MINISTRY OF EDUCATION AND RESEARCH OF THE REPUBLIC OF MOLDOVA

Technical University of Moldova

Faculty of Computers, Informatics, and Microelectronics

Department of Software Engineering and Automation

Approved for defense
Department head:
Ion FIODOROV, phd, associate professor

„___” _________ 2025

ADVANCED METHODS OF DATA ANONYMIZATION IN MEDICAL RESEARCH

Master's project

Student:	_____________	**Siminiuc Sergiu, SI-231M**
Coordinator:	_____________	**Țurcanu Dinu, PhD, Associate Professor**
Consultant:	_____________	**Cojocaru Svetlana, university assistant**

Chisinau, 2025

ABSTRACT

The increasing importance of data-driven healthcare, particularly in rare disease research, necessitates robust methods for data anonymization. This thesis explores advanced anonymization techniques to ensure patient privacy while maintaining the utility of medical data. Rare disease datasets pose unique challenges due to their small sample sizes and the high risk of re-identification. Traditional methods, such as k-anonymity and pseudonymization, are often insufficient in addressing these complexities. This research evaluates three primary anonymization techniques—k-anonymity, differential privacy, and pseudonymization—through a structured methodology applied to synthetic and real-world datasets.

The study begins with an in-depth literature review, identifying gaps in current practices and regulatory challenges, particularly under the European Union's General Data Protection Regulation (GDPR). The methodology includes the preparation of datasets reflecting the complexity of rare disease registries, implementation of anonymization techniques, and evaluation based on metrics such as re-identification risk, data utility, and computational efficiency. Results demonstrate that k-anonymity effectively reduces re-identification risks but at the cost of significant data generalization. Differential privacy provides formal guarantees against re-identification but introduces variability that can impact statistical analyses. Pseudonymization, while useful for longitudinal studies, does not fully address re-identification risks without additional measures.

This thesis highlights the trade-offs between privacy and utility inherent in anonymization techniques and emphasizes the need for a hybrid approach tailored to the specific requirements of rare disease research. Personal contributions include the development of a comprehensive framework for evaluating anonymization methods, integration of ethical considerations into the methodology, and practical recommendations for enhancing data sharing practices. The findings not only advance the understanding of anonymization in medical research but also align with global initiatives to promote ethical data use and privacy-preserving technologies.

REZUMAT

Importanţa tot mai mare a îngrijirii medicale bazate pe date, în special în cercetarea privind bolile rare, impune utilizarea unor metode robuste de anonimizare a datelor. Această teză explorează tehnici avansate de anonimizare pentru a asigura confidenţialitatea pacienţilor, menţinând în acelaşi timp utilitatea datelor medicale. Seturile de date privind bolile rare prezintă provocări unice, datorită dimensiunilor reduse ale eşantioanelor şi riscului ridicat de re-identificare. Metodele tradiţionale, cum ar fi k-anonimitatea şi pseudonimizarea, sunt adesea insuficiente în abordarea acestor complexităţi. Această cercetare evaluează trei tehnici principale de anonimizare—k-anonimitatea, confidenţialitatea diferenţială şi pseudonimizarea—printr-o metodologie structurată aplicată pe seturi de date sintetice şi reale.

Studiul începe cu o revizuire literară detaliată, identificând lacunele din practicile curente şi provocările reglementare, în special în contextul Regulamentului General privind Protecţia Datelor (GDPR) al Uniunii Europene. Metodologia include pregătirea unor seturi de date care reflectă complexitatea registrelor pentru bolile rare, implementarea tehnicilor de anonimizare şi evaluarea acestora pe baza unor metrici precum riscul de re-identificare, utilitatea datelor şi eficienţa computaţională. Rezultatele arată că k-anonimitatea reduce în mod eficient riscurile de re-identificare, însă cu preţul unei generalizări semnificative a datelor. Confidenţialitatea diferenţială oferă garanţii formale împotriva re-identificării, dar introduce variabilitate care poate afecta analizele statistice. Pseudonimizarea, deşi utilă pentru studii longitudinale, nu abordează complet riscurile de re-identificare fără măsuri suplimentare.

Această teză evidenţiază compromisurile dintre confidenţialitate şi utilitate inerente tehnicilor de anonimizare şi subliniază necesitatea unei abordări hibride, adaptate cerinţelor specifice ale cercetării privind bolile rare. Contribuţiile personale includ dezvoltarea unui cadru cuprinzător pentru evaluarea metodelor de anonimizare, integrarea considerentelor etice în metodologie şi recomandări practice pentru îmbunătăţirea practicilor de partajare a datelor. Rezultatele obţinute nu doar că avansează înţelegerea anonimizării în cercetarea medicală, dar se aliniază şi iniţiativelor globale de promovare a utilizării etice a datelor şi a tehnologiilor care protejează confidenţialitatea.

CONTENTS

INTRODUCTION

Ever since joining a Dutch Data Foundation on rare neuromuscular diseases as a Technical Officer, I have always been involved in maintaining, developing, and extending software solutions in the field of IT and medicine. While working with some of the most amazing researchers, professors and scientists I have yet met in my career as an engineer, I became more familiar and aware of the needs & necessities of affected patients, caretakers, guardians & clinicians. Throughout my experience in collecting, managing & processing sensitive data, I have encountered numerous legal & technical challenges in my line of work on the management of patient data. This led me to start studying more about the available tools, techniques, and algorithms to safeguard the data that the Foundation holds and operates with. But the more I got into data management, the more aware I became of the great risks involved in dealing with sensitive data, especially for rare neuromuscular diseases. These data are very valuable for our research, but have many privacy connotations. The data is very rare, as well as very identifiable, as these diseases only affect a very small number of people worldwide. As a result, my focus became critically important in ensuring that these data could be shared and used by researchers without compromising patient confidentiality.

On top of this came the complexity of keeping up with compliance with regulations like the General Data Protection Regulation (GDPR) in Europe. Not only did I have to make sure that the data was effectively anonymized or pseudonymized, I had to find ways to do so without diminishing the analytics use of the data. As a result, we soon realized that traditional anonymization techniques may not always be sufficient for our unique data sets, where, despite anonymization, these data could still be uniquely re-identifiable because of the rarity of the conditions. Realizing this, I began researching more advanced anonymization techniques such as encryption methods and de-identification strategies that would provide the right hand of data privacy and research utility. At the Foundation I experience a perpetual momentum between that innovative madness that shapes our technological future, and the moral imperative to ensure that we use these technologies in responsible ways.

This journey eventually led me to choose the focus of my master's thesis in data anonymization. It occurred to me that further understanding the comparative strengths and pitfalls of different anonymization techniques would help the Foundation, but they would also help the field of medical research as a whole. As healthcare becomes increasingly data-driven, especially with respect to biological medical research and clinical trials, the ability to securely access, share, and manage sensitive patient information is now more needed than ever. I hope to add to this ongoing conversation, bringing together its technical and ethical aspects regarding anonymization in medical research.

In modern medical research, data are fundamentally indispensable, fueling innovation in the diagnosis, treatment, and understanding of diseases. The role is even more critical in the field of rare diseases. Gathering robust data sets is a unique challenge in that small populations are affected by these conditions.

BIBLIOGRAPHY

[1] A. L. Solebo, P. Hysi, L. A. Horvat-Gitsels, and J. S. Rahi, “Data saves lives: Optimising routinely collected clinical data for rare disease research,” *Orphanet Journal of Rare Diseases*, vol. 18, no. 1, p. 285, Sep. 2023, ISSN: 1750-1172. DOI: 10.1186/s13023-023-02912-1. [Online]. Available: https://doi.org/10.1186/s13023-023-02912-1 (visited on 10/06/2024).

[2] S. Courbier, R. Dimond, and V. Bros-Facer, “Share and protect our health data: An evidence based approach to rare disease patients’ perspectives on data sharing and data protection - quantitative survey and recommendations,” *Orphanet Journal of Rare Diseases*, vol. 14, no. 1, p. 175, Jul. 2019, ISSN: 1750-1172. DOI: 10.1186/s13023-019-1123-4. [Online]. Available: https://doi.org/10.1186/s13023-019-1123-4 (visited on 10/06/2024).

[3] A. Thorogood, “International Data Sharing and Rare Disease: The Importance of Ethics and Patient Involvement,” en, in *Rare Diseases*, IntechOpen, Mar. 2020, ISBN: 978-1-83880-024-6. DOI: 10.5772/intechopen.91237. [Online]. Available: https://www.intechopen.com/chapters/71101 (visited on 10/06/2024).

[4] *The Importance of RWD in Rare Disease Research*. [Online]. Available: https://www.appliedclinicaltri com/view/importance-rwd-rare-disease-research (visited on 10/06/2024).

[5] R. Raycheva *et al.*, “Challenges in mapping European rare disease databases, relevant for ML-based screening technologies in terms of organizational, FAIR and legal principles: Scoping review,” English, *Frontiers in Public Health*, vol. 11, Sep. 2023, Publisher: Frontiers, ISSN: 2296-2565. DOI: 10.3389/fpubh.2023.1214766. [Online]. Available: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2023.1214766/full (visited on 10/06/2024).

[6] *Anonymizing Rare Disease Data - Real Life Sciences*, en-US, Oct. 2021. [Online]. Available: https://rlsciences.com/protect-resource/anonymizing-rare-disease-data/ (visited on 10/06/2024).

[7] M. Mesarčík and O. Hamuľák, “General Data Protection Regulation: Current Challenges and Future Directions,” en, in *E-Governance in the European Union: Strategies, Tools, and Implementation*, D. Ramiro Troitiño, Ed., Cham: Springer Nature Switzerland, 2024, pp. 117–133, ISBN: 978-3-031-56045-3. DOI: 10.1007/978-3-031-56045-3_9. [Online]. Available: https://doi.org/10.1007/978-3-031-56045-3_9 (visited on 10/06/2024).

[8] *Uncovering and overcoming common data sharing challenges in the Rare Disease landscape*, en-GB. [Online]. Available: https://www.ga4gh.org/news_item/uncovering-and-overcoming-common-data-sharing-challenges-in-the-rare-disease-landscape/ (visited on 10/06/2024).

[9] Z. He, “From Privacy-Enhancing to Health Data Utilisation: The Traces of Anonymisation and Pseudonymisation in EU Data Protection Law,” en, *Digital Society*, vol. 2, no. 2, p. 17, May 2023, ISSN: 2731-4669. DOI: 10.1007/s44206-023-00043-5. [Online]. Available: https://doi.org/10.1007/s44206-023-00043-5 (visited on 10/06/2024).

[10] E. Shamsinejad, T. Banirostam, M. M. Pedram, and A. M. Rahmani, “A Review of Anonymization Algorithms and Methods in Big Data,” en, *Annals of Data Science*, Jul. 2024, ISSN: 2198-5812. DOI: 10.1007/s40745-024-00557-w. [Online]. Available: https://doi.org/10.1007/s40745-024-00557-w (visited on 10/06/2024).

[11] R. Aufschläger *et al.*, “Anonymization Procedures for Tabular Data: An Explanatory Technical and Legal Synthesis,” en, *Information*, vol. 14, no. 9, p. 487, Sep. 2023, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2078-2489. DOI: 10.3390/info14090487. [Online]. Available: https://www.mdpi.com/2078-2489/14/9/487 (visited on 10/06/2024).

[12] S. Monteiro *et al.*, “Data Anonymization: Techniques and Models,” en, in *Marketing and Smart Technologies*, J. L. Reis, M. Del Rio Araujo, L. P. Reis, and J. P. M. dos Santos, Eds., Singapore: Springer Nature, 2024, pp. 73–84, ISBN: 978-981-9903-33-7. DOI: 10.1007/978-981-99-0333-7_6.

[13] A. Sepas, A. H. Bangash, O. Alraoui, K. El Emam, and A. El-Hussuna, “Algorithms to anonymize structured medical and healthcare data: A systematic review,” English, *Frontiers in Bioinformatics*, vol. 2, Dec. 2022, Publisher: Frontiers, ISSN: 2673-7647. DOI: 10.3389/fbinf.2022.984807. [Online]. Available: https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.984807/full (visited on 10/06/2024).

[14] Z. Weingarden, *Data Anonymization Techniques in a Clinical Trial*, en, Sep. 2021. [Online]. Available: https://www.trialassure.com/resources/blog/data-anonymization-techniques-in-a-clinical-trial/ (visited on 10/06/2024).

[15] B. Bucci, *What’s the Difference between Anonymization and Redaction of Clinical Trial Data?* en-US, Oct. 2022. [Online]. Available: https://www.certara.com/blog/anonymization-redaction-clinical-trial-data/ (visited on 10/06/2024).

[16] I. Ortega-Fernandez, S. E. K. Martinez, and L. A. Orellana, “Large Scale Data Anonymisation for GDPR Compliance,” en, in *Big Data and Artificial Intelligence in Digital Finance: Increasing Personalization and Trust in Digital Finance using Big Data and AI*, J. Soldatos and D. Kyriazis, Eds., Cham: Springer International Publishing, 2022, pp. 325–335, ISBN: 978-3-030-94590-9. DOI: 10.1007/978-3-030-94590-9_19. [Online]. Available: https://doi.org/10.1007/978-3-030-94590-9_19 (visited on 10/06/2024).

[17] S. C. Advisors, *Anonymization and GDPR compliance; an overview*, en-US, Jul. 2020. [Online]. Available: https://www.gdprsummary.com/anonymization-and-gdpr/ (visited on 10/06/2024).

[18] E. M. Weitzenboeck, P. Lison, M. Cyndecka, and M. Langford, “The GDPR and unstructured data: Is anonymization possible?” *International Data Privacy Law*, vol. 12, no. 3, pp. 184–206, Aug. 2022, ISSN: 2044-3994. DOI: 10.1093/idpl/ipac008. [Online]. Available: https://doi.org/10.1093/idpl/ipac008 (visited on 10/06/2024).

[19] S.-C. Li, Y.-W. Chen, and Y. Huang, “Examining Compliance with Personal Data Protection Regulations in Interorganizational Data Analysis,” en, *Sustainability*, vol. 13, no. 20, p. 11 459, Jan. 2021, Number: 20 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2071-1050. DOI: 10.3390/su132011459. [Online]. Available: https://www.mdpi.com/2071-1050/13/20/11459 (visited on 10/06/2024).

[20] S.-N. Vulpe, R. Rughiniş, D. Ţurcanu, and D. Rosner, “AI and cybersecurity: A risk society perspective,” *Frontiers in Computer Science*, vol. 6, p. 1 462 250, Oct. 2024, ISSN: 2624-9898. DOI: 10.3389/fcomp.2024.1462250. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2024.1462250/full (visited on 01/10/2025).

[21] J. Mehtälä *et al.*, “Utilization of anonymization techniques to create an external control arm for clinical trial data,” *BMC Medical Research Methodology*, vol. 23, no. 1, p. 258, Nov. 2023, ISSN: 1471-2288. DOI: 10.1186/s12874-023-02082-5. [Online]. Available: https://doi.org/10.1186/s12874-023-02082-5 (visited on 10/06/2024).

[22] I. C. Hageman, I. A. van Rooij, I. de Blaauw, M. Trajanovska, and S. K. King, “A systematic overview of rare disease patient registries: Challenges in design, quality management, and maintenance,” *Orphanet Journal of Rare Diseases*, vol. 18, no. 1, p. 106, May 2023, ISSN: 1750-1172. DOI: 10.1186/s13023-023-02719-0. [Online]. Available: https://doi.org/10.1186/s13023-023-02719-0 (visited on 10/06/2024).

[23] *The “What” and “Why” of Health Data Anonymization and how Pharmaceutical Sponsors and Contract Research Organizations need to prepare - Real Life Sciences*, en-US, Section: Uncategorized, Sep. 2021. [Online]. Available: https://rlsciences.com/the-what-and-why-of-health-data-anonymization-and-how-pharmaceutical-sponsors-and-contract-research-organizations-need-to-prepare/ (visited on 10/06/2024).

[24] *Protecting Privacy in Large Datasets—First We Assess the Risk; Then We Fuzzy the Data | Cancer Epidemiology, Biomarkers & Prevention | American Association for Cancer Research.* [Online]. Available: https://aacrjournals.org/cebp/article/26/8/1219/283054/Protecting-Privacy-in-Large-Datasets-First-We (visited on 10/06/2024).

[25] E. Bran, R. Rughiniş, D. Ţurcanu, and G. Nadoleanu, “Technical Innovations and Social Implications: Mapping Global Research Focus in AI, Blockchain, Cybersecurity, and Privacy,” en, *Computers*, vol. 13, no. 10, p. 254, Oct. 2024, ISSN: 2073-431X. DOI: 10.3390/computers13100254. [Online]. Available: https://www.mdpi.com/2073-431X/13/10/254 (visited on 01/10/2025).

[26] C. Contasel, A.-V. Pălăcean, D. Ţurcanu, and V.-V. Stoica, “Increasing e-Health systems security and availability by using noSQL databases,” in *2024 23rd RoEduNet Conference: Networking in Education and Research (RoEduNet)*, Bucharest, Romania: IEEE, Sep. 2024, pp. 1–6, ISBN: 9798331540388. DOI: 10.1109/RoEduNet64292.2024.10722437. [Online]. Available: https://ieeexplore.ieee.org/document/10722437/ (visited on 01/10/2025).

[27] E. Bran, R. Rughiniş, D. Ţurcanu, and A. R. Stăiculescu, “Decoding National Innovation Capacities: A Comparative Analysis of Publication Patterns in Cybersecurity, Privacy, and Blockchain,” en, *Applied Sciences*, vol. 14, no. 16, p. 7086, Aug. 2024, ISSN: 2076-3417. DOI: 10.3390/app14167086. [Online]. Available: https://www.mdpi.com/2076-3417/14/16/7086 (visited on 01/10/2025).