the information transfer within our protocol for GLM parameter estimation based on block coordinate descent. In line with previous work on this topic [6, 8, 49, 14, 50, 51], we take the viewpoint of semi-honest parties: *Alice* and *Bob* follow the protocol accurately, though they may be curious and aim to recover the other party’s data. Here, we identify how well *Bob* can approximate *Alice*’s data using a *model inversion attack* [52, 53].

Information about features cannot only leak through full dataset sharing, but also via sharing statistics based on this data. For example, a simple method for regression without explicitly sharing the full dataset is that by [7], who compute the covariance matrix of \mathbf{X} using secure inner-product methods and share it between *Alice* and *Bob*. This covariance matrix allows even a semi-honest *Alice* to (a) know how many features are used by *Bob* and – in the case of categorical predictors – know how many categories there are, (b) predict the values of the features held by *Bob* based on the values of the features held by *Alice*, (c) compute standard errors around this prediction, and (d) compute an R^2 value for this prediction. In other words, in a shared covariance matrix setting *Alice* can know up to a certain degree the values on each of *Bob*’s features for each row in the dataset, and *Alice* can know how good this prediction is. Moreover, each additional feature entered by *Alice* improves the prediction of features at *Bob* by definition.

Thus, sharing the full covariance matrix is undesirable for privacy-preserving regression. Newer methods [5, 8] result in additive shares of $\text{cov}(\mathbf{X})$ at *Alice* and *Bob*, without either of them possessing the full covariance matrix. Afterwards, separate secure multiparty matrix inversion protocols or linear system solvers are used to compute the regression parameters. This generally requires complex protocols involving multiple parties, where it is clear that information transfer does occur (because the full-data estimates are obtained) but its extent is not made explicit: it is unclear how the additive shares of the covariance matrix (the “statistics”) relate to the collaborator’s data – and thus it is unclear whether that data can be reconstructed.

Conversely, in our protocol the covariance matrix of the combined data is never explicitly computed. Our method uses a different “statistic”: predictions $\hat{\mathbf{y}}$ over R iterations. How this information transfer relates to *Alice*’s data is thus explicit:

$$\hat{\mathbf{y}}_a^{(r)} = \mathbf{X}_a \hat{\boldsymbol{\beta}}_a^{(r)} \tag{5.15}$$

As a result, clear conclusions can be made as to the potential for data recovery. The exposition in section 5.3.5 and equations 5.11 to 5.14 show that the predictions sent to *Alice* only contain information about a transformation \mathbf{R}_b of *Bob's* data (\mathbf{X}_b) such that the parameter estimates and standard errors of *Alice* are adequately adjusted (Theorems 1 and 2). Reconstruction of X_b by *Alice* would require finding the correct \mathbf{R}_b , which could be any invertible $P_b \times P_b$ matrix (an infinite number of possibilities). Therefore, we argue that the basic protocol is secure against reconstruction of the data in the case of semi-honest parties.

In the special case where *Alice* enters only a single feature x_a in the analysis protocol, the information contained in \hat{y}_a is sufficient for *Bob* to reproduce the values of this feature up to a multiplicative constant: $\hat{y}_a = x_a \cdot \hat{\beta}_a$. With more than one feature per party, $\hat{\beta}_a$ becomes a vector, meaning the problem of recovering the values of any feature at *Alice* is underidentified, model inversion impossible through this route.

Sharing the final parameter estimates – likely to be a goal of the analysis in the first place – does transfer extra information, which can be combined with the transferred predictions to approximate the collaborator's data. We show empirically in Appendix 2 that this approximation is limited: if *Alice* shares the final parameter estimates with *Bob*, she reveals a proportion $1/P_a$ of the variance in \mathbf{X}_a .

It is possible to further limit the information shared with the collaborator in several ways. For example, in each iteration *Alice* may add noise to the computed parameter estimates or to the predictions sent to *Bob* – a technique from the differential privacy literature [54]. Another method is to put an upper bound on the number of iterations based on the number of features in the data. This has two effects: (a) it shrinks (regularizes) the parameter estimates towards the marginal estimates and (b) it creates an upper bound on the information shared, depending on the allowed number of iterations.

In conclusion, the information transfer between parties is insufficient for model inversion in the case of semi-honest collaborators. We leave more in-depth privacy analysis to future research. In the next section, we show how our implementation of BCD with vertically partitioned data performs in comparison to full-data generalized linear modeling (GLM) in three real-world datasets.

Table 5.1: Properties of the datasets used from the UCI machine learning repository after dataset cleaning and pre-processing. Reg represents Regression, Clf represents Classification in the table.

Dataset	Features	Instances	Task	Parties
Forest fire	13	517	Reg	Weather & Fire dept.
HCC	49	165	Clf	Lab & Clinic
Diabetes	43	15 000	Clf	Clinic & Pharmacy

5.5 Experiments

Our implementation of the BCD algorithm for vertically partitioned data is provided as an R package in the supplementary materials. In this section, we compare this implementation (version 0.9.5) to standard GLM methods on three real-world datasets with multiple parties from the UCI (University of California at Irvine) Machine Learning repository [13]. The datasets were chosen because they can be naturally partitioned into two sources, and their size and targets are different (Table 5.1). The full preprocessing and analysis code for this section is available in the supplementary materials. Analyses were run on two separate computers (an Intel Core i7-8750H at 2.20 GHz and an Intel Xeon E5-2650 v4 at 2.20GHz) connected via a gigabit Ethernet connection on a university network.

5.5.1 Forest fires data

The forest fire data comes from the Montesinho natural park in Portugal [55]. It contains several weather observations by a meteorological station (e.g. wind speed, temperature, relative humidity, etc) as well as fire department risk assessments. The target is to predict the area of forest burned by a particular fire using the features from the aforementioned parties.

We performed linear regression where the target was log-transformed to normalize the residuals. Continuous features were standardized before they were entered into the analysis. The analysis took 450 BCD iterations in the privacy-preserving regression case. Including encryption and networking overhead, estimation took 14.51 seconds and computing standard errors took 0.61 seconds. The coefficients of the full-data analysis and the privacy-preserving procedure are exactly equal (Figure 5.2), the model's performances are equal (MSE = 11.7375), and the standard errors exhibit only very small differences (mean absolute bias of 0.2%). Several months show a significant positive effect on the log-area, meaning that fires in these months (e.g., August and December) burn larger areas of forest – conditional on the ratings of the fire department.

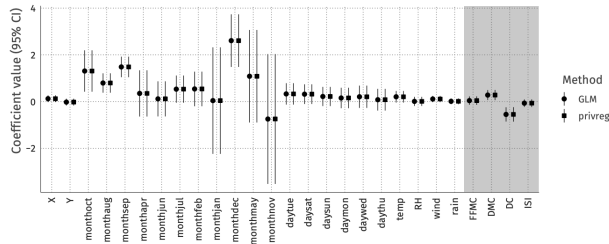


Figure 5.2: The coefficients for the forest fire analysis are exactly the same for the GLM and our privacy-preserving regression estimation methods, and the standard errors exhibit very small differences (mean absolute bias of 0.2%). The shading indicates data partitioning into the weather service (light) and fire department (dark).

5.5.2 Hepatocellular carcinoma data

This dataset was collected by Coimbra Hospital and University Centre in Portugal for studying an epithelial cell cancer of the liver called hepatocellular carcinoma (HCC) [56]. It contains heterogeneous data on demographics, risk factors, laboratory and overall survival features from HCC patients. The goal of the analysis is to use lab results for a tissue sample as well as clinical data for the patient to predict survival after diagnosis.

Since survival is a binary target, a binomial family GLM (logistic regression) was performed. For this analysis, continuous features were standardized before the analysis, which improved the convergence characteristics. The privacy-preserving GLM converged in 1636 iterations. Including encryption and networking overhead, estimation took 3 minutes and 16 seconds and computing standard errors took 0.63 seconds. The results of the analysis (Figure 5.3) show that the estimates are exactly equal across the full-data and the privacy-preserving analyses, meaning survival probability predictions for new incoming patients based on these models will be the same, hence the model’s performances are identical (AUC = 0.9590725). Despite slight deviations in the width of the confidence intervals (mean absolute bias of 5.8%), conclusions about the effects of the features on survival are also the same in this dataset.

5.5.3 Diabetes

The diabetes dataset is an extract representing 10 years (1999-2008) of clinical diabetes care at 130 hospitals and integrated delivery networks throughout

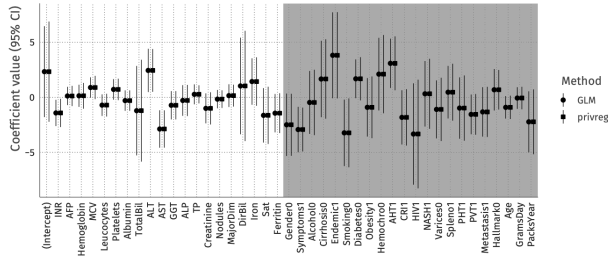


Figure 5.3: The coefficients for the carcinoma analysis are exactly the same for the GLM and our privacy-preserving regression estimation methods, and the standard errors exhibit small differences (mean absolute bias of 5.8%). The shading indicates data partitioning into the lab results (light) and clinic (dark).

the United States [57]. It is a large and also heterogeneous data set including encounter data (emergency, outpatient, and inpatient), provider speciality, demographics, laboratory data, pharmacy data, in-hospital mortality, and hospital characteristics. In this dataset, we predict readmission to the hospital using both administrative features and pharmaceutical features. To keep the computation of the standard errors for this analysis possible, 15000 patients were randomly selected from the dataset. Features were re-coded where necessary, and categorical features with only a single category in the sample were excluded from the analysis.

Since readmission is a binary target, a binomial family GLM (logistic regression) was performed. The diabetes data analysis required 284 iterations of the BCD algorithm. Including encryption and networking overhead, estimation took 1 minute and 37 seconds and computing standard errors took 42 seconds. The coefficients for the diabetes analysis are exactly the same for standard GLM and our privacy-preserving regression estimation methods (Figure 5.4), the model’s performance is identical in both cases (AUC = 0.6510909), and the standard errors exhibit very small differences (mean absolute bias of 0.5%). This analysis is particularly interesting with respect to the effect of insulin (*insulinYes*) on the readmission probability. In the analysis of only the medication data, insulin has a significant positive effect on readmission (OR = 1.20, $p < .001$), whereas conditional on the administrative data, insulin significantly reduces the readmission probability (OR = 0.88, $p < .001$). This is a strong argument for including the data of both parties in the analysis.

In this section, we have shown that our approach is a viable implementation of GLM for analyzing data with varied characteristics. The time constraints

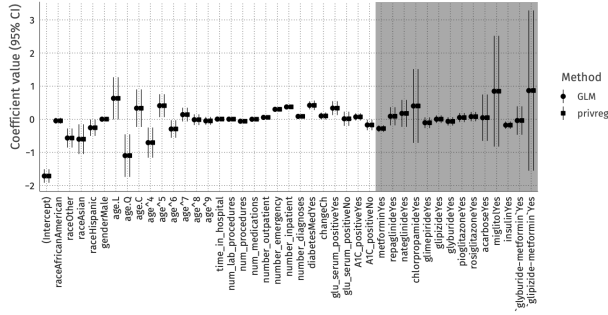


Figure 5.4: The coefficients for the diabetes analysis are exactly the same for the GLM and our privacy-preserving regression estimation methods, and the standard errors exhibit very small differences (mean absolute bias of 0.5%). The shading indicates data partitioning into the clinical data (light) and pharmaceutical data (dark).

on the real-world analyses are manageable, with all example analyses converging in under 4 minutes. We have shown that the parameter estimates exactly match those of the existing reference methods, and that our novel estimation method for the standard errors generally agrees with its full-data counterpart – and where it did not the difference was so small that it led to the same conclusions in the analysis.

To further validate our procedure, we have performed Monte-Carlo simulations (available as part of the supplementary materials) to compare our method with the full-data solutions. This shows that our implementation’s bias in the parameter estimates is almost exactly 0, and its bias in the standard errors is within 0.5% with $P = 10$, increasing slightly to a 95th percentile of around 3% when $P = 200$.

5.6 Discussion

In this paper, we have argued that block coordinate descent is a general method for estimating conditional parts of a generalized linear model (GLM) in a vertically partitioned data situation. Using this approach, two or more data parties can collaboratively estimate a GLM without sharing their features. This is useful when the features are not allowed to be shared, for example when there are privacy issues.

Our method falls within the category of federated learning algorithms. This means it can be implemented for situations when data mining is to be performed over remote devices or siloed data centers [24], where aggregating

the data tables is prohibitively expensive in terms of time, computation, or storage costs. This work aligns with several recent contributions that seek to exploit the privacy-preserving aspects of federated learning algorithms [see, e.g., 58, 59].

Due to the accessibility of our protocol and its similarity to existing regression estimation methods, extensions are relatively simple to implement. First and foremost, our framework can be extended to multiple parties as coordinate descent naturally extends to multiple blocks (see, for example, Appendix 5.6). In addition, our algorithm could include penalties for regularized estimation of the regression parameters through thresholding [40]. Through further research into combining coordinate descent with missing data methods such as full information maximum likelihood [60], our protocol could even be extended for a hybrid partitioning situation where data is both horizontally and vertically partitioned.

Our novel approach is a natural modification of the familiar linear modeling framework – without changes in the assumptions. We argue that our protocol restricts statistical information sharing as much as possible, while being explicit in how the shared information relates to the original data. Because of this, data parties know how much information they share, and the protocol could even incorporate methods from the differential privacy literature – such as additive noise or early stopping – to put a restriction on the amount of information shared with the partner institution [54].

The main tradeoff of this flexibility compared to existing methods is relatively high communication cost: each iteration requires N prediction values to be sent to the partner institution. In addition, like other methods for this situation the block coordinate descent assumes (probabilistic) linkage of the individual records – both parties need to have their records in the same order. Lastly, this method is possible only when the target can be shared, although in absence of a shareable target collaborators could still perform some form of transfer learning, e.g., by predicting a shareable feature *related* to the true target.

Considering the prospect of these extensions and the availability of an accessible open-source implementation, we believe the proposed block coordinate descent protocol can be a springboard for future developments in the privacy-preserving distributed data mining field.

References

- [1] Stephen Kaisler, Frank Armour, J Alberto Espinosa, and William Money. “Big data: Issues and challenges moving forward”. In: *2013 46th Hawaii International Conference on System Sciences*. IEEE, 2013, pp. 995–1004. DOI: 10.1109/HICSS.2013.645.
- [2] Commission on Social Determinants of Health. *Closing the gap in a generation : health equity through action on the social determinants of health : final report : executive summary*. World Health Organization, 2008, p. 33. URL: https://www.who.int/social_determinants/final_report/csdh_finalreport_2008.pdf (visited on 01/11/2022).
- [3] Suranga N Kasthurirathne, Joshua R Vest, Nir Menachemi, Paul K Halverson, and Shaun J Grannis. “Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services”. In: *Journal of the American Medical Informatics Association* 25.1 (2017), pp. 47–53. DOI: 10.1093/jamia/ocx130.
- [4] Jessica S Ancker, Min-Hyung Kim, Yiye Zhang, Yongkang Zhang, and Jyotishman Pathak. “The potential value of social determinants of health in predicting health outcomes”. In: *Journal of the American Medical Informatics Association* 25.8 (2018), pp. 1109–1110. DOI: 10.1093/jamia/ocy061.
- [5] Wenliang Du, Yunghsiung S Han, and Shigang Chen. “Privacy-preserving multivariate statistical analysis: Linear regression and classification”. In: *Proceedings of the 2004 SIAM international conference on data mining*. SIAM, 2004, pp. 222–233. DOI: 10.1137/1.9781611972740.21.
- [6] Sébastien Gambs, Balázs Kégl, and Esmá Aimeur. “Privacy-preserving boosting”. In: *Data Mining and Knowledge Discovery* 14.1 (2007), pp. 131–170. DOI: 10.1007/s10618-006-0051-9.
- [7] Alan F Karr, Xiaodong Lin, Ashish P Sanil, and Jerome P Reiter. “Privacy-preserving analysis of vertically partitioned data using secure matrix products”. In: *Journal of Official Statistics* 25.1 (2009), p. 125. URL: <http://www2.stat.duke.edu/~jerry/Papers/jos09a.pdf>.
- [8] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. “Privacy-preserving distributed linear regression on high-dimensional data”. In: *Proceedings on Privacy Enhancing Technologies* 2017.4 (2017), pp. 345–364. DOI: 10.1007/978-3-540-71701-0.
- [9] Peter McCullagh and John A Nelder. *Generalized linear models*. Routledge, 2019.

- [10] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2008. DOI: 10.1007/s00362-011-0375-4.
- [11] Yuan Wu, Xiaoqian Jiang, Jihoon Kim, and Lucila Ohno-Machado. “Grid Binary LOGistic REgression (GLORE): building shared models without sharing data”. In: *Journal of the American Medical Informatics Association* 19.5 (2012), pp. 758–764. DOI: 10.1136/amiajnl-2012-000862.
- [12] R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2018. URL: <https://www.r-project.org/>.
- [13] Catherine L Blake and Christopher J Merz. *UCI repository of machine learning databases, 1998*. 1998. URL: <https://archive.ics.uci.edu/ml/index.php>.
- [14] Jaideep Vaidya and Chris Clifton. “Privacy-preserving decision trees over vertically partitioned data”. In: *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer. 2005, pp. 139–152. DOI: 10.1145/1409620.1409624.
- [15] Rakesh Agrawal and Ramakrishnan Srikant. “Privacy-preserving data mining”. In: *ACM Sigmod Record*. Vol. 29. ACM. 2000, pp. 439–450. DOI: doi.org/10.1145/342009.335438.
- [16] Wenliang Du and Mikhail J Atallah. “Privacy-preserving cooperative scientific computations”. In: *csfw*. Citeseer. 2001, p. 0273. DOI: 10.1109/CSFW.2001.930152.
- [17] Artak Amirbekyan and Vladimir Estivill-Castro. “Privacy-preserving regression algorithms”. In: *Proceedings of the 7th WSEAS International Conference on simulation, modelling and optimization*. World Scientific, Engineering Academy, and Society (WSEAS). 2007, pp. 37–45. DOI: 10.5555/1353862.1353869.
- [18] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. “Secure Linear Regression on Vertically Partitioned Datasets.” In: *IACR Cryptology ePrint Archive* 2016 (2016), p. 892. URL: <https://eprint.iacr.org/2016/892.pdf>.
- [19] Weiwei Fang, Changsheng Zhou, and Bingru Yang. “Privacy preserving linear regression modeling of distributed databases”. In: *Optimization Letters* 7.4 (2013), pp. 807–818. DOI: 10.1109/ACCESS.2020.3000764.
- [20] Andrew Chi-Chih Yao. “How to generate and exchange secrets”. In: *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*. IEEE, 1986, pp. 162–167. DOI: 10.1109/SFCS.1986.25.

-
- [21] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. “Advances and open problems in federated learning”. In: *arXiv preprint arXiv:1912.04977* (2019).
- [22] Charu C. Aggarwal and Philip S. Yu, eds. *Privacy-Preserving Data Mining: Models and Algorithms*. Advances in Database Systems. Springer US, 2008. DOI: 10.1007/978-0-387-70992-5.
- [23] Benny Pinkas. “Cryptographic techniques for privacy-preserving data mining”. In: *ACM Sigkdd Explorations Newsletter* 4.2 (2002), pp. 12–19. DOI: 10.1145/772862.772865.
- [24] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. “Federated learning: Challenges, methods, and future directions”. In: *arXiv preprint arXiv:1908.07873* (2019).
- [25] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. “Federated learning”. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 13.3 (2019), pp. 1–207. DOI: 10.2200/S00960ED2V01Y201910AIM043.
- [26] Edgar Dobriban and Yue Sheng. “Distributed linear regression by averaging”. In: *The Annals of Statistics* 49.2 (2021), pp. 918–943. DOI: 10.48550/arXiv.1810.00412.
- [27] Ashish P. Sanil, Alan F. Karr, Xiaodong Lin, and Jerome P. Reiter. “Privacy Preserving Regression Modelling via Distributed Computation”. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, WA, USA: ACM, 2004, pp. 677–682. DOI: 10.1145/1014052.1014139.
- [28] Michael JD Powell. “An efficient method for finding the minimum of a function of several variables without calculating derivatives”. In: *The computer journal* 7.2 (1964), pp. 155–162. DOI: 10.1093/comjnl/7.2.155.
- [29] Irene Giacomelli, Somesh Jha, Marc Joye, C David Page, and Kyonghwan Yoon. “Privacy-preserving ridge regression with only linearly-homomorphic encryption”. In: *International Conference on Applied Cryptography and Network Security*. Springer. 2018, pp. 243–261. DOI: 10.1007/978-3-319-93387-0_13.
- [30] Payman Mohassel and Yupeng Zhang. “Secureml: A system for scalable privacy-preserving machine learning”. In: *2017 IEEE symposium on security and privacy (SP)*. IEEE. 2017, pp. 19–38. DOI: 10.1109/SP.2017.12.

- [31] Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and Lihua Wang. “Scalable and secure logistic regression via homomorphic encryption”. In: *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*. 2016, pp. 142–144. DOI: 10.1145/2857705.2857731.
- [32] Payman Mohassel and Peter Rindal. “ABY3: A mixed protocol framework for machine learning”. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 35–52. DOI: 10.1145/3243734.3243760.
- [33] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton: CRC Press, 2015, p. 362. DOI: 10.1201/b18401-1.
- [34] Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*. Vol. 23. Prentice hall Englewood Cliffs, NJ, 1989. DOI: 10.5555/59912.
- [35] Peter Richtárik and Martin Takáč. “Distributed coordinate descent method for learning with big data”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 2657–2681. DOI: 10.5555/2946645.3007028.
- [36] Nora Méray, Johannes B Reitsma, Anita CJ Ravelli, and Gouke J Bonsel. “Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number”. In: *Journal of clinical epidemiology* 60.9 (2007), 883–e1. DOI: 10.1016/j.jclinepi.2006.11.021.
- [37] Jialei Wang, Mladen Kolar, and Nathan Srerbo. “Distributed multi-task learning”. In: *Artificial Intelligence and Statistics*. 2016, pp. 751–760. DOI: 10.48550/arXiv.1510.00633.
- [38] Yu Zhang and Qiang Yang. “A survey on multi-task learning”. In: *arXiv preprint arXiv:1707.08114* (2017). DOI: 10.48550/arXiv.1707.08114.
- [39] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. “Federated multi-task learning”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 4424–4434. DOI: 10.5555/3294996.3295196.
- [40] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *Journal of Statistical Software* 33.1 (2010), p. 1. DOI: 10.1163/ej.9789004178922.i-328.7.
- [41] Paul Tseng. “Dual coordinate ascent methods for non-strictly convex minimization”. In: *Mathematical programming* 59.1 (1993), pp. 231–247. DOI: 10.1007/BF01581245.

-
- [42] Paul Tseng. “Convergence of a block coordinate descent method for nondifferentiable minimization”. In: *Journal of optimization theory and applications* 109.3 (2001), pp. 475–494. DOI: 10.1023/A:1017501703105.
- [43] Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Mingyi Hong. “On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6741–6764. DOI: 10.5555/3122009.3242041.
- [44] Robert WM Wedderburn. “Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method”. In: *Biometrika* 61.3 (1974), pp. 439–447. DOI: 10.2307/2334725.
- [45] Peter J Green. “Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.2 (1984), pp. 149–170. DOI: 10.1111/j.2517-6161.1984.tb01288.x.
- [46] Robert WM Wedderburn. “On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models”. In: *Biometrika* 63.1 (1976), pp. 27–32. DOI: 10.2307/2335080.
- [47] KB Petersen and MS Pedersen. “The matrix cookbook”. In: *Technical Univ. Denmark, Kongens Lyngby, Denmark, Tech. Rep 3274* (2012). URL: <https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>.
- [48] Algebraic Pavel. *Decompose projection matrix into a matrix and its pseudoinverse*. Mathematics Stack Exchange. 2019-08-12. 2019. URL: <https://math.stackexchange.com/q/3317493>.
- [49] Jaideep Vaidya and Chris Clifton. “Privacy-preserving k-means clustering over vertically partitioned data”. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. Association for Computing Machinery, 2003, pp. 206–215. DOI: 10.1145/956750.956776.
- [50] Jaideep Vaidya, Hwanjo Yu, and Xiaoqian Jiang. “Privacy-preserving SVM classification”. In: *Knowledge and Information Systems* 14.2 (2008), pp. 161–178. DOI: 10.1007/s10115-007-0073-7.
- [51] Alan F Karr. “Secure statistical analysis of distributed databases, emphasizing what we don’t know”. In: *Journal of Privacy and Confidentiality* 1.2 (2010). DOI: 10.29012/jpc.v1i2.573.
- [52] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures”. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS ’15. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677.

- [53] Yue Wang, Cheng Si, and Xintao Wu. “Regression model fitting under differential privacy and model inversion attack”. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015. URL: <http://www.csce.uark.edu/~xintaowu/publ/ijcai15.pdf>.
- [54] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. “Calibrating noise to sensitivity in private data analysis”. In: *Theory of cryptography conference*. Springer. 2006, pp. 265–284. DOI: 10.1007/11681878_14.
- [55] Paulo Cortez and Aníbal de Jesus Raimundo Morais. “A data mining approach to predict forest fires using meteorological data”. In: *Proceedings of 13th Portuguese Conference on Artificial Intelligence*. 2007, pp. 512–523. URL: <http://www3.dsi.uminho.pt/pcortez/fires.pdf>.
- [56] Miriam Seoane Santos, Pedro Henriques Abreu, Pedro J García-Laencina, Adélia Simão, and Armando Carvalho. “A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients”. In: *Journal of biomedical informatics* 58 (2015), pp. 49–59. DOI: 10.1016/j.jbi.2015.09.012.
- [57] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. “Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records”. In: *BioMed research international* 2014 (2014). DOI: 10.1155/2014/781670.
- [58] Kallista A. Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. *Practical Secure Aggregation for Federated Learning on User-Held Data*. 2016. URL: <http://arxiv.org/abs/1611.04482>.
- [59] Robin C. Geyer, Tassilo Klein, and Moin Nabi. *Differentially Private Federated Learning: A Client Level Perspective*. 2017. URL: <http://arxiv.org/abs/1712.07557>.
- [60] Craig K Enders. “The performance of the full information maximum likelihood estimator in multiple regression models with missing data”. In: *Educational and Psychological Measurement* 61.5 (2001), pp. 713–740. DOI: 10.1177/0013164401615001.

Supplementary

Proof for recovery of parameter estimates and standard errors

Let $X \in \mathbb{R}^{N \times P}$ be a full-rank data matrix with P_a columns X_a held by *Alice* and P_b columns X_b held by *Bob* such that $P_a + P_b = P$. Let $A = X^T X$, a symmetric positive definite matrix partitioned into four blocks $A_{11} \in \mathbb{R}^{P_a \times P_a}$ (held by *Alice*), $A_{22} \in \mathbb{R}^{P_b \times P_b}$ (held by *Bob*), and $A_{12} \in \mathbb{R}^{P_a \times P_b}$ and $A_{21} \in \mathbb{R}^{P_b \times P_a}$ (unknown to either). Let $B = A^{-1}$ be partitioned in the same way into $B_{11} \in \mathbb{R}^{P_a \times P_a}$, $B_{22} \in \mathbb{R}^{P_b \times P_b}$, $B_{12} \in \mathbb{R}^{P_a \times P_b}$, and $B_{21} \in \mathbb{R}^{P_b \times P_a}$.

Following the procedure outlined in Section 5.3.5, *Alice* replaces X_B with $V_b = X_b R_b$, where R_b is an invertible matrix unknown to *Alice*. This gives a new data matrix, $Z_a = [X_a, V_b]$, and a new cross-product matrix $A^{(a)} = Z_a^T Z_a$, and its inverse, $B^{(a)} = (Z_a^T Z_a)^{-1}$.

Alice needs to compute B_{11} to obtain the asymptotic covariance matrix (ACOV) of her regression parameters $\hat{\beta}_a$.

Theorem 1. B_{11} can be obtained by *Alice* by replacing X_b with a transformed version $V_b = X_b R_b$. Specifically, $B_{11}^{(a)} = B_{11}$.

Proof. The inverse of the partitioned, positive definite symmetric matrix A is

$$A^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} (A_{11} - A_{12} A_{22}^{-1} A_{12}^T)^{-1} & -A_{11}^{-1} A_{12} (A_{22} - A_{12}^T A_{11}^{-1} A_{12})^{-1} \\ -A_{22}^{-1} A_{12}^T (A_{11} - A_{12} A_{22}^{-1} A_{12}^T)^{-1} & (A_{22} - A_{12}^T A_{11}^{-1} A_{12})^{-1} \end{pmatrix} \quad (.16)$$

Following $A^{(a)} = Z_a^T Z_a$, note that

$$\begin{aligned} A_{12}^{(a)} &= X_a^T X_b R_b \\ A_{22}^{(a)} &= R_b^T X_b^T X_b R_b \end{aligned} \quad (.17)$$

So that

$$\begin{aligned}
 B_{11}^{(a)} &= \left(A_{11}^{(a)} - A_{12}^{(a)} (A_{22}^{(a)})^{-1} (A_{12}^{(a)})^T \right)^{-1} \\
 &= \left((X_a^T X_a) - (X_a^T X_b R_b) (R_b^T X_b^T X_b R_b)^{-1} (X_a^T X_b R_b)^T \right)^{-1} \\
 &= \left((X_a^T X_a) - X_a^T X_b R_b R_b^{-1} (X_b^T X_b)^{-1} R_b^{-T} R_b^T X_b^T X_a \right)^{-1} \\
 &= \left(A_{11} - A_{12} A_{22}^{-1} A_{12}^T \right)^{-1} \\
 &= B_{11}
 \end{aligned} \tag{.18}$$

□

This shows that even if R_b is unknown to Alice, the part of the ACOV to do with $\hat{\beta}_a$ can be estimated correctly, and therefore the standard errors are available: $\text{ACOV}(\hat{\beta}_a) = \sigma^2 B_{11}$.

Theorem 2. *The parameter estimates created by Alice $\hat{\beta}_a^{(a)}$ are equal to their full-data counterparts $\hat{\beta}_a$ when replacing X_b with a transformed version $V_b = X_b R_b$.*

Proof. Following the same setup of Theorem 1, note that

$$\begin{aligned}
 B_{21}^{(a)} &= -(A_{22}^{(a)})^{-1} (A_{12}^{(a)})^T B_{11}^{(a)} \\
 &= -(R_b^T X_b^T X_b R_b)^{-1} (X_a^T X_b R_b)^T B_{11} \\
 &= -R_b^{-1} (X_b^T X_b)^{-1} R_b^{-T} R_b^T X_b^T X_a B_{11} \\
 &= -R_b^{-1} A_{22}^{-1} A_{12}^T B_{11} \\
 &= R_b^{-1} B_{21}
 \end{aligned} \tag{.19}$$

Note that, for Alice:

$$\begin{pmatrix} \hat{\beta}_a^{(a)} \\ \hat{\beta}_b^{(a)} \end{pmatrix} = \begin{pmatrix} B_{11}^{(a)} & B_{12}^{(a)} \\ B_{21}^{(a)} & B_{22}^{(a)} \end{pmatrix} \begin{pmatrix} X_a^T y \\ (X_b R_b)^T y \end{pmatrix} \tag{.20}$$

Following this, for the parameter estimates belonging to the variables held by

Alice:

$$\begin{aligned}
\hat{\beta}_a^{(a)} &= B_{11}^{(a)} X_a^T y + B_{21}^{(a)T} (X_b R_b)^T y \\
&= B_{11} X_a^T y + (R_b^{-1} B_{21})^T (X_b R_b)^T y \\
&= B_{11} X_a^T y + B_{21}^T R_b^{-T} R_b^T X_b^T y \\
&= B_{11} X_a^T y + B_{21}^T X_b^T y \\
&= \hat{\beta}_a
\end{aligned} \tag{.21}$$

□

This shows that even if R_b is unknown to Alice, the parameters β_a can be estimated correctly.

Theorem 3. *If R_b is unknown to Alice, then Bob's regression parameters $\hat{\beta}_b$ are unavailable to her when replacing X_b with a transformed version $V_b = X_b R_b$.*

Proof. Following the same setup of Theorem 1, note that

$$\begin{aligned}
B_{22}^{(a)} &= R_b^T X_b^T X_b R_b - R_b^T X_b^T X_a (X_a^T X_a)^{-1} X_a^T X_b R_b \\
&= R_b^T B_{22} R_b
\end{aligned} \tag{.22}$$

Then, following the partitioning of Theorem 2, but focusing on the estimates for Bob's data:

$$\begin{aligned}
\hat{\beta}_b^{(a)} &= B_{21}^{(a)} X_a^T y + B_{22}^{(a)} (X_b R_b)^T y \\
&= R_b^{-1} B_{21} X_a^T y + R_b^T B_{22} R_b R_b^T X_b^T y \\
&= R_b^T B_{21} X_a^T y + R_b^T B_{22} X_b^T y \\
&= R_b^T \hat{\beta}_b
\end{aligned} \tag{.23}$$

□

Theorem 4. *If R_b is unknown to Alice, then the ACOV of the full parameter vector $\hat{\beta} = (\hat{\beta}_a, \hat{\beta}_b)$ is unavailable to her when replacing X_b with a transformed version $V_b = X_b R_b$.*

Proof. Remember that $ACOV(\hat{\beta}) = \sigma^2 B$. Following Theorems 1 to 3, note that

$$(Z_a^T Z_a)^{-1} = \begin{pmatrix} B_{11}^{(a)} & B_{12}^{(a)} \\ B_{21}^{(a)} & B_{22}^{(a)} \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12}R_b \\ R_b^T B_{21} & R_b^T B_{22}R_b \end{pmatrix} \quad (.24)$$

□

By symmetry, the proofs for Theorems 1, 2, 3, and 4 can be given for Bob in the same way.

MSE of rank-R data approximation

The goal of this appendix is to show empirically the amount of explained variance when a set of parameters and their associated predictions are shared with another party. From Equation 5.15, but assuming all in-between parameter estimates $\hat{\beta}_a^{(r)}$ are shared, Bob can create the following approximation:

$$\begin{aligned} \hat{Y}_a &= X_a \hat{B}_a \\ \hat{X}_a &= \hat{Y}_a \hat{B}_a^+ \end{aligned} \quad (.25)$$

where $\hat{Y}_a \in \mathbb{R}^{N \times R}$, $B_a \in \mathbb{R}^{P \times R}$, $X_a \in \mathbb{R}^{N \times P}$, all matrices are full rank, and A^+ indicates the Moore-Penrose pseudoinverse of A . For simplicity, but without loss of generality, we assume here that the variance of all the features in X_a is the same, σ_a^2 , and these features are uncorrelated.

The relation between P , R , and the accuracy of the approximation \hat{X}_a is as follows: as $R \rightarrow P$, the MSE improves linearly, with perfect approximation being achieved when $R = P$. As mentioned in-text, when $P = 1$, sharing one set of parameters ($R = 1$) means the data can be recovered completely. Empirical simulations show that the relation between R , P , and expected mean square error of approximation is $\text{MSE} = \sigma_a^2(1 - R/P)$, where σ_a^2 is the variance of the features in X_a (see Figure .5).

Phrasing the above in terms of information sharing and privacy preservation: in sharing R sets of parameter estimates $\hat{\beta}_a^{(r)}$ with their associated predictions $\hat{y}_a^{(r)}$, Alice reveals a proportion of at least R/P of variance in the data. This proportion is a lower bound: in case there are correlations among the features of Alice, this proportion increases. When $R = P$ the data of Alice can be

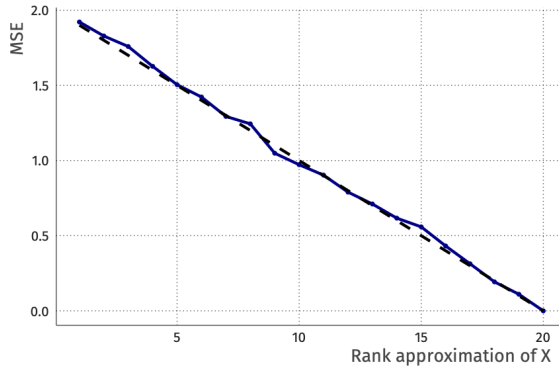


Figure .5: Mean square error (MSE) of the approximation of the data \mathbf{X}_a at Alice by Bob if $\hat{\mathbf{B}}_a$ is known. \mathbf{X}_a was simulated as having $P = 20$ uncorrelated features with variance $\sigma_a^2 = 2$. Note that the approximation linearly improves as the rank of $\hat{\mathbf{B}}_a$ increases, with a perfect approximation reached when $R = P$. Dashed line indicates expected MSE, using the formula $E[\text{MSE}] = \sigma_a^2(1 - R/P)$.

reconstructed by Bob. When either of a pair $(\hat{\beta}_a^{(r)}, \hat{\gamma}_a^{(r)})$ are shared but not the other, no information is revealed.

Extension of distributed algorithm to multiple parties

The distributed estimation algorithm can be readily extended to multiple parties. This appendix describes an extension for linear regression. It can subsequently be further extended to generalized linear models in the same way as the two-party algorithm (Section 5.3.4). One possible multiparty extension is to update and transfer a running linear prediction $\hat{\mathbf{y}}$ in a circular way. For three parties (Alice, Bob, and Carol) the algorithm is as follows:

Algorithm 3: Secure block coordinate descent with three parties

1. Initialize $\hat{\mathbf{y}} \leftarrow \mathbf{0}$
2. Initialize $\hat{\mathbf{y}}_a, \hat{\mathbf{y}}_b, \hat{\mathbf{y}}_c \leftarrow \mathbf{0}$
3. *Alice:*
 - a) $\hat{\mathbf{y}}_{-a} \leftarrow \hat{\mathbf{y}} - \hat{\mathbf{y}}_a$
 - b) $\hat{\mathbf{e}}_{-a} \leftarrow \mathbf{y} - \hat{\mathbf{y}}_{-a}$
 - c) $\hat{\beta}_a \leftarrow (\mathbf{X}_a^T \mathbf{X}_a)^{-1} \mathbf{X}_a^T \hat{\mathbf{e}}_{-a}$
 - d) $\hat{\mathbf{y}}_a \leftarrow \mathbf{X}_a \hat{\beta}_a$
 - e) $\hat{\mathbf{y}} \leftarrow \hat{\mathbf{y}}_{-a} + \hat{\mathbf{y}}_a$
 - f) Send $\hat{\mathbf{y}}$ to *Bob*
4. *Bob:*
 - a) $\hat{\mathbf{y}}_{-b} \leftarrow \hat{\mathbf{y}} - \hat{\mathbf{y}}_b$
 - b) $\hat{\mathbf{e}}_{-b} \leftarrow \mathbf{y} - \hat{\mathbf{y}}_{-b}$
 - c) $\hat{\beta}_b \leftarrow (\mathbf{X}_b^T \mathbf{X}_b)^{-1} \mathbf{X}_b^T \hat{\mathbf{e}}_{-b}$
 - d) $\hat{\mathbf{y}}_b \leftarrow \mathbf{X}_b \hat{\beta}_b$
 - e) $\hat{\mathbf{y}} \leftarrow \hat{\mathbf{y}}_{-b} + \hat{\mathbf{y}}_b$
 - f) Send $\hat{\mathbf{y}}$ to *Carol*
5. *Carol:*
 - a) $\hat{\mathbf{y}}_{-c} \leftarrow \hat{\mathbf{y}} - \hat{\mathbf{y}}_c$
 - b) $\hat{\mathbf{e}}_{-c} \leftarrow \mathbf{y} - \hat{\mathbf{y}}_{-c}$
 - c) $\hat{\beta}_c \leftarrow (\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{X}_c^T \hat{\mathbf{e}}_{-c}$
 - d) $\hat{\mathbf{y}}_c \leftarrow \mathbf{X}_c \hat{\beta}_c$
 - e) $\hat{\mathbf{y}} \leftarrow \hat{\mathbf{y}}_{-c} + \hat{\mathbf{y}}_c$
 - f) Send $\hat{\mathbf{y}}$ to *Alice*
6. Repeat step (3.), (4.), and (5.) for R iterations until convergence.

6

Generating Synthetic Tabular Data using Conditional GANs combining with Differential Privacy

Chang Sun, Johan van Soest, and Michel Dumontier. "Improving Correlation Capture in Generating Imbalanced Data using Differentially Private Conditional GANs". Submitted to *Information Sciences*. (2022) DOI (pre-print): 10.48550/arXiv.2206.13787

Abstract

A large amount of personal data that is highly valuable for the scientific community is still not accessible or requires a lengthy request process because of privacy concerns and legal restrictions. As a solution, synthetic data has been studied and proposed to be a promising alternative to this issue. But, generating realistic and privacy-preserving synthetic data still retains challenges. In this paper, we propose DP-CGANS, a conditional GAN model combining differential privacy to generate synthetic tabular data. DP-CGANS framework consists of four components including transformation, sampling, conditioning, and networking training with differential privacy to generate realistic and privacy-preserving synthetic data. DP-CGANS distinguishes categorical and continuous variables and maps them into latent space using different transformations. We structure a conditional vector as an additional input to not only presents the minority class in the imbalanced data, but also capture the dependency between variables. Moreover, we handle the model collapse to ensure the generator creates diverse and representative synthetic data points. To prevent the synthetic data from leaking any sensitive information from the source data, we apply differential privacy by injecting designed statistical noise to the gradients in the training process of DP-CGANS to provide a differential privacy guarantee. We extensively evaluate our model with state-of-the-art generative models on three public datasets and two real-world personal health datasets in terms of the statistical similarity, machine learning performance, and privacy measurement. We demonstrate that our model outperforms other comparable models, especially in capturing dependency between variables. Finally, we present the balance between data utility and privacy in synthetic data generation considering the different data structure and characteristics of real-world datasets such as imbalance variables, abnormal distributions, and sparsity of data.

6.1 Introduction

Data from individuals such as personal health or behavior data have proven to be highly valuable for health research such as enhancing our understanding of disease and delivering high-quality patient-centered care [1, 2]. This data is sensitive and requires special attention and protection. Due to disclosure limitations and legal requirements, such data is not always accessible for the scientific community [3, 4, 5]. Even if some data is accessible by request, researchers need to invest enormous time and effort in the requesting process. To comply with the legal and organizational regulations, researchers are required to prepare lengthy documentation describing the research questions, specify the data subjects and variables, illustrate a detailed analysis plan to prove the necessity of the request with strong evidence [6, 7, 8]. Access to this data may take months or years without knowing if the data is sufficiently useful for the research studies. This can cause a severe delay and inordinate costs for research projects [9, 10]. Early access to samples of data is useful in the exploratory research phase to determine the usability of the data for answering specific research questions.

Personal data may be distributed and held by multiple parties. Sharing and analyzing this data among multiple parties holds the potential for new insights and a wide variety of applications [11, 12]. State-of-the-art technologies to analyze distributed data such as Federated Learning (e.g., Personal Health Train [13], Privacy-Preserving Federated Neural Network Learning [14]) or Swarm Learning [15]) have two core concepts - 1) keep original data with the data owner, 2) construct machine learning models at each data party. However, in practice, these infrastructures remain challenges such as low data interoperability, inconsistent data standards, and uneven data quality from different data parties. For example, since the source data is unrevealed, the coverage or relevance of the data from each party, the completeness and systematic errors of each variable are unknown. These challenges hinder researchers from constructing accurate and reliable machine learning models using these infrastructures [16, 17, 18].

One approach to eliminate direct learning from personal data is to use synthetic data. In this study, synthetic data is defined as the generated data which is structurally and statistically similar to real data at the population level (i.e., distributions of single variables, correlations between variables), and machine learning utility level (i.e., the analysis results on synthetic data are comparable to the results on real data). An example is that data parties provide researchers with realistic synthetic data to construct machine learning models. Afterwards, the built models are sent to data parties to be executed on the source data and only return the results to the researchers. The realistic synthetic data offers the possibility for researchers to i) assess whether

the data are relevant for their studies and ii) obtain statistically valid insights without access to the underlying data or before starting the data requesting process. To protect the privacy of the source data, the synthetic data should offer strong privacy guarantees to prevent adversaries from extracting any sensitive information about the source data [6, 7].

In this paper, we propose a DP-CGANS framework (Differentially Private - Conditional Generative Adversarial Networks), consisting of four components including transformation, sampling, conditioning, and networking training with differential privacy, to generate realistic and privacy-preserving synthetic data. DP-CGANS constructs conditional vectors and an extra penalty to enforce the generator to capture the under-represented classes in the imbalanced variables and simulate the correlations and dependencies between these imbalanced variables. To motivate the model to generate diverse and representative synthetic data, we apply Wasserstein distances with gradient penalty and then group the training samples to the discriminator. Finally, we provide a privacy guarantee through a differential privacy approach that injects Gaussian noise to the penalty gradients in the training process. Under a certain differential privacy threshold, DP-CGANS prevents the synthetic data from leaking sensitive information originating in the source data. We conduct experiments on three public datasets and two real-life personal health data comparing with the other three state-of-the-art generative models. The performance is evaluated on statistical similarity, machine learning performance, and privacy risks in attribute and identity disclosure under varying differential privacy budgets. Results indicate that DP-CGANS outperforms other comparable models for most datasets and captures the most dependencies between imbalanced variables. We observe the offer a trade-off between data utility and privacy in synthetic data generation.

We summarize the following key contributions of this study:

- We present a comprehensive summary of existing challenges and previous work on generating synthetic data using a variety of GAN frameworks;
- We propose a GAN-based framework to generate realistic and privacy-preserving synthetic tabular data consisting of four components including transformation, sampling, conditioning, and networking training with differential privacy;
- We add a conditional vector and an extra penalty to the generator and apply Wasserstein GAN with gradient penalty to the discriminator to address imbalanced variables issues and to capture correlations and dependencies between variables;

-
- We deploy differential privacy techniques in the training process of the discriminator of DP-CGANS by carefully adding Gaussian noise to the penalty gradients;
 - We evaluate the performance of DP-CGANS on three baseline datasets and two real-life datasets by measuring the statistical similarity, machine learning performance, and privacy.

This paper is structured as follows: Section 6.2 describes background knowledge on conditional GANs and differential privacy, followed by the state-of-art methods. Section 6.3 elaborates on our proposed methods and supporting theories. Section 6.4 presents a set of experiments and results on five datasets, followed by findings and discussion in Section 6.5. Finally, we conclude the study in section 6.6.

6.2 Background

Generative Adversarial Network (GAN) contains two neural networks - a generator and a discriminator competing with each other. The generator aims to create realistic synthetic data points that cannot be indistinguishable by the discriminator, while the discriminator is trained to accurately classify real and synthetic data created by the generator. A number of GAN frameworks have been successfully developed to generate synthetic image, text, music, health and financial data with promising performance [19, 20, 21, 22, 23]. However, using GANs to generate tabular data poses exclusive challenges such as modeling data with mixed types (categorical and continuous), preventing model collapse, handling imbalanced variables, and capture the dependencies among variables [24, 25, 26]. Several variants of GANs have been proposed to overcome these challenges [6]. MedGAN [27] transforms the binary and discrete variables to a continuous space by combining an auto-encoder with a GAN. MedGAN is one of the earliest GAN variants to generate synthetic Electronic Health Records. It handles binary and continuous variables in separate models but not multi-categorical variables. TableGAN [28] adds a third neural network as a classifier in addition to the generator and the discriminator to increase the semantic integrity of the synthetic data. TableGAN has good performance on handling discrete and continuous variables but suffers from model collapse with categorical data.

6.2.1 Handling Mode Collapse

Model collapse occurs when the generator discovers some data points that are classified as real data by the discriminator with high confidence and then

replicates them to all the data points. In this case, the discriminator fails to provide useful gradients to the generator anymore. To address this challenge, Martin et al. [29] proposed a WGAN which used Wasserstein distance to measure the minimal cost of transforming random data points from an arbitrary distribution into the other target distribution. Further, Ishaan et al. [30] introduced a gradient penalty to penalize the discriminator, called WGAN-GP, which stabilized WGAN training and better-prevented vanishing gradients. In addition to WGANs which adjust the objective functions, PacGAN [31] restructures the discriminator from mapping one data sample to a class (real or synthetic) to mapping a set of independent samples to a class. The packed discriminator can effectively detect mode collapse when there is a lack of diversity in a set of data samples. The objective function of WGAN-GP is constructed as equation 6.1. The coefficient λ is defined as the weight of gradient penalty term in the training. $P_{\hat{x}}$ is the distribution uniformly sampled between the real (P_r) and generator model distribution (P_g).

$$L = \underbrace{\mathbb{E}_{\hat{x} \sim P_g} [D(\hat{x})] - \mathbb{E}_{x \sim P_r} [D(x)]}_{WGANLoss} + \underbrace{\lambda \mathbb{E}_{\hat{x} \sim P_{\hat{x}}} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]}_{GradientPenalty} \quad (6.1)$$

6.2.2 Handling Imbalanced Data

When generating imbalanced data, the major category is likely to dominate the training of the discriminator so that the discriminator fails to detect the absence of the minor category. Conditional GAN (CGAN) append an additional vector to the input of the generator and discriminator to address this concern. Engelmann and Lessmann [32] proposed CW-GAN, a CGAN using the WGAN-GP objective function, as an oversampling method for imbalanced datasets with both continuous and categorical variables. With a similar goal, CTGAN [24] invents a training-by-sampling method to handle imbalanced categorical variables in addition to the conditional vector. Based on CTGAN and TableGAN, CTAB-GAN [26] combines two frameworks to solve the challenges in industrial datasets such as variables with mixed data types and long-tail distributions.

6.2.3 Handling Privacy Concerns

GANs could elicit privacy concerns when the training data is personal and/or sensitive [33, 34]. To protect the source data from malicious privacy attacks, recent work shows the promising application of combining Differential Privacy (DP) into GANs [35, 36, 37]. DP uses a solid mathematical formulation

to measure the privacy and provide theoretical privacy guarantees by typically adding noise when training the models [38, 39]. A model is considered to be (ϵ, δ) - differentially private if for any two datasets D and D' differing in a single data point and for any subset of outputs S :

$$\mathbb{P}(M_p(D) \in S) \leq e^\epsilon \cdot \mathbb{P}(M_p(D') \in S) + \delta \quad (6.2)$$

where $M_p(D)$ and $M_p(D')$ are the outputs of the model for input datasets D and D' , \mathbb{P} is the randomness of the noise, ϵ reflects the privacy level. A small ϵ (≤ 1.0) indicates the small difference of model's output probabilities on D and D' which results in a high privacy guarantee. Differential privacy can protect the participation of individual data points in the datasets, which means replacing or removing one data point (data instance) with another one will not make an observable change in the analysis results. Xie et al. [37] added noise on the gradient of Wasserstein distance during the discriminator training, while Chen et al. [40] uses WGAN-GP framework and inject noises in the generator training. However, both models were designed and experimented on image data.

6.2.4 Remaining challenges

Based on these recent studies working on GANs for tabular data, there remains challenges in generating more realistic synthetic data from imbalanced categorical variables [24, 26]. One challenge is that the correlations and dependencies among imbalanced variables are typically not well-preserved in the generated synthetic data. It is crucial to transfer such information from real data to synthetic data in many domains such as healthcare and social sciences. For instance, we would expect the preservation of a positive relationship between daily physical activity and mobility in synthetically generated health data. Another challenge in working with personal data lies in the possibility of using GANs to security attacks that accurately reveal missing characteristics of real individuals, which could compromise their privacy [41, 42, 43]. Optimizing the trade-off between the privacy of the source data and the quality of the synthetic data remains an open challenge.

6.3 Method

We propose DP-CGANS, which generates tabular synthetic data with continuous and categorical variables. DP-CGANS appends conditional vectors to both the generator and discriminator in a combination of a differential privacy technique. The development of DP-CGANS is based on the strengths

of prior studies including [44, 24, 37] and further extended to address the remaining challenges, including better handling imbalanced categorical variables and capturing the correlations and dependencies between variables. This section will elaborate the structure of the DP-CGANS framework covering the transformation of variables, the construction of the conditional vector, sampling imbalanced variables for training, network training method, and applied differential privacy in DP-CGANS to protect the input data.

6.3.1 DP-CGANS Framework

The overall framework of DP-CGANS is illustrated in Figure 6.1 including four main steps which are transformation, sampling, conditioning, and training. DP-CGANS separates input data to categorical and continuous variables to apply different transformation methods and activation functions. Categorical variables are encoded using one-hot encoding method, while continuous variables are transformed using mode-specific normalization proposed by Xu et al. [24]. In real-world datasets, continuous variables commonly have multimode distribution such as heights of males and females. Instead of forcing the values of continuous variables to $[-1, 1]$ using a min-max transformation, mode-specific normalization estimates the number of the distributions of continuous variables by a variational Gaussian mixture model with Dirichlet Process. The values of each continuous variable are normalized according to its estimated distributions. After training, the synthetic data produced by the generator is inversely transformed to the original scales.

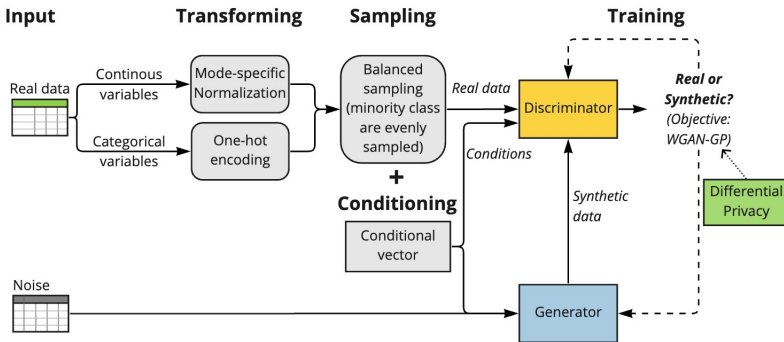


Figure 6.1: The overall structure of DP-CGANS Framework

6.3.2 Conditional vectors and data samples for training

The key process in DP-CGANS to capture the dependencies between imbalanced variables is sampling and conditioning. Our method is leveraged from

the training-by-sampling method and the design of the conditional generator in [24]. The primary idea is to encode the values of each categorical variable and present them as an additional conditional vector to train the generator and discriminator. The input training data is resampled based on this conditional vector to ensure the minority categories can also be observed in training.

However, the existing approach and its variants [26, 45] treat each variable independently, thereby potentially losing the dependencies between variables in the synthetic data. DP-CGANS addresses this issue by conditioning the generator with an extensive vector representing the dependencies between variables. Figure 6.2 shows the construction of the conditional vectors in DP-CGANS. After transformation, each value in the categorical variables is encoded into a one-hot vector. For each row in the sampled data, we randomly select and pair two categorical variables with equal probability. In the example, they are variables of Sex and Diabetes. Then, the probability mass of every possible combination of categorical values is calculated from these two variables and one pair out of all possible combinations is sampled. The sampled values are Female and No diabetes and are represented as 01 100 in the example. The occurrence of two rare categorical values is relatively low and difficult to capture in this case. The sampling is based on the logarithm of the probability which increases the chance of picking up the dependency between two rare categorical values.

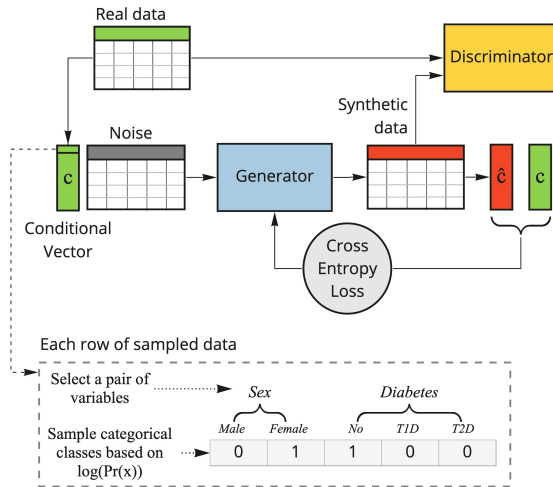


Figure 6.2: Construction of the conditional vector and generator training.

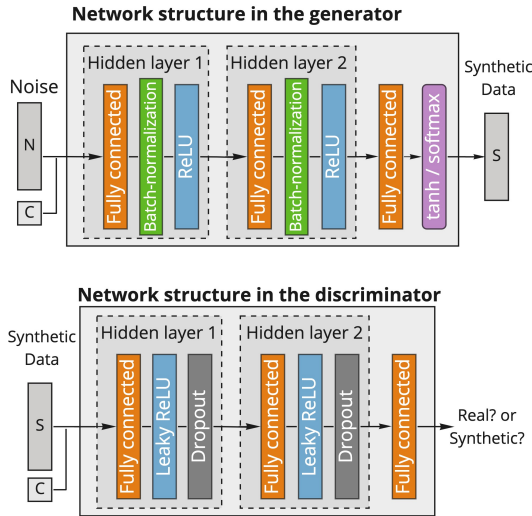


Figure 6.3: Network Structure of the generator and the discriminator.

6.3.3 Network structure and training method

Figure 6.3 illustrates the network structures of the generator and the discriminator of DP-CGANs. The **generator** uses a fully connected network with two hidden layers. Each hidden layer applies the batch-normalization and ReLU activation function for efficiency and stability purposes to address the vanishing gradient and data sparse problems. To generate the mix of categorical and continuous features, *tanh* and *softmax* activation functions are applied on the output layer [46].

The generator is trained to produce more realistic synthetic data by learning the loss based on the discriminator’s classification of the real and synthetic data. Figure 6.2 shows the generator training process in DP-CGANs. In addition to learning from the loss output from the discriminator, the generator takes an extra penalty to present the variables and values which are sampled in the conditional vector (\hat{C}) and to maximally mimic the conditional vector from the real data (C). As the conditional vector includes multiple variables to capture their dependency, we introduced a Binary Cross-Entropy Loss combined with a Sigmoid layer to penalize the generator loss in DP-CGANs.

The **discriminator** is a fully connected network with two hidden layers. The hidden layers apply Leaky ReLU functions which can handle negative input values and has better performance than ReLU in the discriminator and dropout on each layer [47, 48]. To mitigate mode collapse, the discriminator

is constructed following the PacGAN framework, which is an augmented discriminator mapping 10 samples to a single class. DP-CGANS applies the objective function in equation (3) following the WGAN-GP (Wasserstein GAN with gradient penalty) structure with gradient penalty coefficient 10. The more the discriminator is trained, the more useful gradient of the Wasserstein will be obtained. We run 5 iterations of the discriminator in each generator iteration. Lastly, the discriminator and generator both use Adam optimization with learning rate (α) 1×10^{-4} , the exponential decay rate for the first and second moment estimates (β_1, β_2) 0.5 and 0.99.

6.3.4 Differential Privacy in DP-CGANS

DP-CGANS takes real data as input to train the discriminator. To protect these data, we inject Gaussian noise to the penalty gradient of the Wasserstein distance while training the discriminator [37]. A post-processing property of differential privacy has been shown that operations after a differentially private output will not violate the privacy [49]. Therefore, privatizing the discriminator can impose the generator to become differentially private in that the generator is trained based on the differentially private discriminator's output [33, 37].

In each iteration, the discriminator calculates the gradients of loss to optimize its training objective. We clip the gradients to $[-C_p, +C_p]$ where C_p is the parameter clip constant and inject Gaussian noise ($N(0, \sigma^2(C_g)^2 I)$) to the clipped gradient where σ is the noise scale, C_g is the bound on the gradient of Wasserstein distance. To monitor the spent privacy budget (ϵ, δ), the model tracks and checks the privacy budget every time the noise is added to the gradient. Different from existing methods using moment accountant technique, we applied Rényi Differential Privacy (RDP) Accountant [39] which calculates a tighter estimation of privacy budget. At every iteration step, the privacy budget is bounded and accumulated. When the total privacy budget exceeds the initial target, the training process will be terminated and DP-CGANS is able to generate differentially private synthetic data.

6.4 Experiments and results

The experiment includes three public datasets that are commonly used by the machine learning community from UCI Machine Learning Repository [50] and two real-world personal health datasets (Table 6.1). All datasets contain multiple data types continuous, binary, and categorical. The Adult dataset [51] and Census dataset [52] contain socio-economic data from individuals, while Census has a majority of categorical variables. Intrusion

Table 6.1: Description of experimented datasets. #Cat represents the number of multi-class categorical variables, #Con represents the number of continuous variables, and #Bi represents the number of binary variables in the datasets.

Datasets	#Rows	#Cat	#Con	#Bi	Source	Access
Adult	30162	7	6	2	UCI	Public
Census	12000	31	7	3	UCI	Public
Intrusion	123000	4	20	0	UCI	Public
Diabetes	2257	8	10	6	DMS	On request
Cancer	365	5	2	2	Maastro	On request

dataset [53] is about network intrusion detections with most continuous variables. To reduce the computation time, 12k rows and 123k rows of data were randomly sampled from Census and Intrusion datasets in a stratified way with respect to the target variables. The Diabetes dataset is requested from the Maastricht Study, an observational prospective population-based cohort study focusing on Type 2 Diabetes [54]. The diabetes dataset includes demographic, socioeconomic, lifestyle, T2DM data of individuals. The cancer dataset is the clinical outcome data of non-small cell lung cancer (NSCLC) patients collected by the Maastricht Clinic [55, 56].

6.4.1 Experiment Setting

We compared the performance of DP-CGANS with other three well-known GAN frameworks for generating tabular data - CTGAN [24], MedGAN [27], and TableGAN [28]. We applied the comprehensive benchmarking suite developed by Synthetic Data Gym Framework¹ where all these models were programmed in python using PyTorch library. We keep the original structure of their framework and use the same model parameters as they stated in their published studies. All models share the same number of epochs (2000) and batch size (500). We present the key hyperparameters of DP-CGANS for reproducibility purposes in Table 6.2. The experiments were conducted using one 32GB GPU (Nvidia DGX1 8x Tesla V100) in an OKD 4.6 cluster under the Data Science Research Infrastructure (DSRI) at Maastricht University².

6.4.2 Evaluation Metrics

A set of metrics are applied to comprehensively evaluate the performance of DP-CGANS and compared with other state-of-the-art models. The metrics are grouped to test the data utility of the synthetic data and measure the privacy cost of the generative model. The data utility metrics measure the

¹SDGym Github Repository: <https://github.com/sdv-dev/SDGym>

²DSRI: <https://maastrichtu-ids.github.io/dsri-documentation/>

Table 6.2: Key hyperparameters in DP-CGANs. The discriminator step represents the number of updates (iterations) of the discriminator per generator update.

Step	Model	Hyperparameter	Value
Transformation	Gaussian Mixture	Prior type for the weights' distribution	Dirichlet Process
		Max. num of Gaussian distribution	10
		Weight concentration prior	1×10^{-3}
		Weight threshold	1×10^{-3}
		Num of mixture components	≤ 10
Network training	-	Epochs	2000
	-	Batch size	500
	Adam Optimizer	Learning rate (α)	1×10^{-4}
		First exponential decay (β_1)	0.5
		Second exponential decay (β_2)	0.99
	PacGAN	Pac	10
	WGAN-GP	Gradient penalty factor (λ)	10
		Discriminator step*	5
	Softmax	Non-negative scalar temperature (τ)	0.2
	LeakyReLU	Negative slope	0.2
Dropout	Probability of an element is 0	0.5	
Differential Privacy	-	Clip constant (C_p)	0.01
	-	Probability of information leakage(δ)	1×10^{-5}
	-	Privacy budget (ϵ).	0.1, 1, 10, 100, ∞

statistical similarity between real and synthetic data and compare the machine learning performance. The privacy cost metrics measure how much information from the real data may be disclosed by the synthetic data and the generative models. An overview of evaluation metrics applied in this study is reported in Table 6.3.

Table 6.3: An overview of evaluation metrics for synthetic data.

Metrics	Level	Method	Data Type
Statistic similarity	Single variable	Chi Square (CS)	Categorical
		Kolmogorov-Smirnov (KS)	Continuous
		KL Divergence	Categorical
	Continuous		
	Variable pairs	Pearson correlation	Continuous
		Cramer's V coefficient	Categorical
ML performance	Whole dataset	Logistic regression	-
		Decision tree	-
		Random forest (Adaboost)	-
		Multi-layer perceptron	-
Privacy cost	Identity (Rows)	Hamming distance	Categorical
		Euclidean distance	Continuous
	Attributes (Columns)	Linear regression	Categorical
		K-Nearest Neighbor	Continuous

Data Utility Evaluation Metrics

Statistical Similarity. We measured the statistical similarity by comparing the distribution of each variable independently and the correlation between variables. We include Kullback Leibler (KL) Divergence [57], Pearson's Chi-Square (CS) test [58], Kolmogorov Smirnov (KS) test [59], and pairwise correlation difference (PCD) [6]. The KL divergence calculates the marginal probability mass functions (PMF) for each variable independently of the real and synthetic data and measures the similarity of the PMFs of the two variables. It is an information-theory based and asymmetric distance measurement to observe the information change between distributions before and after inferring. We normalized the score to $[0, 1]$ by calculating $1 / (1 + \text{KL divergence})$. When the distributions of two variables are similar, the normalized scores approach 1. The final score is the average of the scores of all measured variables in the data. We apply CS and KS statistical tests on categorical and continuous variables respectively. Different from KL divergence, which measures information loss from one distribution to another, CS and KS tests are null hypothesis statistical tests. CS test checks if the frequencies of categorical values in synthetic data match the frequencies in real data. KS test measures a symmetric distance between two empirical cumulative distributions of the continuous variables.

The difference of dependencies between each pair of variables is measured by the Pearson correlation matrices for continuous variables and Cramér's V Coefficient for categorical variables [60]. Cramér's V Coefficient is based on Pearson's chi-square test to measure how strongly two categorical variables are associated. The difference score is scaled between 0 to 1. The smaller the score, the less difference between synthetic and real data.

Machine Learning Performance. The motivation of this study is to enable researchers to build their data analysis model based on synthetic data. The analysis of the synthetic data is expected to be the same or similar to the analysis of the real data. Therefore, the experimental datasets are split to training sets (75%) and test sets (25%). The training sets are fed into the generative models to produce the synthetic data. Then, a set of machine learning models (Logistic regression (LR), decision tree (DT), random forest (RF), and multilayer perceptron (MLP) models) are trained on the real training data and generated synthetic data separately. Last, the trained machine learning models are evaluated on the real test data using AUC and F1 scores. The better and more realistic synthetic data is, the smaller the difference in its machine learning performance from the real data.

Privacy Cost Evaluation Metrics

The privacy metrics cover two main classes of information disclosure that may happen in the synthetic data identity disclosure and attribute disclosure [6]. **Identity disclosure** means an attacker can exactly identify an individual (data sample) in the training data, which can be understood as if we can find one or more synthetic data with a certain distance to a real data sample which is used to generate the synthetic data [27]. Hamming distance for the categorical variables and Euclidean distance for the continuous variables are calculated on each sample from the synthetic dataset. The attacker may identify the data sample which is indeed used for training (TP), identify the sample but the sample is not used for training (FP), correctly identify the sample which is indeed not used for training (TN), wrongly identify the sample which is not used for training (FN). The final identity disclosure is measured using the precision and recall of the above scores.

Attribute disclosure can be interpreted as if an attacker can predict the original values of the synthesized variables (sensitive variables) from an individual level based on some other variables of the real data that are known to the attacker (known variables). We observe the average posterior probabilities of the attacker correctly predicting the sensitive variables on the real test data. The risk of attribute disclosure is affected by the number of known variables from the source data, the size of the synthetic data, and the attack model setting. A linear regression model is applied to the continuous variables, and a K-nearest-neighboring model is for the categorical variables. We experiment on different sets of known variables to predict other original variables on the same size of the synthetic data from different generative models.

6.5 Results and discussion

This section presents the experiment results of DP-CGANS, CTGAN, MedGAN, and TableGAN on five datasets. Each experiment was conducted 3 times and the results are the average of them. Then, we present the model performance of DP-CGANS using different privacy budgets on Adult and Diabetes datasets and describe the observed limitations of DP-CGANS.

6.5.1 Statistical Similarity

Evaluation results of statistical similarity are presented in Table 6.4. KL divergence, CS, and KS tests measure the similarity of each individual variable independently (the higher the score, the more similar between synthetic and real data), while Cramr's V coefficient and Pearson correlation measure the

difference of dependencies between a pair of variables (the lower the score, the more dependencies in real data captured by synthetic data). DP-CGANS outperforms other models in the CS, KS, and KL Divergence test on categorical variables in most datasets. The conditional vector and the additional penalty in the generator of DP-CGANS successfully capture the underrepresented categories and the dependencies between them. CTGAN performs similarly to DP-CGANS because of its conditional GAN structure and sampling method. Both models can handle the datasets with imbalanced variables better than MedGAN and TableGAN.

Table 6.4: Results of measuring statistical similarity between real and synthetic data.

	Adult	Census	Intrusion	Diabetes	Cancer
KL Divergence (Categorical)					
DP-CGANS	0.921	0.933	0.737	0.982	0.918
CTGAN	0.894	0.834	0.708	0.957	0.881
MedGAN	0.785	0.746	0.605	0.905	0.620
TableGAN	0.746	0.856	0.603	0.941	0.814
KL Divergence (Continuous)					
DP-CGANS	0.887	0.828	0.906	0.801	0.552
CTGAN	0.929	0.791	0.922	0.736	0.560
MedGAN	0.115	0.083	0.198	0.184	0.180
TableGAN	0.752	0.460	0.820	0.860	0.565
CS Test (Categorical)					
DP-CGANS	0.997	0.995	0.982	0.983	0.984
CTGAN	0.988	0.989	0.984	0.960	0.967
MedGAN	0.976	0.975	0.968	0.949	0.839
TableGAN	0.987	0.987	0.981	0.979	0.964
KSTest (Continuous)					
DP-CGANS	0.820	0.796	0.873	0.932	0.910
CTGAN	0.794	0.819	0.870	0.889	0.896
MedGAN	0.127	0.199	0.442	0.176	0.314
TableGAN	0.627	0.440	0.540	0.925	0.896
Cramer's V (Dependency between categorical variables)					
DP-CGANS	0.017	0.024	0.068	0.018	0.029
CTGAN	0.014	0.031	0.085	0.030	0.041
MedGAN	0.061	0.130	0.148	0.063	0.116
TableGAN	0.024	0.031	0.102	0.011	0.030
Pearson correlation (Correlation between continuous variables)					
DP-CGANS	0.025	0.043	0.045	0.066	0.020
CTGAN	0.033	0.064	0.050	0.132	0.056
MedGAN	0.487	0.718	0.324	0.279	0.542
TableGAN	0.077	0.058	0.046	0.092	0.051

The results of Cramer's V Coefficient and Pearson correlation show that DP-CGANS is outstanding in simulating the dependencies and correlations between variables. Figure 6.4 shows the differences of dependencies between categorical variables from the Census dataset and the synthetic data generated by different models. The darker the blue of the cell, the greater the

difference in dependence between two variables in the real and synthetic data. DP-CGANS simulates the most dependencies between variables followed by TableGAN and CTGAN, while MedGAN fails to transfer the most of dependencies. DP-CGANS outperforms TableGAN on the variables that have multiple major classes and many different minor classes. The reason is the additional penalty in the generator and the sampling method of training enable DP-CGANS to transfer the underrepresented dependencies of the minor classes in the imbalanced variables. DP-CGANS outperforms CTGAN in the variables that have one or two extreme dominant classes and several minor classes. The reason is that the construction of the conditional vector of DP-CGANS aims to capture the dependencies between imbalanced variables, but this is not presented in CTGAN.

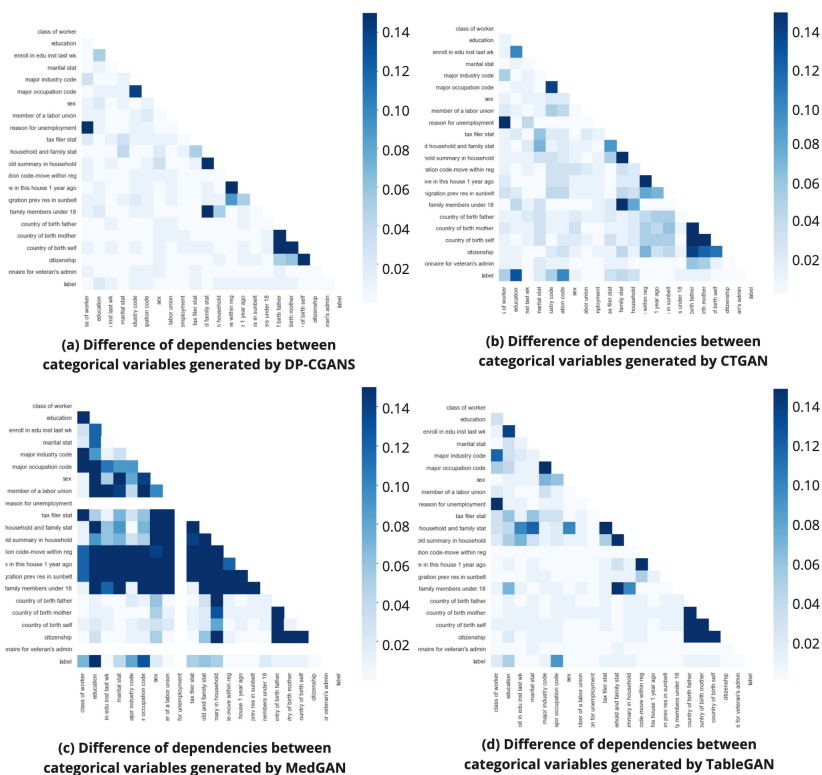


Figure 6.4: Differences of dependencies between categorical variables from real to synthetic data generated by different models. The darker the cell, the greater the difference in dependence between two variables between real and synthetic data.

All models have a common challenge, which is to accurately maintain the de-

dependencies as strong as in real data. For instance, *Age* and *Retirement*, or *Employment status* (*employed*, *unemployed*) and *Occupational class* (*high class*, *intermediate*, *low*, *not working*) have strong dependencies between them. All models capture the dependencies to a different extent. TableGAN simulates the most similar strength of these dependencies to the real data, because the third neural network model in addition to the generator and discriminator in TableGAN captures the dependencies and classifies if the generated data is realistic or not.

For the continuous variables, the inserted conditional vector in DP-CGANS helps in shaping the multimode distributions of the continuous variables and capturing the correlations. However, we found DP-CGANS suffers from oversampling the number of modes in the distributions and handling the variables with a heavy-tailed probability distribution whose tails are not exponentially bounded. This can be observed from the results of the KL divergence test on continuous variables. KL divergence and KS test are both used to observe the difference of two distributions, but only the KL divergence test shows that DP-CGANS does not have a competitive performance. This is because the differences in the tails of the distributions get amplified in KL divergence but not in the KS test.

6.5.2 Machine Learning Performance

Figure 6.5 reports the evaluation results of four machine learning models trained by generated synthetic data. Given that the AUC and F1 score are more reliable in evaluating model performance on the imbalanced datasets, we used these two scores compared with the baseline which results from the real training datasets. The better and more realistic synthetic data is, the smaller the difference in its machine learning performance from the real data. In *Adult*, *Census*, and *Diabetes*, DP-CGANS generates the synthetic data which have the most similar machine learning performance to the real data compared to other models. In these datasets, most variables are imbalanced categorical or binary variables which are handled by inserting the conditional vectors in both DP-CGANS and CTGAN. The advantage of capturing the dependencies between categorical variables in DP-CGANS is reflected on *Census* and *Diabetes*. The dependencies of variables in these two datasets have an obvious positive impact on the final classification.

CTGAN shows close performance to DP-CGANS in some experiments and slightly outperformed in the *Intrusion* dataset. *Intrusion* dataset has a few extremely imbalanced categorical variables and many continuous variables including heavy-tailed variables. The results show all included models have difficulties generating synthetic data that simulates such

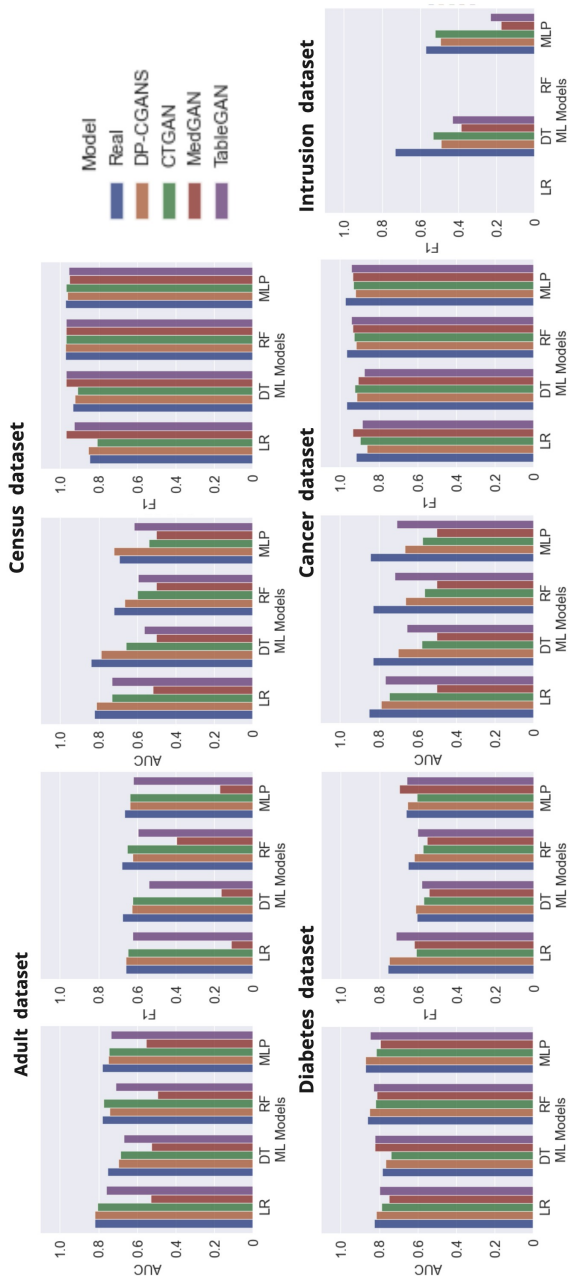


Figure 6.5: Evaluation results of training machine learning models on synthetic data using different generative models. DT and MLP are conducted to solve the multi-class classification in Intrusion dataset and evaluate the model performance using F1 macro score.

extreme distributions from real data. The added conditional vector and extra penalty to the generator strongly encourage DP-CGANs to balance the under-sampling classes and dependencies across categorical variables which weakens the precise mapping of continuous variables. Furthermore, we found TableGAN has comparable performance with CTGAN and DP-CGANs in the Cancer dataset. Cancer dataset has a much smaller size and a simpler structure compared to the other four datasets. As the only one supervised synthetic data generator in the experiment, TableGAN benefits from its third neural network as an auxiliary classifier and its convolutional GAN structure. Other included models are unsupervised synthetic data generators which typically require more data instances to train.

6.5.3 Privacy Cost

Table 6.5 reports the privacy costs in **identity disclosure** with certain threshold distances between synthetic and real data instances. Each dataset applies a different threshold of similarity as the shortest accepted distances between synthetic and real data instances. A lower precision indicates a smaller proportion of real data instances labeled by an attacker are presented in the training data. A lower recall indicates fewer real data instances can be detected by an attacker. A low precision and recall achieve a higher level of privacy. MedGAN has the least utility of the synthetic data but holds the greatest privacy guarantee. Note that the privacy score in identity disclosure is under a certain distance threshold (D). For example, MedGAN has no synthetic data instances close to any real data instances with a distance threshold at 0.05. DP-CGANs outperforms other models in Census and Diabetes datasets regarding data utility but has the most privacy costs in identity disclosure. Similar results are observed in Cancer dataset where TableGAN has best utility scores but lowest privacy level.

Table 6.5: Privacy measurement in identity disclosure on five datasets. Pre represents precision score, while Rec represents recall score. Both scores are 0 to 1.

Dataset	Adult (D=0.1)		Census (D=0.05)		Intrusion (D=0.01)		Diabetes (D=0.2)		Cancer (D=0.2)	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DPCGANs	0.518	0.254	0.576	0.193	0.489	0.389	0.495	0.188	0.585	0.088
CTGAN	0.552	0.272	0.497	0.078	0.496	0.783	0.469	0.127	0.683	0.102
MedGAN	0	0	0	0	0	0	0	0	0	0
TableGAN	0.618	0.110	0.486	0.123	0	0	0.488	0.179	0.895	0.162

The **attribute disclosure** measurement was calculated as $1 - P_{attr}$ where P_{attr} is the average posterior probabilities of correctly predicting the unknown (sensitive) variables in real test data. Table 6.6 reports the average score of experiments on 3, 6, and all rest known variables to predict the unknown variables. A greater score presents a higher level of privacy. Cancer dataset

Table 6.6: Privacy measurement of attribute disclosure with using three, six, and all rest known variables. The average scores are reported.

Dataset	Adult		Census		Intrusion		Diabetes		Cancer	
	Cat	Con	Cat	Con	Cat	Con	Cat	Con	Cat	Con
Real	0.281	0.065	0.118	0.066	0.046	0.004	0.427	0.080	0.126	-
DPCGANS	0.300	0.081	0.127	0.081	0.052	0.016	0.448	0.081	0.181	-
CTGAN	0.301	0.151	0.137	0.104	0.049	0.003	0.510	0.082	0.158	-
MedGAN	0.372	0.266	0.361	0.263	0.319	0.196	0.501	0.216	0.258	-
TableGAN	0.305	0.090	0.138	0.142	0.057	0.036	0.492	0.084	0.137	-

has only 2 continuous variables which is not sufficient to conduct an evaluation test. Note that the privacy measurement in attribute disclosure is calculated with respect to the real data. This means the probability of predicting unknown variables in the synthetic data is close to (typically higher than) the probability in the real data. DP-CGANS, which generates the most realistic synthetic data among other models, has relatively low privacy levels in the Adult, Census and Diabetes datasets. CTGAN and TableGAN which have better performance in the Intrusion and Cancer datasets respectively have the least privacy guarantees in these datasets.

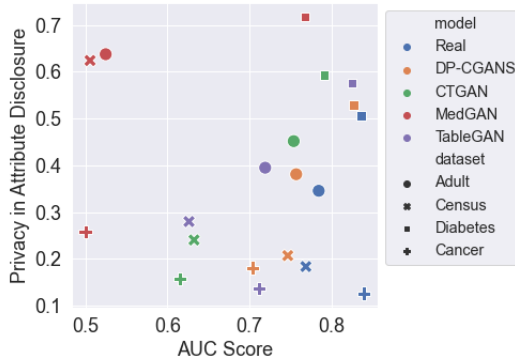


Figure 6.6: AUC scores and privacy level in attribute disclosure of four generators.

The privacy measurements in both attribute and identity disclosure show the trade-off between synthetic data utility and real data privacy. Figure 6.6 plots the AUC scores as the indicator of data utility and privacy level against attribute disclosure as the indicator of data privacy. In all datasets, we found the model that obtains a higher AUC score has a lower level of privacy preservation. On the one hand, we define a generative model is good if it can produce synthetic data as similar as possible to the real data. This means the generator in GAN is motivated to minimize the distance between real and synthetic data. On the other hand, privacy measurement shows a generator

outputs data that has a smaller distance to the real data takes more privacy risk in revealing sensitive information from the real data. Therefore, it is an inevitable trade-off between data privacy and data utility in generating synthetic data. Finally, it is found that all included models obtain a relatively low level of privacy in the experiments. It explains the essentiality and necessity of enhancing the privacy guarantee to the construction and training process of generative models.

6.5.4 DP-CGANS with different privacy budget

The previous experiments were conducted on DP-CGANS without privacy restrictions ($\epsilon = \infty$) to evaluate the capability in generating realistic synthetic data with high data utility. Then, we adjust the privacy budget of DP-CGANS ($\epsilon = 0.1, 1, 10, 100, \text{ and } \infty$) to enhance the privacy guarantee. Figure 6.7 shows the overall changes of model performance using different privacy budgets on Adult and Diabetes datasets. The average scores are plotted in the figure. Figure 6.7(a) and (b) shows the statistical similarity and ML performance are climbing up as the privacy budget (ϵ) increases. A larger privacy budget indicates a smaller scale of noises are added into the model training process which means a lower the level of privacy is preserved in the synthetic data. This is proven by the privacy measurement in attribute and identity disclosure under increasing privacy budget as Figure 6.7(c) and 6.7(d) show.

Both datasets demonstrate the trade-off between model performance and privacy level with similar overall changes under increasing privacy budgets. However, the privacy budgets show different impacts on the learning performance of the model on different datasets. When adding ϵ from 0.1 to 1, the model has an obvious improvement in statistical similarity on the Adult dataset (figure 6.7(a)). When $\epsilon > 1$, this increase becomes slower. The corresponding changes are observed in figure 6.7(c) and 6.7(d) that the level of privacy drops steeply when ϵ increases from 0.1 to 1. The model has a relatively stable increase on the Diabetes dataset with a turning point at $\epsilon=10$. Figure 6.7(d) shows the impact of privacy budget on the model performance becomes dramatic when ϵ is between 10 and 100. Although the Adult and Diabetes datasets are both imbalanced datasets and have the same ratio of categorical and continuous variables, the model reacts on the same privacy budget with different sensitivities in different datasets. Obtaining the most optimal balance between model performance and privacy guarantee depends on the data structure and characteristics of each dataset (such as imbalance variables, abnormal distributions, sparsity of data).

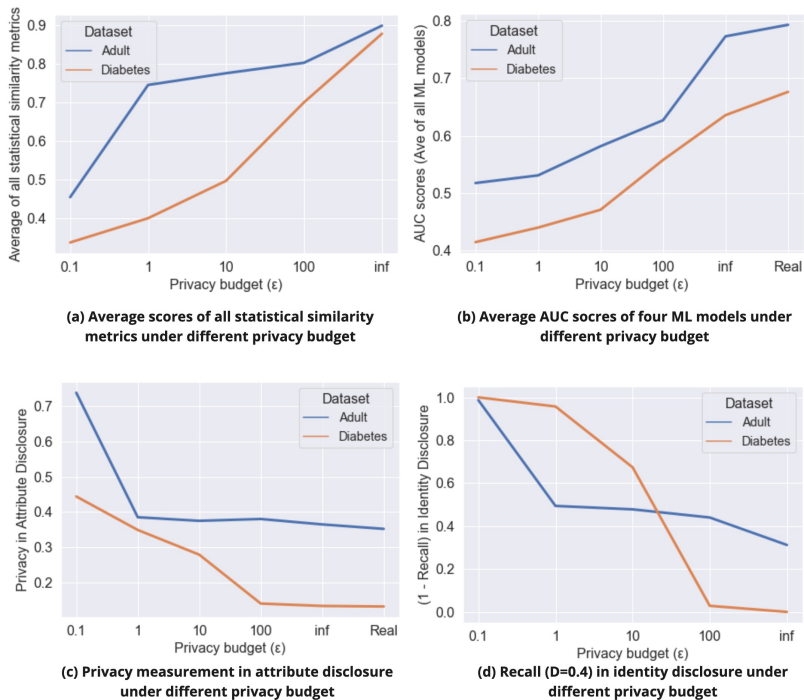


Figure 6.7: Statistical similarity, ML performance, privacy measurement in attribute and identity disclosure of DP-CGANS under different DP budgets.

6.5.5 Limitations

The empirical results in this study should be considered in the light of limitations. First, in the experiments, we sampled 25% of data instances from the Census and Intrusion datasets in a stratified random manner owing to the computation resources and time. Using subsamples of the data may limit the performance of the generative models. However, given the proportion of our samples from the original dataset and the model performance reported by other comparable studies, we do not expect that the sampled data would have a significant impact on the final experiments results.

Second, the conditional vector in the generator of DP-CGANS can successfully capture the dependencies between each pair of variables. However, the generator does not learn to maintain the relations among more than two variables. Extending the construction of the conditional vector to three or more variables dramatically increases the dimension of the vector at the expense of training efficiency. A potential solution can be training the generative model

in a semi-supervised or supervised manner such as selectively including the variables and categories in the conditional vectors or introducing a classifier which is trained with the generator. Further research can be conducted to improve the capture of dependencies among multiple variables.

DP-CGANS can output differential private (DP) synthetic data with imbalanced variables and keeping the dependencies between variables mainly because of the generator structure. However, the differential privacy budget (noise) is added to the discriminator and indirectly affects the generator. The data utility of the synthetic data might drop dramatically when tuning the differential privacy budget. Therefore, to obtain the optimal balance between data utility and privacy, DP-CGANS costs computation, time, and effort to carefully find the best suitable privacy budget for each dataset. In the future work, we intend to control the impact of adding DP to the network training on the generator. The solution could be to apply DP directly on the generator instead of discriminator or stabilize the effect of adding noise on the loss which the generator receives from the discriminator.

6.6 Conclusion

We proposed DP-CGANS, a differentially private conditional GAN, consisting of four main components - transforming, sampling, conditioning, and network training to generate realistic and privacy-preserving synthetic data. DP-CGANS handles data with mixed types and imbalanced variables and captures the correlations and dependencies between variables with privacy guaranteed. We compared our model with state-of-the-art generative models on three public datasets and two real-world personal health datasets using a set of extensive evaluation matrices focusing on the statistical similarity, machine learning performance, and privacy measurement. The evaluation results show that our model outperforms other comparable models, especially in capturing dependency between variables. Meanwhile, we measured the privacy risks in different generative models regarding attribute disclosure and identity disclosure. Our experiments prove the trade-off between output data utility (synthetic data) and input data privacy (real data) and our model can reduce privacy risks to a certain extent while maintaining data quality. Finally, we discussed the limitations of DP-CGANS and provided future directions to improve the generation of synthetic data using the developed framework.

References

- [1] Sharyl J. Nass, Laura A. Levit, Lawrence O. Gostin, and Institute of Medicine (US) Committee on Health Research {and} the Privacy of Health Information: The HIPAA Privacy Rule. “The Value, Importance, and Oversight of Health Research”. In: *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. National Academies Press (US), 2009, pp. 111–152. URL: <https://www.ncbi.nlm.nih.gov/books/NBK9571/>.
- [2] Shona Kalkman, Johannes van Delden, Amitava Banerjee, Benoît Tyl, Menno Mostert, and Ghislaine van Thiel. “Patients’ and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence”. In: *Journal of Medical Ethics* 48.1 (Jan. 1, 2022). Publisher: Institute of Medical Ethics Section: Clinical ethics, pp. 3–13. DOI: 10.1136/medethics-2019-105651.
- [3] David B Resnik. “Openness versus secrecy in scientific research”. In: *Episteme* 2.3 (2006), pp. 135–147. DOI: 10.3366/epi.2005.2.3.135.
- [4] European Commission. Directorate General for Communications Networks, Content and Technology., CEPS., ICF., and Wavestone. *Study to support an impact assessment of regulatory requirements for Artificial Intelligence in Europe: final report*. LU: Publications Office, 2021. URL: <https://data.europa.eu/doi/10.2759/523404>.
- [5] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. “Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record”. In: *Journal of the American Medical Informatics Association* 25.3 (Mar. 1, 2018), pp. 230–238. DOI: 10.1093/jamia/ocx079.
- [6] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. “Generation and evaluation of synthetic patient data”. In: *BMC Medical Research Methodology* 20.1 (2020), p. 108. DOI: 10.1186/s12874-020-00977-1.
- [7] Bill Howe, Julia Stoyanovich, Haoyue Ping, Bernease Herman, and Matt Gee. “Synthetic Data for Social Good”. In: *arXiv:1710.08874 [cs]* (Oct. 24, 2017). URL: <http://arxiv.org/abs/1710.08874> (visited on 11/29/2021).
- [8] Leho Tedersoo et al. “Data sharing practices and data availability upon request differ across scientific disciplines”. In: *Scientific Data* 8.1 (July 27, 2021), p. 192. DOI: 10.1038/s41597-021-00981-0.

- [9] Fiona V Lugg-Widger, Lianna Angel, Rebecca Cannings-John, Kerensa Hood, Kathryn Hughes, Gwenllian Moody, and Michael Robling. "Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: Managing the morass." In: *International Journal of Population Data Science* 3.3 (2018). DOI: 10.23889/ijpds.v3i3.432.
- [10] Nirupa Dattani, Pia Hardelid, Jonathan Davey, Ruth Gilbert, and on behalf of the Working Group of the Research {and} Policy Directorate of the Royal College of Paediatrics {and} Child Health. "Accessing electronic administrative health data for research takes time". In: *Archives of Disease in Childhood* 98.5 (May 1, 2013), pp. 391–392. DOI: 10.1136/archdischild-2013-303730.
- [11] Chang Sun et al. "A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario". In: *Studies in Health Technology and Informatics* 264 (2019), pp. 373–377. DOI: 10.3233/SHTI190246.
- [12] Suranga N Kasthurirathne, Joshua R Vest, Nir Menachemi, Paul K Halverson, and Shaun J Grannis. "Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services". In: *Journal of the American Medical Informatics Association* 25.1 (2017), pp. 47–53. DOI: 10.1093/jamia/ocx130.
- [13] Oya Beyan et al. "Distributed Analytics on Sensitive Medical Data: The Personal Health Train". In: *Data Intelligence* 2.1-2 (2020), pp. 96–107. DOI: 10.1162/dint.a.00032.
- [14] Sinem Sav, Apostolos Pyrgelis, J. Troncoso-Pastoriza, David Froelicher, Jean-Philippe Bossuat, João Sá Sousa, and J. Hubaux. "POSEIDON: Privacy-Preserving Federated Neural Network Learning". In: *NDSS* (2021). DOI: 10.14722/NDSS.2021.24119.
- [15] Stefanie Wornat-Herresthal et al. "Swarm Learning for decentralized and confidential clinical machine learning". In: *Nature* 594.7862 (2021), pp. 265–270. DOI: 10.1038/s41586-021-03583-3.
- [16] Juan González-García et al. "Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations from the JA-InfAct". In: *Archives of Public Health* 79.1 (2021), p. 221. DOI: 10.1186/s13690-021-00731-z.
- [17] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. "Federated Learning for Healthcare Informatics". In: *Journal of Healthcare Informatics Research* 5.1 (2021), pp. 1–19. DOI: 10.1007/s41666-020-00082-4.

-
- [18] David Roschewitz, Mary-Anne Hartley, Luca Corinzia, and Martin Jaggi. "IFedAvg: Interpretable Data-Interoperability for Federated Learning". In: *arXiv:2107.06580 [cs]* (2021). URL: <http://arxiv.org/abs/2107.06580> (visited on 02/12/2022).
- [19] Jean-Pierre Briot. "From artificial neural networks to deep learning for music generation: history, concepts and trends". In: *Neural Computing and Applications* 33.1 (2021), pp. 39–65. DOI: 10.48550/ARXIV.2004.03586.
- [20] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. "Quant GANs: deep generation of financial time series". In: *Quantitative Finance* 0.0 (2020), pp. 1–22. DOI: 10.1080/14697688.2020.1730426.
- [21] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. "Synthesizing electronic health records using improved generative adversarial networks". In: *Journal of the American Medical Informatics Association* 26.3 (2019), pp. 228–241. DOI: 10.1093/jamia/ocy142.
- [22] Lakshmi Kurup, Meera Narvekar, Rahil Sarvaiya, and Aditya Shah. "Evolution of Neural Text Generation: Comparative Analysis". In: *Advances in Computer, Communication and Computational Sciences*. Springer, 2021, pp. 795–804. DOI: 10.1007/978-981-15-4409-5_71.
- [23] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 5908–5916. DOI: 10.1109/ICCV.2017.629.
- [24] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling Tabular data using Conditional GAN". In: *arXiv:1907.00503 [cs, stat]* (Oct. 27, 2019). URL: <http://arxiv.org/abs/1907.00503> (visited on 10/04/2021).
- [25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014). DOI: 10.48550/arXiv.1406.2661.
- [26] Zilong Zhao, Aditya Kumar, Hiek Van der Scheer, Robert Birke, and Lydia Y. Chen. "CTAB-GAN: Effective Table Data Synthesizing". In: *arXiv:2102.08369 [cs]* (May 31, 2021). URL: <http://arxiv.org/abs/2102.08369> (visited on 10/26/2021).

- [27] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. “Generating multi-label discrete patient records using generative adversarial networks”. In: *Machine learning for healthcare conference*. PMLR. 2017, pp. 286–305. DOI: 10.48550/arXiv.1703.06490.
- [28] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. “Data synthesis based on generative adversarial networks”. In: *arXiv preprint arXiv:1806.03384* (2018).
- [29] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein GAN”. In: *arXiv:1701.07875 [cs, stat]* (Dec. 6, 2017). URL: <http://arxiv.org/abs/1701.07875> (visited on 11/29/2021).
- [30] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. “Improved Training of Wasserstein GANs”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017, pp. 5769–5779. URL: <https://papers.nips.cc/paper/2017/hash/892c3b1c6dccb52936e27cbd0ff683d6-Abstract.html> (visited on 11/29/2021).
- [31] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. “PacGAN: The Power of Two Samples in Generative Adversarial Networks”. In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 324–335. DOI: 10.1109/JSAIT.2020.2983071.
- [32] Justin Engelman and S. Lessmann. “Conditional Wasserstein GAN-based Oversampling of Tabular Data for Imbalanced Learning”. In: *Expert Syst. Appl.* (2021). DOI: 10.1016/J.ESWA.2021.114582.
- [33] Lucas Rosenblatt, Xiaoyan Liu, Samira Pouyanfar, Eduardo de Leon, Anuj Desai, and Joshua Allen. “Differentially Private Synthetic Data: Applied Evaluations and Enhancements”. In: *arXiv:2011.05537 [cs]* (Nov. 10, 2020). URL: <http://arxiv.org/abs/2011.05537> (visited on 11/29/2021).
- [34] Aditya Kumar. “Effective and Privacy preserving Tabular Data Synthesizing”. In: *arXiv:2108.10064 [cs]* (Aug. 11, 2021). URL: <http://arxiv.org/abs/2108.10064> (visited on 11/29/2021).
- [35] Amirsina Torfi, Edward A. Fox, and Chandan K. Reddy. “Differentially private synthetic medical data generation using convolutional GANs”. In: *Information Sciences* 586 (2022), pp. 485–500. DOI: 10.1016/j.ins.2021.12.018.

-
- [36] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. “DP-CGAN: Differentially Private Synthetic Data and Label Generation”. In: *arXiv:2001.09700 [cs, stat]* (Jan. 27, 2020). URL: <http://arxiv.org/abs/2001.09700> (visited on 11/29/2021).
- [37] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. “Differentially Private Generative Adversarial Network”. In: *arXiv:1802.06739 [cs, stat]* (Feb. 19, 2018). URL: <http://arxiv.org/abs/1802.06739> (visited on 11/29/2021).
- [38] Cynthia Dwork. “Differential privacy: A survey of results”. In: *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19. DOI: 10.1007/978-3-540-79228-4_1.
- [39] Ilya Mironov. “Rényi Differential Privacy”. In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)* (2017), pp. 263–275. DOI: 10.1109/CSF.2017.11.
- [40] Dingfan Chen, Tribhuvanesh Orekondy, and Mario Fritz. “Gs-wgan: A gradient-sanitized approach for learning differentially private generators”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12673–12684. DOI: 10.48550/arXiv.2006.08265.
- [41] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. “Membership Inference Attacks Against Machine Learning Models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. 2017, pp. 3–18. DOI: 10.1109/SP.2017.41.
- [42] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. “Logan: Membership inference attacks against generative models”. In: *Proceedings on Privacy Enhancing Technologies (PoPETs)*. Vol. 2019. 1. De Gruyter, 2019, pp. 133–152. DOI: 10.2478/popets-2019-0008.
- [43] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures”. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS ’15. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677.
- [44] Manhar Walia, Brendan Tierney, and Susan McKeever. “Synthesising Tabular Data using Wasserstein Conditional GANs with Gradient Penalty (WCGAN-GP)”. In: *AICS*. 2020. DOI: 10.21427/E6WA-SZ92.

- [45] Jaeuk Moon, Seungwon Jung, Sungwoo Park, and Eenjun Hwang. "Conditional Tabular GAN-Based Two-Stage Data Generation Scheme for Short-Term Load Forecasting". In: *IEEE Access* 8 (2020). Conference Name: IEEE Access, pp. 205327–205339. DOI: 10.1109/ACCESS.2020.3037063.
- [46] Lei Xu and Kalyan Veeramachaneni. "Synthesizing Tabular Data using Generative Adversarial Networks". In: *CoRR abs/1811.11264* (2018). DOI: 10.48550/arXiv.1811.11264.
- [47] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. "Empirical evaluation of rectified activations in convolutional network". In: *arXiv preprint arXiv:1505.00853* (2015).
- [48] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. icml*. Vol. 30. 1. Citeseer. 2013, p. 3. DOI: 10.1109/IJCNN.2016.7727219.
- [49] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407. DOI: 10.1561/04000000042.
- [50] Catherine L Blake and Christopher J Merz. *UCI repository of machine learning databases, 1998*. 1998. URL: <https://archive.ics.uci.edu/ml/index.php>.
- [51] Ronny Kohavi and Barry Becker. *Adult Data Set*. 1996. URL: <https://archive.ics.uci.edu/ml/datasets/adult>.
- [52] Terran Lane and Ronny Kohavi. *Census-Income (KDD) Data Set*. 2000. URL: <https://archive.ics.uci.edu/ml/datasets/Census-Income+%5C%28KDD%5C%29>.
- [53] *KDD Cup 1999 Data Data Set*. 1999. URL: <https://archive.ics.uci.edu/ml/datasets/kdd+cup+1999+data>.
- [54] Miranda T Schram, Simone JS Sep, Carla J van der Kallen, Pieter C Dagnelie, Annemarie Koster, Nicolaas Schaper, Ronald Henry, and Coen DA Stehouwer. "The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities". In: *European journal of epidemiology* 29.6 (2014), pp. 439–451. DOI: 10.1007/s10654-014-9889-0.
- [55] Hugo J. W. L. Aerts et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach". In: *Nature Communications* 5.1 (June 3, 2014), p. 4006. DOI: 10.1038/ncomms5006.
- [56] Hugo J. W. L. Aerts et al. *Data From NSCLC-Radiomics*. Version Number: 4 Type: dataset. 2019. DOI: 10.7937/K9/TCIA.2015.PF0M9REI.

-
- [57] John R Hershey and Peder A Olsen. "Approximating the Kullback Leibler divergence between Gaussian mixture models". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Vol. 4. IEEE. 2007, pp. IV-317. DOI: 10.1109/ICASSP.2007.366913.
- [58] Karl Pearson. "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (July 1900), pp. 157-175. DOI: 10.1080/14786440009463897.
- [59] Frank J Massey Jr. "The Kolmogorov-Smirnov test for goodness of fit". In: *Journal of the American statistical Association* 46.253 (1951), pp. 68-78. DOI: 10.2307/2280095.
- [60] Bruce B Frey. *The SAGE encyclopedia of educational research, measurement, and evaluation*. Sage Publications, 2018. DOI: 10.4135/9781506326139.n169.

7

ciTizen-centric Data pLatform (TIDAL): Using Distributed Personal Data in a Privacy-Preserving Manner for Health Research

Adapted from: Chang Sun, Marc Gallofré, Johan van Soest, and Michel Dumontier. "ciTizen-centric Data pLatform (TIDAL): Using Distributed Personal Data in a Privacy-Preserving Manner for Health Research". Submitted in: *Semantic Web Journal* (2022)

Abstract

Developing personal data sharing tools and standards in conformity with data protection regulations is essential to empower citizens to control and share their health data with authorized parties for any purpose they approve. This can be, among others, for primary use in healthcare, or secondary use for research to improve human health and well-being. Ensuring that citizens are able to make fine-grained decisions about how to use and share their personal health data will significantly encourage citizens to participate in more health-related research. In this paper, we propose ciTizen-centric Data pLatform (TIDAL) to give individuals ownership of their own data, and includes mechanisms to provide fine-grained access to external parties. The TIDAL platform integrates a set of components for requesting subsets of RDF (Resource Description Framework) data stored in personal data vaults based on Social LInked Data (SOLID) technology and analyzing them in a privacy-preserving manner. We demonstrate the feasibility and efficiency of the TIDAL platform by conducting a set of simulation experiments using three different pod providers (*Inrupt.net*, *Solidcommunity.net*, Self-hosted Server). On each pod provider, we evaluated the performance of TIDAL by querying and analyzing personal health data from an increasing number of participants and variables. The performance evaluation of TIDAL shows the execution time has a linear correlation between the number of pods on all pod providers. Platforms such as TIDAL can play an important role to connect citizens, researchers, and data organizations to increase the trust placed by citizens in the processing of their personal data.

7.1 Introduction

Giving individuals more control over who can access their personal data for what purpose and to make their data available using privacy-preserving and transparent methodologies will significantly encourage their engagement in health research [1, 2]. Personal health data is needed to improve evidence-based healthcare research and empower healthcare authorities to optimize the accessibility and effectiveness of the healthcare services [3]. However, personal health data is largely collected and managed by various healthcare service providers. In Europe, many citizens still have limited electronic access to their own health data, often scattered among service providers [4]. As a result, citizens have limited control over their own data, or need to control this data at various locations.

Since the General Data Protection Regulation (GDPR) and ePrivacy legislation have been released, European Union's citizens increasingly value their data rights and information privacy [5]. However, there is no mature technology and standards that enable individuals to fully exercise their data rights in a simple way. The public consultation on the European strategy for data showed that almost 88% of all respondents (806 contributors) would like to have more access and control over the data they generate [6]. A large proportion of respondents would be willing to share their data, especially for health-related research, but a majority of them considered that there are no sufficient tools and mechanisms to "donate" their data. For example, at present, if individuals are willing to donate their health data to help chronic disease research, they need to look for an ongoing research study that is recruiting new participants and has requirements that are applicable. Meanwhile, individuals need to trust and be willing to share their data with this research study. However, sharing personal data often raises concerns about privacy, security, ownership, and accountability. Examples of these concerns are: who will have access to the data and study results, how the individuals can change/revoke the permissions to access the data, and whether the data is used for other purposes.

In this paper, we address the research challenge of *how to engage individuals to "donate" their personal data for health-related research with maximal control in data access, storage, and analysis?* The current personal data management technologies are mostly research-driven and in their early stages. Given the gap in the existing personal data platforms, we propose a new citizen-centric data platform (called TIDAL) that gives individuals fine-grained access to their data and ensures citizen controlled data are processed in a predefined manner. We designed a prototype as proof-of-concept following an exploratory technology development process in light of our experience in the development of a privacy-preserving distributed data analysis infrastructure in the

previous studies [7, 8, 9, 10]. TIDAL consists of an integrated set of components for requesting subsets of data stored in personal data vaults using SOLID technologies [11] and analyzing them in a privacy-preserving manner. SOLID, standing for SOcial LIinked Data, is a set of technologies that facilitates users to create decentralized applications using Linked Data and the W3C standards and protocols. We evaluated the performance of TIDAL by executing simulation experiments on various sizes of simulation datasets. The experiments proved the feasibility and efficiency of TIDAL using three different pod providers (*Inrupt.net*, *Solidcommunity.net*, Self-hosted Server). We believe the TIDAL platform will increase the trust placed by individuals and the transparency of the processing of their personal data.

We summarize the main contributions of this paper:

- proposing a new citizen-centric data platform (TIDAL) that facilitates individuals to store and access to their personal data using personal data vault technologies (such as SOLID) and provide direct consent to health-related research;
- applying Data Privacy Vocabulary [12, 13] to structure the personal data requests as digital consents in TIDAL to meet the requirements of GDPR;
- formulating data requests into RDF format with integrating vocabulary services and standards to improve the interoperability of personal data use;
- executing privacy-preserving data mining algorithms automatically using parameters and configurations promised in the data request and only the results are sent to the researchers; and
- evaluating the feasibility and efficiency of TIDAL in different experimental settings.

The article below is structured as follows: section 7.2 introduces the recent related work. Section 7.3 describes the SOLID technology we applied in the TIDAL platform. Section 7.4 presents the architecture of TIDAL, and demonstrates how it works for researchers and participants. Section 7.5 describes the experimental setup and results of TIDAL in different user scenarios. Section 7.6 discusses our discovery and limitations of the current version of TIDAL. Finally, conclusions and future work are outlined in Section 7.7.

7.2 Related work

Researchers and companies have developed several tools with different emphases and features to enable individuals to be in control of their data. We identified the following ten projects and tools which have been applied in practice and provided a comparison table including detailed information and additional resources in the supplementary material¹. DEcentralised Citizen Owned Data Ecosystems (DECODE) [14], MyHealthMyData (MHMD) [15] and OwnYourData [16] are based on distributed-ledger technologies such as Blockchain to provide traceable and transparent data-access control. MIDATA [17] and MedMij [18] are national programs in Switzerland and Netherlands that provide citizens with new data ecosystems to use their medical data for healthcare services and research. Digi.me [19] and CozyCloud [20] are commercial products providing mobile applications and cloud services to share personal data. The Hub of All Things (HAT) [21] – a foundation –, MyDex [22] – a community interest company –, and openPDS [23] – a research project – utilize the Personal Data Store (PDS) [24] technology to provide users with servers to store and share their personal data and execute on-device computations.

DECODE and MHMD were both funded by the European Union's Horizon 2020 research and innovation programme [25]. DECODE enables individuals to keep personal information private or share it for the public good using peer-to-peer networks and Blockchain technologies. This ecosystem, from an operating system to an interactive dashboard, has been developed and piloted in Barcelona and Amsterdam. However, it focuses on individual control over data sharing rather than data process and analysis. Individuals can specify the "smart rules" for their personal data to pre-define under what conditions their data can be used. Since DECODE relies on its own operating system and tools, it lacks the interoperability and extensibility that would be required for data mobility across healthcare systems and national borders.

MHMD [15] is another Blockchain-based software that connects organizations and individuals to make anonymised data available for open research. It enables individuals to provide dynamic consent for different types of potential data usage and monitor the usage. Similar to the DECODE, the MHMD consent determines under what conditions the data can be used. MHMD supports data analysis algorithms combined with secure multiparty computation and asymmetric encryption for preserving privacy. However, individuals' data is still hosted at organizations (e.g., hospitals), which are the only ones empowered to give permission to researchers requesting data. OwnYourData [16], developed by a non-profit organization, is another personal

¹<https://doi.org/10.6084/m9.figshare.19111508>

data management product that uses Blockchain technology to make data immutable. The key feature of OwnYourData is creating insights into users' personal data and providing certain algorithms to analyze their own data.

MIDATA [17], a nonprofit cooperative in Switzerland, operates a data platform that enables Swiss citizens to selectively share their data with medical research and clinical studies. MIDATA shares the same limitations as DECODE on the interoperability and extensibility of their data ecosystem. MedMij [18] is established as a standard in the Netherlands for the secure exchange of health data between Dutch residents and healthcare providers. MedMij, serving as a high-level guideline, proposes a set of information standards to structure the health data from different sources and standardize the data exchange. However, MedMij does not yet include researchers in the network nor facilitates citizens to voluntarily share their health data for research studies or any other purposes.

Digi.me [19] and CozyCloud [20] deliver commercial products to give people control of their data when using web or mobile applications, but both host data centralized in their own cloud servers. Similarly, MyDex [22] and HAT [21] offer PDS as cloud hosted servers to store personal data and connect it with other web or mobile applications and services. Different from the previous tools that host the PDS in their own servers, openPDS [23] allows users to self-host the PDS and use it as a service. OpenPDS also applies the SafeAnswers framework which executes the queries inside the PDS rather than sending anonymized data and returns and aggregates results from more than one PDS. It allows users to manage data access and monitor data usage. However, SafeAnswers presents a computational challenge for complex data analysis and does not consider the scenario of conducting research studies in a large populations.

The existing solutions often focus on one particular aspect such as personal data storage and overview, personal data sharing with healthcare services or with mobile applications and personal data access control. To the best of our knowledge, there is no platform that enables individuals to connect with researchers to donate their personal data for research while being in control of the whole data life cycle including data access, storage, and analysis. Only a few tools support personal data analysis over a number of participants. These tools face challenges such as the permissions to data are from the data organizations rather than individuals and the analysis algorithms are relatively simple. We also see an urgent need for more investments in data quality and interoperability to improve the feasibility and sustainability of personal data management platforms [2]. Therefore, we propose TIDAL to fill the gaps that we have recognized from the existing work.

7.3 Background

7.3.1 SOLID - Decentralized data management

SOLID (SOcial LIInked Data) is a decentralized data management platform based on W3C standards, Resource Description Framework (RDF), and Semantic Web technologies, initiated by Tim Berners-Lee [26, 27, 28]. Rather than the tech giants storing and controlling personal data from their users, SOLID technologies enable users to store and manage their own data independently from the applications so that users can retain sovereignty over their data. SOLID is composed of three core components - the data pod (i.e., where the data is stored), the application (i.e., the services that users can use and grant access to), and providers (i.e., where the pod and application are technically hosted).

Each SOLID user is assigned with a WebID² as a unique global ID for identification and authentication. SOLID data pods are web-based storage services and databases where various types of data can be stored such as RDF triples, free text, images, videos, or even webpages. However, SOLID is featured by its capability to parse and serialize structured data using RDF in syntaxes like Turtle and JSON-LD. Data in SOLID pods can be accessed and managed by a decentralized authentication³ and Web Access Control (WAC)⁴ mechanism [26] which is a decentralized cross domain access control system. WAC in SOLID provides the pod owners with a fine-grained access control for every single data element in their data pod by granting other SOLID users and applications the permissions to read, modify, and write the stored data elements. The Access Control List ontology⁵ [29] is applied to SOLID to describe the different operations over the target data elements in the pods.

SOLID applications are developed on top of the aforementioned technology stack. Most applications are developed for web or mobile. Users are able to grant and revoke permissions to both SOLID applications and other users at any time. SOLID allows multiple applications to access and reuse the same data from a pod, thereby potentially minimizing data duplication and staleness. SOLID pods can be hosted on public servers by pod providers which play a similar role as the cloud storage providers. SOLID pods can also be self-hosted on personal servers, and migrated from pod providers to self-hosted. A single SOLID user can own more than one data pod which is hosted by one or multiple pod providers. Users are able to select and change their

²<https://www.w3.org/wiki/WebID>

³<https://solid.github.io/authentication-panel/solid-oidc>

⁴<https://solid.github.io/web-access-control-spec>

⁵<https://www.w3.org/ns/auth/acl>

pod providers at any time based on providers' geographical locations, responsibilities, different degrees of privacy protection and legislation. Thus, SOLID presents a distributed scenario that challenges the communication between SOLID applications and data pods, but provides fine-grained data control to users.

7.3.2 Personal Health Train - Distributed data analysis initiative

The Personal Health Train (PHT) initiative was designed for healthcare innovators and researchers to access heterogeneous data sources and learn from the distributed data in a privacy-preserving manner [30], [10]. The essence of this approach is to transfer the research questions and analysis algorithms (from researchers) to data rather than centralizing data and moving them to researchers. Only the analysis results are sent back to the researchers.

The PHT technology has been developed and implemented in several real-life use cases in the healthcare domain. In our previous studies, we have developed the PHT infrastructure to address horizontally and vertically partitioned data problems [9], [8], [31], [32]. In this study, we further extend the PHT infrastructure from the level of information sharing among organizations to information sharing by individuals themselves.

7.4 Overview and implementation of TIDAL

The primary use case of TIDAL is for researchers (data requesters) who want to analyze personal data and participants (data subjects) who are willing to donate their data for research. In this section, we will present the overview and implementation of TIDAL by describing a use case between two users.

The participants and researchers need SOLID accounts and data pods so that they can be authenticated on TIDAL. TIDAL facilitates them to create new data elements or files, modify or delete existing RDF data elements, and query data elements from their own pods. An example of fetching RDF data from a solid pod is shown in Figure 7.1.

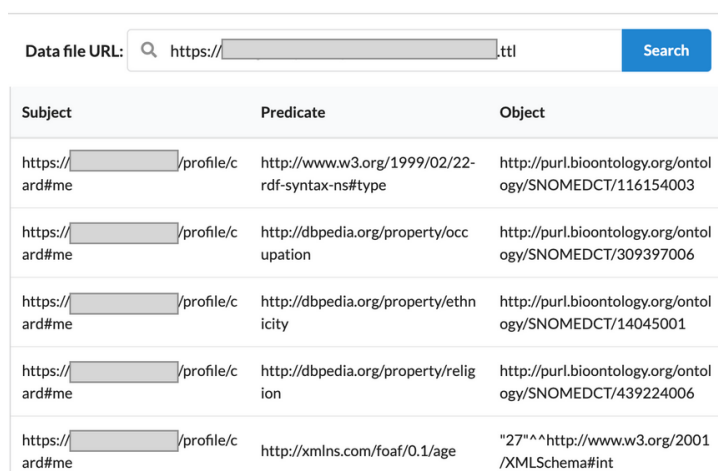
TIDAL authenticates and interacts with SOLID pods with a Javascript package - `solid-node-client` (V2.0.2) [33]. `solid-node-client` enables pod owners to access their pods, create or modify data in their pods, and grant or revoke the permissions via a web application. To store, parse, and query RDF data from SOLID pods, TIDAL uses the `rdflib.js` (V2.1.6) [34] and `tripleDoc` (V4.4.0) [35] library. Similar libraries such as `solid/query-ldflex` can also be used to access data in Solid pods through LDFlex expressions [36].

7.4.1 Researcher Posts Participation Request

In the first phase, a participation request is crafted by the researcher. The content of the request is only stored at the researcher’s SOLID pod, while the Uniform Resource Identifier (URI) of the request is stored in an index file on TIDAL. Subsequently, TIDAL reads the index file to find all the existing requests and presents them to the participants for approval (Figure 7.2).

Fetch all triples from the file

Input the URL of the data file that you want to fetch.



Subject	Predicate	Object
https://[redacted]/profile/c-ard#me	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://purl.bioontology.org/ontology/SNOMEDCT/116154003
https://[redacted]/profile/c-ard#me	http://dbpedia.org/property/occupation	http://purl.bioontology.org/ontology/SNOMEDCT/309397006
https://[redacted]/profile/c-ard#me	http://dbpedia.org/property/ethnicity	http://purl.bioontology.org/ontology/SNOMEDCT/14045001
https://[redacted]/profile/c-ard#me	http://dbpedia.org/property/religion	http://purl.bioontology.org/ontology/SNOMEDCT/439224006
https://[redacted]/profile/c-ard#me	http://xmlns.com/foaf/0.1/age	"27"^^http://www.w3.org/2001/XMLSchema#int

Figure 7.1: An example of query RDF data file from a SOLID pod on TIDAL.

To post a participation request, the researcher is required to register as a “researcher role” by providing basic information such as job position, affiliation, and research topics. The researcher is issued a public-private key pair that will be used to verify the identity of the researcher and the integrity of the request. When the researcher publishes a request, the URI and content of the request will be automatically signed by the researcher’s private key. Any changes to the request will cause a verification failure when the request is executed to retrieve participants’ data. TIDAL uses the Ed25519 algorithm [37, 38], a high-speed and high-security signature scheme, for public-key signature encryption. Ed25519 is an implementation of the Elliptic Curve Digital Signature Algorithm (EdDSA) using SHA-512 (SHA-2) and Curve25519 with Twisted Edwards Curve [39]. It has been widely used in protocols such as TLS 1.3 and SSH [40]. TIDAL uses Ed25519 from the TweetNaCl (v1.0.3) package [41], a port of Networking and Cryptography library (NaCl) [42] to Javascript.

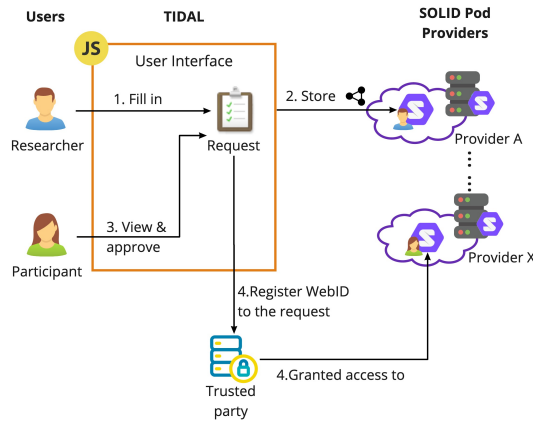


Figure 7.2: Interaction between researchers and participants on TIDAL. The researcher completes and stores the participation request form in their SOLID pod. A participant can view and approve the request on TIDAL web application. The participation record is stored at the participants SOLID pod and at the trusted party.

The researcher creates and publishes a participation request by completing a participation request form (Figure 7.3). The request form was designed as a digital consent to be informed and specific. Complying with the GDPR requirements on consents, the request form describes the identifiers of the requester (researcher) and controller (trusted party – a certificated organization compliant with GDPR that executes the requests and analyses), what data elements in which personal data categories will be processed in what time frame, how the requested data will be processed for what purposes, and the possible risks and consequences of data processing such as for participants in relation to automated decision making. The request is represented using Schema.org vocabulary (<https://schema.org/>) and the Data Privacy Vocabulary (DPV, <http://www.w3.org/ns/dpv>). The DPV specifically captures the nature of data processing in relation to EU General Data Protection Regulation. The overall schema of an example participation request is illustrated in Figure 7.4 and the stored RDF format of the example instance is shown in Listing 7.1.

The request form includes fields to specify the following:

Requested field (RF) 1: the purpose of the research where researchers clearly indicate the purpose of processing personal data in their research. Researchers can select one or more from a list of data processing purposes described in DPV such as *dpv:Security*, *dpv:ResearchAndDevelopment*. These elements will be described and stored as <http://www.w3.org/ns/dpv#Security> and <http://www.w3.org/ns/dpv#ResearchAndDevelopment> in the request form.

RF 2: description of the specific purpose where researchers elaborate the purpose with more details in human readable text. Researchers can fill in answers in free text such as “Learn association between the status of Type 2 diabetes and patients’ dietary patterns using linear regression”.

RF 3: the category of requested data elements where researchers indicate which personal data category best describes the requested data elements. Researchers can select one or more from a list of personal data categories described in DPV such as *dpv:Health* or *dpv:Income* and stored them in the researcher’s SOLID pod.

Please Note: This request form is structured using the [Data Privacy Vocabulary \(DPV\)](#). DPV provides terms (classes and properties) to describe and represent information related to processing of personal data based on established requirements such as GDPR.

Purpose of your research ⓘ *

Research and Development ✕

Description of your purpose: ⓘ

Learn association between diabetes status and dietary pattern Recommender

Personal data categories ⓘ *

Medical Health [Special] (hysical Health, Mental Health, DNA Code, Disability, Health History) ✕

Demographic (Physical Trait, Income Bracket, Geographic) ✕

Data elements (URI) ⓘ *

diagnosis +

Expiry Time *

01/01/2022

Number of instances (minimal) *

100

Data Processing Category ⓘ *

> Analyse ✕

Analysis Model *

Linear Regression

Consequences of data processing and impact of your research:

Help diabetes patients understand the impact of their diet pattern

Publish

Searching terms from BioPortal ontologies

NCIT	Diagnosis http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C15220 The investigation, analysis and recognition of the presence and nature of disease, condition, or injury from expressed signs and symptoms; also, the scientific determination of any kind; the concise results of such an investigation.
PREMEDONTO	Diagnosis http://purl.obolibrary.org/obo/NCIT_C15220 The investigation, analysis and recognition of the presence and nature of disease, condition, or injury from expressed signs and symptoms; also, the scientific determination of any kind; the concise results of such an investigation.
CRISP	diagnosis http://purl.bioontology.org/ontology/CSP/4000-0159 general term for detecting and classifying diseases.
IOBC	Diagnosis http://purl.jp/bio/4/id/200906001611549035

Figure 7.3: A participation request form on TIDAL completed by a researcher.

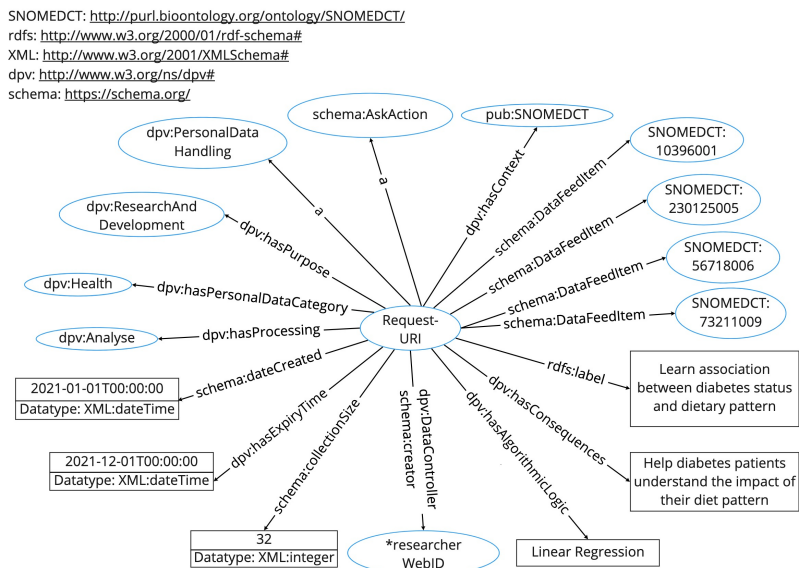


Figure 7.4: Schema of an example of a participation request.

RF 4: the data elements where researchers indicate the data elements (URI) are requested from the participants. Researchers can fill in one or more URIs of the requested data elements. Researchers can also search for the existing URIs from the existing ontologies and select the ones for the requested data elements. For example, instead of requesting the “Age” in plain text, researchers can set the URI of Age in SNOMED CT <http://purl.bioontology.org/ontology/SNOMEDCT/397669002> as requested data element in the form.

RF 5: the expiration date of consent where researchers specify an exact date when the consent will be no longer valid. Researchers can only give future dates as answer in this field such as 2025-01-23.

RF 6: the number of individuals who agree to participate in the study where researchers specify a minimal number of participants required to initiate the data processing. Researchers can only give integer numbers as the answer in this field such as 1000.)

RF 7: the categories of data processing where researchers indicate which category or a chain of data processing will be performed on the requested data. Researchers can select one or more from a list of data processing categories described in DPV such as *dpv:Copy*, *dpv:Anonymise*, and *dpv:Analyse*.

RF 8: the methods or algorithms in data processing where researchers specify how the requested data will be processed. Researchers can select one or more from a list of predefined algorithms such as *Linear or logistic regression*.

(Optional) RF 9: the consequences and impact where researchers communicate the possible risks and consequences of data processing to the participants such as for participants in relation to automated decision making. Researchers can answer in free text that is human-readable and understandable for the general public.

```
@prefix : <http://exampleresearcher.solidprovider.com/public/request.ttl#>.
@prefix schema: <https://schema.org/>.
@prefix exre: <http://exampleresearcher.solidprovider.com/profile/card#>.
@prefix XML: <http://www.w3.org/2001/XMLSchema#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix dpv: <http://www.w3.org/ns/dpv#>.
@prefix SNOMEDCT: <http://purl.bioontology.org/ontology/SNOMEDCT/>.

:161964062096710764675982245664
  a schema:AskAction, dpv:PersonalDataHandling;
  rdfs:label "Learn association between T2DM and dietary pattern";
  schema:collectionSize 32;
  schema:creator exre:me;
  schema:DataFeedItem SNOMEDCT:10396001, SNOMEDCT:230125005, SNOMEDCT:56718006,
    SNOMEDCT:73211009;
  schema:dateCreated "2021-01-18T00:00:00Z"^^XML:dateTime;
  dpv:hasAlgorithmicLogic "Linear Regression";
  dpv:hasConsequences "Help diabetes patients understand the impact of their
    dietary pattern";
  dpv:hasContext SNOMEDCT;
  dpv:hasDataController exre:me;
  dpv:hasExpiryTime "2021-12-31T00:00:00Z"^^XML:dateTime;
  dpv:hasPersonalDataCategory dpv:Health;
  dpv:hasProcessing dpv:Analyse;
  dpv:hasPurpose dpv:ResearchAndDevelopment.
```

Listing 7.1: An example of generated RDF triples of a request stored in researcher's pod

To improve the interoperability of the requested data elements, we have integrated BioPortal API [43] in TIDAL to help researchers use standardized ontologies and terminologies for specific information elements. Bioportal is the most comprehensive ontology repository for biomedical ontologies including more than 800 ontologies. TIDAL supports researchers to search the existing biomedical ontologies and terminologies provided by Bioportal and apply them to the requested data elements. For example, instead of using “*diagnosis*” as a requested data element, researchers can look for the terms from well-established ontologies such as “*http://purl.obolibrary.org/obo/NCIT_C152*” or “*http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C15220*”, by searching the keyword “*diagnosis*” in Data Elements (URI) in the request form.

Each request form is assigned with a URI when it gets published. All the information in the form is structured in RDF format as a *schema:AskAction* and

dpv:PersonalDataHandling and stored in the researchers' SOLID pods (Step 2 in Figure 7.2). The request is signed with the researcher's private encryption key while it is published in order to prevent any subsequent changes. The URI and the signature of the request are stored on TIDAL, while the content of the request is only stored in the researcher's SOLID pod.

7.4.2 Participant Views and Approves Requests

All published requests that are in the valid period (i.e., before the expiration date of the request) are visible on TIDAL. TIDAL queries RDF data from the request files and displays them in a human readable manner in a card view. We assume the participants have their personal health data (e.g., medical records, medications, lifestyle and behavior data) structured and stored using RDF in their own SOLID pods. Each card is linked to the original request file from the researcher's SOLID pod. Figure 7.5 shows an example of the published participation requests view on TIDAL from a participant's perspective. The research purpose, personal data category, data processing category, and data elements are linked with the valid URIs of the terms.

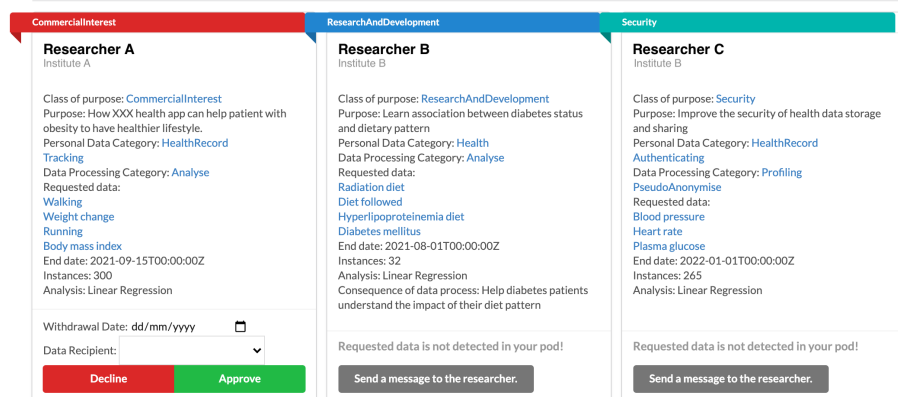


Figure 7.5: An example of viewing published participation requests on TIDAL.

If TIDAL detects the requested data elements in the participant's pod, the participant will be able to voluntarily join the data request by setting up a preferable withdrawal time (earlier than the request expiry date) and selecting the party they trust to process their personal data. TIDAL generates an instance adhering to the *schema:JoinAction* and *dpv:Consent* in RDF format describing which request has been approved at what time and until when this approval is valid. The statement is structured by using DPV and stored in a private

folder in the participant's SOLID pod. Figure 7.6 and Listing 7.2 shows the schema of an example of participation and generated consent statements.

By approving the request, the participant gives the trusted (or authenticated) party access to the requested data elements in the pod. The participant's WebID will be registered at the trusted party under the analysis request URI (Step 4 in Figure 7.2). TIDAL generates logging information in participant's pod including the participant at what time giving whom (WebID) access to what data elements in which data request (request ID) and the valid period of the permission. The logging is readable by the participants but not editable by anyone. Until now, data elements have not been accessed and retrieved by any parties.

If TIDAL fails to detect the requested data elements in the pod, the participant is not able to join the research. It is possible that the participant does not have the requested data or the researcher and participant use different standards or ontologies to describe the data elements. In this case, the participant can send messages to the researcher anonymously on TIDAL to report this issue.

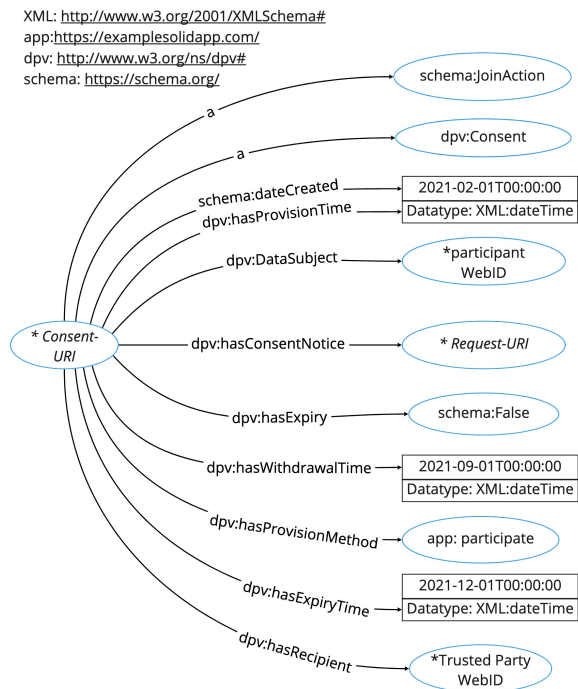


Figure 7.6: Schema of an example of participating in a request.

```
@prefix : <http://participant.solidProvider.com/private/participation#>.
@prefix dpv: <http://www.w3.org/ns/dpv#>.
@prefix part: <https://participant.solidprovider.com/profile/card#>.
@prefix req: <https://researcher.solidprovider.com/public/request.ttl#>.
@prefix extp: <http://trustedparty.solidprovider.net/profile/card#>
@prefix app: <https://solidapp.com/

:16197041266295299657542155198
  a schema:JoinAction, dpv:Consent;
  schema:dateCreated "2021-02-18T00:00:00Z"^^XML:dateTime;
  dpv:DataSubject part:me;
  dpv:hasConsentNotice req:161964062096710764675982245664;
  dpv:hasExpiry schema:false;
  dpv:hasExpiryTime "2021-12-31T00:00:00Z"^^XML:dateTime;
  dpv:hasProvisionMethod app:participate;
  dpv:hasProvisionTime "2021-02-18T00:00:00Z"^^XML:dateTime;
  dpv:hasWithdrawalTime "2021-09-18T00:00:00Z"^^XML:dateTime;
  dpv:hasRecipient extp:me.
```

Listing 7.2: An example of generated participation statements in a RDF format in the participant's SOLID pod (when approving the request).

7.4.3 Data Retrieval and Analysis Execution

To process the request, the following conditions need to be satisfied: (1) the request being in the inclusion period, and (2) the number of participants exceeding the minimum number set in the request. When the request meets both conditions, the researcher can communicate with the trusted party on TIDAL to trigger the data retrieval and analysis. The trusted party hosts the data analysis component including verifying the request, querying data from participants' pods, and executing the predefined analysis algorithms. The data analysis component was built using Javascript and Docker Containers. Docker Container has similar resource isolation and allocation benefits to virtual machines, creating temporary and secure sandboxes. We used the node-docker-api package (version-1.1.22) [44] in a combination of solid-node-client and rdflib.js libraries to access SOLID pods from a Docker container.

Figure 7.7 shows the workflow of data retrieval and analysis after the researcher triggers the execution of a request. TIDAL will first generate and send a *schema:ActivateAction* message (Listing 7.3) to the trusted party. The request file is retrieved from the researcher's pod, parsed, and verified using the public key. The data must specify the docker image identifiers (*dpv:hasAlgorithmicLogic*), requested data elements (*schema:DataFeedItem*), valid period of the request (*dpv:hasExpiryTime*) and other input parameters for the trusted party to retrieve the Docker image from the central repository and execute the analysis. TIDAL can manage multiple data retrieval and analysis request from researchers simultaneously.

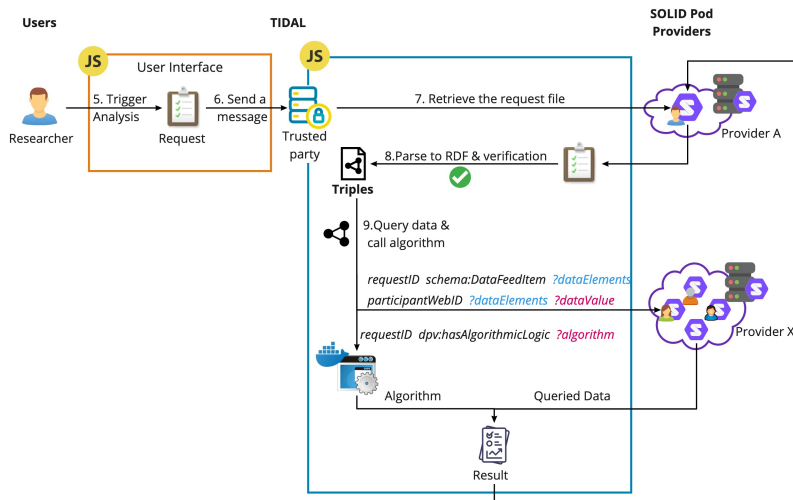


Figure 7.7: The workflow of data retrieval and analysis triggered by the researcher. The analysis execution occurs at the trusted party, and only results are returned to the researcher.

```

@prefix : <http://exampletp.solidprovider.net/inbox/triggermessage#>.
@prefix schema: <https://schema.org/>.
@prefix req: <http://researcher.solidprovider.com/public/request.ttl#>.
@prefix exre: <http://researcher.solidprovider.com/profile/card#>.

:160622932739325095672093710975
  schema:actionStatus schema:ActivateAction;
  schema:creator exre:me;
  schema:dateCreated "2021-04-20T09:37:57.499Z"^^XML:dateTime;
  schema:target req:161964062096710764675982245664.
  
```

Listing 7.3: A generated trigger message (Activate Action) sent by the researcher.

If the integrity of the request is verified, the trusted party fetches the requested data elements from each participant’s pod (adhering to participation constraints such as participation time period) without storing their identifiers (i.e., WebIDs). This fetching process includes querying full RDF files from participants’ pods, parsing them using the `rdflib.js` library, and extracting the requested data elements. When any data are being retrieved from the participants, TIDAL writes logging records in participants pods about what data elements are extracted by whom (WebID) at what time for which data request (request ID), and whether the analysis is executed. The queried data is then fed into the data analysis model which is pre-defined in the Docker image. Finally, the results of the analysis will be generated automatically and sent back to the researcher’s pod. All received and created information at the trusted party such as queried data and intermediate results will be destroyed.

7.5 Experiments and results

7.5.1 Experiment Setting

At the moment of implementing TIDAL (Dec 2020), there were two public SOLID pod providers: *Inrupt.net* and *Solidcommunity.net*. We tested the feasibility and efficiency of TIDAL using two public pods providers and one self-hosted server. Each pod provider hosts 256 SOLID pods, corresponding to 256 participants. Each participant has a data file containing 128 generated variables and values structured by SNOMED CT [45] vocabularies in RDF/turtle format in their SOLID pods. A simplified example of the data file is presented in Listing 7.4.

```
@prefix : <https://exampleparticipant.solidprovider.com/profile/card#>.
@prefix SNOMEDCT: <http://purl.bioontology.org/ontology/SNOMEDCT/>.

:me a SNOMEDCT:116154003; # Patient
    SNOMEDCT:397669002 "27"^^xsd:int; # Age
    SNOMEDCT:50373000 "165"^^xsd:int; # Height
    SNOMEDCT:726527001 "55"^^xsd:int; # Weight
    SNOMEDCT:263495000 SNOMEDCT:248152002; # Gender, Female
    SNOMEDCT:271649006 "110"; # Systolic blood pressure
    SNOMEDCT:271650006 "90"; # Diastolic blood pressure
    SNOMEDCT:405751000 SNOMEDCT:44054006. # Type 2 diabetes
```

Listing 7.4: An example of the RDF data file in a participant's SOLID pod.

Using each pod provider, we conducted a set of experiments using an increasing number of participants and variables. We started with requesting 4 variables from 4 to 256 participants, and ended with requesting 128 variables from 4 to 256 participants. The experiments focused on the steps after the researcher collects enough responses from the participants and triggers the analysis. The execution time has been measured from:

1. querying data request URI;
2. querying signature and verification key of data request;
3. verifying signature to ensure the request has not been modified;
4. (if the verification succeeds) querying the content of data request and WebIDs of participants; and
5. querying RDF data from all participants' pods.

The web interface of TIDAL was developed using the Semantic User Interface Framework (V2.4.2) [46] with responsive and scalable layout. We tested the web interface in the recent versions of Safari, Chrome, and Firefox. Data retrieval and analysis is performed on a 2.3 GHz PC using Dual-Core Intel Core i5 with 16GB RAM and 500GB hard disk running MacOS 10.15.7. To

run the simulation experiment, we created 256 SOLID pods, generated and stored simulation data in each pod, and granted permission to the requests in an automatic way.

7.5.2 Results

Figure 7.8 shows how TIDAL scales for querying and analyzing data from individual pods as we increase the number of variables from 4 to 128 and the number of pods from 4 to 256 hosted by *Inrupt.net*, *Solidcommunity.net*, and the self-hosted server. The pod providers limit the number of requests that can be responded to at one time by the servers. Considering the scalability, we enable TIDAL to access participants' pods in a concurrent way using HTTP requests. TIDAL only queries the required data elements from SOLID pods of 64 participants simultaneously. Once a task gets finished, a new task is scheduled in the execution queue. We ran each experiment 10 times and presented the average time of the 10 experiments to avoid possible network latency fluctuations.

Figure 7.8 shows that the total time costs in querying 4 and 8 pods is approximately 4 to 5 seconds with a negligible increase as the number of variables increases. When we query data from a large number of pods, the time costs in fetching data from participants' pods becomes substantial. It rises linearly when we increase the number of pods using all pod providers. In the case of querying data from 256 pods, a gradual rise in time costs is observed as the number of variables increases. In all experiments, the time costs of the first 4 execution steps are constant and independent of how many variables and pods are required because they query information from a fixed number of pods from researchers or trusted parties.

Figure 7.9 shows the total time cost when querying the number of variables from 4 to 128 and the number of pods from 4 to 256 on three pod providers. From the experiments on all pod providers, the total time cost linearly scales when the number of pods is increased. The more variables are queried from each pod, the steeper the increase in time cost is presented. By contrast, the *Inrupt.net* server has a more stable rate and the least time consumption than the other two pod providers when querying data from more than 64 pods.

Chapter 7. *ciTizen-centric Data pLatform (TIDAL): Using Distributed Personal Data in a Privacy-Preserving Manner for Health Research*

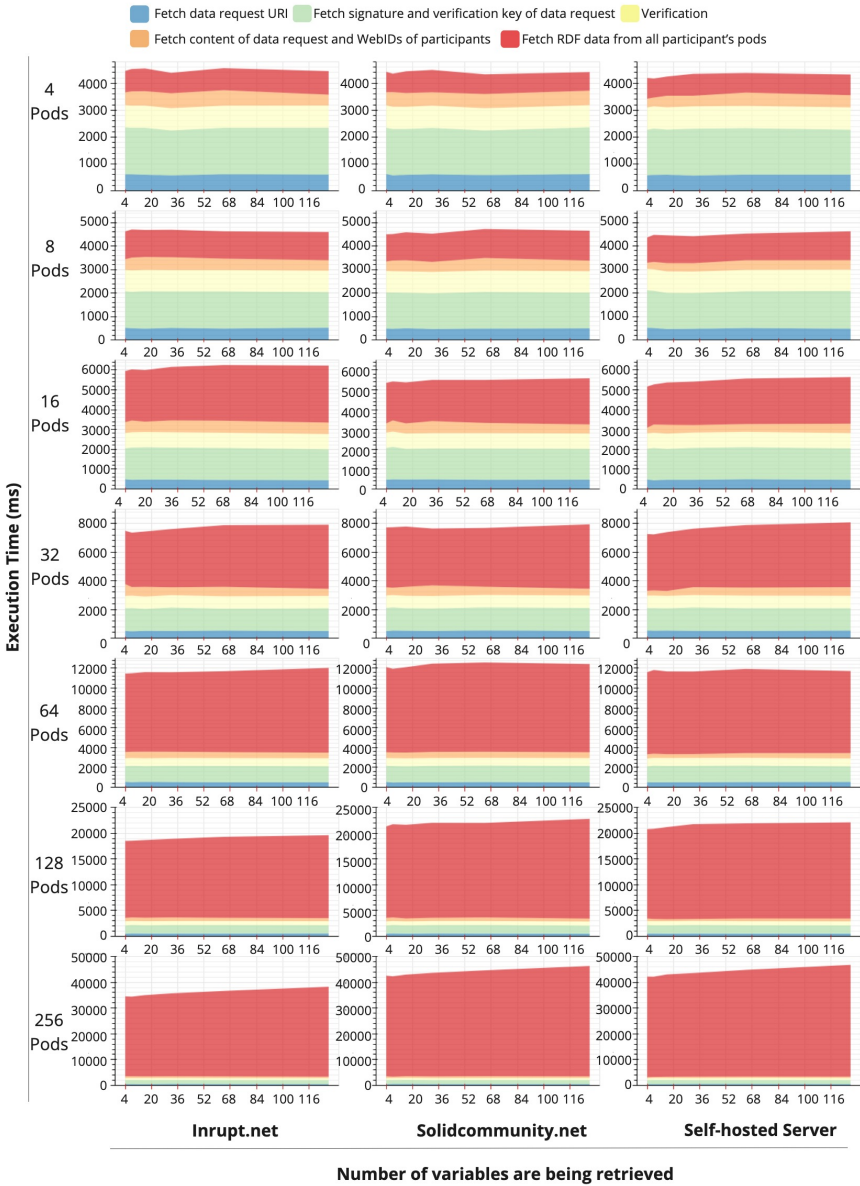


Figure 7.8: Time costs in each execution steps in querying and analyzing data from SOLID pods with increasing the number of variables and pods hosted by Inrupt.net, Solidcommunity.net, and self-hosted server respectively.

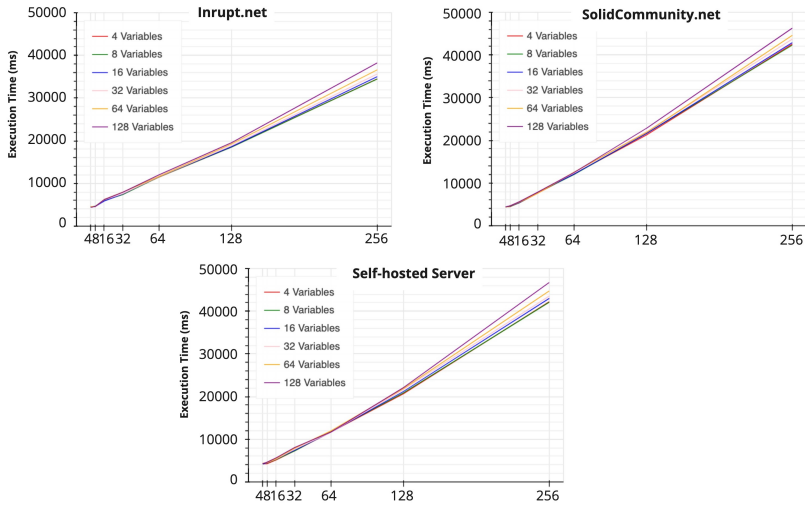


Figure 7.9: Total time costs in querying the increasing number of variables and pods on three pod providers Inrupt.net, Solidcommunitynet, and self-hosted server.

7.6 Discussion

We have demonstrated and tested a ciTizen-centric DatA pLatform (TIDAL) using an increasing number of requested data elements retrieved from an increasing number of SOLID pods. From the performance evaluation of TIDAL, the execution time shows a linear correlation between the number of pods and the number of variables. The process expends the most of the time in querying data from all the participants. However, it only requires an average of 40 seconds to query 128 variables from 256 participants' SOLID pods. For a limited set of participants, this can be considered as an acceptable time for a batch process for use cases which do not demand instant results. In the future, we will improve the workflow and reduce the processing time.

When querying data from enormous pods, the number of variables being queried influences the total querying time (Figure 7.9). A possible solution to improve the query performance would be the provision of SPARQL support on SOLID pods which is missing in the current SOLID specification. A SPARQL endpoint would facilitate the execution of complex queries on pods instead of retrieving full RDF files and post-processing them on the client side to extract the requested data elements, decreasing applications performance. The increase in time of the Fetch RDF data from all participants pods (Figure 7.8) and the execution (Figure 7.9) is also influenced by the limited number

of simultaneous requests being handled by the SOLID server. Additionally, the processing capability of the experimental hardware also created a bottleneck on the querying and analyzing data processes. Therefore, for a practical application we advise the allocation of sufficient computational resources at key architectural locations to reduce the potential bottle-neck when querying and analyzing data from a large number of participants' pods.

TIDAL supports users to store and request personal data in a structured RDF format using well-established ontologies and terminologies by integrating the Biopartial API. These structured data are human and machine readable, provide language neutrality, unambiguous definitions, and clear relationships. It will contribute to enrich and improve the quality of personal data in the SOLID pods by linking data from multiple data sources. Furthermore, to align with the data protection laws such as EU General Data Protection Regulation (GDPR), TIDAL applied the Data Privacy Vocabulary in the participation request to describe and represent information related to requesting and processing of personal data. Data protection laws grant data subjects (participants) rights to withdraw or modify their data anytime they want. On TIDAL, these rights are respected. After the participants approve the data request, they can still update the data elements or withdraw the approval decision anytime. The analysis can be done on the updated value of the data element or without the data elements which have been withdrawn. This process can also enhance reproducibility in research, as researchers can expand and scale their research, both in participants as future long-term effects/follow-up studies.

Furthermore, participants' data are queried and analyzed only at the trusted party. Researchers can only formulate the request, define the parameters of the algorithm, and receive the final analysis results but never have access to the data. In our current approach, the trusted party can be a separate, independent entity in comparison to the researcher, SOLID provider and/or participant. However, if the SOLID provider hosts the trusted party, this trusted party can become a node in a Personal Health Train (PHT) or Federated Learning (FL) infrastructure. In such an infrastructure, the research question travels to the data rather than data being transported to the research question. PHT or FL methodologies connect multiple distributed data sources (e.g., hospitals, clinics) and enable researchers to send analysis models to each data source (e.g. SOLID providers) and get the final learning results. However, in addition to strengthening the binary connection between data sources and researchers, TIDAL emphasizes on engaging individuals in health research and connecting them with both researchers and data sources. This is currently still missing in most PHT/FL implementations.

However, our development has to be seen in light of some potential limita-

tions. First, we assume participants have their personal data structured and stored in their own SOLID pods. In practice, people who do not have enough knowledge about the data or the technologies will face challenges to structure and store their data correctly. To tackle this limitation, one solution can be encouraging data collectors such as hospitals or pharmacies to help participants structure their own data. For example, if the patients' medical records have been structured and linked with some international terminologies by the hospital, then the hospital can request to store the structured medical records data to patients' SOLID pods directly.

Furthermore, the current version of TIDAL presents every published data request that is in the valid period to all participants. In this case, participants receive some data requests that are not relevant to them. As researchers do not know which participants have the relevant data for their research, they are not able to send the data request to the target cohort instead of the general public. Therefore, to improve TIDAL, we are investing in generating privacy-preserving metadata of each SOLID pod. The privacy-preserving metadata is supposed to describe sufficient information about one pod but not reveal any sensitive information. One of the potential solutions is to employ Bloom Filter which is a probabilistic data structure for efficient set membership querying [47]. Bloom filter tests whether the participant has the requested data elements in their data pods and return two possible answers - "probably in the pod" or "definitely not in the pod". With this method, we can prevent the participants in a specific study (e.g. for psychological disorders) from being identified that they are diagnosed with a specific disease or disorder. Another approach is that TIDAL asks participants to indicate their preference on the type of the research and data request. For example, if the participant is only interested in diabetes research, then TIDAL will only present data requests that are related to diabetes research in order to decrease the complexity of using TIDAL for general users.

7.7 Conclusion

In this paper, we presented a novel citizen-centric data platform (called TIDAL) to give individuals fine-grained access to their data and facilitate health research. We demonstrated the feasibility and efficiency of TIDAL by running a set of simulation experiments using different numbers of variables and SOLID pods hosted on three different providers (*Inrupt.net*, *Solidcommunity.net* and a self-hosted server). TIDAL is not only limited to health research, it can be used in other fields such as social sciences (e.g., demographic and anthropology studies), economics and finance studies, political, marketing and education research.

To improve the user experience, we intend to recruit a group of users to assess the human interaction of TIDAL and collect their feedback. In the future, we will evaluate TIDAL in a real-life use case with real participants and health researchers. We will evaluate how usable the request form is for researchers, and how long it will take researchers to complete the entire request form. Meanwhile, we will also investigate how understandable the data request cards are for general participants, and how easy they feel to approve and withdraw the permissions.

The current version of TIDAL allows researchers to only perform a predefined set of analysis models. More complex analysis models will be designed in future work to facilitate researchers to perform experiments according to their scientific questions. Researchers can apply the needed model and tune the parameters instead of coding or modifying the entire model. The risk of hacking or data leak in the analysis process can be minimized. Another future work can be considered is to improve the logging process. The logging files in the current version of TIDAL stores the data access records in participants SOLID pods when the participants grant permission or anyone access to their data. Next, we intend to investigate in applying Blockchain technologies for handling loggings in a more transparent and secure manner. Several studies have developed tools integrating SOLID and Blockchain [48], [49].

Furthermore, the current version of TIDAL only handles static data. In the further development, we consider extending TIDAL to also handle streams of RDF data (RDF triples or graphs with temporal annotations) or real-time data processing [50]. For example, TIDAL users can synchronize their health or fitness data from their wearable devices such as mobile phones or fitness watches to their SOLID pods. These data are first converted to RDF stream data and stored in the users' pods. Then, we consider integrating with RDF Stream processing engines in TIDAL to handle the long-standing query, which is continuously executed, over RDF stream data from the distributed SOLID data pods.

References

- [1] Jie Chen, C Daniel Mullins, Priscilla Novak, and Stephen B Thomas. "Personalized strategies to activate and empower patients in health care and reduce health disparities". In: *Health Education & Behavior* 43.1 (2016), pp. 25–34. DOI: 10.1177/1090198115579415.
- [2] Tim Hulsen. "Sharing Is Caring—Data Sharing Initiatives in Healthcare". In: *International Journal of Environmental Research and Public Health* 17.9 (2020). DOI: 10.3390/ijerph17093046.

-
- [3] European Commission, Directorate-General for Communications Networks, Content, and Technology. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European strategy for data*. 2020. URL: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020DC0066>.
- [4] European Commission. *White paper: on enabling the digital transformation of health and care in the Digital Single Market; empowering citizens and building a healthier society*. Tech. rep. COM(2018) 233 final. European Commission, 2018. URL: <https://ec.europa.eu/digital-single-market/en/news/communication-enabling-digital-transformation-health-and-care-digital-single-market-empowering>.
- [5] European Parliament. *Understanding EU data protection policy*. Tech. rep. European Commission, 2020. URL: [https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651923/EPRS_BRI\(2020\)651923_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651923/EPRS_BRI(2020)651923_EN.pdf).
- [6] European Commission. *Summary report of the public consultation on the European strategy for data*. Tech. rep. COM(2018) 233 final. European Commission, 2020. URL: <https://ec.europa.eu/digital-single-market/en/news/summary-report-public-consultation-european-strategy-data>.
- [7] Dutch Techcentre for Life Sciences (DTL). *Personal Health Train*, <https://www.dtls.nl/fair-data/personal-health-train/>. Access on 12-8-2021.
- [8] Chang Sun, Lianne Ippel, Johan van Soest, Birgit Wouters, Alexander Malic, Onaopepo Adekunle, Bob van den Berg, Ole Mussmann, Annemarie Koster, Carla van der Kallen, et al. "A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario." In: *MEDINFO 2019: Health and Wellbeing E-Networks for All: Proceedings of the 17th World Congress on Medical and Health Informatics*. Vol. 264. IOS Press, 2019, pp. 373–377. DOI: 10.3233/SHTI190246.
- [9] Johan van Soest, Chang Sun, Ole Mussmann, Marco Puts, Bob van den Berg, Alexander Malic, Claudia van Oppen, David Townend, Andre Dekker, and Michel Dumontier. "Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data." In: *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*. Vol. 247. IOS Press, 2018, pp. 581–585. DOI: 10.3233/978-1-61499-852-5-581.

- [10] Arthur Jochems, Timo M Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. “Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital—a real life proof of concept”. In: *Radiotherapy and Oncology* 121.3 (2016), pp. 459–467. DOI: 10.1016/j.radonc.2016.10.002.
- [11] Essam Mansour, Andrei Vlad Sambra, Sandro Hawke, Maged Zereba, Sarven Capadislis, Abdurrahman Ghanem, Ashraf Aboulnaga, and Tim Berners-Lee. “A demonstration of the solid platform for social web applications”. In: *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee. 2016, pp. 223–226. DOI: 10.1145/2872518.2890529.
- [12] Harshvardhan J. Pandit et al. “Creating a Vocabulary for Data Privacy”. In: *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*. Cham: Springer International Publishing, 2019, pp. 714–730. DOI: 10.1007/978-3-030-33246-4_44.
- [13] *Data Privacy Vocabulary (DPV) - version 0.2*. <https://dpvcg.github.io/dpv/>. Accessed on 12-08-2021.
- [14] Tom Symons and Theo Bass. *Me, my data and I: The future of the personal data economy*. Tech. rep. DECODE (DEcentralised Citizen Owned Data Ecosystems), 2017.
- [15] Mirko Koscina, David Manset, Claudia Negri, and Octavio Perez. “Enabling Trust in Healthcare Data Exchange with a Federated Blockchain-Based Architecture”. In: *IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume*. WI ’19 Companion. Thessaloniki, Greece: Association for Computing Machinery, 2019, pp. 231–237. DOI: 10.1145/3358695.3360897.
- [16] *Own Your Data*. <https://www.ownyourdata.eu/>. Accessed on 12-08-2021.
- [17] *MIDATA: My Data - Our Health*. <https://www.midata.coop/en/home/>. Accessed on 12-08-2021.
- [18] *MedMij: Personal health data in the palm of your hand*. <https://www.medmij.nl/en>. Accessed on 12-08-2021.
- [19] *Digi.me*, <https://digi.me/>. Access on 12-08-2021.
- [20] *Cozy Cloud*, <https://cozy.io/>. Access on 12-08-2021.
- [21] *Hub of All Things (HAT)*, <https://www.hubofallthings.com/>. Access on 06-10-2021.
- [22] *MyDex*, <https://mydex.org>. Access on 06-10-2021.

-
- [23] Yves-Alexandre de Montjoye, Erez Shmueli, Samuel S. Wang, and Alex Sandy Pentland. "openPDS: Protecting the Privacy of Metadata through SafeAnswers". In: *PLOS ONE* 9.7 (2014). DOI: 10.1371/journal.pone.0098790.
- [24] Heleen Janssen, Jennifer Cobbe, Chris Norval, and Jatinder Singh. "Decentralised data processing: Personal data stores and the GDPR". In: *International Data Privacy Law* 10.4 (2020), pp. 356–384. DOI: 10.1093/idpl/ipaa016.
- [25] *The European Union's Horizon 2020 research and innovation programme*, <https://ec.europa.eu/programmes/horizon2020/>. Access on 12-08-2021.
- [26] Andrei Vlad Sambra, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, Abdurrahman Ghanem, Dmitri Zagidulin, Ashraf Abounaga, and Tim Berners-Lee. "Solid: A platform for decentralized social applications based on linked data". In: *MIT CSAIL & Qatar Computing Research Institute, Tech. Rep.* (2016). URL: http://emansour.com/research/lusail/solid_protocols.pdf.
- [27] *Solid: Your data, your choice*, <https://solidproject.org/>. Access on 12-08-2021.
- [28] Steve Lohr. "He Created the Web. Now He's Out to Remake the Digital World." In: *The New York Times* (Jan. 10, 2021). URL: <https://www.nytimes.com/2021/01/10/technology/tim-berners-lee-privacy-internet.html> (visited on 08/13/2021).
- [29] Fausto Giunchiglia, Rui Zhang, and Bruno Crispo. *Ontology driven community access control*. Tech. rep. University of Trento, 2008. URL: <http://ceur-ws.org/Vol-447/paper3.pdf>.
- [30] Timo M. Deist et al. "Distributed learning on 20 000+ lung cancer patients – The Personal Health Train". In: *Radiotherapy and Oncology* 144 (2020), pp. 189–200. DOI: 10.1016/j.radonc.2019.11.019.
- [31] Zhenwei Shi, Ivan Zhovannik, Alberto Traverso, Frank JWM Dankers, Timo M Deist, Petros Kalendralis, René Monshouwer, Johan Bussink, Rianne Fijten, Hugo JWL Aerts, et al. "Distributed radiomics as a signature validation study using the Personal Health Train infrastructure". In: *Scientific data* 6.1 (2019), pp. 1–8. DOI: 10.1038/s41597-019-0241-0.
- [32] Oya Beyan et al. "Distributed Analytics on Sensitive Medical Data: The Personal Health Train". In: *Data Intelligence* 2.1-2 (2020), pp. 96–107. DOI: 10.1162/dint_a_00032.
- [33] *Solid access to Pods, local file systems, and other backends via nodejs.*, <https://www.npmjs.com/package/solid-node-client>. Access on 12-08-2021.

- [34] *Javascript RDF library for browsers and Node.js*, <https://github.com/linkedata/rdfliib.js/>. Access on 28-02-2021.
- [35] *Tripledoc - The easiest way to get started writing Solid apps.*, <https://vincenttunru.gitlab.io/tripledoc/>. Access on 28-02-2021.
- [36] Ruben Verborgh and Ruben Taelman. "LDflex: A Read/Write Linked Data Abstraction for Front-End Web Developers". In: *International Semantic Web Conference*. Springer. 2020, pp. 193–211. DOI: 10.1007/978-3-030-62466-8_13.
- [37] Daniel J Bernstein, Niels Duif, Tanja Lange, Peter Schwabe, and Bo-Yin Yang. "High-speed high-security signatures". In: *Journal of cryptographic engineering* 2.2 (2012), pp. 77–89. DOI: 10.1007/s13389-012-0027-1.
- [38] Daniel J Bernstein, Simon Josefsson, Tanja Lange, Peter Schwabe, and Bo-Yin Yang. "EdDSA for more curves". In: *Cryptology ePrint Archive* (2015). URL: <https://eprint.iacr.org/2015/677>.
- [39] Simon Josefsson and Ilari Liusvaara. "Edwards-curve digital signature algorithm (eddsa)". In: *Internet Research Task Force, Crypto Forum Research Group, RFC*. Vol. 8032. 2017, pp. 257–260. URL: <https://www.rfc-editor.org/rfc/pdf/rfc8032.txt.pdf>.
- [40] Jacqueline Brendel, Cas Cremers, Dennis Jackson, and Mang Zhao. "The Provable Security of Ed25519: Theory and Practice". In: *2021 IEEE Symposium on Security and Privacy (SP)*. 2021, pp. 1659–1676. DOI: 10.1109/SP40001.2021.00042.
- [41] *TweetNaCl.js - a port of TweetNaCl / NaCl to JavaScript.*, <https://www.npmjs.com/package/tweetnacl>. Access on 28-02-2021.
- [42] *NaCl: Networking and Cryptography library.*, <http://nacl.cr.yp.to/>. Access on 28-02-2021. 2016.
- [43] Natalya F. Noy et al. "BioPortal: Ontologies and integrated data resources at the click of a mouse". English (US). In: *Nucleic Acids Research* 37.SUPPL. 2 (2009). Funding Information: National Center for Biomedical Ontology, under roadmap-initiative from the National Institutes of Health [grant U54 HG004028]. Funding for open access charge: National Institutes of Health [grant U54 HG004028]., W170–W173. DOI: 10.1093/nar/gkp440.
- [44] *Docker Remote API driver for node.js.*, <https://www.npmjs.com/package/node-docker-api>. Access on 28-02-2021.
- [45] Kevin Donnelly. "SNOMED-CT: The advanced terminology and coding system for eHealth". In: *Studies in health technology and informatics* 121 (2006), p. 279. URL: <https://pubmed.ncbi.nlm.nih.gov/17095826/>.

-
- [46] *Semantic - a UI framework designed for theming.*, <https://semantic-ui.com/>. Access on 20-02-2021. 2013.
- [47] Burton H. Bloom. "Space/Time Trade-Offs in Hash Coding with Allowable Errors". In: *Commun. ACM* 13.7 (July 1970), pp. 422–426. DOI: 10.1145/362686.362692.
- [48] Allan Third and John Domingue. "Decentralised Verification Technologies and the Web". In: *Media, Technology and Education in a Post-Truth Society*. Emerald Publishing Limited, 2021. DOI: 10.1108/978-1-80043-906-120211018.
- [49] Marc Eisenstadt, Manoharan Ramachandran, Niaz Chowdhury, Allan Third, and John Domingue. "COVID-19 antibody test vaccination certification: there's an app for that". In: *IEEE Open Journal of Engineering in Medicine and Biology* 1 (2020), pp. 148–155. DOI: 10.1109/OJEMB.2020.2999214.
- [50] Sherif Sakr, Marcin Wylot, Raghava Mutharaju, Danh Le Phuoc, and Irini Fundulaki. "Processing of RDF Stream Data". In: *Linked Data*. Springer, 2018, pp. 85–108. DOI: 10.1007/978-3-319-73515-3_5.

8

General Discussion

This chapter starts with a review of each chapter in this thesis. Then, we discuss the scientific challenges that are addressed by this thesis when applying theoretical privacy-preserving distributed data mining methods to practical applications. Furthermore, we present the development and potential values of synthetic data followed by a discussion of necessary ethical-legal support in the generation and use of synthetic data in practice. Moreover, we highlight the importance of interdisciplinary collaborations with ethical-legal experts in developing new technologies for privacy and data protection. Finally, we picture a possible future in which citizens take more control over their data access and how their data is processed. We identified what is needed to achieve this future by changing the current centralized data ecosystem to a fully decentralized personal data network.

8.1 Review of chapters

Chapter 2 presents a systematic overview of the development and the existing technologies for analyzing distributed data in a privacy-preserving manner. In this chapter, we analyzed the field, compared existing technologies, identified several remaining issues, and provided recommendations for improving and maturing the field of analyzing distributed data in a privacy-preserving manner. **Chapters 3 to 5** focused on developing a secure infrastructure and methods to combine and analyze distributed data without revealing the source data. Instead of centralizing the data, the infrastructure enables researchers to send data-processing applications to each involved data organization (organization that collects, maintains, or provides data for primary or secondary use). These applications can query data, pre-process data, and execute machine learning or statistical analysis. The data organizations only return the results of the analysis instead of the source data (**Chapter 3**). As a use case, this infrastructure has been installed and tested successfully in collaboration with the Maastricht Study and Statistics Netherlands, to study annual healthcare costs in relation to the incidence of Type 2 Diabetes (**Chapter 4**). Based on the experience and the limitations of the infrastructure, **Chapter 5** proposes the privacy-preserving generalized linear models to solve the challenge of agreeing on a trusted third party by all the organizations in practice and improving the efficiency of the learning process.

While addressing the privacy concerns from the data organization perspective, we observed that the new privacy-preserving distributed learning approaches bring new challenges to researchers, such as building accurate analysis models without using the source data. **Chapter 6** tackles this challenge by proposing a conditional generative adversarial network combined with

differential privacy techniques. The proposed generator can handle the imbalanced issue in the source data and capture the correlations between variables. Its performance is proven to be better than other state-of-the-art generative models using extensive evaluation metrics. Finally, in **Chapter 7**, we developed a citizen-centric data platform for individuals to consent (or withdraw consent) for their data to be shared for health research, and get informed about the research results. The objective of this platform is to shift data access control from data organizations to individuals, and give individuals the means to decide at a granular level how their data is shared and used. Citizens, as being a custodian of their own data, can be connected with researchers, and data organizations to increase the trust placed by citizens in the processing of their personal data. This thesis gradually addresses the challenges of analyzing distributed personal data from the aspects of data organizations, researchers, and citizens.

8.2 Challenges in applying theories to practice

With the widespread implementation of data protection laws such as the General Data Protection Regulation (GDPR) and ePrivacy legislation, the attention on combining and analyzing distributed personal data has been dramatically increased. Our systematic review in Chapter 2 shows the total number of publications in this domain vastly increased in the last decade. In the past five years, the theoretical methods from the research community have been increasingly developed in practical use cases in healthcare [1, 2, 3, 4], finance [5], and technology companies [6, 7, 8]. The interests of both public and private sectors accelerate the transformation of theoretical methods into practical applications. However, there are some challenges hindering the application of privacy-preserving technologies in practice.

8.2.1 Data linkage in vertically partitioned data

In the case of analyzing vertically partitioned data, one point which is important for practice, but easily overlooked by the theory, is how to accurately link data records across distributed datasets without revealing sensitive identifiable information. Data parties must link their data and/or order them in an identical manner prior to data analysis. However, many privacy-preserving methods assume this correspondence between data records exists by default. In fact, this assumption is not valid in the most practical cases. The secure infrastructure we developed in Chapter 3 and Chapter 4 addresses this issue

by using one-way pseudonymization on the selected linking features and securely computing the set intersection across two parties before the data analyses start. The practical application of the infrastructure in Chapter 4 presents a successful data linkage with 97% matching accuracy on a dataset of 1 million data records.

8.2.2 A trusted third party in reality

Our approach, in chapter 3 and chapter 4, like many other privacy-preserving methods, requires a third party or a secure environment (in our case) that all participating organizations can fully trust and agree on. It is challenging to find a fully trusted third party while complying with restrictive legislation and organizational regulations. In our application, we established an ethical-legal framework to support the trusted secure environment and adapted the technical development to comply with the organizational regulations. Implementing such a privacy-preserving data analysis infrastructure requires close interdisciplinary collaboration between people from the technical, ethical-legal backgrounds, and domain-specific experts of the research questions. The idea of our infrastructure which is the researcher sending the algorithm to the data, rather than receiving a copy of the data, shows an entirely new paradigm for most organizations. Stakeholders need the time and effort to carefully evaluate the new infrastructure in terms of their legal and technical requirements. We will elaborate more on the importance of ethical-legal support for the privacy-preserving technologies in Section 8.4.

An alternative solution to this challenge is modifying the analysis model from a mathematical level to eliminate the need of using a third party. In Chapter 5, we investigated a generalized linear model using distributed block coordinate descent to analyze vertically partitioned data without requiring a third party. We restrict the statistical information shared between data organizations as much as possible and measure privacy by proving how the shared information relates to the source data. Our model can be extended to more data organizations and to handle arbitrarily partitioned data (a hybrid of horizontally and vertically partitioned data). Although the generalized linear model resolves the restraints of a third party, it is only applicable when the target features can be shared by all participating organizations, which is not required by the secure infrastructure in chapter 3-4.

8.2.3 The optimal privacy-preserving technology

The studies in Chapter 2 to Chapter 5 show at the current development stage of privacy-preserving technologies, there is no single optimal method that

can fit all data situations in practice, preserve the most sufficient privacy, and reach the best model performance at the same time. In our opinion, to choose the most suitable privacy-preserving technology, we recommend data organizations and researchers collaboratively consider the following items:

1. agree on an overall goal of distributed data analysis such as to answer a specific research question (from a scientific aspect) and the desired level of data protection (from a legal-ethical aspect),
2. investigate the data characteristics such as quality and completeness of data and if participating parties have relevant or linkable data elements, technical feasibility, and the possibilities and challenges from the legal domain of each data organization (from a technical perspective),
3. balance the trade-off between the desired privacy level and the acceptable performance of data analysis (such as if researchers can still answer the research question based on the analysis result with a certain level of data protection),
4. consider complexity and cost of computation and communication and adaptability for future data analysis collaboration to increase the reusability of the privacy-preserving technology.

8.2.4 Privacy measurement and proof

Our experience in applying theoretical privacy-preserving methods to practical applications discovers two highlights that are critical but underestimated in the previous research. First, providing sufficient quantifiable proof of privacy is key for the data organization(s) to decide whether they are allowed to and willing to apply such technology. There are privacy aspects that can be preserved by the technology from multiple perspectives such as law, security (including cryptography, differential privacy), statistics, and information disclosure. For example, in information disclosure, privacy proof has been given by measuring the probability of predicting the source data using the known information [9] or if the participation of a certain person can be detected from the analysis results [10, 11]. Chapter 3, 5, and 6 apply three different ways to prove and guarantee privacy (from legal, information disclosure, and differential privacy angles, respectively) based on different requirements of the project. In our opinion, the optimal measurement or proof method is the one that is accepted and understood by the stakeholders and communities who use or are affected by privacy-preserving technologies. This privacy measurement and proof should align with the legal and organizational requirements and be agreed upon by the ethical-legal experts in the project.

8.2.5 Explainability and transparency

The explainability and transparency of privacy-preserving technology play an important role when applied to the practice. The privacy-preserving technologies can address the current privacy violation resulting from sharing source data at a centralized site. However, applying one single privacy-preserving technology will not solve all the privacy concerns. These new technologies may create new forms of privacy and data protection violations from, for example, the log information, error messages, and intermediate results shared between organizations which require different corresponding solutions to mitigate the risks of privacy violation. Additionally, researchers need to ensure the reliability and integrity of the scientific results from the privacy-preserving technologies since they do not experiment on the source data directly. Therefore, the design and development of the technologies should be well-explained and transparent to the stakeholders. Typically, these technologies consist of technical specifications, mathematical details, and/or cryptography protocols that require background knowledge to understand. Therefore, we encourage researchers and developers in the field to lower the entry barriers for beginners and people with other backgrounds by providing clear documentation and technical specifications, explaining background knowledge, and importantly publishing open-source tools and algorithms for other researchers and developers to implement and validate the technologies independently.

8.3 Generation and use of synthetic data

The most privacy-preserving distributed learning infrastructures such as federated learning [12, 13, 14], Swarm Learning [15] or frameworks using cryptographic technologies share two key concepts 1) keep the source data with the data owner, 2) execute analysis models at each data organization without transferring the source data. In this setting, new outstanding challenges are brought to researchers. The first challenge is to deliver an accurate analysis model without knowing the quality and insights of the source data. Second, when the final analysis results are abnormal or out of the expected range, it is impossible to check the source data to recognize the causes such as abnormal or unusual data points or errors in the dataset.

We tackled these challenges in Chapter 6 by developing a conditional generative adversarial network model to generate synthetic data which contains similar statistical characteristics and correlations between variables to the real data. Our generative model handles imbalanced data with a mixed type of

variables and transfers the correlations between variables from real to synthetic data. We prove the similar performance of machine learning models on synthetic and real data. Researchers can build and improve their analysis models using the synthetic data and execute the model on the real data. The synthetic data can provide researchers an insight into the data before they start the lengthy data request process. However, we highlight that the use of synthetic data is not to fully replace the real data. Although the analysis results on synthetic data can be similar to those on real data, the research experiments should still be conducted on the real data so that the scientific discovery is reliable.

8.3.1 Privacy preservation in synthetic data

In Chapter 6, we evaluated the model performance and privacy risk by measuring the identity and attribute disclosure of the synthetic data. Combining the state-of-the-art differential privacy method, our model guarantees that no individual records in the source data can be distinguished. We prove that privacy is preserved in the definition of differential privacy.

As we discussed previously, it is essential to get ethical-legal support such as a proper legal framework for the privacy-preserving technologies and their applications. From a legal perspective, synthetic data is regarded as anonymous data. As synthetic data is generated from personal data, a legal basis and an anonymous data assessment are required by GDPR to generate and use synthetic personal data. One of the strict interpretations of the anonymous data assessment states when the data controller does not delete the source data, and the data controller hands over part of this dataset (for example after removing or masking identifiable data), the resulting dataset is still personal data. Briefly speaking, synthetic data does not completely sidestep restrictions of legal compliance. Even though no record from the source data can be identified from the synthetic data, the synthetic data could still be considered as personal data.

In addition to the research community, the generation and use of synthetic data have gradually emerged in the industry in the past 5 years. Companies such as Syntho (<https://www.syntho.ai/>), Stalice (<https://www.stalice.ai/>) in Europe, Replica Analytics (<https://replica-analytics.com/>) in Canada, and Mostly AI (<https://mostly.ai/>) in US and many more are claiming they generate realistic and privacy-preserving synthetic data to tackle the data privacy issues. The growing interest from industry accelerates the transformation of theoretical methods to practical applications. However, before the widespread use of synthetic data in practice, some key issues should be carefully considered and addressed to prevent potential harmful consequences of

misuse of synthetic data or re-identification of source data. First, every country or area has its own legislation for data or personal information protection. When the generated synthetic data is claimed to be privacy-compliant, the meaning or definition of privacy is mostly not defined or referred to. Some EU companies state the generated synthetic data is non-identifiable or impossible to re-identify individuals so that the synthetic data is excluded from GDPR. However, the measurement of privacy or relevant evidence of privacy preservation is missing to support the privacy-compliant claims. Additionally, as we discussed before, if the data controller does not delete the source data, the synthetic data could be still seen as personal data. Chapter 6 presents some possible attacks that may reveal the source data from identity disclosure or attribute disclosure with certain probabilities. When the synthetic data is published or disseminated for secondary use, data rights from the data providers may be violated if individuals' data are used to generate synthetic data without sufficient data protection. Therefore, in our opinion, a sound legal framework is required to supervise the generation and use of the synthetic personal data in collaboration with technical experts who have sufficient knowledge of generative models.

8.4 Indispensability of ethical-legal support

In the previous discussion, we have highlighted the impact of legal support in developing new technologies to protect data privacy, applying existing methods to practical applications, and generating synthetic personal data. In Chapter 2, we conclude many existing methods are not able to be implemented in real-life use cases because their technical solutions lack ethical-legal support. From another perspective, the new privacy-preserving technologies pose distinct challenges to the current legislative framework [16]. Therefore, we would advocate for the coordination and collaboration across scientific, technical, and ethical-legal expertise for exploring viable solutions for privacy-preserving data sharing and analysis in practice.

When implementing the infrastructure and application in Chapter 3 and 4, we have encountered several challenges which could only be solved through cooperation between legal and technical experts. For instance, due to the data protection regulations from one of the data organizations in our project, the data (including encrypted data, anonymized data, pseudonymized data, aggregated data, or data in any format) cannot be transferred out of their system without a manual information disclosure check by their experts. Any information that is based on less than a certain number of data subjects is not allowed to be exported. Based on these regulations, many existing privacy-preserving technologies cannot be directly applied, such as secure multiparty

computation or homomorphic encryption. These technologies either require heavy communications or information exchange on data subject level between data organizations, which do not comply with the organizational restrictions. To address this issue, our ethical-legal experts proposed a legal framework and established a joint controller agreement between data organizations to support the trusted secure environment. Only encrypted and pseudonymized data can be executed in the trusted secure environment and no data organizations can intervene in the execution by digital signatures of the executed code.

Furthermore, as we mentioned at the beginning of this section, vertically partitioned data must be linked across multiple organizations before the analysis starts. Unique identifiers, such as national identification numbers or social security numbers are needed to link data. However, the GDPR leaves it up to the national governments to determine the use of the national identification number. The Netherlands has adopted a very restrictive approach regarding the use of the national identification number (Burgerservicenummer - BSN). As the most reliable identifier, BSN (including encrypted, anonymized, pseudonymized, or in many other formats) cannot be used for scientific purposes such as the project in Chapter 3 and Chapter 4. To continue the research and comply with Dutch law, together with our legal experts, we selected a set of personal features (gender, date of birth, zip code, house number) that are allowed to use for linking purposes and pseudonymized the combination of these features. Finally, the data was successfully linked with an accuracy of 97%. The ethical-legal support played a vital role in the success of implementing the privacy-preserving technology using personal data in a practical setting. Meanwhile, the accurate data linkage from our work gives us a necessary reflection on the Dutch legislation on using BSN for scientific research. Unable to use the BSN (only one data element), researchers need multiple reliable personal or demographic data to have a relatively accurate link. These personal or demographic data are often considered to be more privacy-sensitive than the BSN numbers. Using multiple instead of one data element does not comply with one of the GDPRs key principles - data minimization - which requires data collection to be limited to what is necessary for its purpose [17]. Therefore, our work shows the use of the BSN number (in a modified form, e.g. one-way hashing) for scientific research is necessary and can improve the protection and trust from individuals.

8.5 Privacy-preserving methods and (re)consenting

Combining and analyzing personal data through various sources may lead to privacy and data protection breaches. However, the algorithm and analy-

ses themselves are not designed to harm individuals. Instead, individual data rights are violated by how these analyses are being executed during the different phases of data science without a legal basis and without complying with the data protection laws. Under GDPR, consent is one of the most common legal bases to process personal data. Therefore, the overall aim of developing and applying privacy-preserving technologies is to protect individuals privacy and data rights. Using these privacy-preserving technologies does not dismiss researchers from individuals giving (re-)consent. This is important for researchers and developers who develop privacy-preserving technologies to acknowledge and for data organizations to understand.

The GDPR requires that consent must be informed and freely given and can be withdrawn at any time. Giving and withdrawing informed consent for data collection and analysis faces practical challenges such as the way in which consent is obtained is insufficient for the posed research question, the consent is not fully informed (e.g., without background information about the research), or freely given by the individuals. We address the challenge of facilitating individuals to fully exercise their data rights such as giving digital consent to health research by developing a citizen-centric data platform.

TIDAL connects individuals with researchers and provides them a simple way to give or withdraw their consent for donating personal data for health research. In our opinion, the value of data does not lie in its collection and storage, but in the data flow and (re-)use of data. Therefore, TIDAL enables individuals to monitor and control the whole life cycle of their data including the access, storage, and analysis of data. Personal data management tools such as TIDAL are promising starts of building a completely new paradigm of personal data storage and (re-)use. Data stewardship, access control, and the responsibilities of managing data can be shifted from organizations to individuals. In this promising future, the new data paradigm may bring new challenges such as how we educate the general public to manage their own data, and how we ensure the majority of our citizens understand the consequences of giving access?

8.6 Citizen control over their data

We envision in the future citizens will have real-time data access and more control over their own data and make decisions on who can access, store, and process data that are generated and directly based on these individuals. Individuals can easily store or move their data between personal data storage providers or using their self-hosted data storage. Ideally, these data are generated and stored in a human- and machine-readable format and are

structured by well-established data standards. Hence, these personal data are interoperable and can be shared and (re-)used by broader applications.

To shape this future, a joint effort is needed from technology developers, ethical-legal experts, policymakers, data organizations (who are collecting or storing personal data), and the general public. New technologies should be developed and applied to implement data protection regulations and facilitate our citizens to practice their rights in a collaboration between technology developers and ethical-legal experts by, for example, having dialogues to obtain a common understanding across domains and find out what is possible from an ethical-legal and technical perspective. When introducing such a new personal data platform to society, we expect new challenges and discussions to be brought from the general public, data organizations, and other stakeholders of our society. Policymakers and regulators may need to conceptualize the corresponding legislation and regulations in response to these challenges and discussions about the new personal data use paradigm. Furthermore, the general public should be provided with sufficient guidance and education to be able to make their own decisions on the access and use of their personal data and understand the benefits and consequences of those decisions. In our opinion, this is not the only optimal strategy for using personal data in the future, but this thesis pictured a potential solution. It is up to society and/or individuals to decide on the preferable way to manage and use their personal data.

8.7 Future perspectives

From the current development of privacy-preserving federated learning infrastructures that share and analyze personal data across multiple organizations, one of the crucial challenges is data quality and interoperability from different organizations. The case studies from the existing infrastructures are usually in a controlled context such as limited data requirements, basic data models, or simplified legal or organizational restrictions. Data quality and interoperability have not been well-studied as big challenges in a federated learning scenario. In our opinion, data quality and interoperability have a significant impact on the utility of the infrastructure and the performance of the analysis results. We applied FAIR guiding principles in the infrastructure to enhance data interoperability (Chapter 3). Future work is needed to automate the process of making data FAIR, and recognize and select the proper data standards at the source.

Furthermore, the trade-off between privacy preservation and data utility is a key to selecting the most suitable privacy-preserving technique in different use case scenarios (Chapter 5-6). However, finding the optimal balance of this

trade-off in each particular use case depends on the different legal constraints and the purposes or use of the data. Therefore, future work should extend the current studies (Chapter 3, 5, 6) to adjustable approaches that users can customize this trade-off between privacy and data utility based on their scientific goal, legal restrictions, and technical requirements.

Finally, we envision the future of storage, use, and sharing of personal data will be shifted from a centralized system to a decentralized network. More personal data will be generated by citizens' smart objects, such as wearable devices, home appliances, manufacturing robots, and computing facilities close to citizens [18]. These data will be in the control of the citizens. The personal data management tool - TIDAL (Chapter 7) - has demonstrated the possibility of storing, controlling, and analyzing individual data in a decentralized network using SOLID. More work will be done to systematically fill the pods with existing data from its current locations (e.g., data organizations) and store new data from its source (e.g., wearable devices) in compliance with FAIR principles. Ideally, we can make these personal data FAIR directly in citizens' data pods. Not only does it enable citizens to control their own data, but it also increases the availability and utility of personal data for research and keeps the free flow of data across organizations and across borders.

References

- [1] Timo M. Deist et al. "Distributed learning on 20 000+ lung cancer patients – The Personal Health Train". In: *Radiotherapy and Oncology* 144 (2020), pp. 189–200. DOI: 10.1016/j.radonc.2019.11.019.
- [2] Arthur Jochems, Timo M Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. "Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital—a real life proof of concept". In: *Radiotherapy and Oncology* 121.3 (2016), pp. 459–467. DOI: 10.1016/j.radonc.2016.10.002.
- [3] Jin Li, Yu Tian, Yan Zhu, Tianshu Zhou, Jun Li, Kefeng Ding, and Jingsong Li. "A multicenter random forest model for effective prognosis prediction in collaborative clinical research network". In: *Artificial intelligence in medicine* 103 (2020), p. 101814. DOI: 10.1016/j.artmed.2020.101814.

-
- [4] Hiroaki Kikuchi, Chika Hamanaga, Hideo Yasunaga, Hiroki Matsui, Hideki Hashimoto, and Chun-I Fan. "Privacy-preserving multiple linear regression of vertically partitioned real medical datasets". In: vol. 26. Information Processing Society of Japan, 2018, pp. 638–647. DOI: 10.2197/ipsjip.26.638.
- [5] Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. "Federated learning for privacy-preserving AI". In: *Communications of the ACM* 63.12 (2020), pp. 33–36. DOI: 10.1145/3387107.
- [6] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. "Protection against reconstruction and its applications in private federated learning". In: *arXiv preprint arXiv:1812.00984* (2018).
- [7] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. "Federated learning: Strategies for improving communication efficiency". In: *arXiv preprint arXiv:1610.05492* (2016).
- [8] Yaochen Hu, Di Niu, Jianming Yang, and Shengping Zhou. "Fdml: A collaborative machine learning framework for distributed features". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2019, pp. 2232–2240. DOI: 10.1145/3292500.3330765.
- [9] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. "Generating multi-label discrete patient records using generative adversarial networks". In: *Machine learning for healthcare conference*. PMLR. 2017, pp. 286–305. DOI: 10.48550/arXiv.1703.06490.
- [10] Marius Bozga, Radu Iosif, and Yassine Lakhnech. "Flat parametric counter automata". In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2006, pp. 577–588.
- [11] Cynthia Dwork. "Differential privacy: A survey of results". In: *International conference on theory and applications of models of computation*. Springer. 2008, pp. 1–19. DOI: 10.1007/978-3-540-79228-4_1.
- [12] Chang Sun et al. "A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario". In: *Studies in Health Technology and Informatics* 264 (2019), pp. 373–377. DOI: 10.3233/SHTI190246.
- [13] Oya Beyan et al. "Distributed Analytics on Sensitive Medical Data: The Personal Health Train". In: *Data Intelligence* 2.1-2 (2020), pp. 96–107. DOI: 10.1162/dint_a.00032.

- [14] Sinem Sav, Apostolos Pyrgelis, J. Troncoso-Pastoriza, David Froelicher, Jean-Philippe Bossuat, João Sá Sousa, and J. Hubaux. “POSEIDON: Privacy-Preserving Federated Neural Network Learning”. In: *NDSS* (2021). DOI: 10.14722/NDSS.2021.24119.
- [15] Stefanie Warnat-Herresthal et al. “Swarm Learning for decentralized and confidential clinical machine learning”. In: *Nature* 594.7862 (2021), pp. 265–270. DOI: 10.1038/s41586-021-03583-3.
- [16] European Commission. Directorate General for Communications Networks, Content and Technology., CEPS., ICF, and Wavestone. *Study to support an impact assessment of regulatory requirements for Artificial Intelligence in Europe: final report*. LU: Publications Office, 2021. URL: <https://data.europa.eu/doi/10.2759/523404>.
- [17] Birgit Wouters et al. “Putting the GDPR into Practice: Difficulties and Uncertainties Experienced in the Conduct of Big Data Health Research”. In: *European Data Protection Law Review* 7.2 (2021). DOI: 10.21552/edpl/2021/2/9.
- [18] European Commission, Directorate-General for Communications Networks, Content, and Technology. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European strategy for data*. 2020. URL: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020DC0066>.

Summary

An ever-increasing amount of data is generated by our citizens and used in our daily life every single day. These massive amounts of data can be used to improve digital technologies and develop data-driven innovations that can impact every aspect of peoples lives. However, lack of sharing, accessing to and reusing from multiple organizations hinders the analysis possibilities and hence potential insights from the data. A number of challenges have been recognized such as technical barriers, security, data protection compliance to one or more legal jurisdictions, privacy concerns, and trust issues. The overall aim of this thesis is to develop new privacy-preserving data sharing and analysis techniques that strengthen and extend the (re-)use of personal data while maximally protecting individuals privacy. To achieve this aim, this thesis addressed the research challenges on personal data sharing and use from the perspectives of data organizations, the research community, and individuals (data providers).

This thesis first presents a systematic literature review (**Chapter 2**) on privacy-preserving distributed data mining (PPDDM) techniques which considers the issue of executing data mining algorithms on private, sensitive, and/or confidential data from multiple data organizations while maintaining privacy. This chapter draws an overview of existing PPDDM methods to help researchers better understand the development of this domain and assist practitioners to select suitable solutions for their practical cases. We discussed the highlights and remaining challenges in the field including a lack of standard evaluation criteria for new PPDDM techniques, the ambiguous definition of privacy, and the gap between theoretical solutions and practical applications. Finally, we provided a list of recommendations for future research in the field.

Chapter 3 presents an innovative infrastructure, which supports secure and privacy-preserving analysis of personal health data from multiple independent organizations with different governance policies. Instead of centralizing the data, the infrastructure enables researchers to send data-processing applications to each involved data organization. This chapter describes an optimal solution accounting for scientific, technical, and ethical/legal challenges in a practical use case. **Chapter 4** proves the feasibility of the proposed privacy-preserving infrastructure using real-life patient data from The Maastricht Study and Statistics Netherlands to study the association between Type 2 Diabetes and annual healthcare expenses. We handled challenges that have

not been adequately studied by previous works such as data linkage in vertically partitioned data, privacy definition and measurement corresponding to technical and legal requirements, and the indispensability of ethical-legal support in the development of new privacy-preserving technology. Based on the work in Chapter 3-4, **Chapter 5** solves the limitations of using a third party and decreases the costs of communication and computation. The proposed privacy-preserving generalized linear model is based on a distributed block coordinate descent algorithm to obtain parameter estimates, and appended an extension to compute accurate standard errors without additional communication cost. We critically evaluate the information transfer of our model and prove the security and privacy against data reconstruction.

The motivation of **Chapter 6** comes from the experience of requesting data and building up a data analysis model without accessing the source data using the privacy-preserving infrastructure. Chapter 6 presents DP-CGANS, a conditional GAN model combining differential privacy to generate realistic and privacy-preserving synthetic tabular data that is structurally and statistically similar to the real data. DP-CGANS tackles two outstanding challenges in generating synthetic (tabular) data - 1) capturing the correlations and dependencies between variables in an imbalanced dataset, 2) addressing privacy concerns when training DP-CGANS on sensitive private data using a differential privacy technique. We extensively evaluate DP-CGANS compared with three other state-of-the-art generative models. We demonstrate that DP-CGANS outperforms other comparable models and shows the trade-off between data utility and privacy in synthetic data generation.

The focus of **Chapter 7** lies on a citizen-centric data platform (TIDAL) which can give individuals ownership of their own data, and includes mechanisms to provide fine-grained access to external parties. Combined with the previous development, the TIDAL integrates a set of components for requesting subsets of RDF (Resource Description Framework) data stored in personal data vaults based on Social Linked Data (SOLID) technology and analyzing them in a privacy-preserving infrastructure. We demonstrate the feasibility and efficiency of the TIDAL platform by querying and analyzing personal health data from an increasing number of data pods and variables. This chapter shows platforms such as TIDAL play an increasingly important role to connect citizens, researchers, and data organizations to increase the trust placed by citizens in the processing of their personal data.

Chapter 8 describes the scientific challenges addressed by this thesis in applying theoretical privacy-preserving distributed data mining methods to practical applications such as the data linkage across sources, trusted party in reality, privacy measurement, optimal choice of privacy-preserving methods, and explainability and transparency of the methods. This chapter highlights the

generation and use of synthetic data that needs support from a sound legal framework, followed by a discussion on the importance of interdisciplinary collaborations between technical and ethical-legal experts in developing new privacy-preserving technologies. Last but not least, we envision a new personal data paradigm for citizens to take more control over their data access and how their data is processed. We believe the future personal data use and sharing will be in a fully decentralized network. The changes from now to the future require efforts from all the stakeholders such as individuals (data providers), policymakers, researchers and scientists, data organizations, and our society.

Samenvatting

Data wordt continue gegenereerd door de mens, en gebruikt in het dagelijks leven. Deze stroom aan data blijft groeien, en kan worden gebruikt voor het ontwikkelen en verbeteren van data-gedreven digitale technieken en die het dagelijks leven kunnen beïnvloeden. Helaas leidt het niet delen van (of toegang geven tot) persoonlijke data vanuit meerdere organisaties tot een groot obstakel in het analyseren en ontwikkelen van deze digitale technieken. Hierbij zijn een aantal obstakels bekend, zoals technisch, beveiliging, gegevensbescherming binnen verschillende wetgevingen, zorgen rondom privacy, en vertrouwen in verschillende belanghebbenden.

Het doel van deze thesis is om nieuwe privacy-beschermende data-deel en analyse technieken te ontwikkelen, en op deze manier het data delen (of beschikbaar stellen) te bevorderen waarbij de privacy van het individu zo goed mogelijk wordt beschermd. Om dit doel te bereiken worden een aantal onderzoeksvraagstukken geadresseerd aangaande persoonlijke data deling en hergebruik vanuit hieronder benoemde perspectieven:

1. Data verzamelende organisaties: het ontwikkelen van een veilige infrastructuur voor het combineren en analyseren van data uit meerdere bronnen, zonder het openbaren van gevoelige persoonlijke informatie
2. Onderzoekers:
 - Het ontwikkelen en toepassen van privacy-beschermende gedistribueerde data analyse methoden voor verticaal gepartitioneerde data, met en zonder gebruik van een onafhankelijke derde partij
 - Het bouwen van een synthetische data generator voor het simuleren van persoonlijke data, zodat onderzoekers inzichten op simulatie data kunnen ontwikkelen, voordat een data aanvraag procedure nodig is. Hierdoor kan het analyse model gebouwd worden voordat de brondata beschikbaar is, waardoor de doorlooptijd van een project wordt verkort.
3. Voor het individu: het ontwerpen van een nieuwe technologie, waarbij het individu controle heeft en zelf toestemming en toegang kan geven voor het (her-)gebruik van hun eigen data.

In deze thesis wordt allereerst een systematisch literatuur onderzoek (Hoofdstuk 2) uitgevoerd, met als doel het in kaart brengen van privacy-beschermende gedistribueerde data mining technieken (PPDDM)

waarbij een analyse tussen meerdere data-organisaties wordt uitgevoerd, terwijl de privacy van deelnemers zo goed mogelijk wordt gewaarborgd.

Dit hoofdstuk beschrijft een overzicht van bestaande PPDDM methoden om inzicht te krijgen in de huidige status van het onderwerp, en om onderzoekers te helpen een keuze te maken in methoden, passend bij de (onderzoeks-)vraag die wordt gesteld. Verder beschrijft dit hoofdstuk de hoogtepunten en openstaande uitdagingen in het veld, inclusief de afwezigheid van evaluatie criteria voor nieuwe PPDDM technieken. De onduidelijke definitie van privacy, en het gat tussen theoretische oplossingen en praktische toepassingen zijn hierbij de grootste uitdagingen. Dit hoofdstuk eindigt met een lijst van aanbevelingen voor toekomstig onderzoek in dit veld.

Hoofdstuk 3 beschrijft een innovatieve infrastructuur, waarbij veilige en privacy-beschermende maatregelen centraal staan. Hierbij wordt de casus van persoonlijke gezondheidsdata uit meerdere (onafhankelijke) bronnen/organisaties met verschillend beleid besproken. De infrastructuur biedt de mogelijkheid om data bij de bron (de organisatie) te laten staan, en data-analyse algoritmen te versturen naar de betrokken organisaties. In deze oplossing komen wetenschap, techniek en ethisch-juridische uitdagingen in een praktische casus bij elkaar. Hoofdstuk 4 sluit hierbij aan, en laat de daadwerkelijke uitwerking van deze infrastructuur zien, waarbij data van de Maastricht Studie en het Centraal Bureau voor de Statistiek (CBS) wordt geanalyseerd voor het onderzoeken van de relatie tussen Diabetes type 2 en gezondheidszorg uitgaven. Uitdagingen die voorheen niet specifiek zijn geadresseerd, zoals het linken van persoonlijke informatie over meerdere bronnen, de definitie van privacy en tests aangaande de technische en juridische vereisten, worden in dit onderzoek geadresseerd. Gebaseerd op hoofdstukken 3 en 4 beschrijft Hoofdstuk 5 een oplossing waarbij geen onafhankelijke derde partij meer nodig is, en waarbij de benodigde communicatie en rekenkracht wordt verminderd. Het voorgestelde algoritme is gebaseerd op een gedistribueerde block coordinate descent algoritme om schattingen voor parameters te verkrijgen. Hierop is een uitbreiding gemaakt om de standaardfout nauwkeurig te berekenen zonder extra communicatie. Deze methode is kritisch geëvalueerd met betrekking tot de informatie uitwisseling om beveiliging en privacy te waarborgen, en om brondata reconstructie te voorkomen.

Hoofdstuk 6 bouwt voort op de ervaring rondom het aanvragen van data, en de bijbehorende analyses, zonder de daadwerkelijke data zelf te kunnen benaderen. Hierbij wordt DP-CGANS beschreven: een conditioneel generatieve adversarial network (GAN), gecombineerd met differentiele privacy, voor het genereren van realistische en privacy-beschermende synthetische

tabel data. Deze synthetische data is zowel in structuur en statistisch vergelijkbaar met de daadwerkelijke data. Deze methode is getest in vergelijking met drie andere moderne generatieve modellen. Hierbij laten we zien dat het voorgestelde model beter werkt dan de bestaande modellen, en laat de afweging tussen bruikbaarheid van data en privacy in synthetisch data genereren zien.

Hoofdstuk 7 beschrijft een burgergericht data platform (TIDAL) waarbij het individu het eigenaarschap krijgt over zijn/haar eigen data. Hierbij worden ook mechanismen voor fijnmazige toegang door derden beschreven. TIDAL bouwt voort op voorgaande ontwikkelingen rondom het beschrijven van data (subsets) in het Resource Description Framework (RDF) formaat, en het opslaan in het SoOcial LInked Data (SOLID) platform, waarbij de analyse met een privacy-beschermende infrastructuur plaatsvinden. In dit hoofdstuk laten we de haalbaarheid en efficiëntie van het TIDAL platform zien, door middel van het bevragen en analyseren van persoonlijke gezondheidsdata in een toenemend aantal data pods (lees: een persoonsgebonden datakluis) en gegevens per pod.

Hoofdstuk 8 beschrijft de discussie rondom de wetenschappelijke uitdagingen die in deze thesis aan bod zijn gekomen. Van het toepassen van theoretische privacy-beschermende gedistribueerde data analyse methoden, tot aan de praktische uitvoering zoals het linken van informatie over meerdere bronnen, vertrouwde derde partijen in de praktijk, privacy toetsing, optimale keuzes van privacy-beschermende methoden, en de verklaarbaarheid en transparantie van deze methoden. Dit hoofdstuk beschrijft de noodzaak voor het genereren en gebruik van synthetische data. Verder wordt ingegaan op de noodzaak van een duidelijk juridisch framework, gevolgd met een discussie rondom de noodzaak van interdisciplinaire samenwerkingen tussen technisch en ethisch-juridische experts in de ontwikkeling van nieuwe privacy-beschermende technieken. Verder gaan we in op een mogelijk toekomst-scenario, waarbij het individu meer controle kan nemen over data die (over hen wordt) verzameld, wie hier toegang tot heeft, en wie deze mag verwerken.

Impact Paragraph

Digital technologies have advanced rapidly and applied broadly in our society and affect everyone's life. By using digital technologies, our citizens generate a massive amount of personal data every single day. These distributed personal data are collected and used to improve digital technologies and enhance data-driven innovations. The potential values and benefits of sharing and (re-)use of distributed personal data in a responsible manner are significant for our society and the scientific community. However, these data are collected and maintained by different independent organizations. Sharing personal data across multiple organizations faces challenges from technical barriers, security and privacy concerns, legal restrictions, and trust issues. Moreover, citizens, whose data have been collected and used, highly value their data rights and privacy. However, our citizens currently have very limited control over their own data. Technical tools and standards are lacking to facilitate citizens to make their own decision for their data and shift data control from the data organizations to individual data providers.

The overall goal of this thesis is to develop new privacy-preserving data sharing and analysis techniques so as to enable new possibilities for (re-)use of personal data while maximally protecting individual privacy. To achieve this, this thesis makes contributions of interest to three key stakeholders:

1. Data organizations: we developed a secure infrastructure that can combine and analyze personal data from multiple sources without revealing sensitive private information.
2. Scientific community: 1) we developed and applied privacy-preserving distributed data mining methods to analyze vertically partitioned data with and without a third party; 2) built a synthetic data generator to simulate the personal data so that researchers can have an insight into data before the lengthy data request process or build-up analysis model without accessing the source data.
3. Individuals: we designed a novel citizen-controlled technology that enables individuals to access and control their personal data and monitor the (re)use of their data.

1 Scientific Impact

The highlighted scientific contribution of this thesis is creating and experimenting new data paradigms for sharing and using personal data with respect to privacy from the organizational to the individual levels. Among data organizations, we proposed a new infrastructure to transfer the analysis models to vertically partitioned data. It is a scalable and secure solution to analyze personal data across multiple sources. Significantly, it unlocks research questions that could not be answered before due to the restrictions on data access and privacy concerns. Unlike other theoretical methods, our infrastructure has been successfully implemented and tested in practice using a large size of real-life data with the support of an ethical-legal framework. We demonstrated the feasibility of our infrastructure by studying the association between diabetes and annual healthcare costs from a Dutch cohort.

The second new data paradigm presented in this thesis is for researchers to use synthetic data to design accurate analysis algorithms without accessing the source data. Our generative model (DP-CGANS) creates realistic and privacy-preserving synthetic tabular data that are structurally and statistically similar to the source data. DP-CGANS tackles two remaining scientific challenges in generating synthetic (tabular) data - 1) capturing the correlations and dependencies between variables in an imbalanced dataset, 2) addressing privacy concerns when training on sensitive private data using a differential privacy technique. We prove DP-CGANS outperforms other state-of-the-art generative models in extensive experiments.

Another innovation lies in the TIDAL citizen-centric data platform, which makes it easier for individuals to store and access their personal data using personal data vault technologies and provide direct consent to health-related research using SOLID (SOcial LIInked Data) and Personal Health Train architecture. TIDAL integrates vocabulary services and standards to 1) structure digital consents to meet the requirements of GDPR and 2) address a scientific challenge in improving the interoperability of personal data use. We believe TIDAL is a start to shift the control and use of personal data from a centralized system to a decentralized network.

The datasets, experiments, algorithms, and intermediate and final results in this thesis are all uploaded to public data or code repositories with descriptive documentation following FAIR principles (Findable, Accessible, Interoperable, Reusable). The accessible links to these materials are provided in each chapter. The manuscripts in this thesis are or will be published in open-access scientific journals or conference proceedings. The FAIR data, open-source code, and open-access manuscripts ensure the works in this thesis are reproducible for other researchers.

2 Social Impact

Advancing privacy-preserving data sharing and analysis techniques is a key to achieving responsible use of personal data. The privacy-preserving infrastructure that we developed to securely share data between organizations uncovers more potential use of personal data to improve public and social services, deliver timely healthcare treatments, and other potential benefits to society. This infrastructure protects individual data rights and privacy, which may increase confidence and trust from the data providers (e.g., citizens) in data organizations and how their data is being used by and between organizations.

The generation and use of synthetic data uncover the possibility of mining the value of the data even when the data are inaccessible or unavailable. Like the digital twin can accurately reflect a physical object and simulate its life cycle, our synthetic data generator can generate realistic synthetic personal data that can be used to build and test the analysis models as a replacement for real data. We found that the higher the quality of synthetic data we generate, the more data privacy is sacrificed. This may accelerate research projects which suffer from data access issues. However, it opens new challenges and discussions to the public and our society on the proper generation and responsible use of synthetic personal data.

The citizen-centric data platform (TIDAL) gives individual citizens fine-grained access to their personal data and provides digital consent to use their data for health research. Citizens can monitor and control the whole life cycle of their data including the access, storage, and analysis. TIDAL connects citizens, researchers, and data organizations and facilitates citizens to contribute to health research in a simple way that will improve our society. TIDAL shifts data stewardship and access control from organizations to individuals and encourages citizens to take more responsibility for managing their own data. We believe that TIDAL can start a completely new personal data paradigm that can gain more trust placed by citizens and the transparency of the processing of personal data.

Acknowledgments

I would like to express my gratitude to my promoter and mentor, Michel Dumontier. I sincerely appreciate that he offered me this position five years ago and believed in my potential and skills. He is always supportive and patient in guiding me to be a good researcher and scientist. He leads me to explore new knowledge, taking the lead, and mastering my expertise.

I sincerely thank my co-supervisor Johan van Soest who is always with me to address challenges and difficulties during my Ph.D. His support, kindness, and valuable advice immensely helped me in the past years.

I am grateful to Amrapali Zaveri, Claudia van Oppen, Pedro Hernandez, Kody Moodley, Nadine Rouleaux, and Alex Malic who gave me support and a warm feeling when I joined the team. Special thanks go to Amrapali Zaveri and Lianne Ippel who guided and inspired me in my research and life! Many thanks to Dorina Claessens, Remzi Celebi, Lars Jacobs, Vincent Emonet, Linda Rieswijk, Carlos Utrilla Guerrero, Anda Iamnitchi, Christopher Brewster, Visara Urovi, Seun Adekunle, Jinzhou Yang, Thales Costa Bertaglia, Nina Stahl, Hannah Schmitt and many previous and current colleagues from my group and the university. Thank Bob van den Berg, Ole Mussmann, Annemarie Koster, David Townend, Marc Gallofre Ocana, Erik-Jan van Kesteren, and other collaborators in my Ph.D. projects.

Thank the members of the assessment and defense committee for their effort and time in reviewing and assessing my dissertation thesis. I would also like to express my appreciation to the people who inspired and led me to pursue this Ph.D. journey, Piet Daas, Marco Puts, and Sofie de Broe.

Importantly, I would like to thank my parents, grandparents, and all family members for supporting, understanding, and believing in me during my Ph.D. My most sincere thanks go to my grandparents who give me the confidence, faith, and strength to face the difficulties and challenges in my life. Many thanks go to my significant one and his family for their respect and support for my decisions.

I would also like to send my love and thanks to my essential friends - Nan Yang, Bingxi Yuan, Salil Bhat, Nishita Kapoor, Dewi Peerlings, Roy Engelen, Enhui(Jessica) Zhou, Qi Zhang, Yao Zheng, Jiayao Yang, Runting Ma, and many more. It would become a very long list to thank all who make remarkable contributions to my work and life. Nevertheless, I thank you all for making me become the current me who is confident, strong, brave, kind, and ready to face every challenge and opportunity in the future of my life.

List of manuscripts

Published:

Chang Sun, Lianne Ippel, Johan van Soest, Birgit Wouters, Alexander Malic, Onaopepo Adekunle, Bob van den Berg, Ole Mussmann, Annemarie Koster, Carla van der Kallen, Claudia van Oppen, David Townend, Andre Dekker, and Michel Dumontier. "A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario". In: *Studies in Health Technology and Informatics* 264 (2019), pp. 373–377. DOI: 10.3233/SHTI190246

Johan van Soest, Chang Sun, Ole Mussmann, Marco Puts, Bob van den Berg, Alexander Malic, Claudia van Oppen, David Townend, Andre Dekker, and Michel Dumontier. "Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data." In: *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*. Vol. 247. IOS Press, 2018, pp. 581–585. DOI: 10.3233/978-1-61499-852-5-581

Chang Sun, Lianne Ippel, Andre Dekker, Michel Dumontier, and Johan van Soest. "A systematic review on privacy-preserving distributed data mining". English. In: *Data Science* 4.2 (Oct. 2021), pp. 121–150. DOI: 10.3233/DS-210036

Birgit Wouters, David Shaw, Chang Sun, Lianne Ippel, Johan van Soest, Bob van den Berg, Ole Mussmann, Annemarie Koster, Carla van der Kallen, Claudia van Oppen, Andre Dekker, Michel Dumontier, and David Townend. "Putting the GDPR into Practice: Difficulties and Uncertainties Experienced in the Conduct of Big Data Health Research". In: *European Data Protection Law Review* 7.2 (2021). DOI: 10.21552/edpl/2021/2/9

Mara Houbraken, Chang Sun, Evgueni Smirnov, and Kurt Driessens. "Discovering hidden course requirements and student competences from grade data". In: *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*. 2017, pp. 147–152. DOI: 10.1145/3099023.3099034

Ananya Choudhury, Chang Sun, Andre Dekker, Michel Dumontier, and Johan van Soest. "Privacy-Preserving Federated Data Analysis: Data Sharing, Protection, and Bioethics in Healthcare". In: *Machine and Deep Learning in Oncology, Medical Physics and Radiology*. Springer, 2022, pp. 135–172. DOI: 10.1007/978-3-030-83047-2_8

Chang Sun, Vincent Emonet, Johan van Soest, Annemarie Koster, Andre Dekker, and Michel Dumontier. "Transformation and Integration of Heterogeneous Health Data in a Privacy-preserving Distributed Learning Infrastructure". In: *12th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences*. 2019. URL: <http://ceur-ws.org/Vol-2849/paper-21.pdf>

Chang Sun, Vincent Emonet, and Michel Dumontier. "A comprehensive comparison of automated FAIRness Evaluation Tools". In: *13th International Conference on Semantic*

Web Applications and Tools for Health Care and Life Sciences. 2022. URL: <http://ceur-ws.org/Vol-3127/paper-6.pdf>

Chang Sun, Federico Igne, Gianmarco Spinaci, Glenda Amaral, John Domingue, Kabul Kurniawan, and Marc Gallofré Ocana. "Privacy-Protection within the Evolution and Preservation of Knowledge Graphs: The VAD2ER approach towards Medical Citizen Science". In: (2019). DOI: 10.13140/RG.2.2.20536.47360

Chang Sun, Johan van Soest, Annemarie Koster, Simone J.P.M. Eussen, Miranda T. Schram, Coen D.A. Stehouwer, Pieter C. Dagnelie, and Michel Dumontier. "Studying the association of diabetes and healthcare cost on distributed data from the Maastricht Study and Statistics Netherlands using a privacy-preserving federated learning infrastructure". In: *Journal of Biomedical Informatics* (2022). DOI: 10.1016/j.jbi.2022.104194

Submitted/Pre-printed:

Erik-Jan van Kesteren, Chang Sun, Daniel L Oberski, Michel Dumontier, and Lianne Ippel. "Privacy-Preserving Generalized Linear Models using Distributed Block Coordinate Descent". In: *International Journal of Data Science and Analytics* (2019). URL: <https://arxiv.org/abs/1911.03183>

Chang Sun. "Knowledge Graph for Microdata of Statistics Netherlands". In: *arXiv preprint arXiv:2101.07622* (2021)

Peter Eigenschink, Stefan Vamosi, Ralf Vamosi, Chang Sun, Thomas Reutterer, and Klaudius Kalcher. "Deep Generative Models for Synthetic Data". In: *ACM Computing Surveys* (2021)

Chang Sun, Johan van Soest, and Michel Dumontier. "Generating Synthetic Tabular Data using Conditional GANs combining with Differential Privacy". In: *Information Sciences* (2022)

Chang Sun, Marc Gallofré Ocana, Johan van Soest, and Michel Dumontier. "ciTizen-centric DatA pLatform (TIDAL): Using Distributed Personal Data in a Privacy-Preserving Manner for Health Research". In: *Semantic Web* (2022). URL: <http://www.semantic-web-journal.net/content/citizen-centric-data-platform-tidal-sharing-distributed-personal-data-privacy-preserving>

Christian Esposito, Olaf Hartig, Ross Horne, and Chang Sun. "Assessing the Solid Protocol in Relation to Security & Privacy Obligations". In: *SEMANTICS Conference* (2022)

About the author

Chang Sun was born in 1992 in Yanji City, Jilin Province, south-east of China. She grew up in a mixed cultural environment - Chinese, Korean, and Russian. In 2008, she moved to Shanghai for her high school education and pursued her bachelor degree in Electrical Engineering and Intelligence Control at Shanghai Maritime University in 2015.

Then, she studied in Artificial Intelligence Master Program at Department of Knowledge and Engineering at Maastricht University. In 2017, she did her thesis internship at Statistics Netherlands on the topic of Classification of Social Media Accounts. September 2017, she started her Ph.D. at Institute of Data



Science at Maastricht University under the supervision of Michel Dumontier and Johan van Soest. Her research focus covers privacy-preserving distributed data mining technologies, federated learning, personal data vault technologies, decentralized networks, synthetic data generation, knowledge graph, and semantic web technologies. During her PhD, Chang is motivated to collaborate with researchers across different organizations, backgrounds, and domains. She has contributed to multiple national and international programs such as VWData (Responsible Value Creation with Big Data, <https://www.nwo.nl/en/projects/40017605>), ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations, <https://odissei-data.nl/en/>), CMyGuideline (<https://www.cmylife.nl/projecten/cmyguideline>).

Chang is highly motivated to achieve her research goal which is to find ways to balance the emerging importance of responsible data practices with the social and scientific value of research. She attaches great importance to communicating her scientific outcomes not only to the academic community but also to the general public by presenting at conferences and in the public media. She is passionate about supporting females and junior researchers in the Data Science domain and inspiring the young generations to enter the field. She has organized three editions of Women in Data Science (WIDS) Maas-

tricht events since 2020 as a WIDS Ambassador. She is also a project leader of the Data Scientist MINDSETS Podcast which aims to promote diversity and inclusivity in data science. Chang is eager to contribute to building a more diverse, dynamic and gender-balanced environment in the STEM field.

Award and Grants

- Women in Data Science Maastricht and Netherlands Ambassador (2020 - Until now)
- Horizon Europe Grant - Real-world-data Enabled Assessment for heaLth regulatory decision-Making (2022)
- Small projects for NWA (Dutch National Research Agenda) Route Grant - Big Data Route (2021)
- Maastricht University Diversity and Inclusivity Grants for Data Scientist MINDSETS Podcast (2021)
- PhD Representative (Chair) of Faculty of Science and Engineering (2020-2021)
- Best Poster award at ICT.OPEN 2020 Conference (2020)
- Research Mobility Awards from Young European Research (2020)
- Best Student Paper award at 17th World Congress of Medical and Health Informatics (2019)
- Universities 2019-2020 University Fund Limburg (SWOL) Grant (2019)
- International Semantic Web Summer School Scholarship (2019)