FACULDADE DE MEDICINA DA UNIVERSIDADE DO PORTO

Knowledge Discovery in Healthcare: Exploring the role of real-world data to leverage clinical practice

João Filipe Coutinho de Almeida

Programa Doutoral em Ciência de Dados de Saúde

Supervisor: Pedro Pereira Rodrigues

Second Supervisor: Ricardo Correia

October 17, 2024

Cover page

ii

Integrity Declaration

I declare on my honour that what is written in this work has been written exclusively by me and that, excluding quotations, no part has been copied from scientific publications, Internet (any type of programs, set of tools and others included) or research works - or, more generally, any other source - already presented in the academic field (but not only) by me, other students or third parties.

Reproducibility

The code for all the experiments conducted in this thesis is available online on GitHub at the following link: https://github.com/joofio/heads-thesis. From there, you can access the list of all repositories involved in this thesis. Data is also available where possible; however, most of the data used was directly retrieved from Electronic Health Records and is restricted from sharing with third parties by ethical committees.

To my parents, for always supporting me. To Ana Lia, for your unlimited patience.

List of Publications

Core Research Papers

The 8 papers described below are the core structure of this thesis (6 were already published and 2 are under review). The manuscripts are listed by order of appearance in the thesis.

- Coutinho-Almeida, J., Rodrigues, P., & Cruz-Correia, R. (2021). GANs for Tabular Healthcare Data Generation: A Review on Utility and Privacy. In Discovery Science (pp. 282-291). Springer International Publishing.

- Coutinho-Almeida, J., Cruz-Correia, R., & Rodrigues, P. (2022). Dataset Comparison Tool: Utility and Privacy. Stud Health Technol Inform, 294, 23-27.

- (in review) Using Machine Learning Models' feature importance to assess dataset similarity

- Coutinho-Almeida, J., Saez, C., Correia, R., & Rodrigues, P. P. (2024). Development and initial validation of a data quality evaluation tool in obstetrics real-world data through HL7-FHIR interoperable Bayesian networks and expert rules. JAMIA Open, 7(3), ooae062. https://doi.org/10.1093/jamiaopen/ooae062

- Coutinho-Almeida, J., Cruz-Correia, R. J., & Rodrigues, P. P. (2024). Evaluating distributed-learning on real-world obstetrics data: Comparing distributed, centralized and local models. Scientific Reports, 14(1), 11128. https://doi.org/10.1038/s41598-024-61371-1

- (in review) Benchmarking institutions' health outcomes with clustering methods

- Coutinho-Almeida, J., Silva, A. S., Redondo, P., Rodrigues, P. P., & Ferreira, A. (2024). CDK4/6 inhibitors and endocrine therapy in the treatment of metastatic breast cancer: A real-world and propensity score-adjusted comparison. Cancer Treatment and Research Communications, 40, 100818. https://doi.org/10.1016/j.ctarc.2024.100818

- Coutinho-Almeida, J., Cardoso, A., Cruz-Correia, R., & Pereira-Rodrigues, P. (2024). Fast Healthcare Interoperability Resources-Based Support System for Predicting Delivery Type: Model Development and Evaluation Study. JMIR Formative Research, 8, e54109. https://doi.org/10.2196/54109

Other Publications and activities

In addition, during the duration of this thesis conduction, the candidate was also the author and co-author of other papers. Although these studies were not part of the thesis core structure, they were important to improve the researcher's knowledge of the field and/or to present the results to the community. They are listed below:

- Coutinho-Almeida, J., & Cruz-Correia, R. (2022). Developing a Process Mining Tool Based on HL7. Procedia Computer Science, 196, 501-508.

- Holmgren, A., Esdar, M., Hüsers, J., & Coutinho-Almeida, J. (2023). Health Information Exchange: Understanding the Policy Landscape and Future of Data Interoperability. Yearbook of Medical Informatics, s-0043-1768719.

- Costa, P., Almeida, J., Araujo, S., Alves, P., Cruz-Correia, R., Saranto, K., & Mantas, J. (2023). Biomedical and Health Informatics Teaching in Portugal: Current Status. Heliyon, 9(3).

- Gazzarata, R., Almeida, J., Lindsköld, L., Cangioli, G., Gaeta, E., Fico, G., & Chronaki, C. E. (2024). HL7 Fast Healthcare Interoperability Resources (HL7 FHIR) in digital healthcare ecosystems for chronic disease management: Scoping review. International journal of medical informatics, 189, 105507. https://doi.org/10.1016/j.ijmedinf.2024.105507

Abstract

This thesis delves into the intricate process of extracting knowledge from healthcare data, a task fraught with challenges yet brimming with potential. Central to this investigation is the acknowledgment, inspired by Richard P. Feynman, that absolute certainty is elusive in scientific inquiry; instead, this journey is marked by continual learning and improvement. We confront various obstacles, including data accessibility, quality concerns, and the integration of real-world evidence into clinical practice. The main issues affecting this process can be explained as follows:

- Van der Lei's First Law of Medical Informatics: Data shall be used only for the purpose for which they were collected.

- Effectiveness of Routine Data Sources and Analytical Innovations: Examining the extent to which routine data sources and innovations in analytical methods alleviate the need for randomized clinical trials.

- Governance, Privacy, and Trust Issues: Addressing governance, privacy, and trust questions when routine health data are made available for research.

A portion of this work is dedicated to addressing data quality. The quality of healthcare data emerges as a complex and elusive concept, demanding extensive data preprocessing to manage missing values, outliers, and inconsistencies across different health information systems. The thesis emphasizes the criticality of clear functional and clinical data descriptions, advocating for comprehensive data dictionaries and governance tools to facilitate effective data utilization. On the one hand, we assessed how machine learning can help support the quality of health data. On the other hand, we explored synthetic data as a potential method for creating more data, while also offering a secure and legal avenue for data analysis, as well as algorithm development and testing.

We also explored the requirements of ethics committees and Data Protection Officers , which, while designed to safeguard patient privacy, often impede timely data access or even prevent access altogether. The thesis evaluates the application of distributed data analysis, allowing for secure, location-based data analysis, thereby enhancing the timeliness and security of the process. We assess the potential of distributed machine learning models to support decision-making and compare different institutions.

Furthermore, this work investigates how the seamless integration of real-world evidence into clinical practice can drive innovation and improve patient outcomes. We explore the hurdles and potential of AI-based clinical decision support systems and how observational data can be used to create knowledge and trust in real-world settings. Emphasizing the need for a trust framework that ensures transparency and explainability in evidence production, we argue that this is crucial for building clinician and patient trust in the data and decision-making processes.

With this work, we underscore the necessity of a collaborative approach with clinicians, who are the end-users of the developed tools. Understanding their needs and workflows is paramount,

requiring user-friendly tools that clinicians can seamlessly integrate into their practice without needing extensive data science training. Additionally, we realized that information preprocessing is vital, as is the fundamental cataloguing of existing data in institutions for rigorous data analysis. We also understood that distributed methods can facilitate access to information and provide greater protection than traditional methods of data usage and analysis.

In conclusion, this thesis contributes to the field of healthcare data science by highlighting the multifaceted challenges and proposing innovative approaches for effective knowledge extraction from healthcare data. It underscores the importance of cross-disciplinary collaboration, robust data infrastructures, and a balanced legal and technical framework to harness the full potential of healthcare data, ultimately driving innovation and improving patient outcomes.

Resumo

Esta tese debruça-se sobre o intrincado processo de extração de conhecimento a partir de dados de saúde, uma tarefa repleta de desafios mas também cheia de potencial. Central para esta investigação é a frase, inspirada por Richard P. Feynman, de que a certeza absoluta é ilusória na investigação científica; ao invés disso, esta jornada é marcada por aprendizagem e melhoria contínuas. Há vários obstáculos a ultrapassar, incluindo acessibilidade dos dados, assim como a sua qualidade e a integração de evidência do mundo real na prática clínica. Os principais problemas que afetam este processo podem ser explicados da seguinte forma:

- Primeira Lei de Informática Médica de Van der Lei: Os dados devem ser usados apenas para os fins para os quais foram recolhidos.

- Efetividade das Fontes de Dados de saúde primárias e inovações analíticas: Explorar até que ponto as fontes de dados primárias e as inovações em métodos estatísticos reduzem a necessidade de realizar ensaios clínicos randomizados.

- Questões de Governança, Privacidade e Confiança: Abordar questões de governança, privacidade e confiança quando os dados de saúde primários são disponibilizados para investigação.

Uma parte deste trabalho é dedicada às várias dimensões que compreendem o conceito de qualidade dos dados. Por um lado, avaliámos como a aprendizagem automática pode ajudar a melhorar a representatividade da realidade dos dados de saúde. Por outro lado, explorámos os dados sintéticos como um método potencial para criar maior volume de dados, assim como oferecendo uma via mais segura e legal para análise de dados, desenvolvimento e teste de algoritmos ou software. Sabemos que a qualidade dos dados de saúde é um conceito complexo e difícil de definir, exigindo um extenso pré-processamento para gerir valores ausentes, *outliers* e inconsistências entre diferentes sistemas de informação de saúde. A tese enfatiza a importância de descrições claras de dados funcionais e clínicos, defendendo dicionários de dados abrangentes e ferramentas de governança para facilitar a utilização eficaz dos dados.

Também explorámos os requisitos dos comités de ética e dos Encarregados de Proteção de Dados, que, embora concebidos para proteger a privacidade dos pacientes, muitas vezes dificultam ou impedem o acesso atempado aos dados. A tese avalia a aplicação da análise distribuída de dados, permitindo uma análise de dados segura e circunscrita ao local onde os dados estão guardados, melhorando assim a celeridade e segurança do processo. Avaliámos o potencial dos modelos distribuídos de aprendizagem automática para apoiar a tomada de decisões e comparar diferentes instituições sem nunca retirar os dados do seu local de origem.

Além disso, este trabalho investiga como a integração da evidência do mundo real na prática clínica pode impulsionar a inovação e melhorar os *outcomes* clínicos. Explorámos os obstáculos e o potencial dos sistemas de apoio à decisão clínica baseados em Inteligência Artificial e como os dados observacionais podem ser utilizados para criar conhecimento com confiança em ambiente

real. Enfatizando a necessidade de uma estrutura robusta e resiliente que garanta transparência e explicabilidade na produção de evidência, argumentamos que isso é crucial para fomentar e manter a confiança dos profissionais de saúde e doentes nos dados e nos processos de tomada de decisão.

Com este trabalho, sublinhamos a necessidade de uma abordagem colaborativa com os clínicos, que são os utilizadores finais das ferramentas desenvolvidas. Compreender as suas necessidades e fluxos de trabalho é primordial, exigindo ferramentas fáceis de usar que os clínicos possam integrar perfeitamente na sua prática sem precisar de formação extensa em ciência de dados. Adicionalmente, percebemos que o processamento da informação é vital, assim como é fundamental a catalogação dos dados existentes nas instituições para uma análise rigorosa dos dados. Percebemos também que métodos distribuídos que podem facilitar e agilizar os acessos à informação e conseguem garantir maior proteção que os métodos tradicionais de uso e análise de dados.

Em conclusão, esta tese contribui para o campo da ciência de dados em saúde ao destacar os desafios multifacetados e propor abordagens inovadoras para a extração eficaz de conhecimento a partir de dados de saúde. Ela salienta a importância da colaboração interdisciplinar, infraestruturas robustas de dados e um enquadramento legal e técnico equilibrado, assim como o uso dos métodos analíticos e estatísticos mais adequados para aproveitar todo o potencial dos dados de saúde, impulsionando a inovação e melhorando os resultados dos pacientes.

Palavras-chave: Dados de mundo real, Dados sintéticos, Aprendizagem distribuida, Qualidade de dados, Aprendizagem automática, Inteligência Artificial

Acknowledgements

When I began envisioning and preparing for this journey, even before my official enrollment, I could not have imagined what it would be like. João from 2019 eagerly anticipated the future, picturing beauty and excitement in delving deep into a field I had long worked in and dearly loved. *Oh, sweet Summer child. . .*

The path from undergraduate to now has been difficult, perilous, and an emotional roller-coaster. No one can progress from naive enthusiasm to a completed dissertation without significant support, both technical and emotional. I was fortunate to be supported by a group of people with diverse perspectives, tastes, and wisdom, who kept me afloat when things looked grim and the hurdles seemed too high. That said, I would like to acknowledge Professor Pedro and Ricardo. Thank you for letting me do my thing, even when I was wrong. Having the liberty to explore my ideas and motivations are the best takeaway from this journey. Thank you!

To the Virtual Care team for always answering my questions and often going out of their way to help me. Feeling so well-supported was special and definitely allowed me to focus more on my research. My special thanks to Tiago, Eliana, and Pedro.

To my willing colleagues that were always happy to discuss my research with me and brainstorm. My special thank you to Paulo Costa, Daniela Ferreira, Rafael Vieira, Inês Ribeiro-Vaz, João Viana and Priscila Maranhão.

To the Outcomes Research Lab from IPOP, especially to Ana Sofia, Patricia Redondo and Ana Ferreira for providing support and being always available to discuss all of my ideas. Working with such a team is easier and reassuring. To Carlos Saez, for providing invaluable insights to my papers. Your availability to meet with me and answer even the simplest questions was priceless. To Alexandrina Cardoso, for the help with sometimes cryptic issues with data. Your drive and willingness to help were key lessons for me.

To my friends, especially Amarílis, André, Catarina, Diogo, Fiama, João Ribeiro, João Teixeira, Mafalda, Maria, Mariana, Miguel Maia, Miguel Ribeiro and Sofia. The possibility to vent and forget about the problems while we are together was vital in helping me stay grounded and continue pushing forward.

Last but not least, to my family, for always being there for me, even though they did not quite understand why I wanted this so much.

João

"If you ain't aim too high,
Then you aim too low."

Jermaine Lamarr Cole

Outline

This thesis is structured as follows: Chapter 1 synthesizes the aim and specific objectives of this thesis. Chapter 2 presents a brief introduction to core concepts for the thesis, such as Knowledge Discovery in Databases (KDD), Evidence Based Medicine (EBM) or privacy and ethical concerns.

In order to enable a better structure of the thesis, we split the different works across the data science methodology: KDD. Focusing on the steps of data cleaning/generation, then data acquisition and analysis and finally on the usage and application of knowledge created from the data in order to have impact in real-world. With this, chapter 3 refers to work done on the generating and preprocessing of data. Chapter 4 focus on data acquisition methods and chapter 5 focus on enabling decisions in healthcare practice based on data, through clinical research and/or Clinical Decision Support System (CDSS).

Chapter 3 delves into innovative methods for enhancing data quality, a critical component of the data preparation phase in KDD. This chapter presents findings from research focused on the development and utilization of synthetic data generation and automatic data quality assessment methods. The use of Generative Adversarial Networks (GANs) to create realistic, non-sensitive datasets exemplifies the synthesis of new data collection methodologies that expand data volume while protecting privacy. Furthermore, this chapter explores automatic tools designed to assess data quality, significantly reducing manual effort and enhancing the efficiency and accuracy of data analysis. Such advancements are crucial for ensuring the integrity and reliability of datasets, which form the foundation for effective data mining and subsequent knowledge discovery.

Chapter 4 assesses health data science methods under the constraints of limited data access, aligning with the data access and data mining phases of KDD. This chapter investigates the effectiveness of distributed data approaches and benchmarking strategies that enable data science techniques to operate efficiently without comprehensive data access. The focus on distributed data analysis across multiple locations supports privacy and security, particularly critical in sensitive health data scenarios. Additionally, benchmarking these methods provides a framework to evaluate their effectiveness, ensuring that the insights derived are both accurate and actionable despite limitations in data availability.

Chapter 5 explores the practical application of health data in real-time decision-making processes, particularly in clinical settings like drug evaluation and obstetrics. This chapter relates closely to the interpretation/evaluation phase of KDD, where data insights are translated into actionable knowledge. By applying causality principles and Machine Learning models, this research assesses the effectiveness of breast cancer drug treatments and develops CDSS for obstetrics. The innovative use of explainable ML and Inverse Probability of Treatment Weighting (IPTW) showcases how advanced data analysis techniques can influence policy and decision-making in healthcare, demonstrating the real-world applicability of KDD processes.

Chapter 6 summarizes the findings of the works developed in this thesis and how they can be leveraged.

Chapter 7 communicates the conclusion, limitations, and future work.

Attachments include supplementary data to some papers.

xx

Contents

List of Figures

List of Tables

Abbreviations

Area Under the Precision-Recall Curve is a metric used in binary classification tasks which evaluates the trade-off between precision and recall for different thresholds, particularly useful in datasets with a significant imbalance between classes.

Area Under the Receiver Operating Characteristic Curve is a performance measurement for classification problems at various thresholds settings, representing the degree or measure of separability between classes by plotting the true positive rate against the false positive rate.

Average Treatment Effect is a measure used in statistics and econometrics to estimate the mean effect of a treatment (or intervention) compared to a control condition across an entire population.

Average Treatment Effect on the Treated is a statistical measure that estimates the average effect of a treatment on those individuals who have received the treatment, as opposed to comparing them with a control group who did not receive the treatment.

Bayesian Network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph.

Causality is the relationship between causes and effects, implying that an action or event can produce a specific outcome.

CDK4/6i (Cyclin-Dependent Kinases 4 and 6 Inhibitors) are a class of targeted cancer therapies that inhibit the activity of cyclin-dependent kinases 4 and 6. These kinases play a crucial role in cell cycle progression, and their inhibition can prevent cancer cells from proliferating. CDK4/6 inhibitors are commonly used in the treatment of hormone receptor-positive, HER2-negative advanced breast cancer to slow disease progression and improve outcomes..

Clinical Decision Support System is a health information technology system designed to assist healthcare providers in making evidence-based clinical decisions.

Cox Regression is a statistical method for investigating the effect of several variables on the time a specified event takes to happen, often used in survival analysis.

Eastern Cooperative Oncology Group also known as the ECOG Performance Status, is a standardized measure used to assess a cancer patient's level of functioning and ability to perform daily activities. The scale ranges from 0 (fully active, no restrictions) to 5 (dead). It helps guide treatment decisions and assess the patient's ability to tolerate specific treatments..

Electronic Health Record is a digital version of a patient's medical history, maintained by the provider over time, that includes all key administrative clinical data relevant to that person's care.

Evidence-Based Medicine is a clinical discipline that emphasizes the use of empirical evidence from clinical research to inform medical decision-making.

Explainable AI refers to artificial intelligence and machine learning techniques that provide human-understandable explanations of their operations and decisions.

F1 score is the harmonic mean of the precision and recall. It thus symmetrically represents both precision and recall in one metric..

FHIR is a standard for exchanging healthcare information electronically, focusing on ease of implementation and interoperability.

Generative Adversarial Networks are a class of machine learning frameworks where two neural networks, the generator and the discriminator, are trained simultaneously in a competitive manner, with the generator creating data samples and the discriminator evaluating them.

HER2- refers to a classification of cancer cells that do not overexpress the HER2 protein, a receptor involved in cell growth and differentiation. In HER2-negative cancers, particularly in breast cancer, HER2-targeted therapies are usually ineffective, so treatment options focus on other pathways. HER2 status is a crucial factor in determining treatment strategies and prognosis..

Inverse Probability of Treatment Weighting is a statistical technique used in observational studies to adjust for confounding, where each subject is weighted inversely to the probability of receiving the treatment they actually received.

Overall Survival is the duration from the start of treatment or diagnosis until death from any cause. It is a key endpoint in clinical trials, reflecting the effectiveness of a treatment in extending the life of patients. OS is considered a definitive measure of clinical benefit, providing a comprehensive view of a treatment's impact on patient survival..

Progression-Free Survival is the length of time during and after treatment in which a patient's disease does not worsen or progress. It is an important endpoint in clinical trials, particularly for cancer therapies, as it indicates how well a treatment can delay disease advancement. PFS is used to assess the efficacy of new treatments in controlling the disease..

Root Mean Square Error is a standard way to measure the error of a model in predicting quantitative data, representing the square root of the average squared differences between predicted values and actual values.

Nothing great in the world was accomplished without passion.

Friedrich Hegel

1

Introduction

I first started thinking about the topic of this thesis during a class of the PhD program called precisely HEADS - Health Data Science. The class concerned the gap between the availability and usage of health data for secondary use and how this was a complex issue. At the time, Professor Pedro Rodrigues raised three big questions that were (are) still open in the field. The first was the reuse of data, and if it was possible, ethical, and feasible to use data that was created for another purpose for secondary use - research. The second was about the extent to which routine data sources and innovations in analytical methods alleviate the need to conduct randomized clinical trials. The last was a question related to matters of governance, privacy, and trust when routine health data is made available for research.

I spent some time on this subject and decided to tackle it. This decision was motivated by the fact that these subjects relate to core challenges for research with health data and define an essential research agenda for the health data science community. Furthermore, I also believe that these subjects are related to the slow adoption of innovative digital and informatics solutions in clinical settings, which often results in significant losses in the potential benefits of these technologies. Despite groundbreaking discoveries in the industry and academic settings, there seems to be a substantial gap in translating these advancements into practical industry applications. This realization sparked a question: how could we effectively bridge this gap and fully leverage the potential of healthcare data?

We do know that healthcare practices are deeply intertwined with technological advancements. Technology, in its broadest definition, encompasses *"methods, systems, and devices which are the result of scientific knowledge being used for practical purposes"*. In essence, healthcare and medicine represent applied sciences, utilizing principles from biology, physics, chemistry, and

mathematics to develop treatments, diagnostic methods, and medical procedures. Over the past two to three decades, the fields of computer science and informatics have increasingly integrated into the healthcare domain, significantly influencing its evolution and methodologies [1]. A paper-based industry is now being digitalized and computerized, harnessing its potential. This has been leading to an increase in the amount of data generated by healthcare systems [2, 3].

These data have the potential to greatly improve the current methods and practices in healthcare. However, they are still not being used to their full potential [2, 4]. But why is that?

The questions raised in that class of 2019, published in [5], seem to sum it up quite well.

- Van der Lei's First Law of Medical Informatics [6]: Data shall be used only for the purpose for which they were collected. Is it advisable to reuse data from Electronic Health Records (EHRs) for research?

- Effectiveness of Routine Data Sources and Analytical Innovations: Examining the extent to which routine data sources and innovations in analytical methods alleviate the need for randomized clinical trials. Can statistics and big data on observational data replace Randomized Clinical Trials (RCTs) and provide insight into causality?

- Governance, Privacy, and Trust Issues: Addressing governance, privacy, and trust questions when routine health data are made available for research. Are patient's comfortable with the reuse of their data?

This is especially important when we note that the gold standard for evidence creation is RCTs, which can vary in quality, time, and resources. A RCT may cost no less than 20 million euros to run, and according to a report submitted to the United States of America (USA) Department of Health and Human Services [7] can cost as much as 100 million USA dollars. This is indeed a very steep price for obtaining the information we need to innovate. Parallel to this, usually supported by these RCTs are systematic reviews and meta-analyses, highly supported and promoted by Evidence Based Medicine (EBM) which are estimated to cost approximately 140 thousand dollars each [8]. Additionally, we must take into account the time it takes to create and publish a good paper on evidence synthesis, often making it hard to keep up with the pace of innovation.

So, we are now being faced with huge amounts of clinical data generated by EHRs and Health Information Systems (HISs). But which tools are the most suited for harvesting the potential of this data? The capabilities and assumptions behind modern Knowledge Discovery in Databases (KDD), Machine Learning (ML), and Artificial Intelligence (AI) seem to be good approaches for harvesting this potential. However, they are very different from the traditional statistical methods that are typically used in healthcare. Therefore, to properly use these methods in healthcare and actually provide value to patients, we need to understand the differences between these methods and how they can be used to complement each other. Currently, we already have an idea of what are the major key areas that hinder the adoption of AI in healthcare like problems related to data privacy and security, data quality and integrity, interoperability, ethical considerations, and the fact that the hype of AI is far greater than the AI science, the acceptance, and trust of healthcare

practitioners of AI based systems [9, 10], and how to proper evaluate the potential risks of AI in healthcare, just to mention a few [11]. This is a very complex problem that requires many approaches and solutions. It is a popular assumption that 87% of data science projects never reach production [12]. Even if numbers for the healthcare domain are not available at this time, it is safe to assume that the number is not much different, if not higher. Those that actually do may never actually create any impact owing to the lack of adoption by healthcare practitioners or the lack of trust in the system [13].

There is still a long way to go to harvest all the potential healthcare data has to offer, and our research objectives are focused on powering up this adoption. What can be done to improve these chances? What can be brought to the table to enhance the success rate?

1.1 Research Objectives

This thesis has three main goals:

- Goal 1: Research methods to improve data quality, whether using synthetic data generation to enlarge data volume and protect privacy (sections 3.1, 3.2 and 3.3) or by creating automatic data quality assessment methods (section 3.4)

- Goal 2: Assess health data science methods with limited data access (sections 4.1 and 4.2).

- Goal 3: Explore strategies to transform health data into actionable decisions and policies (sections 5.1 and 5.2).

1.2 Research Questions

- Goal 1: How can we improve the quality of data used in health data science?

    - RQ1.1: Can GANs help create realistic datasets?
    - RQ1.2: How Can we compare two tabular datasets?
    - RQ1.3: Can we use machine learning feature to compare datasets?
    - RQ1.4: Can we use machine learning to create automatic data quality assessments?

For this goal, we aim to improve the quality of data used in health data science. We will start by exploring the use of Generative Adversarial Network (GAN) to create realistic tabular datasets (Section 3.1). The advent of deep learning has led to remarkable progress in generating data, particularly for images, videos, and sounds. From these examples, GANs have produced excellent results. However, can this performance be matched in tabular datasets?

Next, we will explore methods for comparing datasets (Section 3.2), which relates to the first point since we cannot create good synthetic data without having robust metrics to assess their similarity. Therefore, we have attempted to compile the current state-of-the-art metrics for this purpose.

Since we did not find, in our opinion, a comprehensive set of metrics in the second point, we sought to apply ML to create effective evaluation metrics for comparing two tabular datasets (Section 3.3). Finally, we will explore the use of ML to create automatic data quality assessments (Section 3.4). This is crucial because we need to know if the data we are using is of sufficient quality for the task at hand. This is especially important when using ML methods, as they are highly sensitive to data quality.

- Goal 2: How can we assess health data science methods with limited data access?

    – RQ2.1: Leveraging distributed systems in healthcare: is it advisable?

    – RQ2.2: Can institutions share their performance metrics without hesitation of retaliation?

To achieve this goal, we sought to overcome data access limitations. If we do not have access to data, how can we explore its potential? Several limitations are in place to ensure the ethical, safe, and appropriate use of patient data, but sometimes these limitations create obstacles to timely data usage or even prevent access altogether. Therefore, we aimed to evaluate whether distributed learning is a viable option to overcome this limitation (Section 4.1).

Subsequently, we attempted to develop an alternative method for comparing health institutions performance metrics without knowing the true values of each metric, but still enabling them to position themselves on a scale (Section 4.2).

- Goal 3: How can we convert health data into decisions and policies?

    – RQ3.1: How can we leverage data to create clinical decision support systems?

    – RQ3.2: How can we leverage data to assess treatment efficacy?

For the last goal, we tried to actually go from end-to-end. This means that we tried to leverage real world data in its raw form and transform it into actionable insights. Our major objective was to check the challenges that block the development of such tools. We tried to assess the real-world effect of two drugs for breast cancer and compare them among themselves and with the previous gold standard: endocrine therapy (section 5.1). Then we tried to create a Clinical Decision Support System (CDSS) that could be used in real-time clinical environments and be able to provide support for the need of C-sections (section 5.2).

*If we knew what it was we were doing, it would not be
called research, would it?*

Albert Einstein

2

State of the art

In this chapter, we cover some fundamental concepts and connections between some of the most important fields of study that are relevant to this thesis. We will start by covering the basics of AI and its history and how it came to be the powerhouse and almost generic term that it is today. Then we will follow up to EBM and its importance to modern medicine, some barriers and benefits of following such practice and paradigm. Thirdly, we will focus on KDD and how healthcare databases can provide support for new knowledge.

Then we will cover the field of Health Data Science (HEADS) and its connections to AI, EBM and KDD as a way of bringing all of these together. We will then cover the field of Explainable AI (XAI) and its connections and Causal Machine Learning (CausalML) and how these subfields are vital to tackle modern medicine and a cornerstone of building trust in computerized systems and autonomous decision support tools. We conclude this chapter by covering some legal and ethical considerations relevant to this thesis.

2.1 Artificial Intelligence

AI has already been under public focus for a few years now, but its concept is still elusive, mainly because the definition has been changing rapidly as well. From the very beginning, the field of AI focused not only on understanding but also on building intelligent entities [14]. Intelligent entities can be understood as machines that can act according to what is expected in a wide range of situations.

The first work of AI can be credited to Warren McCulloch and Walter Pitts (1943) with the proposed model of artificial neurons. In the 1950s, AI could be associated with the works of Christopher Strachey and two chess-playing programs.

In the 1960s, perceptrons were considered state-of-the-art AI. In the 1980s, expert systems provided advanced reasoning that the so-called weak methods of previous iterations could not compete with.

The 1990s brought probabilistic reasoning and ML, which led to more robust systems that went beyond the Boolean logic used so far. In the 2000s, big data and ML became the focus. Big data was used as a symbol of the increasing amounts of data in some industries [15], and ML as *the study of computer algorithms that improve automatically through experience* [16]. This last definition is especially important since it is currently used as a synonym for AI across several industries but has a different meaning, as discussed below. This era probably peaked around IBM Watson's victory in Jeopardy but yielded far fewer interesting results in healthcare [17], and the 2010s brought deep learning. Nowadays, AI is a buzzword that is used to describe a wide range of systems, from the simplest to the most complex. It is clearly trending, as reports on AI show that papers regarding the subject have seen big increase, up to 20-fold increase from 2010 to 2019 [14, 18, 19, 20] To define AI, we can refer to the description given by a team of specialists whom the European Commission tasked with developing guidelines on AI [21]. This document clarifies that initially, it is essential to distinguish between intelligence and rationality. Intelligence, being more subjective and philosophical, differs from rationality, which is pragmatic and linked to the ability to select the optimal course of action in a given situation to achieve a specific goal. While rationality is a more tangible concept and not identical to intelligence, it should be considered an integral component of it [21, 14].

From these two concepts, we can go even deeper and define that rationality can be achieved in an AI system by perceiving the environment, reasoning with what is perceived, and acting on the environment. From these three elements, we can argue that reasoning is the core functionality, which is related to taking data, understanding it or interpreting it, and reasoning on this data through a model (numerical or symbolical) to reach the best action.

There is also the need to address the current distinction in AI, which is the narrow and general AI. The first is the one that exists nowadays, and it's an AI that is not generic; it is focused on a specific task. The second is the one that is not yet achieved, and it is more generic and can be applied to several tasks. This is the one that is usually associated with the popular or common concept of AI [21, 14]. So, with this in mind, AI can be defined as:

> *Artificial intelligence systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning,*

scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems). [21]

Certainly, given the rapid advancements in this field, this concept might become outdated swiftly. Nonetheless, it serves as an effective starting point for grasping the fundamentals of AI and its implications.

2.2   Evidence Based Medicine

In 1996, David Sackett and colleagues defined EBM as *the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients* [22]. Despite having historical antecedents dating back to at least the 19th century, the term "evidence-based medicine" was first coined by a team at McMaster University in Canada in the 1980s [23]. This was a time when clinical decision-making was mostly based on untested observations and physicians' experience, leading to variability in treatment strategies. The birth of EBM marked a pivotal moment in medical history, aiming to standardize patient care and improve outcomes. So, EBM is still a relatively recent concept in healthcare, which entails integrating the best available research evidence with clinical experience and patient values to make decisions about patient care. EBM can be defined into three major pillars:

- Best available evidence
- Clinical expertise
- Patient values, expectations, and/or wishes.

Clinical expertise refers to the acumen and discernment gained from hands-on clinical experiences and consistent practice. This expertise manifests notably in enhanced diagnostic abilities and in the considerate recognition of a patient's unique circumstances, rights, and wishes when making care decisions. The term "Best available evidence" pertains to pertinent clinical studies, often stemming from epidemiological investigations. This is linked with the ability (and willingness) to challenge current diagnostic methods and treatments, introducing alternatives that are more robust, precise, effective, and safer.

Without experience, clinical practices blindly follow the best available evidence, which is not always the best option for the patient, since sometimes it may be inapplicable to a specific scenario. Without evidence, clinical practice becomes stagnant and unable to evolve [22].

The main concept of EBM is the hierarchy of evidence, which classifies different types of research studies based on their methodological quality and applicability to patients. At the top of this hierarchy are RCTs and systematic reviews of RCTs, which are considered to provide the most robust evidence. Observational studies, case series, and expert opinions are further down the hierarchy due to their inherent limitations (figure 2.1). EBM advocates for the application of the highest level of evidence available in clinical decision-making.

Historically, medical decisions leaned heavily on anecdotal observations and the prevailing beliefs of seasoned practitioners. To underscore the dangers of relying solely on such expert
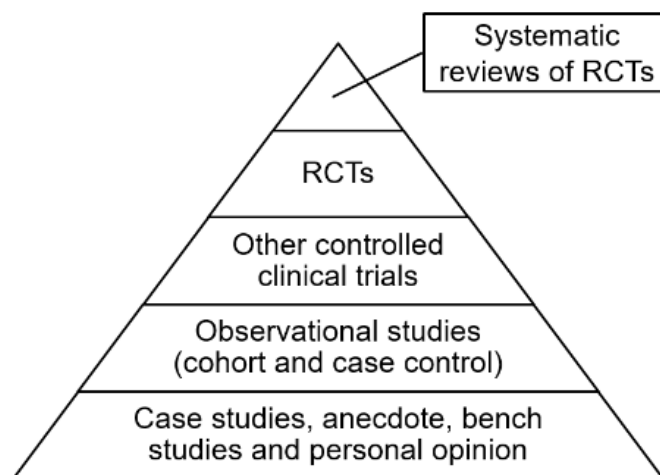
Figure 2.1: Evidence Based Medicine diagram adapted from [24]

opinions, Sackett frequently recounted the circumstances surrounding George Washington's unfortunate end. Despite being in good health at the age of 68, Washington developed epiglottitis. Rather than opting for a tracheostomy, a treatment method known since ancient Greek times, his physicians, guided by the prevailing expert opinion, chose bloodletting as the course of action. Tragically, this decision led to Washington's likely preventable death, highlighting the critical importance of grounding medical decisions in robust evidence.

Nonetheless, EBM faces several criticisms. The primary critique is that EBM merely reflects the core practice of medicine as it is already widely implemented. However, data indicate a different reality [22].

The second criticism targets the seemingly insurmountable challenge of staying current with the vast and ever-growing body of medical literature. While there are instances of clinicians successfully managing this feat, the argument still highlights a crucial issue: how to cope with the burgeoning overflow of evidence in modern times. It raises pertinent questions about not only keeping up with the literature but also ensuring its application in clinical practice. The utility of evidence lies in its application, making this a vital consideration. As we will explore in subsequent sections, this is where KDD and AI may offer significant contributions.

2.3   Extracting Knowledge from Data

KDD is about turning data into knowledge. However, turning data into knowledge or insights is not new in healthcare. The first attempts to use data to improve healthcare date back to the 17th century, when John Graunt used data from the London Bills of Mortality to study the causes of death in the city [25]. This was the first time that data were used to understand the health of a population. Since then, the field of KDD has evolved significantly, and it is now a crucial part of healthcare, helping to improve patient outcomes, enhance clinical decision-making, and optimize healthcare delivery.

Figure 2.2: Knowledge Discovery in Databases Process, adapted from [26]

Additionally, the fact that data are being collected at an unprecedented rate, and the need to extract knowledge from them, has led to the development of several methodologies and frameworks to map low-level data (granular) into short reports, more abstract, or more useful formats [26]. So, it is only natural to see that KDD has become very popular in a wide range of industries nowadays.

Healthcare is no exception, and KDD has been applied to several areas of healthcare, from clinical decision support to disease surveillance and outbreak detection. Reports and papers suggest that [15] the digital data in the healthcare space have been increasing rapidly, due to the adoption of EHR and similar digital tools in the healthcare space.

The complexity and vastness of healthcare data, encompassing electronic health records, genomic data, medical imaging data, and various other types of data, call for the adoption of intelligent systems that can mine this data for useful insights. The KDD process, comprising data cleaning, integration, selection, transformation, data mining, pattern evaluation, and knowledge presentation, can effectively help discover patterns and relationships in healthcare data, which are often not apparent to traditional analysis methods. This process facilitates the prediction of disease outbreaks, the identification of high-risk patient groups, the optimization of treatment plans, and the enhancement of healthcare service delivery. The generic process for KDD is shown in figure 2.2.

Several frameworks have been proposed to implement the KDD process. One such prominent framework is Cross-Industry Standard Process for Data Mining (CRISP-DM), which comprises business understanding, data understanding, data preparation, modelling, evaluation, and deployment. CRISP-DM was conceived in 1996 and became a European Union (EU) project under the ESPRIT funding initiative in 1997 [27]. Sample, Explore, Modify, Model, and Assess (SEMMA) [28] involves five stages: sampling, exploration, modification, modelling, and assessment. It starts

by analysing a subset of data, then seeks patterns and modifies variables. A model is built, and the results are evaluated. While SEMMA covers key data-mining aspects, it misses fundamental components of information system projects like analysis and implementation.

It is important to distinguish, however, that KDD is not the same as Data Mining. As stated in [26], we agree that KDD is a major process of which Data Mining is a part. So, in order to understand the process of KDD, we need to understand the process of Data Mining, which can be understood as the application of algorithms for extracting patterns from data. There are several classes of algorithms, each best suited for different kinds of tasks:

- Classification Algorithms: These are used to predict categorical class labels. Examples include Decision Trees, Naive Bayes, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and various types of Neural Networks. These are used in disease diagnosis, patient risk prediction, and readmission prediction.

- Clustering Algorithms: These are unsupervised methods used to group similar data points together. K-Means, Hierarchical Clustering, DBSCAN, and Self-Organizing Maps are common clustering algorithms used in patient segmentation and anomaly detection.

- Regression Algorithms: These are used to predict continuous output variables. Examples include Linear Regression, Logistic Regression, and Regression Trees. These algorithms find application in predicting disease progression and healthcare costs.

- Association Rule Mining Algorithms: These discover associations or patterns among a set of items in large databases. *Apriori* and FP-Growth are commonly used algorithms in this class, helping in discovering co-occurring health conditions or drug interactions.

- Sequential Pattern Mining Algorithms: These help discover or predict specific sequences of events, which is particularly useful in medical trajectory analysis.

- More sophisticated architectures and algorithms appeared with neural networks, generative AI, and reinforcement learning, among others.

As a result, KDD is the process of applying Data Mining algorithms to data but also the data preparation, selection, cleaning, and most important of all, the incorporation of prior knowledge about the domain along with the proper interpretation of results. This difference is vital to understanding KDD, since blindly applying data mining or ML methods to data will only render results that are not useful or even misleading [26].

In short, KDD can be understood as a multidisciplinary subject that bridges and aggregates knowledge from different areas like ML, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing. On top of all of these subjects and research areas sits the most important of all—which is domain expertise.

2.4 Health Data Science

HEADS is an interdisciplinary field that applies rigorous methods to transform healthcare data into actionable knowledge for improving health outcomes. It involves the collection, interpretation, and application of vast amounts of biological, clinical, population, and health system data to improve patient care and public health. The advent of electronic health records, genomics, mobile health technologies, and other forms of big data have fuelled the growth of this discipline.

In practice, HEADS involves the use of statistical and ML methods to analyse healthcare data. This data can be patient records, genomic data, demographic data, and more. It includes elements from various disciplines like biostatistics, epidemiology, informatics, and health economics. The ultimate goal is to provide a data-driven foundation for health decision-making for clinicians, health administrators, policymakers, and researchers.

An integral part of HEADS is predictive modelling and hypothesis testing. Predictive modelling involves the creation and use of statistical models or ML algorithms to predict future outcomes based on historical data. Hypothesis testing, on the other hand, is used to test the validity of a claim or theory about a population based on sample data. These are crucial for HEADS as they allow us to make educated guesses about health trends and outcomes.

Importantly, HEADS has significant ethical and privacy considerations. Health data is often sensitive and personal, so maintaining privacy and confidentiality is crucial. This requires secure data handling and storage practices, as well as careful consideration of ethical implications when designing studies and algorithms. Health Data Scientists must also be wary of algorithmic bias and must ensure their models do not perpetuate or amplify health disparities. The ultimate goal of HEADS is to improve patient outcomes and health equity using the best available data and methods.

The potential of using systematically created data in healthcare has certainly a lot of potential. However, we have seen in the past as well, that the hype of AI and ML usually are not supported by truth. There are currently six main aspects that hinder the potential of HEADS [29, 5]:

- Interoperability
- Semantic
- Secondary usage
- Data quality
- Privacy and ethics
- Observational data

Interoperability is defined by *the ability of two or more systems or components to exchange information and to use the information that has been exchanged* [30]. In the context of healthcare, this means that different systems should be able to exchange data and interpret the data that has been exchanged. This is a very important aspect of HEADS since the data is usually stored in different systems, with different structures and different purposes. So, if systems are locked inside themselves and no export is possible, data becomes inaccessible. So, it is only natural that interoperability has been a key factor in gathering data. With tens or hundreds of different systems in

every health institution, the possibility of exchanging data between EHRs plays a vital role. The usage of interoperable standards is of extreme importance in order to tackle the need to get data with a predefined structure.

Semantic adds a layer to the previous points, being sometimes related to interoperability as well. The fact that several institutions and EHRs are involved in creating knowledge from data, raises the problem that not all have data coded in clinical terminologies, or if they do, it is seldom the same across systems, since semantics has a very tight relationship with domain, especially in healthcare. So the normalization of uncoded terms is often required and mapping across terminologies is also very common, which is time-consuming and requires expertise in several fields.

Secondary usage is related to the fact that we are aiming to use data for a purpose for which the data was not created. The main goal of the healthcare data is to provide care. It is not meant for analysis and gaining insights. More than that, is already pretty well documented that the usage of EHR is very different from institution to institution and from country to country [31, 32, 5]. This means that the context where data was collected, even the actual person who inserted the information could be key to interpreting the results. To make things more complicated, the degree of precision of the data inserted varies highly on the type of information and context, as reported in [33].

Data Quality stems from the secondary usage. If the data is not reliable, how can we use it to gather useful knowledge from it? In order to, at least, try to counter this, we can apply several statistical methods and ML algorithms to try to clean the data. However, this is not a trivial task, since the data is usually very heterogeneous and the context where it was collected is not always available. So, data quality is a very important aspect of HEADS, since it can be the difference between a good and a bad model.

Privacy and ethics add yet another layer to the problems of HEADS. The fact that we are dealing with sensitive private data, which is not meant to be used for secondary purposes, raises the question of privacy and ethical concerns. Anonymization techniques and privacy-preserving methods are key to tackling this problem. However, they are not problem-free and are often complicated to assess. Moreover, the risks are very high, since the data is very sensitive and the consequences of a breach of privacy can be very serious, undermining public trust in clinicians, healthcare institutions and the healthcare system as a whole.

Observational Data relates to the fact that all HEADS will be based on observational data. This means that the data is not collected in a controlled environment, which is the case for RCTs. Consequently, this data is subject to several biases, which are not always possible to control. The cornerstone of RCTs is simply not possible to apply here, preventing a proper comparison between groups. Even though there are techniques to tackle the unbalance in the measured variables, there is no way to control the unmeasured variables, which can be the cause of the observed effect. This is of particular importance and a major area of research at the moment, as we will see in the sections 2.5 and 2.6.

With this in mind, it is natural to assume that HEADS and EBM are very synergic. If, on the one hand, we could argue that KDD can take EBM even further by using Data Mining and AI to

produce synthetic evidence by analysing, summarizing, or even combining evidence from several sources in order to feed medical practice with the best evidence available in a useful manner. On the other hand, we could also argue that EBM can be used to guide the KDD process, by providing the necessary domain knowledge to interpret the results and to guide the process of data preparation, selection, and contextualization. The domain knowledge mentioned in the KDD section could be applied by EBM.

The synergy of KDD and EBM has the potential to revolutionize healthcare delivery and improve patient outcomes. By leveraging the power of data analysis and advanced algorithms, health data scientists can identify novel biomarkers, develop predictive models, and personalize treatment plans based on individual patient characteristics. This not only enhances clinical decision-making but also enables precision medicine, where treatments can be tailored to the specific needs of each patient. Additionally, the use of HEADS in evidence-based medicine allows for the continuous monitoring of treatment effectiveness and safety, facilitating the identification of best practices and the refinement of clinical guidelines over time.

2.5   Explainable Artificial Intelligence

AI has experienced unprecedented advancements in the last decade, leading to its integration in various domains, including medicine. It has been instrumental in transforming clinical decision-making, drug discovery, patient monitoring, and predicting disease trajectories. Despite these advancements, the "black box" nature of complex AI models poses interpretability challenges, limiting their widespread adoption in healthcare, a field where transparency, reliability, and understanding of decision-making processes are vital. This lack of interpretability, also known as opacity, can lead to misdiagnoses, inappropriate treatment plans, and, most importantly, breaches in trust among clinicians, patients, and AI systems.

As such, the concept of XAI, which aims to create a suite of techniques that produce more explainable models while maintaining a high level of predictive accuracy, has gained significant attention in medical AI research. XAI seeks to bridge the gap between AI opacity and human interpretability, and in doing so, it can enhance the transparency, reliability, and acceptance of AI applications in the healthcare setting.

So, for this to happen, we need a new framework for applying such mechanisms. A new step that could be attached to the ones seen before in section 2.3 will enable human comprehension of the model's output.

Even though several grouping and taxonomies of XAI are available mentioned in [34, 35, 36, 35, 37], a simplified approach based on [37] will be used in order to contextualize this concept.

We can divide it into two main categories. Firstly, the explanation type is divided into global and local. Local and global explanations are methods used to interpret ML models, especially those that are considered "black box" models, such as deep learning networks. These methods help us understand why and how a model makes certain decisions, which can be crucial in many settings for ethical, legal, and practical reasons.

Local Explanations: These involve understanding the prediction of a ML model for a specific individual instance. They help to answer questions like: "Why did the model predict that this particular patient has cancer?" or "Why was this specific transaction flagged as fraudulent?".

Global Explanations: These focus on understanding the model behaviour across all instances, or more broadly on a dataset-wide level. They help to answer questions like: "What features are generally important for prediction in the model?" or "What is the overall logic of the model?".

Secondly, we have the method type, where we have 3 main subcategories related to the stage of the data science process it is applied, *pre*, during, and *post*-model training.

Pre-Model XAI: These methods involve improving the transparency and interpretability of models before they are even trained. This includes thoughtful feature engineering, Exploratory Data Analysis (EDA), and applying domain knowledge to create meaningful variables. The goal is to design a model that will be more interpretable from the onset.

Intrinsic XAI: This involves using ML models that are intrinsically explainable. These models are designed in such a way that their decision-making process is understandable by default. Examples include linear and logistic regression, cox regressions, decision trees, Naïve Bayes, Bayesian Network (BN), and rule-based models. While these models may sometimes lack the predictive power of more complex models, they provide clear interpretability: you can directly examine the impact of the variables and understand how the model makes its predictions.

Linear Regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables. It assumes that the relationship between these variables is linear and can be represented by a straight line. The goal is to fit the best possible line that describes this relationship by minimizing the sum of the squared differences (errors) between the observed values and the values predicted by the line. Linear regression is widely used in various fields for prediction, modelling, and determining the strength and character of the relationship between variables. It forms the basis of many more complex statistical modelling techniques.

Logistic Regression is used to model the probability of a binary outcome that depends on one or more independent variables. Unlike linear regression, which predicts a continuous outcome, logistic regression predicts the probability of a categorical outcome (e.g., success/failure, yes/no, 1/0). The logistic function is applied to the linear combination of independent variables to ensure that the estimated probabilities are between 0 and 1. It's often used in fields like medicine, economics, and social sciences to predict the likelihood of an event occurring based on various factors.

Cox Regression or the Cox proportional-hazard's model, is a statistical technique used for investigating the effect of several variables on the time a specified event takes to happen. In medical research, this often refers to survival times. The model allows for the estimation of hazard ratios, which describe how the hazard changes with a one-unit change in the predictor variable. The Cox model makes an assumption that the hazard ratios are constant over time, known as the proportional hazard's assumption. This model is vital for understanding how different factors influence survival or failure time and is commonly applied in epidemiological and medical research.

Bayesian Networks A BN, also known as a belief network or Directed Acyclic Graph (DAG)

model, is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a DAG.

Given a set of variables $X = \{X_1, X_2, ..., X_n\}$, the joint probability distribution is given by:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

where $Parents(X_i)$ is the set of parent variables of X_i in the network.

This formula represents the factorization of the joint distribution over X, based on the graphical structure of the Bayesian network.

Now, in the Bayesian network, each node is conditional independent of its non-descendants given its parents. If we denote $ND(X_i)$ as the set of non-descendants of X_i and $Pa(X_i)$ as the parents of X_i, the conditional independence is described as:

$$X_i \perp ND(X_i) | Pa(X_i)$$

This means that X_i is conditionally independent of its non-descendants given its parents.

A common task for Bayesian networks is inference, which means computing the posterior probability of a set of query variables Q, given some observed variables E. That is, we want to compute $P(Q|E)$. According to the Bayes rule, we have:

$$P(Q|E) = \frac{P(Q,E)}{P(E)} = \frac{P(Q,E)}{\sum_{q \in Q} P(Q = q, E)}$$

where the denominator is a normalization constant ensuring the result is a valid probability distribution. Note that performing this inference is NP-hard, which is why various approximation algorithms have been developed.

Tree based methods Tree-based ML methods are a subset of algorithms that use a tree-like graph structure for making decisions or predictions. The most basic type is the Decision Tree, where the tree is used to go from observations about an item to conclusions about the item's target value (classification or regression). Each node in the tree represents a feature in the dataset, each branch represents a decision rule, and each leaf node represents the output value. More advanced tree-based methods include Random Forests, which build multiple Decision Trees and average their predictions for better accuracy and generalization, and gradient-boosted trees, which build trees sequentially, each one correcting the errors from the previous one.

The major advantage of tree-based methods is their ease of interpretation and understanding, especially for Decision Trees. However, a single tree is often prone to overfitting, where it performs well on the training data but poorly on unseen data. This is why ensemble methods like Random Forest and Gradient Boosting are popular; they aim to increase robustness and predictive power by combining multiple trees. These methods are widely used in various domains including but not limited to finance, healthcare, and natural language processing for tasks like classification, regression, and even unsupervised learning tasks like clustering.

Post-Hoc XAI: Post-hoc methods are applied after a model has been trained, to try to explain its decisions. This includes techniques like feature importance analysis, partial dependence plots, Local Interpretable Model-Agnostic Explanation (LIME), SHapley Additive exPlanations (SHAP), and counterfactuals. For instance, LIME can be used to create local explanations for individual predictions made by any model, and SHAP values can be used to interpret the impact of features on the model's output both locally and globally. Counterfactuals try to explain a model by example, providing possible changes that would alter the outcome provided by the model.

It is to be noted that a methodology can be classified into two categories. For example, LIME is a local explanation model in a *post-hoc* manner.

Despite all of this, we have to take into account that pre-model and post-hoc methodologies are a proxy for an explanation of the models. That is why we could argue, as stated in [38] that only an intrinsically transparent model can really be the basis of XAI. While *post-hoc* of pre-model methods are only a potentially unreliable proxy for an explanation.

2.6 Causality

Using, once again, the tale of George Washington but now with a different purpose; the medical doctors in that region of the globe, at least, followed the theory of humours, which relied on the fact that a healthy human person was a balance between four humours (blood, phlegm, yellow bile, and black bile). So, the treatment for George Washington was to rebalance those four humours, and so, doctors needed to remove blood, which was the supposed cause of his illness. Since microbiology and its importance would only be discovered sometime after, the idea at the time was inspired by the fact that the imbalance of these four senses of humour and illness was present at the same time.

This is now known as a textbook definition of confounding correlation with causation. And in this subject in particular, it was not the imbalance in the humours that caused illness, but an illness that caused the imbalance.

So, this example shows that evidence without proper causality can lead to misguided results and mistrust.

That is why, nowadays, EBM and XAI can be brought together and expanded through CausalML. But what is causality? We could argue that it is related to something **causing** something else. This causal effect, especially in medicine, can be related to a comparison of the outcome a particular person would exhibit given a particular intervention and the outcome in the same person of the control intervention. This is particularly hard since we cannot do both things at the same time. This is the basis of why RCTs are the gold standard of experimentation, since they are the current best tool to achieve something similar to this [39].

CausalML is a branch of ML that focuses on understanding and quantifying causal relationships from data [40]. Instead of just finding patterns or correlations in data, CausalML aims to uncover the cause-and-effect relationships that explain these patterns. This is especially important

since current or traditional ML and AI methodologies rely heavily on association and not causation. So, CausalML can support traditional algorithms to solve its limitations [41]. There are currently two main frameworks for trying to unveil causality in data: the Structural Causal Model (SCM) and the Potential Outcome Framework (POF) [42]. SCM relies on 1) Causal Graphs and 2) structural equations.

1. Causal Graphs are based on DAGs: These are graphical models used to represent causal relationships between different variables. The nodes in the graph represent variables, and the edges (arrows) between nodes represent causal relationships. For instance, an edge from Node A to Node B signifies that A has a causal effect on B. We should not confuse causal graphs with BN. Even though both rely on DAGs, Causal Graphs represent causal relationships, and BN represent conditional dependencies.

2. Structural Equations refer to a set of mathematical expressions that represent causal relationships between variables. These equations model the way changes in one variable, often termed the "cause," lead to changes in another, termed the "effect." Within a structural equation model (SEM), both observed and latent (unobserved) variables can be incorporated, and the causal pathways between them are explicitly defined. By employing SEM in CausalML, researchers can elucidate intricate relationships among variables, disentangle direct from indirect effects, and infer causal mechanisms. This approach provides a more profound understanding of the underlying data-generating process, enabling better predictions and interventions in complex systems.

POF model centres on the concept of potential outcomes which can be understood as all of the possible outcomes for a patient. Each unit (e.g., a patient or a sample) has a set of potential outcomes, each corresponding to one of the possible treatments the unit could receive. The causal effect is defined as the difference between these potential outcomes. This framework allows for the formal definition and estimation of causal effects. In this approach, we consider the potential outcomes for each unit (for example, a patient in a healthcare context) under each possible treatment or intervention. Each unit has a set of potential outcomes corresponding to each possible intervention. However, we can only observe one of these outcomes for each unit, corresponding to the intervention that was actually received. The other outcomes, which would have occurred had different interventions been implemented, remain latent. These are known as counterfactual outcomes.

The difference between potential outcomes under different treatments represents the causal effect of the treatments. For instance, in a healthcare scenario, if we are studying the effect of a drug, we might consider two potential outcomes for each patient: the outcome if the patient is given the drug, and the outcome if the patient is not given the drug. The difference between these outcomes represents the causal effect of the drug on the patient. However, as we can only observe one of these outcomes for each patient (the one corresponding to the treatment they actually received),

a key challenge in causal inference is estimating the unobserved potential outcomes. Various statistical methods, including randomized experiments, matching methods, and instrumental variable methods, can be used to estimate these unobserved potential outcomes.

1. **Counterfactuals**: This is a concept rooted in the idea of "what-if" scenarios. A counterfactual outcome for a given individual is the outcome that would have occurred had the individual been exposed to a different treatment or condition. Counterfactuals play a pivotal role in the field of CausalML, offering a sophisticated approach to understanding cause-and-effect relationships. In essence, a counterfactual is a conceptual device used to contemplate what would have happened under a different set of circumstances than what actually occurred. This hypothetical scenario is created by altering some aspect of the actual situation, providing a means of comparison to evaluate the effect of a particular variable or intervention.

   For instance, in the context of healthcare, consider a scenario where a patient was given a particular drug and recovered. The counterfactual question here would be: "What would have happened to the patient if they hadn't been given the drug?" Answering this question allows us to estimate the causal effect of the drug on the patient's recovery. While the true counterfactual outcome is unobservable (since we cannot rewind time and alter the decision), various statistical techniques, ML algorithms, and experimental designs are employed in causal inference to estimate this effect as accurately as possible. The ability to make such counterfactual inferences is crucial in numerous fields, including medicine, economics, social sciences, and policy-making, where understanding causal relationships is paramount.

2. **Instrumental Variables**: These are variables that are related to the treatment but not the outcome, except through their effect on the treatment [43, 44]. They can be used to control for unmeasured confounding variables. Instrumental variables (IVs) are a powerful tool used in causal inference to help address the problem of confounding variables, especially in situations where randomization is not feasible. An instrumental variable is a variable that is correlated with the independent variable (the treatment) but does not directly affect the dependent variable (the outcome), except through its effect on the treatment. In other words, it is a variable that induces changes in the explanatory variable but is otherwise unrelated to the outcome of interest. The idea behind using an instrumental variable is to isolate the portion of the variation in the treatment that is independent of the confounders and therefore provides a "natural" form of randomization. The causal effect of the treatment on the outcome can then be estimated based on this variation.

   For example, in a study assessing the impact of education on income, it's challenging to identify causal effects because numerous unobserved factors (like ability or motivation) could affect both education and income, thus confounding the relationship. If we find an instrumental variable – say, distance to the nearest college (which affects the likelihood of getting higher education but doesn't directly affect income) – we can use this to isolate

the part of the variation in education that is unrelated to the unobserved confounders, and thereby get a more accurate estimate of the causal effect of education on income.

It's crucial, however, to remember that the use of instrumental variables relies on certain assumptions, such as the relevance and exogeneity of the IV. The relevance assumption requires that the IV is correlated with the treatment, and the exogeneity assumption requires that the IV affects the outcome only through the treatment and is not related to the unobserved confounders. Violations of these assumptions can lead to biased and inconsistent estimates of causal effects.

3. **Propensity Score**: This is the probability of a unit (e.g., a patient) being assigned to a particular treatment given a set of observed characteristics. Propensity scores are used to balance the characteristics of treatment and control groups, mimicking the conditions of a randomized experiment [45, 46]

The propensity score is a statistical concept widely used in causal inference, particularly in the field of observational studies where random assignment of treatment is not possible. The propensity score for an individual is the probability of receiving the treatment given the observed characteristics of that individual. In other words, it's the likelihood that a particular individual would be assigned to the treatment group based on their observed features.

The key idea behind propensity scores is to create a balance between the treatment and control groups based on these observed characteristics, thus mimicking the conditions of a randomized controlled trial. This balance helps to eliminate bias caused by confounding variables, allowing for a more accurate estimate of the treatment effect. Once propensity scores are calculated, they can be used in several ways including matching, stratification, Inverse Probability of Treatment Weighting (IPTW), and as covariates in regression adjustment.

For example, consider a study investigating the effect of a training program on job outcomes. Individuals might self-select into the training program based on characteristics like motivation or prior education, which are also related to job outcomes, creating confounding. The propensity score, calculated based on these observed characteristics, can be used to match each participant in the training program with a similar non-participant or to weight the observations, such that the distribution of observed characteristics is similar between the groups. This helps to isolate the effect of the training program on job outcomes.

After achieving this balance, it becomes more meaningful and less biased to estimate treatment effects, such as Average Treatment Effect (ATE) and Average Treatment Effect on the Treated (ATT).

The ATE quantifies the difference in mean outcomes between units that are treated and units that are not. Essentially, it calculates the expected difference in outcomes if everyone in a population received a treatment versus if no one received it. Mathematically, the ATE is

represented as:

$$\text{ATE} = E[Y_1 - Y_0]$$

where Y_1 is the potential outcome under treatment and Y_0 is the potential outcome under control. The expectation is taken over the entire population.

After addressing confounding using propensity scores, the ATT narrows its focus to the treated subpopulation. It measures the average effect of a treatment on those units that actually received the treatment, comparing their observed outcomes to what their outcomes would have been without the treatment. The formula for ATT is:

$$\text{ATT} = E[Y_1 - Y_0|D = 1]$$

where Y_1 and Y_0 once more denote potential outcomes under treatment and control, respectively, and D is an indicator for treatment (with $D = 1$ indicating treatment).

However, it's important to note that propensity scores only account for observed confounders. If there are unobserved confounders that influence both treatment assignment and the outcome, propensity score methods may still produce biased estimates of the causal effect.

2.7 Legal and Ethical Considerations

As HEADS, KDD, and AI in healthcare become more and more popular, it is important to consider the words postulated by Francis Bacon in the "Wisdom of the Ancients", *"mechanical arts are of ambiguous use, serving as well for hurt as for remedy."* [47]. This is currently as true for AI as it was at the time. We must consider the good and the bad of such technologies, and how to mitigate the bad and enhance the good. In this section, we will discuss the legal and ethical considerations of AI in healthcare. Ensuring the proper use of healthcare data is key to preserving public trust and ensuring the long-term viability of data-driven health initiatives.

One of the primary legal considerations is data privacy. Laws such as the Health Insurance Portability and Accountability Act (HIPAA) in the USA, and the General Data Protection Regulation (GDPR) in the EU, set tight rules on how healthcare data should be stored, shared, and processed. They require data scientists and healthcare providers to take steps to anonymize data and limit the scope of data usage. Breaching these regulations can lead to severe penalties, including fines and imprisonment. Secondly, there's the matter of data security. With the rise of cyber-attacks, ensuring the robustness of the system against such breaches is both a legal requirement and an ethical obligation. Security breaches could lead to sensitive patient data being stolen, with severe implications for the individuals involved and for the trust in the healthcare system as a whole.

The European Health Data Space (EHDS) refers to a strategic initiative by the EU aimed at creating a unified and secure platform for sharing and accessing health-related data across member states. AI is expected to have a significant impact on the EHDS in several ways [48]:

- Improved Diagnostics and Personalized Medicine: AI can analyse vast amounts of health data, including medical records, imaging, and genetic information, to enhance diagnostic accuracy and tailor treatments to individual patients. This can lead to more effective and efficient healthcare delivery.

- Data Integration and Interoperability: AI can help harmonize data from various sources within the EHDS, including electronic health records, wearable devices, and clinical databases. This promotes interoperability, allowing healthcare professionals to access comprehensive patient information seamlessly.

- Predictive Analytics: AI-powered predictive models can help forecast disease outbreaks, patient admission rates, and healthcare resource utilization. This enables better resource allocation and proactive healthcare planning.

- Drug Discovery and Development: AI can accelerate drug discovery by analysing genetic data, identifying potential drug candidates, and predicting their efficacy and safety profiles. This can expedite the development of new treatments and therapies.

- Enhanced Clinical Decision Support: AI can provide healthcare providers with real-time decision support, offering recommendations based on the latest medical evidence and patient-specific data. This can lead to more informed clinical decisions and better patient outcomes.

- Data Security and Privacy: The EHDS must ensure the privacy and security of health data. AI can help by implementing robust encryption, access controls, and anomaly detection systems to safeguard sensitive information.

- Research and Insights: AI can facilitate large-scale data analysis for medical research, enabling researchers to identify patterns, correlations, and potential breakthroughs in healthcare. This can lead to advancements in medical knowledge and treatments.

- Patient Engagement and Monitoring: AI-driven apps and wearable devices can empower patients to take a more active role in managing their health. These technologies can monitor vital signs, offer health advice, and send alerts to healthcare providers when necessary.

- Reduced Healthcare Costs: By optimizing healthcare processes, improving diagnosis accuracy, and preventing medical errors, AI can contribute to cost savings within the healthcare system, making it more sustainable.

- Regulatory Challenges: Implementing AI in healthcare requires navigating complex regulatory frameworks, ensuring ethical use, and addressing issues related to bias and fairness in AI algorithms. The EHDS will need to establish guidelines and standards to address these challenges.

On the ethical front, considerations include ensuring data fairness and avoiding bias. Given the diversity of patients in terms of age, race, sex, socioeconomic status, etc., algorithms should

be designed and validated to ensure that they do not unintentionally perpetuate or amplify societal biases. For instance, a predictive model for disease risk should not unfairly disadvantage certain demographic groups.

If we use data to derive knowledge and create CDSSs that orient and support clinical practice, they can be biased by the type of data that originated said knowledge [49, 50].

The importance of ethics in AI cannot be overstated, primarily because the decisions that these systems make can have profound implications on individuals and society. These decisions may affect anything from employment opportunities to legal outcomes, and increasingly, health outcomes. As AI models grow in complexity and application, they possess an enormous power that needs to be harnessed responsibly. This necessitates rigorous ethical considerations to ensure fair, unbiased, and transparent operations. Ethical lapses can result in discrimination, loss of privacy, and unjust outcomes, among other issues, which erode public trust in these technologies.

Equally important in the realm of AI is the understanding of why a model works the way it does. This concept, known as "explainability" or "interpretability", is central to AI ethics. It concerns the transparency of AI algorithms and the ability to understand and interpret their inner workings and decisions. Without this understanding, we run the risk of blind reliance on AI's 'black box' that may lead to erroneous or biased outcomes. It is critical to scrutinize AI models' reasoning processes, ensuring they align with human values and principles and are not based on inappropriate or discriminatory features.

In the context of healthcare, these considerations take on an even greater significance. AI applications in healthcare, such as diagnostic tools or treatment recommendation systems, directly impact human lives. They may influence critical decisions such as who gets treatment, what kind of treatment is administered, and when it should be given. These systems must not only be accurate but also transparent, fair, and accountable. They should be designed and implemented in a way that respects patient rights, including privacy, autonomy, and informed consent.

Therefore, in healthcare, the need for ethical AI and model explainability is not just a matter of good practice, it's a matter of life and death. Bias or errors in AI could lead to misdiagnoses or inappropriate treatment recommendations, with potentially fatal consequences. Similarly, if AI-based systems make decisions that healthcare professionals or patients can't understand, it may lead to mistrust and potential harm. The advancement of AI in healthcare must ensure ethical considerations and explainability are at the core of AI model design, development, and deployment. This will build trust in AI systems and ultimately lead to better health outcomes.

Furthermore, the informed consent of patients is another significant ethical consideration. Patients should be fully informed about how their data will be used, and they should have the right to *opt-out* if they wish. Transparency is another crucial aspect that straddles both legal and ethical dimensions. It involves explaining how decisions or predictions are made by complex algorithms, particularly when they have significant implications for patient care. For instance, if an AI model is used to prioritize patients for treatment, it should be transparent about how the model makes its decisions. The explainability of ML models can help achieve this transparency, which aids in maintaining accountability and trust.

Finally, at the moment of this writing, there are in the EU several proposals that could impact AI in general and in healthcare in specific. The Medical Device regulation could impact the deployment of AI based systems and the AI act could also impact the development of AI based systems in healthcare.

An expert is a person who has made all the mistakes that
can be made in a very narrow field.

Niels Bohr

3

Research Methods to Improve Data Quality

This chapter focuses on enhancing data quality through innovative research methods. Firstly, the utilization of synthetic data generation, as detailed in sections 3.1, 3.2, and 3.3, serves a dual purpose: expanding the volume of available data while simultaneously safeguarding privacy. This approach focuses on techniques such as GANs to create realistic, yet non-sensitive data sets. Secondly, the development of automatic data quality assessment methods, explored in section 3.4, marks a significant stride in ensuring the integrity and reliability of data. These methods aim to automate the process of evaluating data quality, thus reducing manual effort and increasing the efficiency and accuracy of data analysis.

3.1 Can GANs Help Create Realistic Datasets?

This section is based on the paper entitled "GANs for Tabular Healthcare Data Generation: A Review on Utility and Privacy". It focuses on a review of the GAN framework for creating synthetic data for healthcare. Tries to compile the metrics used for comparing and assessing synthetic data in terms of utility - or how similar they are to the original data and privacy - how protective of the patient's data it is.

3.1.1 Introduction

With the growing technological advances, the quantity of healthcare-related data produced around the world increased exponentially [51, 52]. Consequently, the potential for harvesting this data also increases. The value locked within this data could help provide better healthcare with new information about diseases, drugs, and preventive therapies. It can also help create better HISs,

meaning an overall better clinical practice [53]. But for this to happen, data must reach capable hands at the right time. But the release of clinical data has several barriers attached and rightly so. The leakage of patient's privacy can break the confidence of the population in healthcare professionals and institutions. Patient safety and privacy should be kept at all costs. However, the current mechanisms for privacy maintenance are very long, bureaucratic, and time-consuming, nationally [54], and internationally [55]. The current scenario and general methods for privacy safeguards are related to pseudo-anonymisation techniques. The removal of certain attributes, identifier modification, code grouping, or discretization are some methodologies. But not even these are totally safe [56]. Synthetic data appear as an alternative for clinical data sharing, promising great data utility with minimal privacy concerns. Synthetic data is data that is generated automatically through programmatic processes. This is especially impactful for the case at hand since synthetic data has no explicit connection with the original data. There are several mechanisms for data synthesis postulated by [57], there are process-driven methods and data-driven methods. Process-driven methods generate data through pre-determined models inputted into the generator. Data-driven methods produce new data based on inputted source data. With this, it is possible to create new patient data that has no relation to reality while providing the same statistical relations between variables. This provides the basis for quality clinical research on top of this new data. Even though these techniques are still new and in rapid development, the results seem interesting [57], but not without questions and doubts [58]. Creating a thorough survey based on the generation of synthetic data is seldom a simple task when compared to other surveys since synthetic data is present across several domains and has several uses, like software testing, assessing methods, or generating hypotheses. Moreover, synthesis has the double meaning of summing up information and generating something, easily wielding hundreds of results per query. Finally, trying to filter algorithms aimed at tabular data is also burdensome, since not always it is easy to discriminate input types. These factors make the survey interesting to focus on the state-of-the-art mechanisms of generating tabular data.

3.1.2   Theoretical background

First introduced in 2014, GANs [59] have been under the scope and have been proven very good for generating complex data. Images, text, and video have been successfully generated with very good performances. The original architecture is based on two artificial neural networks trained simultaneously in a competitive manner. One of them, the generator, has the objective of generating the most realistic possible data, while the second network – the discriminator, has the opposite aim of aiming to distinguish the realistic data from the synthetic data the best it can. So, the elegance of this architecture is that each network tries to make the other perform better every time. The GAN architecture is shown in 3.1.

The generator is represented by G_θ where the parameter θ represents the weights of the neural network. It takes as input, a Gaussian random variable, and outputs $G_\theta(Z)$. Distribution of $G_\theta(Z)$ is denoted by P_θ. The goal of the generator is to choose θ such that the output $G_\theta(Z)$ has a distribution close to the real data. The discriminator is represented by D_ω, parametrized by weights
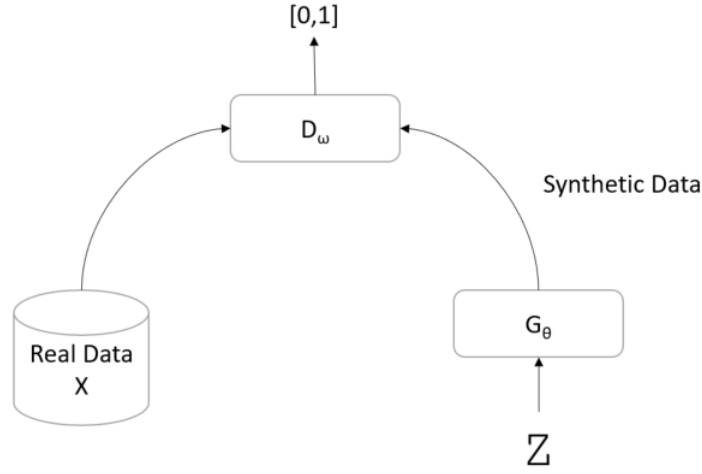
Figure 3.1: GAN framework

ω. The goal of the discriminator is to assign 1 to the samples from the real distribution P_X and 0 to the generated samples (P_θ). So, GANs can be mathematically represented by a *MinMax* game identified by:

$$\min_{G} \max_{D} E[log(D_\omega(X)) + log(1 - D_\omega(G_\theta(Z)))] \tag{3.1}$$

So, G must minimize this equation and D must maximize it, each one tweaking the weights of its network (θ and ω) to do so. This is the loss function on the initial GAN architecture. After the classification of D, the G is trained again with the error signal from D through backpropagation. This equation is the log of the probability of D predicting that the real data is genuine and the log probability of D classifying synthetic data as not genuine. The equation is essentially the same as minimising the *Jensen-Shannon Divergence* (JSD) [59]:

$$\min_{G} JS(P_x || P_\theta) \tag{3.2}$$

Where the JS means the *Jensen-Shannon Divergence* between the probability of the real data and the probability of the generated data. The JS divergence provides a measure of the distance between two probability distributions. Therefore, the minimization over θ means, choosing the P_θ that is closest to the target distribution P_X in the JS divergence distance. Despite the significant results provided by GANs with continuous real values, categorical values still seem to be a problem for this approach [60], since it is not directly applicable for calculating the gradients of latent categorical variables in order to train these networks through backpropagation. This happens since the output of the generator, even though can be transformed into a multinomial distribution with a *softmax* layer, sampling from it is not a differentiable operation, limiting the backpropagation process of the GAN.

3.1.3   Methods

This search was made between December 2020 and January 2021. It was made on "Web of Science", IEEE, PubMed, Arxiv and finally GitHub. The terms searched were related to GANs, synthetic data generation, electronic health records, patient data, or tabular data. Applications of GANs to non-tabular data were filtered, like image, sound, video, or graphs. Time series and text data were also removed since the methodology for synthesizing this type of data has specific functions related to the nature of the data. The filter for date was after 2014 since GANs were introduced at that time. The queries used were similar to the one below, adapted for the search mechanics for each website.

("generation" OR "creation" OR "synthesis" OR "synthesizing" OR "generating" OR "creating") AND ("synthetic data" OR "synthetic patient" OR "synthetic electronic health record" OR "synthetic EHR" OR "realistic patient data" OR "realistic health record" OR ("synthetic" AND "privacy" AND "utility")) AND ("GAN" OR "Generative Adversarial Network")

From the total articles found (1165) with all the queries, 100 articles were chosen for full text and in the end, 22 papers with GAN implementations that were tested on tabular data were selected.

3.1.4   Results

The selected papers ranged from 2017 to 2020. Being that 2 are from 2017, 4 from 2018, 8 from 2019 and 8 from 2020. All authors showed original GAN implementations, apart from 2 papers. Beaulieu-Jones et al. [61] used a GAN architecture that was originally published with usage on image datasets [62]. Additionally, Vega-Marquez et al. [63] used an already known implementation of conditional GANs [64]. We classified papers regarding 3 metrics: utility, privacy and clinical. For utility, we looked for methods for measuring the generated data's quality. As for privacy, we aimed for some mechanism for measuring the privacy loss of the new data. Concerning clinical metrics, any kind of evaluation from healthcare professionals was considered. This can be seen in table 3.1.

The metrics the authors used are exhibited in table 3.2. Regarding privacy, 15 papers assessed it or included some kind of mechanism to improve data protection. The most common was including Differential Privacy (DP) in the generation process. Other mechanisms for measuring privacy loss were Membership Inference (Member. Inf.), Attributes Disclosure (Attrib. Disc.), Euclidean distance (Eucl.), record-linkage (R. Linkage) and K-Nearest Neighbours (KNN). As for utility, all papers assessed it. There were 3 major areas of utility assessment: Dimension-wise (DW) probability, cross-testing, and distance metrics. The most basic one was dimension-wise probability, which is important for making sanity checks for the generated data, comparing the distributions of each column between real and synthetic. In this category, we can find *Bernoulli* (Bern.), cumulative distributions (Cumul. Dist.), *Pearson* correlation (Pearson) and *Spearman* correlation (Spearman), correlation coefficients (CCS), chi-squared test (χ^2), *Kolmogorov-Smirnov* (KS) or Correlation Matrices (Corre. Mat.). Cross-testing was about training machine-learning algorithms

Table 3.1: Summary of the articles selected. The year, acronym, reference and type of metrics mentioned are indicated. Code repository is mentioned when such information was provided.

ID	year	Acronym	Article	Metric	Code
1	2017	medGAN	[51]	Utility, Privacy, Clinical	[65]
2	2017	POSTER	[66]	Utility, Privacy	[67]
3	2018	table-GAN	[68]	Utility, Privacy	[69]
4	2018	dp-GAN	[70]	Utility, Privacy	[71]
5	2018	mc-medGAN	[72]	Utility	[73]
6	2018	TGAN	[74]	Utility	[75]
7	2019	PATE-GAN	[76]	Utility, Privacy	–
8	2019	SPRINT-GAN	[61]	Utility, Privacy, Clinical	[77]
9	2019	GAN-based	[78]	Utility, Privacy	–
10	2019	CTGAN	[79]	Utility	[80]
11	2019	WGAN-DP	[81]	Utility, Privacy	[82]
12	2019	PPGAN	[83]	Utility, Privacy	[84]
13	2019	medBGAN	[53]	Utility	–
14	2019	medWGAN	[85]	Utility	[86]
15	2020	ADS-GAN	[87]	Utility, Privacy	–
16	2020	corGAN	[88]	Utility, Privacy	[89]
17	2020	CGAN	[63]	Utility	–
18	2020	DPAutoGAN	[90]	Utility, Privacy	[91]
19	2020	GAN Boosting	[92]	Utility, Privacy	[93]
20	2020	RDP-CGAN	[94]	Utility, Privacy	[95]
21	2020	WCGAN-GP	[96]	Utility, Privacy	–
22	2020	SMOOTH-GAN	[97]	Utility	[98]

with both datasets in order to compare the results. The key factor is generating a synthetic dataset based on the training set and then training models on the original training set and the generated dataset. Then the models are compared regarding their predictive capability on the (real) test set. This was a way of assessing if the generator models were capturing inter-variable relationships. The authors applied different metrics from Area Under the Receiver Operating Characteristic Curve (AUROC), F1, Area Under the Precision Recall Curve (AUPRC), Accuracy (Acc.) to Mean Relative Error (MRE). Finally, there was also the application of distance metrics, for measuring the difference between column distribution in both datasets. *Jensen-Shannon Divergence* (JSD), *Wasserstein* Distance (WD), *Bhattacharyya* Distance (BD) or Generate Scores (GS) that was a metric implemented by the authors of [83] that creates a metric based on the sum of the mean of *kullblack-leibler* distance of all columns. Other less used methods were Principal Component Analysis (PCA), propensity score mean squared error ratio (pMSE). Normalised Mutual Information (NMI), which is the ability to capture correlations between columns by computing the pairwise mutual information and MMD (Maximum Mean Discrepancy), which is similar to distance metrics were also used. Regarding datasets utilized, the most used was MIMIC-III [99] (9 times). The papers used 27 different datasets, being 16 healthcare-related and 11 non-healthcare related. Finally, regarding clinical evaluation, only two papers assessed it, as it is possible to see in table 3.1. Both had a group of clinicians assessing a sample of both real and synthetic information and evaluating from 0 to 10, where 10 is the most realistic. One major point preventing a larger

comparison is that despite some papers using the same dataset and same methodologies, the presented values are different, making it difficult for a clear comparison of results. One example is a dimension-wise prediction with F1 score for MIMIC-III. CorGAN presents the mean difference between the two classifications (real on real and synthetic on synthetic), while medBGAN presents the correlation coefficients of the two, and medGAN only presents the visual comparisons. Regarding code availability, 16 papers had the code publicly available in some form. As of January 2021, papers pointed in table 3.1 have public code.

Table 3.2: Summary of different metrics utilised for evaluating synthetic data. Grouped by utility and privacy metrics. Acronyms indicate the source paper

Acronym	Utility	Privacy
medGAN	1. Bern. 2. Pred F1	1. Attrib. disc. 2. Memb. inf. 3. KNN
POSTER	1. Pred Acc. 2. Corre. Mat. 3. BD	DP
table-GAN	1. Cumul. Dist. 2. Pred F1|MRE	1. Eucl. 2. Member. inf.
dp-GAN	1. Pred AUC 2. Bern.	DP
mc-medGAN	1. Pred F1|AUC 2. Bern. 3. ME F1|Acc	–
TGAN	1. KNN 2. NMI 3. Pred F1	–
PATE-GAN	1. Pred AUC|AUPRC	DP
SPRINT-GAN	1. Pred AUC 2. Corre. Mat.	DP
GAN-based	1. Pred Acc. 2. Corre. Mat.	1. Hit. Rate 2. R. Linkage 3. Eucl.
CTGAN	1. Pred F1|R2|Acc.	–
WGAN-DP	1. Corre. Mat. 2. PCA 3. Pearson RMSE 4. Pred F1|RMSE|1-MAPE(F1)	1. Eucl. 2. Dupl. 3. DP
PPGAN	1. GS	DP
medBGAN	1. Assoc. Rul. 2. CCS Pred F1 3. KS	–
medWGAN	1. Assoc. Rul. 2. CCS Pred F1 3. KS	–
ADS-GAN	1. χ^2 2. JSD 3. WD 4. t-test 5. Pred AUROC 6. Corre. Mat.	DP
CorGAN	1. Pred F1 2. Bern.	Member. Inf.
CGAN	1. Pearson 2. Spearman 3. Pred F1|AUC|Acc	–
DPAutoGAN	1. Pred AUROC|R^2 2. Bern.	DP
GAN Boosting	1. pRMSE 2. Pred AUROC|AUPRC|Acc.	DP
RDP-CGAN	1. Pred F1|AUROC|AUPRC 2. MMD	DP
WCGAN-GP	1. Corre. Mat. 2. Pred F1	1. Dupl. 2. Eucl.
SMOOTH-GAN	1. DW MAE 2. Pearson 3. Pred AUROC|AUPRC	–

3.1.5   Implications for future research

From the work done in this paper, it is clear that synthetic data generation is a growing field. The increasing number of papers through the years as the growing quality in the mechanisms of generating data and assessing its quality is clear proof. It also became apparent that privacy and utility in synthetic data represent a delicate balance. The very same definition of differential privacy represents it. The compromise between privacy and utility is real and should be taken into account when creating privacy-demanding datasets. Creating statistically good tabular datasets is already possible, but that task becomes increasingly difficult if privacy concerns are added. However, privacy is also a complex subject, and the context of the setting is important for privacy assessment, which explains the different approaches for evaluating privacy protection of synthetic data. From this review, we believe that a proper evaluation of synthetic data generators in the healthcare setting with privacy concerns should at least include utility and privacy evaluations. For utility, we believe that evaluating column-wise is a nice first check but insufficient alone. For table-wise, since there is no fundamental metric for assessing the inter-column correlations between mixed-type variables, cross-testing is the best next thing. Distance metrics are nice to have and seem to have the potential for creating a table-wise metric [100], so presenting them is important. Second, for privacy evaluation, we believe that Differential Privacy in itself is not a guarantee of protection for real patients. More research and depth should be employed when presenting results for such generators; record-linkage and attribute disclosure can provide extra guarantees. Thirdly, a clinical evaluation should be done as well to understand if the synthetic patients are a reality in the clinical setting. Since the correlations could be correct but clinically (or biologically) they might not make sense. Finally, in the scope of this paper, only GANs were assessed, but there are more mechanisms for generating data, and could be interesting to assess how all of them perform on the same datasets. There are other methods for handling the mixed data types that regularly appear in clinical settings, like Variational Autoencoders (VAEs) Gaussian Mixtures, BN, and imputation mechanisms, making them excellent candidates for this assessment.

3.1.6   Conclusion

In this paper, we had the opportunity to survey the current framework for generating tabular data using GANs and which ones were already tested in the healthcare setting. We summarised the utility and privacy metrics employed, and the datasets used to measure them. We analysed the code availability and made suggestions for further work on cataloguing, comparing, and assessing synthetic health data generators. A survey with a global benchmark of methodologies, despite being arduous, could yield great results for the community and take the aim of this paper further.

3.2   How Can We Compare Two Tabular Datasets?

This section is based on the paper entitled "Dataset Comparison Tool: Utility and Privacy". This work followed the work on section 3.1, where we compiled ways of assessing the utility of syn-

thetic data. We understood that the mechanisms were far from consensual and a tool could be of use to merge all of this into a single file and report about data. Our purpose was to facilitate health data owners and legal responsible to understand how similar and protective a dataset was regarding the original one.

3.2.1 Introduction

Synthetic data can be defined as data that has no connection with a real-world phenomenon or event. It did not originate from a process in the real world, but rather a synthetic one. The idea is that synthetic data can have similar properties with real data, without needing to have an independent process for its generation. Synthetic data has been used over the years for several usages, but in healthcare is still not very used. However, this scenario seems to be changing. It can be used for several use cases namely [101]; i) Software testing, ii) educational purposes, iii) ML, iv) regulatory, v) retention, vi) secondary and vii) enhanced privacy.

Software testing relates to using synthetic data to create use cases for software testing. This can be used for the development or pre-production stages for example. Often the data needed is not available on-demand and a synthetic generator of reliable data could be useful. Educational purposes relate to, at least, two different scenarios. One is for onboarding of employees [101], the other is related to healthcare students for using health information systems and creating mechanisms for providing reliable data on-demand. ML is one of the areas where synthetic data has more widespread usage, where data augmentation through data synthesis is already common. It can be used for class imbalance, sample-size boosting, or machine-learning algorithm testing. Regulatory purposes could be important as well, with the rise of AI as medical device systems and synthetic data could be used to properly evaluate these systems under controlled environments. Retention can be an important case for synthetic data as well, since personal data must not be kept more than it would be necessary. Synthetic data generators can be of use, where the original data can be deleted and a generator kept for further usage, given that the privacy mechanisms are properly employed. Secondary uses relate to using synthetic data to share data with academia or industry. Simulacrum [102] is a nice example of how the National Health Service (NHS) uses these mechanisms to help scientists get a better grasp of data before having to fill in documentation to query the real data. The same occurs for Integraal Kankercentrum Nederland (IKNL), which has a synthetic version of the cancer registry for scientific purposes [103] and the Healthcare Products Regulatory Agency (MHRA) that uses synthetic data as well for its CPRD real-word evidence [104].

Finally, an aspect that is underlying all these applications is the promise that synthetic data can be used to improve privacy. Even though specially tweaked data generators can be used to create more privacy-aware datasets, it will be inherently always at the cost of some utility [58]. So, even though synthetic data is not the silver bullet as primarily thought, synthetic data generation can be undeniably used to help create more private data for all the use cases seen above at the cost of its utility. As for proper methods of evaluating security and utility, there are, for now, open research questions. At the present time, it is still complicated to properly assess the utility of the generated data. We have qualitative and quantitative methods. Qualitative methods are related to plots, and

quantitative are related to some value that defines an evaluation metric. These quantitative metrics can be applied to equal columns from each data set, pair of columns from each dataset or applied over the whole dataset. As for privacy metrics, the metrics rely on duplicates. Full duplicates or membership inference related.

So in this paper, we developed a data pipeline for data analysis in order to create a report for providing several metrics for data utility and privacy.

3.2.2 Methods

Table 3.3: Metrics Assessed

Metric	Method	Context
Bar Plot	visual	utility
KDE Plot	visual	utility
Heat-map	visual	utility
Pair-plot	visual	utility
KS test	column-quantitative	utility
ChiSquared Test	column-quantitative	utility
Kullback–Leibler divergence	column-quantitative	utility
Jensen-Shannon Divergence	column-quantitative	utility
Wasserstein distance	column-quantitative	utility
Entropy	column-quantitative	utility
DiscKLD	table-quantitative	utility
ContinuousKLD	table-quantitative	utility
BNLikelihood	table-quantitative	utility
BNLogLikelihood	table-quantitative	utility
GMLogLikelihood	table-quantitative	utility
Same dataset ratio	table-quantitative	utility
Support rules	table-quantitative	utility
Different dataset validation	table-quantitative	utility
Duplicates	quantitative	privacy
Duplicate minus 1	quantitative	privacy
Record Linkage	quantitative	privacy
Matrix distance	quantitative	privacy/utility
Cosine distance	quantitative	privacy/utility
Euclidean distance	quantitative	privacy/utility

The pipeline relies on Python and latex for creating the document. It relies also on several packages that implemented methods for evaluating data, namely *scipy* [105], *sdmetrics* [106] and *scikit-learn* [107] and *mlxtend* [108]. Its basis is related to uploading 2 datasets, and a report in PDF is produced. Being that is based on programmatic code, it can be easily converted into Application Programming Interface (API). The report has a section for dataset description, columns removed due to high-null, and a brief variable overview. Then a null comparison is done by column and dataset. Following this is the utility subsection. Firstly, by visual methodologies: heat maps for the correlation, bar plots for categorical, density plots for continuous, and a pair plot for

an overview. As for the quantitative utility evaluation, we divided it column-wise, pair-wise, and table-wise. The first comprehends the KS test for continuous and χ^2 test for categorical variables. Distance metrics were also applied to categorical columns. First, they are transformed into distributions and then distance metrics are applied. The results are a descriptive overview of the distance metrics, having minimum value, average, max value, and standard deviation. The distance metrics selected were JSD, *Wasserstein distance*, *Kullback–Leibler divergence*, and entropy. As for pairwise metrics, we used a discrete and continuous *Kullblack-Leibler divergence*. In this, two pairs of continuous columns are compared using *Kullback–Leibler divergence*. For this, they are put into bins for further application. The same is applied to categorical columns without binning. As for table-wise metrics, first, we used likelihood metrics. We fitted several Gaussian Mixture models or BN models to the real data and then calculated the likelihood of the synthetic data belonging to the same distribution. The metrics are likelihood for the Gaussian mixture and Bayesian models and log-likelihood for the Bayesian model as well.

Then we used machine-learning models (linear regression and decision trees) to assess how similar models behave on both datasets. First, we tested on the same dataset in order to compare evaluation metrics. Then we cross-tested, meaning that the training set was drawn from one dataset and the test set was drawn from the second dataset. Finally, data privacy constraints duplicate evaluation, duplicate existence by removal of a single column and a record linkage approach. With the record linkage, we define a record linkage blocking ("age" in the example) and then try to match rows from the synthetic dataset to the real, with varying known attributes. Then matrix, Euclidean and cosine distance was also calculated. Even though it is used for privacy evaluation, by definition, we could also use it for utility assessment. For proper assessment, the continuous and categorical variables should be indicated at the start of the code. The metrics are listed in the table 3.3.

3.2.3 Results

A trial example of comparing data is available for data in the UC Irvine Machine Learning Repository (UCI) repository, namely the heart disease dataset [109]. The synthetic data was created by using the *synthpop* package [110]. The variables evaluated are listed in table below. The code can be seen in https://github.com/joofio/dataset-comparasion-report. As an example, the image for visual analysis for categorical (figure 3.2) and continuous variables (figure 3.3).

3.2.4 Discussion & Conclusion

The data possible to create to evaluate similarities between two datasets is important not only for synthetic vs real datasets. For example, in distributed learning, where different silos exist, with similar or even equal features, a method for evaluating the similarities can be useful for understanding how the populations are similar between them, trying to shed light on the most similarities among them, or different in order to understand the differences in the silos or data
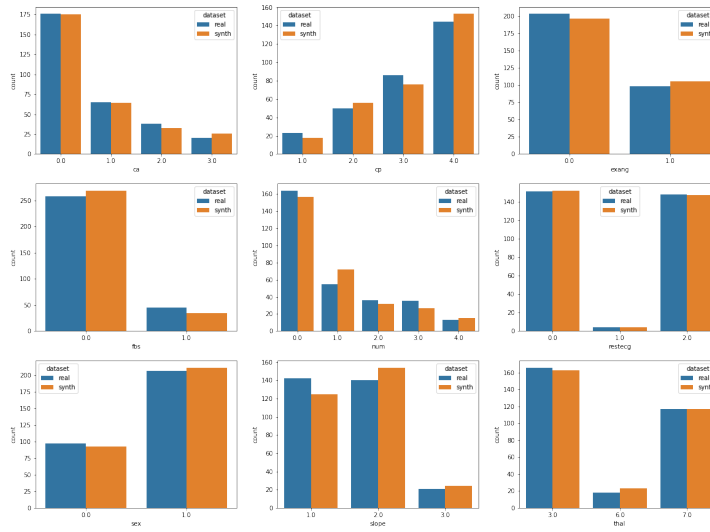
Figure 3.2: Categorical Variables plotted

acquisition inside them. Furthermore, the differences can be assessed on a more granular level. The column-wise similarities can be different from the inter-column similarities and this in itself, can be a metric of interest regarding the quality of the synthetic data and its generator.

With this work, we hope to help institutions and academics get access to a benchmark of the datasets provided in order to leverage synthetic data in the healthcare space. Finally, we hope this work helps other researchers reach an evaluation metric that could be a unique and clear response to the question of how similar two datasets are.

3.3 Can We Use Machine Learning Feature to Compare Datasets?

This section is based on the paper entitled "Using Machine Learning Models' feature importance to assess dataset similarity". The reasoning behind this paper was the results of section 3.1, where we felt that evaluation metrics for synthetic data could be improved. Better yet, we felt that the comparison of two datasets (that shared the same columns) could be done in a more robust way. Being that the current gold-standard was cross-validation which was not bound to any number range and the significance of the result could not be easily interpretable. We used the feature importance of several ML models to compare datasets and concluded that it was a valid alternative to the traditional metrics.

3.3.1 Introduction

In recent years, the use of AI and ML algorithms has gained increasing prominence in healthcare research and practice. One of the key requirements for the successful application of these methods
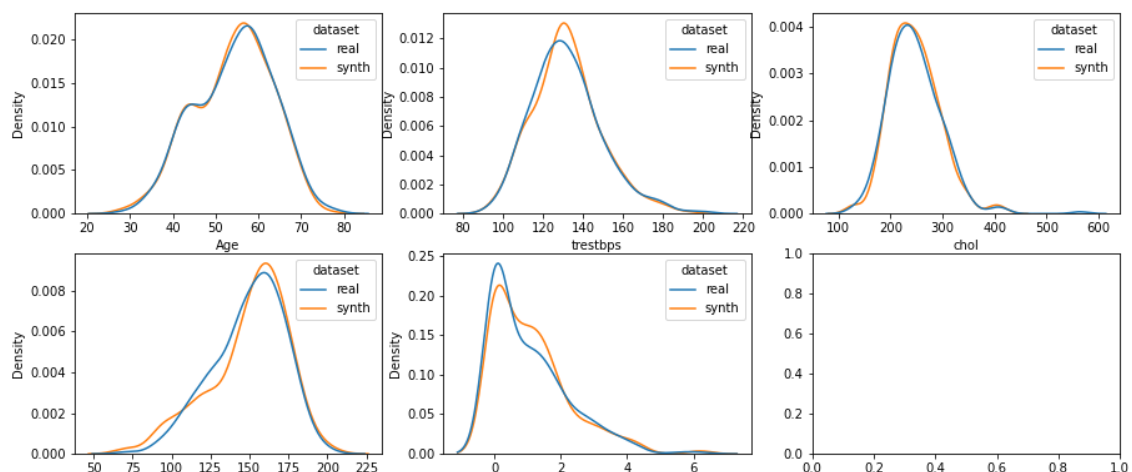
Figure 3.3: Continuous Variables plotted

is access to large, high-quality datasets. However, in many cases, the availability of such datasets can be limited due to issues around data privacy, security, and ethical concerns [111]. To address this challenge, synthetic data has emerged as a promising solution.

Synthetic data refers to artificially generated data that closely mimic the statistical properties and patterns of real-world data [112] and has many applications in healthcare. For instance, in clinical research, synthetic data can simulate patient responses to a drug, enabling researchers to conduct preliminary analyses and identify potential outcomes without risking actual patient health. In training machine learning models, it's used to augment datasets, improving the accuracy and robustness of predictive models without exposing sensitive patient information. Hospitals and healthcare providers use synthetic data for resource planning and management, creating virtual scenarios to optimize staff allocation and equipment usage. Additionally, in medical imaging, synthetic data helps in developing more accurate diagnostic tools by providing a diverse range of images, which might be scarce in real datasets, particularly for rare conditions. Each of these examples showcases the versatility and potential of synthetic data in enhancing healthcare services while safeguarding patient confidentiality.

So, with this, synthetic data offers a promising solution to several challenges inherent in real-world data, including limited data volume and privacy issues. While there are ongoing debates about its effectiveness as a comprehensive solution for privacy concerns [113], its application in data upsampling has been a well-established practice for years. The effectiveness of synthetic data generators varies considerably, making it critical to evaluate the resemblance between synthetic and real data prior to utilization. This notion of similarity, or utility, is pivotal in unlocking the full potential of synthetic data. Without understanding the degree of similarity, it's challenging to gauge its utility. Particularly in healthcare, rigorously assessing synthetic data is essential to confirm its validity in offering insightful contributions and informing decision-making processes.

Yet, the crux lies in guaranteeing the high quality and validation of synthetic data to ensure it provides dependable and meaningful insights. Quality evaluation of synthetic data involves a thor-

ough comparison of its statistical attributes and patterns against those of the original dataset. The current state-of-the-art quality assessment of synthetic data involves a detailed comparison of its statistical characteristics and patterns with those of the original dataset. This includes examining the similarity of columns through various statistical tests and exploring inter-column relationships using methods like cross-classification, where two datasets are divided into training and test sets, cross-tested, and the ratio of their evaluation results is used as a metric [112, 114]. However, this approach is somewhat limited in capturing the nuances of inter-column relationships. The challenge then is to develop a more nuanced metric that better represents the intricacies of these relationships in two distinct datasets. Additionally, since the metric is ratio-based, it is not constrained by a specific range and can exceed 1, which complicates the interpretation of the results.

To address these issues, this paper proposes the use of ML models' feature importance values as a basis for a more comprehensive and robust metric to evaluate the similarity of inter-column relationships and overall dataset congruence.

3.3.2 Rationale and Related Work

Recently there has been a series of works related to assessing how synthetic data generators behave with data like the work of Emam et. al [115] that especially focused on utility metrics for synthetic data generators. At the moment, comparing data is based on intra-columns and inter-columns relationship. The intra-column relationship is assumed as something that compares equal columns between datasets, with highly known statistical methods like chi-squared or *Kolmogorov Smirnoff* like done in the works of [116] among many others, acting more like sanity checks than anything else. Other known metrics are distance-based metrics like *Jensen-Shannon Divergence*, *Wasserstein* Distance, *Bhattacharyya* Distance or *Hellinger* distance, which are based on the calculation of the distance between distributions like seen in the works of several teams [87, 117, 118].

However, regarding inter-column relationships, the metrics applied are often very different across papers. One example of trying to capture inter-column relationship is about the use of propensity score [119, 112] where a classifier is trained to the merged datasets, with the added variable of the original dataset (i.e., 1 for real and 0 for synthetic). The model is trained, and the propensity Mean square error is the mean squared difference of the estimated probability from the average prediction. Most recently, a unified metric appeared as the sum of other metrics known as described in the work of Chundawat et. al, [120], known as TabSynDex. Other examples are likelihood of fitness like in the works of [121], coverage support [114] or very specific metrics implemented for evaluating specific data generators. However, one metric that seems to standout is Cross-Classification (CC), which takes two datasets, one that is real and a second which is synthetic, and we split both into train and test and train a machine learning model on the real data training set, then we test the model on both test sets. Then a ratio is created, rendering the actual value. This methodology is probably the gold-standard at the moment for this type of study, has some liabilities since this value can be a bit erratic, and even above one since the evaluation metric could be better on the second dataset, and we don't have a clear grasp of what that can represent in

terms of dataset similarity. The Figure 3.4 represents this in detail. Several works used this metric as the comparing metric [112].

Figure 3.4: Cross-classification of datasets
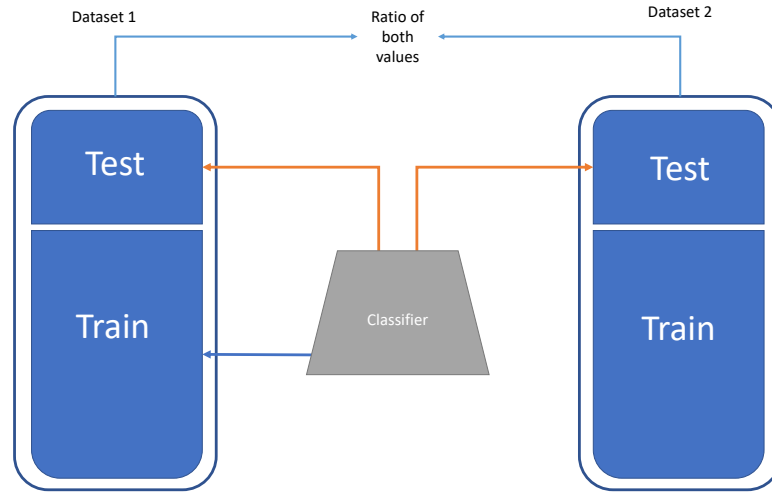
3.3.3 Materials & Methods

3.3.3.1 Materials

We used 5 datasets from the UCI dataset repository. The ones chosen were related to healthcare and were heart disease [109], thyroid disease [122], liver disorders [123], breast cancer [124] and the primary tumour dataset [125]. These datasets were chosen due to three main reasons: 1) being open and facilitating recreation of results; 2) being tabular with mixed datatypes to better represent data in the real world and 3) diversity across domains inside healthcare (different diseases and contexts).

We made minimal preprocessing on the datasets, namely removing the missing variables by imputing the mean on continuous variables and mode on categorical. We also created several synthetic datasets by applying known methods. Descriptive statistics of the datasets used are shown in table 3.4.

3.3.3.2 Method Overview

For this work, our goal is to test several metrics based on the ranking of feature importance of a trained model. Normalized Discounted Cumulative Gain (NDCG) [126] which is the sum of the true scores ranked in the order induced by the predicted scores, after applying a logarithmic discount. Then divide by the best possible score to obtain a score between 0 and 1. It is calculated by

$$\text{NDGC} = \frac{\text{DCG}(P)}{\text{IDCG}(P)}$$

where DCG(P) is the Discounted Cumulative Gain and IDCG(P) is the Ideal Discounted Cumulative Gain.

Table 3.4: Descriptive statistics of datasets used. Mean (Standard Deviation) for continuous variables. Mode [nr categories] for categorical variables.

Dataset	Column	Statistic	% Nulls	Dataset	Column	Statistic	% Nulls
heart	Age	54.4 (9.0)	0.0	liver	gammagt	38.3 (39.3)	0.0
heart	sex	1.0 [2]	0.0	liver	drinks	3.5 (3.3)	0.0
heart	cp	4.0 [4]	0.0	liver	Selector	2 [2]	0.0
heart	trestbps	131.7 (17.6)	0.0	thyroid	Class	1 [3]	0.0
heart	chol	246.7 (51.8)	0.0	thyroid	T3	109.6 (13.1)	0.0
heart	fbs	0.0 [2]	0.0	thyroid	TST	9.8 (4.7)	0.0
heart	restecg	0.0 [3]	0.0	thyroid	TSTRI	2.1 (1.4)	0.0
heart	thalach	149.6 (22.9)	0.0	thyroid	TSH	2.9 (6.1)	0.0
heart	exang	0.0 [2]	0.0	thyroid	TMAX	4.2 (8.1)	0.0
heart	oldpeak	1.0 (1.2)	0.0	tumour	class	1 [21]	0.0
heart	slope	1.0 [3]	0.0	tumour	age	2 [3]	0.0
heart	ca	0.0 [4]	1.3	tumour	sex	2 [2]	0.3
heart	thal	3.0 [3]	0.7	tumour	histologic-type	2 [3]	19.8
heart	num	0 [5]	0.0	tumour	degree-of-diffe	3 [3]	45.7
breast	Clump Thickness	4.4 (2.8)	0.0	tumour	bone	2 [2]	0.0
breast	Uniformity of Cell Size	3.1 (3.1)	0.0	tumour	bone-marrow	2 [2]	0.0
breast	Uniformity of Cell Shape	3.2 (3.0)	0.0	tumour	lung	2 [2]	0.0
breast	Marginal Adhesion	2.8 (2.9)	0.0	tumour	pleura	2 [2]	0.0
breast	Single Epithelial Cell Size	3.2 (2.2)	0.0	tumour	peritoneum	2 [2]	0.0
breast	Bare Nuclei	3.5 (3.6)	2.3	tumour	liver	2 [2]	0.0
breast	Bland Chromatin	3.4 (2.4)	0.0	tumour	brain	2 [2]	0.0
breast	Normal Nucleoli	2.9 (3.1)	0.0	tumour	skin	2 [2]	0.3
breast	Mitoses	1.6 (1.7)	0.0	tumour	neck	2 [2]	0.0
breast	Class	2 [2]	0.0	tumour	supraclavicular	2 [2]	0.0
liver	mcv	90.2 (4.4)	0.0	tumour	axillar	2 [2]	0.3
liver	alkphos	69.9 (18.3)	0.0	tumour	mediastinum	2 [2]	0.0
liver	sgpt	30.4 (19.5)	0.0	tumour	abdominal	2 [2]	0.0
liver	sgot	24.6 (10.1)	0.0				

Cohen's kappa coefficient [127] is a statistic that is commonly used to assess the level of agreement between two or more raters or evaluators who are providing categorical ratings or rankings of a set of items. So, we want to use to assess if it could be of use to check how similar the ranking of the features is, using the numbers as categorical.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed agreement between the two raters and P_e is the expected agreement between the two raters by chance.

We also intend to use the R^2 to check if the explainability changes across datasets.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i are the observed values of the dependent variable, $\hat{y}_i$ are the predicted values of the dependent variable, $\bar{y}$ is the mean of the observed values of the dependent variable and n is the number of data points.

Then we intend to use ranking metrics, namely Kendall tau, weighted Kendall tau and Rank-biased overlap (RBO). Kendall tau is a measure of correlation that measures the similarity between two rankings. It is commonly used in statistics and data analysis to evaluate the agreement or disagreement between two sets of rankings.

The Kendall tau coefficient [128] is defined as the difference between the number of concordant and discordant pairs of observations, divided by the total number of pairs. A concordant pair is a pair of observations that have the same ranking order in both sets, while a discordant pair is a pair of observations that have opposite ranking orders. The Kendall tau coefficient ranges from -1 to 1, where -1 represents perfect negative correlation, 0 represents no correlation, and 1 represents perfect positive correlation.

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\text{total number of pairs}}$$

Weighted Kendall tau [129] is an extension of Kendall tau that takes into account the importance or weight of each observation in the rankings. In some cases, some observations may be more important than others, and their positions in the ranking may have a greater impact on the overall correlation. Weighted Kendall tau assigns a weight to each observation, and the correlation is calculated based on the weighted concordant and discordant pairs.

$$\tau_w = \frac{\sum_{i<j} w_{ij} \cdot sgn(x_i - x_j)}{\sum_{i<j} w_{ij}}$$

where w_{ij} is the weight associated with the pair (x_i, x_j) and $sgn(\cdot)$ is the sign function.

The RBO [130] is a measure used to compare the similarity of two ranked lists, especially when these lists are of different lengths or have only partial overlap. The RBO value ranges from 0 (no overlap) to 1 (complete agreement). One of the key features of RBO is that it gives more weight to the top-ranked items. The formula for RBO at a given depth d is as follows:

$$RBO_d = (1-p) \sum_{k=1}^{d} \left[p^{k-1} \cdot \frac{|S_k \cap T_k|}{k} \right]$$

Where S_k and T_k are the sets of elements in the top k positions of the two ranked lists S and T respectively, $|S_k \cap T_k|$ is the size of the intersection of these top k elements, p is a persistence

parameter (usually between 0 and 1) that determines the weight given to the rankings at different depths. A lower value of *p* gives more weight to the top-ranked items and *d* is the depth to which you are calculating the RBO, which can be up to the length of the longest list. This formula calculates the RBO up to a finite depth *d*. The parameter *p* is crucial as it models the user's persistence in considering the rankings down the list. The higher the value of *p*, the more the metric considers items further down the list.

Finally, we intend to use text-distance metrics. The theory behind this experiment is to treat the ordered columns in a ranking manner and apply text-distance metrics to check the distance between the two. Levenshtein distance [131] is the minimum number of single-character insertions, deletions, or substitutions required to transform one string into another. Damerau-Levenshtein distance [131] is similar to Levenshtein distance but also includes the transposition of two adjacent characters as an allowable operation. The hamming distance [132] is a measure of the difference between two strings of equal length, defined as the number of positions at which the corresponding symbols are different. Jaro-Winkler distance [131] is a string similarity measure that takes into account the number of matching characters, the number of transpositions, and the length of common prefixes, with a higher weight given to the common prefix.

for *dataset in datasets list* **do**

    create two copies of dataset, one that remains the same and a second to be disturbed with permutation

    for *ML algorithms in ML algorithms* **do**

        for *i in number of columns to test* **do**

            for *rep in 10 repetitions* **do**

                create second dataset by permutating values in i columns in the first dataset **for** *target in dataset columns* **do**

- Train-Test Split (95:5) for both

- model fit to train for both

- get feature importance per column for both

- Create an ordered rank of features for both

- Create new metrics values by comparing the results from both

- make cross-classification to compare with feature importance metrics

- aggregate results per metric

                end

            end

        end

    end

end

Algorithm 1: Testing similarity scores in tabular datasets. Dataset list is the 5 datasets used in this work. ML algorithms are the 6 algorithms used in this work. Number of columns to test is the number of columns in the dataset. 10 repetitions is the number of times the columns are permutated.

Like seen in algorithm 1, the CC was performed several times and according to the image 3.4, we trained the model on dataset1 and tested on dataset1 and 2 and compared the results. However,

we applied this twice; 1) where dataset1 is the original dataset and the dataset2 is the permutated/synthetic (RS) and 2) where dataset1 is the permutated/synthetic and dataset2 is the real one or original (SR). Both are examples of Cross-Classification.

The algorithms chosen were decision trees with *gini* entropy function for decision making on splits on classification and squared error for regression; random forests with 100 trees and *gini* criterion for classification and squared error for regression; support vector machines with C-support Vector classification and Epsilon-Support Vector Regression, KNN with 5 neighbours and uniform weights, linear regression/logistic regression and gaussian naive bayes for classification and Bayesian ridge with 300 maximum iterations and α and λ of $1e^{-6}$. All of these were used as implemented in the *scikit-learn* package [107]. The hyperparameters chosen were the default ones. We felt that tuning was not necessary here to test our hypothesis, since it is based on the ratio of results. The text distance metrics were implemented by the text-distance package [133]. Kendall tau, weighted Kendall tau were used as implemented by *scipy* [134] and RBO, as implemented in [135]. The methods chosen for creating several synthetic datasets were the synthpop package [110] with "cart" method, which is rpart implementation of a CART model. We also used the SDV package [136] to leverage their implementation of the CTGAN and Gaussian Copula to create 2 more synthetic datasets to test different methodologies of synthetic data creation.

3.3.4 Results

With the method described in the algorithm 1, we created a figure where the metrics are presented with increasingly different datasets: Figure 3.5.

Figure 3.5: Plot showing the decrease of the metric over increasingly changed datasets. The X axis represents the number of columns mutated. The Y axis represents the value of the metric and the hue represents the algorithm used to calculate the metric.
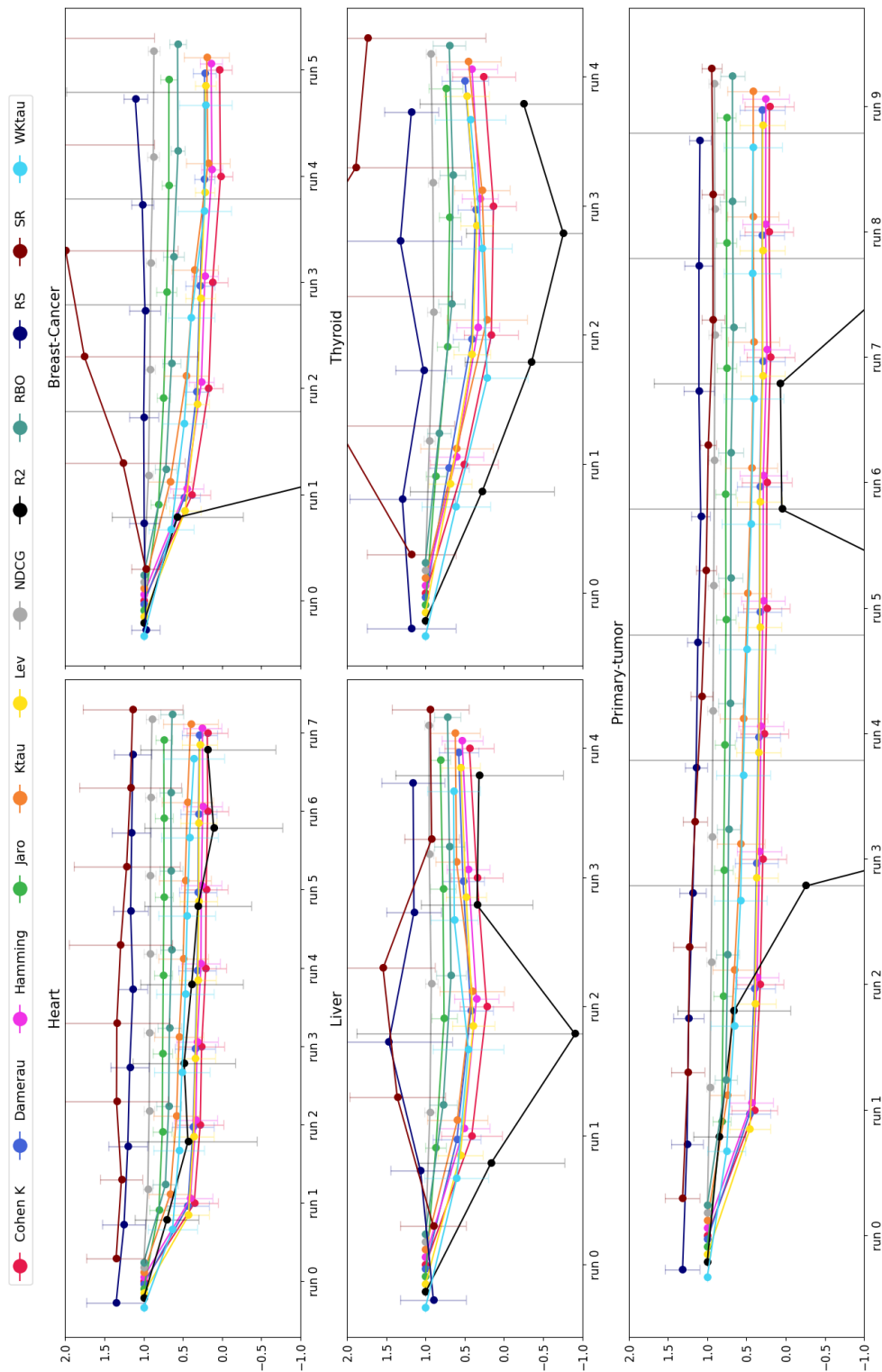
We also checked to see if the number of repetitions and how that impacts the variance of the scores. We found out that variance was low across most metrics, being the ones with higher variance was CC and R^2 with values around 0.1 and Wktau and Ktau with values around 0.04 and 0.08. The others were less than 0.02. The data can be seen in figure 3.6. As for the test for the synthetic and real dataset, the results are displayed in figure 3.8. Values available in figure 3.7. This is the metrics distribution for the comparison of the 5 mentioned datasets and the synthetic counterpart generated as stated in the methods section.

Figure 3.6: Heatmap showing the variance of different repetitions for every metric and the number of different columns changed. X is the metric. Y is the number of repetitions and the number of columns. This was obtained by getting the variance of all values from all datasets.
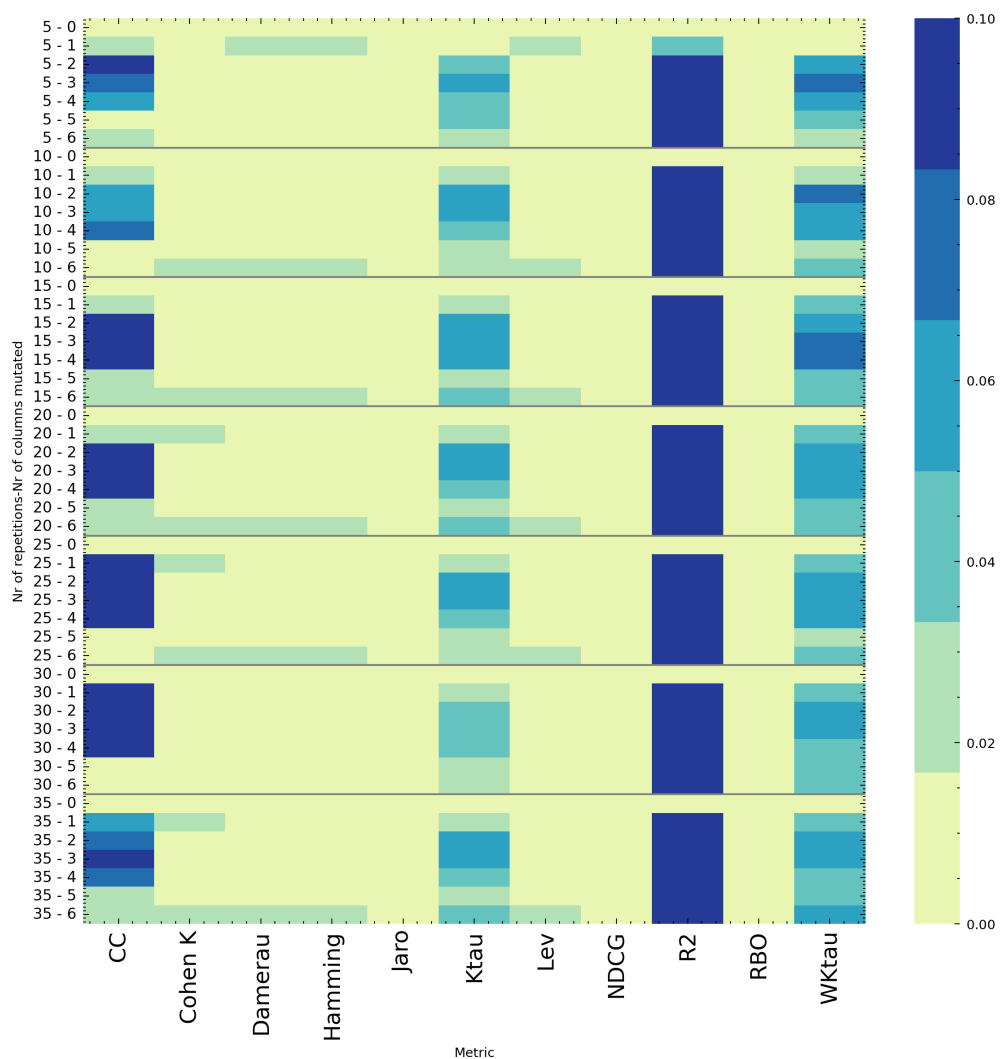
CC metrics per different models and column type was also assessed in order to help understand its advantages and disadvantages. The results can be in table 3.5. Finally, the imbalance of data was also assessed. We used a column with 21 different categories which was ideal to assess that impact. We mapped the percentage of each category per synthetic datasets and how they compare

Figure 3.7: Values and comparison of the metrics results comparing 5 synthetic and real datasets across 3 different generation methods

Method-Dataset	NDCG	Cohen k	R2	Lev	Ktau	WKtau	RBO	Damerau	Jaro	Hamming	RS	SR	CC
Gasussian-heart	0.96	0.13	0.39	0.23	0.58	0.64	0.59	0.23	0.7	0.2	1	0.94	1.6
Gasussian-breast	0.88	0.075	-0.11	0.27	0.022	0.22	0.6	0.28	0.71	0.18	1.2	0.66	3.2
Gasussian-liver	0.95	0.14	-1.5	0.38	0.52	0.56	0.72	0.45	0.82	0.29	1	0.96	0.96
Gasussian-thyroid	0.94	0.21	0.3	0.4	0.47	0.46	0.67	0.43	0.62	0.37	0.89	1.9	0.9
Gasussian-tumour	0.92	0.12	0.4	0.21	0.4	0.47	0.69	0.21	0.76	0.18	0.93	1.1	0.86
CTGAN-heart	0.95	0.065	0.38	0.16	0.55	0.57	0.55	0.16	0.67	0.14	1	0.98	1.6
CTGAN-breast	0.84	0.05	-0.11	0.26	0.15	0.21	0.6	0.27	0.74	0.16	1	0.88	4.6
CTGAN-liver	0.96	0.34	-0.054	0.52	0.49	0.62	0.69	0.52	0.8	0.45	1.1	0.99	2.3
CTGAN-thyroid	0.92	0.17	0.12	0.5	0.27	0.31	0.72	0.53	0.8	0.33	1	0.81	4.4
CTGAN-tumour	0.92	0.15	0.47	0.23	0.42	0.48	0.69	0.23	0.76	0.2	0.96	1	0.86
synthpop-heart	0.96	0.18	0.74	0.29	0.7	0.68	0.62	0.29	0.71	0.25	1	0.99	1
synthpop-breast	0.92	0.16	0.26	0.34	0.47	0.47	0.64	0.36	0.75	0.26	0.97	0.94	1
synthpop-liver	0.96	0.31	0.37	0.52	0.58	0.68	0.74	0.57	0.85	0.43	1.3	1.3	0.83
synthpop-thyroid	0.97	0.46	-0.023	0.57	0.6	0.7	0.84	0.63	0.84	0.57	0.89	2	0.88
synthpop-tumour	0.95	0.14	0.73	0.22	0.55	0.59	0.67	0.23	0.73	0.19	0.99	0.99	1.1

Figure 3.8: Distributions of the metrics results comparing 5 synthetic and real datasets across 3 different generation methods
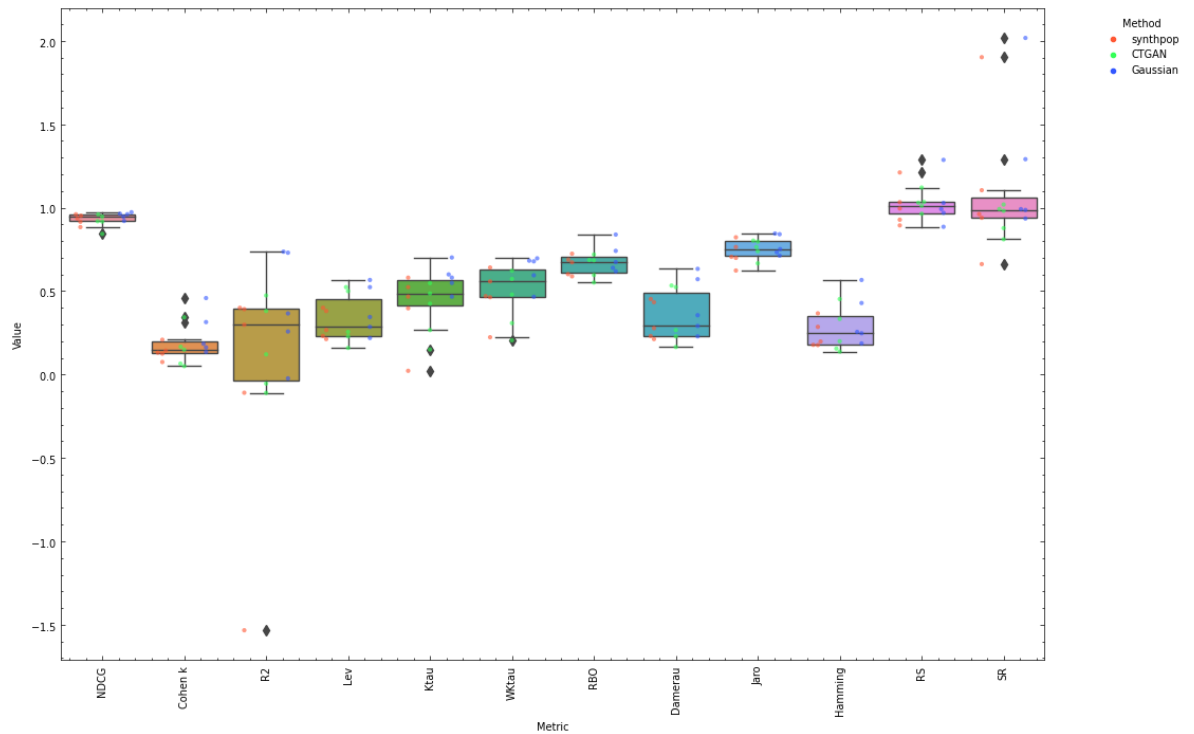
with the real dataset. This can be seen in figure 3.9.

Table 3.5: Metrics per model and variable type.

Column Type	Method	Real on Real	Real on Synth	Synth on Synth	Synth on Real
Categorical	DT	0.673 (0.161)	0.7 (0.143)	0.784 (0.141)	0.746 (0.164)
Categorical	KNN	0.674 (0.161)	0.7 (0.143)	0.783 (0.144)	0.747 (0.164)
Categorical	LM	0.672 (0.161)	0.7 (0.143)	0.783 (0.143)	0.746 (0.165)
Categorical	NB	0.671 (0.165)	0.7 (0.143)	0.783 (0.148)	0.747 (0.163)
Categorical	RF	0.675 (0.158)	0.7 (0.143)	0.782 (0.148)	0.744 (0.164)
Categorical	SVM	0.672 (0.158)	0.7 (0.143)	0.783 (0.147)	0.747 (0.164)
Continuous	DT	275 (719)	324 (840)	236 (571)	269 (612)
Continuous	KNN	316 (796)	324 (840)	244 (595)	277 (637)
Continuous	LM	283 (758)	324 (840)	242 (590)	272 (628)
Continuous	NB	288 (713)	324 (840)	243 (586)	278 (642)
Continuous	RF	324 (907)	324 (840)	242 (588)	276 (632)
Continuous	SVM	283 (727)	324 (840)	244 (594)	278 (638)

3.3.5 Discussion

With the results obtained, we believe that there are better alternatives to CC (RS/SR). RBO seems to be a viable alternative. Firstly, as with Kendall tau and Weighted Kendall tau, it appears to be

Figure 3.9: Class distributions for an highly imbalanced category across different synthetic datasets.
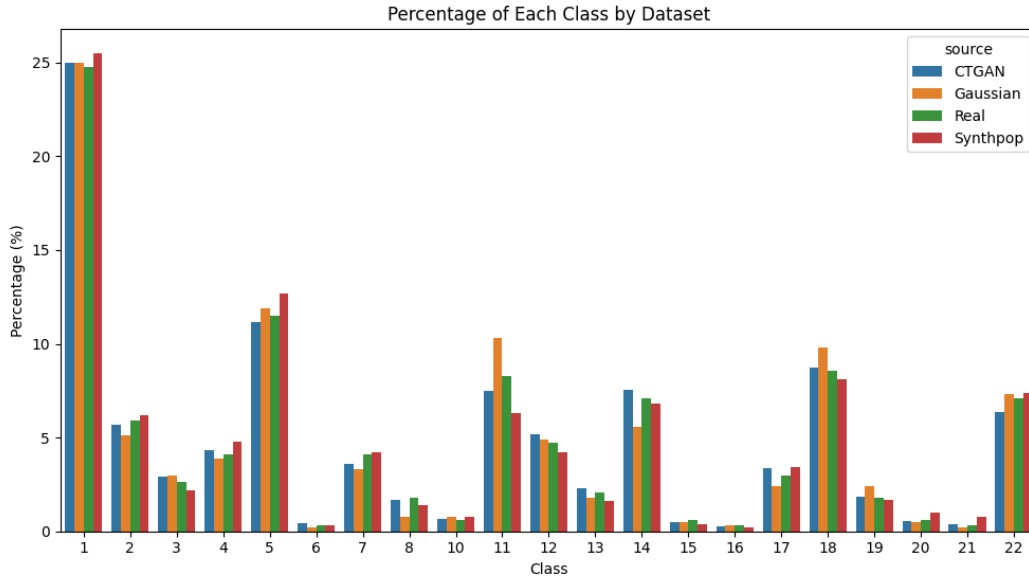
directly linked to differences in the dataset. The metric tends to decrease from 1 (identical datasets) to approximately 50% when half of the dataset's columns are altered, which is an encouraging outcome. Secondly, it is a bounded metric between 0 and 1, which aids interpretation. This is particularly useful, as metrics like CC and R^2 can fall outside these bounds, complicating the interpretation of results in some scenarios. Thirdly, variance across different iterations is also lower. Although the variance across all metrics is generally low (except for R^2), RBO exhibits less variance than Kendall tau and Weighted Kendall tau, which could be a distinguishing factor between these three metrics. CC, despite having a low variance (around 0.01), still shows higher variance compared to the other metrics (except R^2), which have variances close to zero. This is important because it allows for stable results regardless of the number of runs.

Regarding text-based metrics, there is a sharp drop when only one column is mutated (Figure 3.5), but the decline is steady thereafter. As for R^2, it is not a suitable metric for this type of problem, as it is unbounded between 0 and 1 and exhibits high variance across iterations. This is likely due to its design as a regression metric, which is not suited for this particular problem.

While proposed metrics such as RBO and Kendall tau show promise in capturing changes in feature importance, several limitations should be addressed in future research. Firstly, synthetic datasets may differ in ways that a simple value permutation cannot capture. For instance, the distribution of the data or the number of unique values may vary. Nevertheless, we tested three synthetic datasets generated by us as counterparts to the five original datasets, performing a sanity check to ensure that the new approaches align with CC for synthetic datasets (Figure 3.8). The results show that the values are at least close to those of CC, and they seem to exhibit less variability, which is reasonable given that the synthetic data generator was the same. These results suggest

that the new approaches could be good alternatives to CC.

Secondly, the metrics have primarily been tested on synthetic datasets, and their performance on real-world datasets with varying distributions, noise, and missing data remains to be fully explored. Thirdly, these metrics may not fully account for complex interactions between features, which are often present in real-world datasets. Future research could focus on incorporating feature interaction measures or combining these metrics with statistical distance metrics to better capture such complexities. Additionally, although the variance of the proposed metrics is generally low, further research into methods for reducing variance, particularly in high-dimensional datasets, would enhance their robustness. Finally, developing methods to provide confidence intervals or uncertainty estimates for these metrics would improve their interpretability and reliability in practical applications.

Regarding column types and variables, we note in figure 3.9 that at least one category was highly imbalanced throughout the process. Imbalanced data clearly affects the data generators, but more importantly, it introduces challenges when creating CC metrics. This must be considered when training and testing models, especially when encoding—such as making exceptions for unknown categories—and, more importantly, when using these variables as target variables, with particular attention to stratification of categories. Additionally, the metrics applied may also be impacted.

Examining the impact of the columns, Table 3.5 shows that the column type significantly affects the metric. This is expected, as metrics for regression are not bound by the 0-1 range. Even for categorical data, ratios higher than 1 can be observed when the model performs better on synthetic data than on the real dataset. This phenomenon occurs more easily and with greater impact in regression. While this suggests that CC also shares this limitation, it highlights the need for further research to identify which metrics and target columns are most suitable for exploration. Furthermore, considering that most real-world datasets consist of mixed data types, an effective CC method should include regression tasks and encompass several target variables. We also explored the inverse of our main approach, comparing results from testing on synthetic datasets with those from testing on both real and synthetic datasets. The behaviour was similar to the original approach, and no significant differences were observed.

In light of these findings, we believe progress has been made in comparing tabular datasets, which may prove beneficial not only for evaluating synthetic data generators but also for formalising the degree of similarity between datasets that share common features. This approach provides a valuable reference for the development of new data generators and strengthens confidence in benchmarks and comparisons.

3.3.6 Conclusion

Comparing two tabular datasets has been growing in demand in the past year mainly because of the increase in popularity of tabular data synthesis methods which have exhibited the potential in generating valuable synthetic data. However, due to the absence of a uniform metric, evaluating different methods has been inconsistent. This research proposes some alternatives for assessing

synthetic tabular data's utility. RBO seems to have the potential to capture inter-column relationships in a more consistent way than CC. They could become a useful tool for comparing statistical methods of generating synthetic tabular data. Furthermore, this metric can aid in evaluating these generators' training, providing insights into improving synthetic data quality. The proposed metrics open up possibilities for future research to enhance tabular data synthesis methods and compare two datasets overall. Future research could be expanded in new comparison with others evaluation metrics, other datasets and other synthetic data generators.

3.4 Can We Use Machine Learning to Create Automatic Data Quality Assessments?

This section is based on the paper entitled "Development and Validation of a Data Quality Evaluation Tool in Obstetrics Real-World Data through Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) interoperable Bayesian Networks and Expert Rules" This paper focuses on the fact that data quality is a major concern in healthcare. We developed a tool that could be used to assess the quality of data in a EHR and provide a report on the quality of the data. We used a combination of BN and expert rules to assess the quality of the data. Furthermore, we tested the tool on 9 real-world datasets of obstetrics EHRs and concluded that the tool was a valid alternative to the traditional methods of assessing data quality.

3.4.1 Introduction

With the wide spreading of healthcare information systems across all contexts of healthcare practice, the production of health-related data has followed this incremental behaviour. The potential for using this data to create new clinical knowledge and push medicine further is tempting [137]. However, to correctly use the data stored in EHRs, the quality of the data must be robust enough to sustain the clinical decisions made based on this data. Data quality cannot be understood as a straightforward concept; it is highly dependent on the context in which it is evaluated. The quality thresholds and dimensions required to classify the quality of the data depend on the purpose that we intend to use that very same data [138]. These uses can be very distinct and have different impacts as well. For one, we can use data to support day-to-day decisions regarding individual patients' care [139]. These decisions can include ones based on recorded information to understand a patient's history, clinical decision support systems based on this data, or even using the data to help support a more macro, public health-oriented decision. Another area is using information for management purposes. The data can be used by management bodies and regulatory authorities to extract metrics regarding the quality of care or reimbursement purposes. Thirdly, data can be used for research purposes, namely observational studies and, more recently, to support clinical trials through real-world evidence analysis [140, 139, 141]. So, all the EHR data-based decisions can only be as good as the data supporting them. Several studies have already warned about the lack of

data quality in EHRs and how this can be a significant hurdle to an accurate representation of the population and potentially lead to erroneous healthcare decisions [142, 143, 144, 145, 146, 147].

There are several steps in the data lifecycle that can be prone to error, from data generation, where the data is registered by healthcare professionals, passing by data processing, whether inside healthcare institutions or by software engineers aiming to reuse data, to data interpretation and reuse, where investigators try to interpret the meaning of registered data [141]. So, with all the data's possible uses added to the several steps that can introduce errors throughout the data lifecycle, data quality frameworks and sequential implementations can have very distinct approaches and methodologies to assess data quality. Data quality tools for checking data being registered live to support day-to-day decisions will be significantly different from one whose only purpose is to provide quality checks for research purposes. So, methodologies to tackle these issues are necessary for guaranteeing the quality of healthcare practice and the knowledge derived from EHR data.

3.4.2 Background and Related Work

There is already a significant number of papers trying to define data quality assessment frameworks for EHR data, all of them plausible and recommendable, already described in other papers [148]. The literature has over 20 different methods, descriptions, and summaries of different frameworks over the years. Some may be highlighted from the review from Weiskopf et al., [149], where five data quality concepts were identified over 230 papers: Completeness, Correctness, Concordance, Plausibility, and Currency. The work of Saez et al. defined a unified set of DQ dimensions: completeness, consistency, duplicity, correctness, timeliness, spatial stability, contextualization, predictive value, and reliability [150]. Then a review of Bian et al. [148] expanded on the previous ones, categorizing data quality into 14 dimensions and mapping them to the previous most known definitions. These were: currency, correctness, plausibility, completeness, concordance, comparability, conformance, flexibility, relevance, usability, security, information loss, consistency, and interpretability.

Finally, the work of Khan et al. tried to harmonize data quality assessment frameworks, which simplified all previous concepts into three main categories: Conformance, Completeness, and Plausibility, and two assessment contexts: Verification and Validation [151]. Conformance assesses if data values adhere to specified standards and formats. For instance, checking if a data field like 'gender' conforms to accepted values such as 'M', 'F', or 'U'. Completeness focuses on whether all necessary data values are present. An example would be checking for missing values in a critical data field like 'patient ID'. Plausibility evaluates the believability or truthfulness of data values. An example is verifying that the dates in a dataset (like birth date and date of diagnosis) follow a logical order, where the birth date precedes the diagnosis date. Despite all of these comprehensive works, there is still no consensus regarding which one is best or which has taken the lead in usage. Moreover, looking at all of the descriptions related in the literature, a significant portion of concepts are overlapping, and sometimes hard to conceptualize such dimensions in practice.

As for implementations, there are already some available, such as the work from [152] where a tool created by primary care in the Flanders was built to assess completeness and percentage of values within the normal range. The work from Liaw et al. [153] already reviewed some data quality assessment tools, like tools from OHDSI [154] or TAQIH [155]. Additionally, we found some others with similar purposes and characteristics like the work presented data dataquieR [156], an R language-based package that can assess several data quality dimensions in observational health research data. Also, the work from Razzaghi et al. developed a methodology for assessing data quality in clinical data [157], taking into account the semantics of data and their meanings within their context. Furthermore, the work from Rajan et al. [158] presented a tool that can assess data quality and characterize health data repositories. Parallel to this, Kaspner et al. created a tool called DQAStats that enables the profiling and quality assessment of the MIRACUM database, being possible to integrate into other databases as well [159].

Regarding data quality assessment as a whole, the works of [160], focused on outlier detection in large-scale data repositories. The works of [161] focused on the exploration and identification of dataset shifts, contributing to the broad examination and repurposing of large, longitudinal data sets. The works of García-de-Léon-Chocano [162, 163, 164] are the only ones focused on obstetrics data, but aimed to improve the process of generating high quality data repositories for research and best practices monitoring. These are similar and complementary works to this one. Finally, the work of [165] focused on the manipulation of EHR data, including data quality assessment, data cleaning, and data extraction. However, these tools are not meant to be used at the production level, assessing data as it is being registered or outputs reports for human consumption and not a quantitative metric for metric comparison. Furthermore, none of these tools had interoperability in mind. Finally, we have not seen, until the moment of this paper, any implementation that used ML to evaluate the correctness of the value.

3.4.3   Materials

The data was gathered from 9 different Portuguese hospitals regarding obstetric information: data from the mother, several data points about the foetus and delivery mode. The data is from 2019 to 2020. The software for collecting data was the same in every institution, and the columns were the same, even though the version of each software differed across hospitals. Across the different hospitals, data rows ranged from 2364 to 18177. The sum of all rows is 73351 rows. The data dictionary is in appendix A. This study received Institutional Review Board approval from all hospitals included in this study with the following references: Centro Hospitalar São João; 08/2021, Centro Hospitalar Baixo Vouga; 12-03-2021, Unidade Local de Saúde de Matosinhos; 39/CES/JAS, Hospital da Senhora da Oliveira; 85/2020, Centro Hospitalar Tâmega Sousa; 43/2020, Centro Hospitalar Vila Nova de Gaia/Espinho; 192/2020,Centro Hospitalar entre Douro e Vouga; CA-371/2020-0t_MP/CC, Unidade Local de saúde do Alto Minho; 11/2021. All methods were carried out in accordance with relevant guidelines and regulations. Data was anonymized before usage.

For this purpose, we took the Khan harmonized framework since we understood it as simpler to communicate we feel that the three main categories are indeed non-reducible, which makes sense from an organizational standpoint. Furthermore, the work done by Khan et al. with mapping to already existing frameworks could help compare this work with others who felt the need to use other frameworks. With this in mind, we will use three main categories, Completeness, Plausibility and Conformance. Completeness relates to missing data. Plausibility relates to how believable the values are. Conformance relates to the compliance of the data representation, like formatting, computational conformance and other data standards implemented.

With this in mind, we will use three main categories, Completeness, Conformance and Plausibility. Completeness relates to missing data. Conformance relates to the compliance of the data representation, like formatting, computational conformance and other data standards implemented. Plausibility relates to how believable the values are.

Figure 3.10: Dimensions of data quality
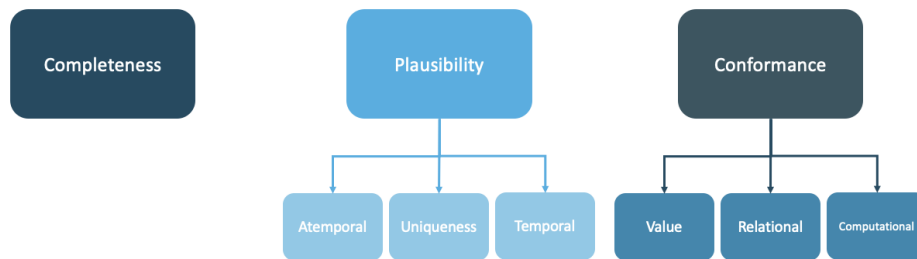
3.4.4 Methods

For completeness, we used the inverse of the percentage of nulls in the training set. For plausibility, several methods were applied. The first was a Bayesian network.

In our approach, Bayesian networks, which are probabilistic graphical models, play a pivotal role in predicting the plausibility of different elements. These networks are structured as directed acyclic graphs, where each node represents a variable and edges denote conditional dependencies among these variables [166]. This structure allows the network to efficiently manage and represent the probabilistic relationships between multiple variables. The core strength of Bayesian networks in our context lies in their ability to predict the plausibility of various elements by analysing these interdependencies. By integrating the conditional probabilities of variables and their dependencies, the network can infer the likelihood of certain outcomes or states, thereby assessing the plausibility of different columns in our dataset, when compared with the registered value.

With this, we hope to capture the heterogeneous essence of the data, as well as possible outliers that are also plausible. We chose this model for its dual advantages: its capability to classify the plausibility of all columns within a single unified framework, and its interpretability, which allows for a clearer understanding of how each variable influences the overall plausibility prediction. The networks were created with the pgmpy package [167].

Secondly, we added the outlier-tree method [168] which tries to integrate a decision tree that "predicts" the values of each column based on the values of each other column. In the process, every time separation is evaluated, it takes observations from each branch as a homogeneous cluster to search for outliers in the predicted 1-d distribution of the column. Outliers are determined according to confidence intervals in this 1-d distribution and need to have large gaps in order to be marked as outliers in the next observation. Because it looks for outliers in the branch of the decision tree, it knows the conditions that make it a rare observation relative to other observation types corresponding to the same conditions, and these conditions are always related to target variables (as predicted by them). As such, it can only detect outliers described by decision tree logic, and unlike other methods such as isolation forests, it can not assign outlier points to each observation, or detect outliers that are generally rare, but will always provide human-readable justification when it recognizes outliers. Therefore, these methods not only identify anomalies based on a single column/variable but also consider the context of the data, providing a more nuanced understanding of what constitutes an outlier. This contextual awareness ensures that the outliers are not merely statistical deviations but are also substantively significant within the specific framework of the target variables.

We added also elliptic envelope and Local Outlier Factor as complementary models to these two. Elliptic envelope is a method that assumes a Gaussian distribution of data, fitting an ellipse to the central data points to identify outliers. It works best with normally distributed data but is less effective in higher dimensions or non-normal distributions. Local Outlier Factor measures the local density deviation of a data point relative to its neighbours, identifying outliers without assuming a specific data distribution. It is versatile for different data structures but sensitive to parameter settings, like the number of neighbours.

An Inter-Quartile Range (IQR) based metric was also added as a supportive metric. This metric used the difference between Q1 and the triple of IQR to define a lower threshold and Q3 + 3IQR to define an upper threshold. We only categorized as outlier the values that fell outside these margins. Finally, a rule system was implemented to leverage domain knowledge in the overall scoring. The system is based on great expectations package [169]. A set of 17 rules was defined by the team, focusing on impossible numbers or relationship between variables or value format. The rules covered plausibility and conformance.

The Conformance-based were related to technical issues like the format of dates (date of birth like d/m/y), and conformance to the value set (i.e. Robson group, bishop scores, or delivery types). Plausibility rules were linked to expected values for BMI, weight, and gestational age (gestational age between 20 and 44). We also added plausibility for the relationship between columns, namely weight across different weeks of gestation (weight week 35 > weight week 25). We have also added a relationship of greatness between ultrasound weights more than 5 weeks apart.

As for preprocessing, all null representations were standardized, we also removed features with high missing rates (> 80%). The imputation process was performed with the median for continuous and a new category (NULLIMP) for categorical variables.

For the usage of the Bayesian network in particular, the continuous variables were discretized

into three bins defined by quantile. We defined three as the number of bins in order to reduce the number of states in each node of the network. The evaluation was done with cross-validation with 10 splits and two repetitions for each column as the target.

The API for serving the prediction models was developed with FastAPI. So, the methods applied in terms of the DQA framework shown in figure 3.10 are described in the table 3.6.

Table 3.6: Implemented Methods in the tool. The first column is the category or data quality dimension. The second is a subcategory of the first column if applicable and the third column is the actual method used to assess such a dimension.

Category	Subcategory	Method
Completeness	N/A	Score by the inverse percentage of missing in the train data
Plausibility	Atemporal Plausibility	Bayesian model prediction based on the other values of row
Plausibility	Atemporal Plausibility	Z-score for column value based on IQR train data
Plausibility	Atemporal Plausibility	Elliptic Envelope
Plausibility	Atemporal Plausibility	Local Outlier Factor
Conformance	Value Conformance	Manual Rule engine
Plausibility	Atemporal Plausibility	Manual Rule engine
Plausibility	Atemporal Plausibility	outlier-tree
Conformance	Value Conformance	Manual Rule engine

For trying to compile all of these models into a single value, that could grasp the quality of the row or patient, a scoring method was created. The method of calculating the final score is stated in figure 3.11.

Figure 3.11: Workflow and weights used for creating the final score and which elements are used to do so.
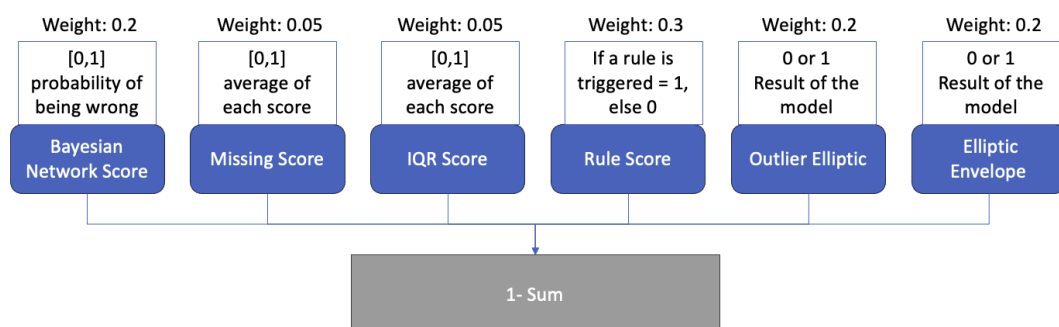
To conduct an initial validation of the tool and assess its usefulness, we implemented it in a

production environment and collected metrics regarding the data being produced. We then presented selected rows (or patient records) to obstetric clinicians, asking them to assess the likelihood that the information was suitable for use and to rank it according to the perceived quality of the record. This was done through a questionnaire, where clinicians ranked every record from 1-10 (one being the best quality one and 10 the worst quality record) and described the most important feature influencing their decision. We then compared the clinicians' rankings with the model's results to perform sanity checks on the model's performance and adequacy.

Firstly, we used Kendall's Tau and the Average Spearman's Rank Correlation Coefficient. Kendall's Tau is a non-parametric statistic that measures the strength and direction of the association between two ordinal variables, normalizing the difference between the number of concordant and discordant pairs of observations to ensure a value between -1 (perfect disagreement) and 1 (perfect agreement). Spearman's rank correlation coefficient is a non-parametric measure that assesses the strength and direction of a monotonic relationship between two ranked variables, producing a value between -1 (perfect inverse relationship) and 1 (perfect direct relationship).

Secondly, we used several thresholds to distinguish bad quality records from good quality records, transforming this into a classification problem. We assessed the AUROC for the model, taking into account the different thresholds. All the code was written in Python 3.10.6, using the *scikit-learn* library for preprocessing and initial validation [107].

3.4.5 Results

Our main result is the tool we developed that we are going to further explore its components. The main one is the Bayesian network developed, and its structure is presented in figure 3.12. The example in the image shows the probability for all the classes in the category of Robson group, taking into account the values of the other categories (both known and unknown). In this case, the probability for Robson group number 3 is 77.41%.

The results of the cross validation can be seen in the table 3.7. The average AUROC was 0.857. In parentheses is the number of non-null rows that were used in the validation for that column as target.

3.4.6 Deployment & Validation

The purpose of this model is to serve as an API for usage within a healthcare institution and act as a supplementary data quality assessment tool. Although a concrete, vendor-specific information model and health information system were initially used, our goal is to develop a more universal clinical decision support system. This system should be usable across all systems involved in birth and obstetrics departments. Therefore, we constructed it using the HL7 FHIR R5 version standard. This approach simplifies the process of API interaction. Rather than utilizing a proprietary model for the data, we based our decision on the use of FHIR resources: *Bundle* and *Observation*. These resources handle the request and response through a customized operation named "$quality_check". We intend to publish the profiles of these objects to streamline API access via

Figure 3.12: Bayesian Network learned. Nodes acronyms are explained in appendix 1. The example shows the inference for the Robson Group (10 categories) and the probability of each category, given a set of other features.
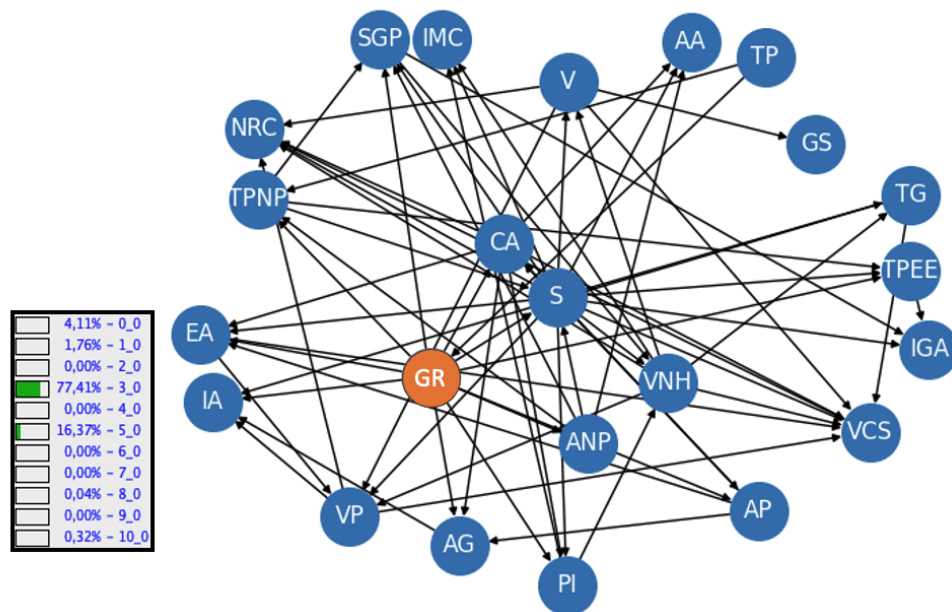
standardized mechanisms and data models. The model then makes use of the customized operation and of several base resources to construct a FHIR message, which are: *Bundle*, *MessageHeader*, *Observation*, *Device*. *Observation* is where the information about the record is contained, Device contains information about the model, and *MessageHeader* is used to add information about the request. Finally, the Bundle is used to group all of these resources together. The current version of the profiles can be accessed here [170].

For validation, we deployed the tool in docker format in a hospital to gather new data. We gathered 3223 new cases and returned a score for quality as exemplified in figure 3.13. Being that the score is from 0 to 1, the average score was 0.75 and IQR was 0.016. The formula gives weights to different dimensions since we feel some are more robust than others. We gave more weight to rule system, and gave less to the missing and IQR score.

As for the clinicians' assessment, we got 4 answers. Figure 3.14 shows the distribution of the perceived ranking of each record.

Figure 3.15 shows the performance of the model with several ranking thresholds to differentiate bad quality record from good quality record. Each line/color is a threshold (3,4,5,6) and the AUROC is shown in the label. The Average *Spearman's* Rank Correlation Coefficient was 0.42 (*P*=.23) and the Kendall's Tau was 0.3 (*P*=.2). Both tests were based on an α of 0.05.

Figure 3.13: Distribution of the trained model's scores for newly seen data retrieved from real-world scenario
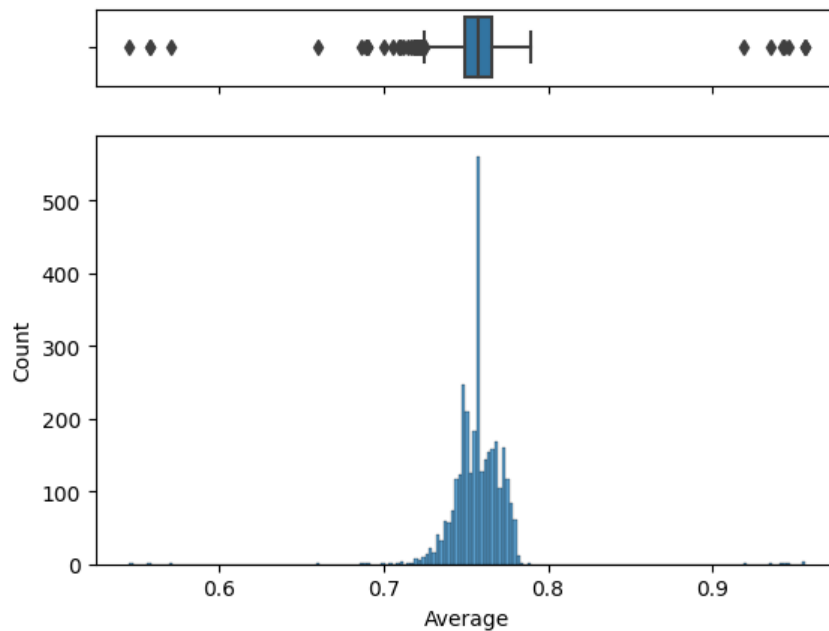
Figure 3.14: Distribution of rankings obtained from the assessment of 10 records by 4 different clinicians. Y is the distribution of clinicians' assessment, X is the patient ID.
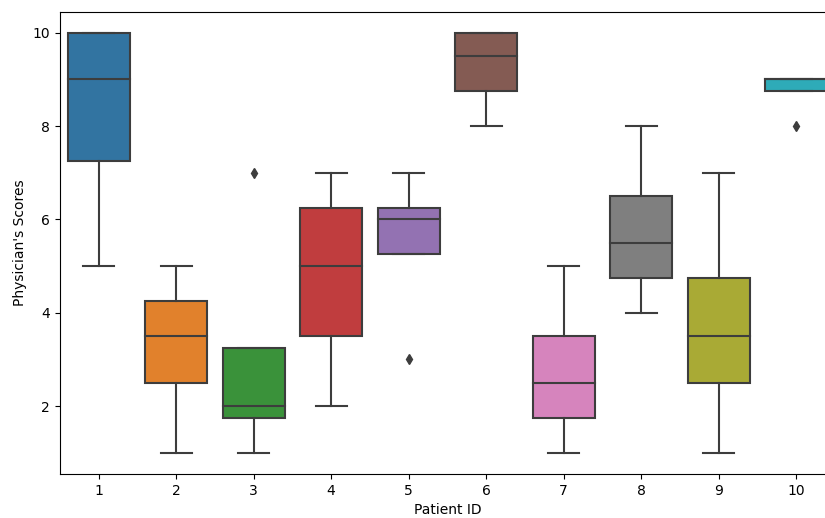
Table 3.7: Repeated Cross-Validation (10x2) Results: Column description with AUROC along with 95% CI. (n) is the number of non null rows.

Name of Variable (n)	Average	95% CI
Nr of previously born babies (44387)	0.944	[0.943, 0.945]
Nr pregnancies (73335)	0.797	[0.778, 0.816]
Nr eutotic deliveries (28809)	0.969	[0.968, 0.969]
Nr Prev. C-section (17879)	0.958	[0.958, 0.958]
Mother's Age (73337)	0.638	[0.637, 0.638]
Mother's weight start (63324)	0.881	[0.88, 0.881]
BMI (62260)	0.881	[0.881, 0.882]
Nr Prenatal Consultations (61388)	0.75	[0.75, 0.75]
Nr Weeks on admission (72715)	0.968	[0.968, 0.969]
Pregnancy weeks on delivery (73217)	0.974	[0.974, 0.974]
Nr deliveries with vacuum (15985)	0.974	[0.974, 0.974]
Pregnancy Type (64517)	0.728	[0.726, 0.73]
If pregnancy was accompanied in the hospital (49738)	0.894	[0.893, 0.895]
If delivery was spontaneous (26360)	0.816	[0.815, 0.816]
Baby's position admission (20166)	0.751	[0.743, 0.758]
Robson Group (69280)	0.931	[0.93, 0.932]
If pregnancy was accompanied (73219)	0.983	[0.982, 0.983]
Delivery Type (73350)	0.866	[0.865, 0.868]
If was accompanied in the primary care setting (49812)	0.79	[0.789, 0.791]
Baby's position delivery (73227)	0.942	[0.938, 0.946]
Blood Group (73132)	0.514	[0.507, 0.52]
Hospital ID (73352)	0.896	[0.896, 0.897]
If accompanied in a private care setting (18049)	0.771	[0.77, 0.772]
Actual Type of delivery (65606)	0.952	[0.951, 0.952]
Average 0.857 [0.846, 0.868]		

3.4.7 Discussion

This work adds several pieces of information to the state of the art of data quality analysis. First we tried to map the output of an automatic assessment tool to the human perception of quality and the issues linked to doing so. Secondly, the fact that we applied XAI methods such as bayesian networks to leverage the potency of advanced data analysis without compromising interpretability and explainability. Furthermore, a single model was able to reach high performance metrics for almost all variables. Thirdly, the fact that interoperability standard such as FHIR can be adopted to facilitate the usage and information exchange of such tools. However, there are also shortcoming and challenges to address. The first is that data quality is still an elusive concept since it has a contextual dimension and the quality of the record depends on the usage of the information,

Figure 3.15: Model Performance in terms of AUROC, depending on the threshold defined on the physician assessed data. The colours show different threshold used to consider a bad quality record given the average ranking. Label shows the threshold and respective AUROC.
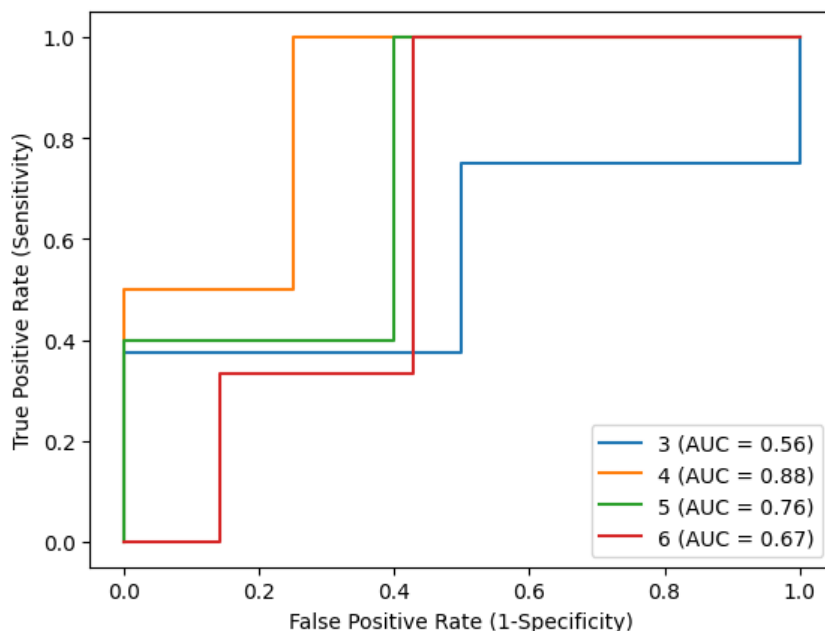
For example, data aimed at primary usage and day-to-day healthcare decisions about a patient will have different requirements regarding the importance of some variable or completeness of information very different from data needed to create summary statistics for key performance indicators extraction. Moreover, the data is still very vendor-specific. Even though we used an interoperability standard, the semantic layer, more connected with terminology is still lacking. This is an issue to be addressed in order to improve the interoperability of the standard. Moreover, we do not know how the training done with this data is generalizable to other vendors. One opportunity arises of mapping all of this data to a widely used terminology like SNOMED CT or LOINC. Nevertheless, the usage of FHIR and the fact that the data is mapped to a standard terminology, makes it easier to use the data in other systems and to compare the results with other studies. Furthermore, being available freely and online makes it easier to understand how to map vendor-specific datasets to the model and use it in other contexts. Regarding the model, the usage of explainable methodologies like outlier-tree and transparent models like Bayesian networks are vital for clinical application. Since we use a single model to classify possible errors in the records, the ability to try to show clinicians why that value was tagged is of uttermost importance in order to get feedback and action from humans. From the experience gathered with the study, we believe that a weaker but transparent model could have more impact than better performant but opaque ones. If explainability and interpretability are important for any ML problem, this need only increases when we are dealing with such subjective concepts as data quality.

Regarding the clinical evaluation, we found that asking clinicians to purely assess the quality of a record in an EHR is not an easy task. We discovered that for a proper assessment, a context

and objective must be defined in order to make the evaluation more objective and manageable. Moreover, the ranking methodology, though very useful for comparison with the model, presents challenges for clinicians who find it difficult to order 10 records when some appear to be of equal quality. This is a very important aspect to consider when designing an evaluation method for data quality. Perhaps a categorical evaluation of yes/no would be more effective than ordering several records. These reasons might explain the great variability between clinicians (figure 3.14) and between clinicians and the model (Spearman and Kendall tau). Despite that, our preliminary results are promising, demonstrating an AUROC curve for categorizing bad quality records as high as 88% and low as 56%. The highest value was achieved by classifying all record with a mean rank of 4 or above as bad quality and the others as good quality records. However, these results rely on very few samples, so more data and research are needed in this area since it is a very subjective decision, and it should take into account the context and the objective of the evaluation. For example, if the objective is research use, the weights given to each dimension can be a set. On the other hand, if the objective is to use the data for day-to-day clinical decisions, another set of weights could be used.

For the next steps, a promising research direction would be identifying contexts for applying data quality checks like primary usage, research purposes, and aggregated analysis for decision-making among others. This could enhance targeting those contexts and understanding the importance of each variable for those use cases. Incorporating this approach into the tool to weigh the different variables according to the context would be beneficial. Finally, gaining access to more data and clinician evaluation of records, although challenging, is important to thoroughly assess the performance of the tool.

3.4.8 Conclusion

We believe the work done is already a valuable insight into how to use data quality frameworks and several statistical tools in order to assess EHR data quality in real time. This is a fundamental process not only to guarantee the quality of data for primary usage but also for securing quality for secondary analysis and usage. We believe the fact that we created an interoperable tool that was trained on real obstetrics data from 9 different hospitals and has the ability to provide a single score for a clinical record can help institutions, academics, and EHR vendors implement data quality assessment tools in their own systems and institutions. With the further evaluation of the score and its relationship with clinical usefulness and a further assessment of a threshold for the score for defining a record that would require human attention would be vital to apply this tool in production with high levels of trust and quality.

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'

Isaac Asimov

4

Assess Health Data Science Methods With Limited Data Access

This chapter delves into the assessment of health data science methods under the constraints of limited data access, as outlined in sections 4.1 and 4.2. It emphasizes the importance of developing and evaluating data science techniques that can operate effectively even when direct access to comprehensive datasets is restricted. In section 4.1, the focus is on distributed data approaches, which allow for the analysis of health data across multiple locations without the need to centralize the information. This method is crucial for maintaining privacy and security, especially in sensitive health data contexts. Meanwhile, section 4.2 discusses benchmarking strategies for these methodologies, providing a framework to evaluate their effectiveness and reliability. These benchmarks are essential to ensure that the methods yield accurate and useful insights, despite the limitations in data accessibility.

4.1 Leveraging Distributed Systems in Healthcare: is it Advisable?

This section is based on the paper entitled "Evaluating distributed-learning algorithms on real-world healthcare data". This paper was focused on the fact that access to healthcare data is often laboursome and time-consuming. So we evaluated the distributed paradigm to its gold-standard, the centralised paradigm. We used 9 real-world datasets of obstetrics EHRs and compared the performance of several ML algorithms in both paradigms. We concluded that the distributed paradigm is a valid alternative to the centralised paradigm, with the added benefit of not requiring heavy data sharing.

4.1.1   Introduction

As the use of AI is increasing in the healthcare space [171], increased demand for ethical usage of personal patient data is occurring as well [172]. This has been happening both on the governmental side, with several regulations passed to protect citizens' data and personal information (such as GDPR in the EU [173] and HIPAA in the USA [174]), and on the public side, with an increased concern with continuous data breaches across institutions [175]. So, we are now faced with a dilemma on a compromise between what is possible to do with the available data and what should be done regarding patient privacy [176]. This is the main reason why health institutions implement burdensome processes and methodologies for sharing patient data, often costing a great deal of time, money, and human resources, seldomly overtaking the ideal time frame for analysing such data. Due to these privacy concerns, the traditional method for using data in healthcare is, nowadays, by focusing on data from a single institution in order to predict or infer something regarding those patients; this could be understood as local learning. This approach has some drawbacks, namely data quantity, data quality and possible class imbalance [177], never quite raising into its full potential for promoting the best healthcare practices [178, 179] with data sharing between institutions. In order to overcome this issue, there are a few, more complex, systems that consolidate data from several institutions, so more robust algorithms could be trained. However, this globally centralised consolidation of data encompasses a very important data breach hazard.

This is the setting where distributed learning could create a greater impact. A halfway point between local and centralised learning is where we train several models, one in each institution (or silo), and where the sole information that leaves the premises is a trained model or its metadata. A distributed model is built as the aggregation of all the local models, consequently aiming to create a model similar to one globally trained with all the data in a centralised server. However, the distributed model never contacted with any data, only the local models did. This provides the opportunity to create better models, improve data protection, reduce training time and cost and provide better scaling capabilities [180].

While numerous multi-institutional initiatives have successfully established integrated data repositories for healthcare research, there remains an incomplete understanding of the performance and scalability of distributed systems when directly compared to traditional, centralised models. Specifically, the nuanced behaviours of these distributed frameworks under real-world data conditions—contrasted against classical models that utilize consolidated data—have yet to be fully delineated. This paper aims to critically evaluate the efficacy and suitability of distributed mechanisms within the healthcare domain, assessing their potential as viable alternatives to conventional machine-learning pipelines. The contributions of this paper include:

- Evaluate a distributed model against its local counterparts;
- Measure the prediction performance difference between a distributed model and a centralised one;

4.1.2 Theoretical background and Related Work

Distributed learning [181] can be understood as training several models in a different setting and then aggregating them as a whole. There are two main branches of these approaches, distinguishable by the existence of a central orchestrator server: federated learning where such an entity exists, and peer-to-peer (or swarm) [176] learning where it does not.

Even though distributed learning has been receiving a lot of attention recently, only some of its concepts have been focused on, mainly distributed-deep learning with a federated learning approach [182, 183]. These methods use the strength of neural networks and several algorithms such as federated averaging to create distributed models capable of handling complex data like text, sound, or image [184]. However, considering that there are great amounts of information, especially in healthcare, stored as tabular data [155, 185, 186, 187, 188] and that neural networks are often not the best tool for such data structures and often outperformed by boosting algorithms and tree based models [189, 190], there is a lack of knowledge in the traditional ML techniques in a distributed manner. This is especially important since tabular datacomes mainly from EHRs and this kind of data is often of lower quality, with missing values, and with a high number of categorical variables and unstructured/semi-structured variables which make the application of classical machine-learning algorithms harder than for example images, which are mainly computer and systematically generated [5].

Nevertheless, there have been some health-related distributed machine-learning projects successfully implemented, such as euroCAT [191] which implemented an infrastructure across five clinics in three countries. SVM models were used to learn from the data distributed across the five clinics. Each clinic has a connector to the outside where only the model's parameters are passed to the central server which acts as a master deployer regarding the model training with the radiation oncology data. Also, ukCAT [192] did similar work, with an added centralised database in the middle, but the training being done with a decentralised system. There are also reports of a study that introduces "confederated machine learning" for modelling health insurance data that is fragmented both horizontally (by individual) and vertically (by data type), without the need for central data consolidation. It showcases the method's efficacy in predicting diseases like diabetes and heart conditions across data silos, achieving notable prediction accuracy, thereby advancing federated learning in healthcare by accommodating complex data separations and enhancing model training without compromising patient privacy or data security [193]. Distributed initiatives have also been covered in a review by Kirienko et al., [194], where we can see very few papers have described a distributed learning approach without federation. However, from these, we can highlight the works of Wang et al. [195] tried to use these approaches to detect re-hospitalization for heart failure and Tuladhar et al., [181] where they used the distributed approach to detect several diseases like diabetes, heart disease, and mild cognitive impairment.

Finally, a few works have explored the evaluation of models in a distributed manner, for example comparing centralised ML and distributed ML on MNIST dataset [196]. Also, works that evaluate federated learning on MNIST, MIMIC-III and PhysioNet ECG datasets, but not in com-

parison with other methods [197]. The work by Tuladhar and colleagues [181] uses healthcare images and/or public and curated datasets. Furthermore, these findings are supported by a scoping review which clearly states that proper evaluation of distributed/federated when compared to local methodologies and models [198]. With this, as far as we know, this is the first time that an evaluation of distributed ML has been conducted using real-world tabular clinical data from several real (9) different sources, in such a large scale of algorithms and outcome variables and compared to centralised and local counterparts, which can be applied to both federated and peer-to-peer approaches.

4.1.3 Materials

Clinical data was gathered from nine different Portuguese hospitals regarding obstetric information, pertaining to admissions from 2019 to 2020. This originated nine different files representing different sets of patients but with the same features associated to them. The software for collecting data was the same in every institution (although different versions existed across hospitals) - ObsCare [199]. The data columns are the same in every hospital's database. Each hospital was considered a silo and summary statistics of the different silos are reported in the tables 4.1 and 4.2. The data dictionary is in appendix A.1. The datasets were anonymized and de-identified prior to analysis and each hospital was assigned a number to ensure confidentiality. Each dataset represents a different hospital, which we will use for this analysis as a isolated silo and the number of patients in each dataset is reported in the last row of the tables 4.1 and 4.2. Dataset comprised of patient's features like age and weight and characteristics as well, like if the patient smoked during pregnancy or had gestational diabetes. The dataset also comprises information about the pregnancy like number of weeks, type of birth, bishop score (pre-labour scoring system used to predict the success of induction of labor), or if the pregnancy was followed by a specific physician in a specific scenario.

This study received Institutional Review Board approval from all hospitals included in this study with the following references: Centro Hospitalar São João; 08/2021, Centro Hospitalar Baixo Vouga; 12-03-2021, Unidade Local de Saúde de Matosinho; 39/CES/JAS, Hospital da Senhora da Oliveira; 85/2020, Centro Hospitalar Tamega Sousa; 43/2020, Centro Hospitalar Vila Nova de Gaia/Espinho; 192/2020, Centro Hospitalar entre Douro e Vouga; CA-371/2020-0t_MP/CC, Unidade Local de saúde do Alto Minho; 11/2021. All methods were carried out in accordance with relevant guidelines and regulations.

Table 4.1: Silos overview. Each hospital is considered a silo. Categorical columns have the number of categories (C) and the percentage of the most frequent (%). Continuous variables have a mean (μ) and standard deviation (σ). The first row is the number of patients. Bold columns were used as target (n=19).

Variable	Silo 1	Silo 2	Silo 3	Silo 4	Silo 5	Total
N (total)	8039	8566	4989	2364	18177	80874
Actual Type of Delivery C (%)	10 (52.6)	3 (51.6)	3 (57.8)	3 (61.8)	9 (61.5)	11 (52.9)
Bishop Score C (%)	15 (98.5)	15 (78.8)	13 (97.4)	16 (86.4)	15 (97.4)	16 (95.3)
Blood Group C (%)	9 (39.9)	10 (39.9)	9 (39.3)	11 (37.9)	10 (40.9)	14 (40.5)
Body Mass Index $\mu(\sigma)$	25.2 (8.6)	25.2 (6.2)	25.0 (5.3)	25.0 (8.9)	24.9 (7.8)	25.1 (7.0)
Cervical Consistency C (%)	4 (98.6)	4 (83.4)	4 (99.3)	4 (87.4)	4 (97.5)	4 (96.5)
Cervical Position C (%)	4 (98.6)	4 (83.3)	4 (99.3)	4 (87.5)	4 (97.6)	4 (96.6)
Delivery Type C (%)	6 (43.4)	6 (53.5)	5 (44.4)	7 (52.2)	7 (49.3)	8 (51.3)
Dilatation C (%)	5 (98.5)	5 (83.1)	5 (99.3)	5 (87.2)	5 (97.5)	5 (96.5)
Effacement C (%)	5 (98.6)	5 (83.2)	5 (99.3)	5 (87.2)	5 (97.5)	5 (96.5)
Fetal Station C (%)	5 (98.6)	5 (83.3)	5 (99.3)	5 (87.9)	5 (97.5)	5 (96.6)
Followed physician C (%)	3 (99.2)	4 (92.2)	3 (99.1)	3 (94.3)	3 (99.0)	4 (97.9)
Followed physician hospital delivery C (%)	2 (87.6)	2 (75.8)	2 (81.4)	2 (52.2)	2 (71.0)	2 (69.0)
Followed physician primary care C (%)	2 (61.3)	2 (52.8)	2 (78.1)	2 (50.4)	2 (70.4)	2 (67.6)
Followed physician private clinic C (%)	2 (81.8)	2 (85.0)	2 (80.6)	2 (78.8)	2 (73.3)	2 (75.8)
Gestational Diabetes C (%)	2 (87.7)	2 (90.0)	2 (90.2)	2 (90.8)	2 (89.8)	2 (89.5)
Induced Delivery C (%)	2 (97.8)	2 (83.9)	2 (93.3)	2 (91.9)	2 (98.5)	2 (92.5)
Mother Age $\mu(\sigma)$	31.1 (5.7)	30.7 (5.6)	31.1 (5.9)	31.1 (6.3)	31.3 (5.6)	31.1 (5.6)
Nr Deliveries forceps C (%)	4 (99.2)	3 (83.3)	4 (94.3)	4 (95.8)	3 (60.1)	5 (82.6)
Nr Deliveries no assistance C (%)	10 (74.7)	9 (60.3)	9 (74.9)	9 (67.3)	11 (45.4)	12 (60.3)
Nr Deliveries vacuum C (%)	5 (90.4)	4 (79.9)	4 (89.0)	4 (93.1)	5 (55.3)	5 (77.4)
Nr of C-sections C (%)	6 (87.9)	6 (72.6)	5 (86.1)	5 (89.5)	6 (62.1)	6 (74.6)
Nr of Pregnancies C (%)	11 (40.9)	11 (43.1)	13 (39.1)	12 (38.7)	16 (42.8)	19 (42.1)
Nr of born babies C (%)	10 (44.8)	10 (41.4)	10 (36.9)	10 (42.0)	12 (35.3)	12 (38.8)
Nr of consultations $\mu(\sigma)$	7.3 (4.7)	7.0 (6.4)	6.4 (3.9)	5.5 (3.6)	10.5 (5.1)	8.4 (5.1)
Pelvis Adequacy C (%)	4 (95.4)	4 (77.7)	4 (90.1)	3 (96.9)	4 (81.2)	4 (82.6)
Position Admission C (%)	5 (88.5)	6 (78.0)	6 (51.8)	3 (95.9)	6 (71.3)	7 (73.1)
Position on Delivery C (%)	5 (91.5)	5 (94.4)	5 (94.7)	5 (95.5)	5 (94.3)	5 (93.9)
Pregnancy Type C (%)	7 (62.1)	7 (90.5)	7 (85.4)	7 (63.0)	7 (89.2)	7 (85.4)
Robson Group C (%)	11 (22.4)	11 (20.1)	10 (23.8)	10 (80.5)	11 (27.7)	11 (24.4)
Rupture amniotic pocket before delivery C (%)	2 (91.1)	2 (93.6)	2 (89.3)	2 (91.6)	2 (84.6)	2 (88.5)
Smoker C (%)	2 (84.4)	2 (85.2)	2 (87.2)	2 (89.7)	2 (87.9)	2 (88.1)
Spontaneous Delivery C (%)	2 (70.3)	2 (74.7)	2 (64.8)	2 (64.3)	2 (59.7)	2 (64.9)
Weeks on Admission C (%)	38.1 (3.5)	38.8 (2.2)	38.9 (1.6)	38.8 (2.4)	38.6 (2.1)	38.7 (2.2)
Weeks on Delivery $\mu(\sigma)$	38.5 (2.8)	38.9 (2.0)	39.1 (1.7)	39.0 (2.3)	38.9 (2.0)	38.9 (2.0)
Weight on Admission $\mu(\sigma)$	81.4 (14.9)	79.5 (14.5)	78.0 (15.2)	79.6 (16.3)	78.3 (14.2)	78.8 (14.5)
Weight start of pregnancy $\mu(\sigma)$	66.4 (14.4)	66.1 (13.5)	65.5 (14.1)	65.5 (14.1)	65.5 (14.4)	66.0 (14.1)

Table 4.2: Silos overview part 2. Each hospital is considered a silo. Categorical columns have the number of categories (C) and the percentage of the most frequent (%). Continuous variables have a mean (μ) and standard deviation (σ). Abbreviation meaning in the appendix. The first row is the number of patients. Bold columns were used as target (n=19).

Variable	Silo 6	Silo 7	Silo 8	Silo 9	Total
N (total)	12002	8258	6693	11786	80874
Actual Type of Delivery C (%)	10 (63.8)	0 (100)	10 (50.1)	9 (64.6)	11 (52.9)
Bishop Score C (%)	14 (99.3)	15 (97.9)	14 (99.2)	15 (95.0)	16 (95.3)
Blood Group C (%)	13 (41.6)	10 (39.2)	10 (40.1)	10 (41.7)	14 (40.4)
Body Mass Index $\mu(\sigma)$	24.9 (5.1)	24.9 (7.0)	24.8 (8.0)	25.7 (5.6)	25.1 (7.0)
Cervical Consistency C (%)	4 (99.5)	4 (99.7)	4 (99.5)	4 (96.9)	4 (96.5)
Cervical Position C (%)	4 (99.5)	4 (99.7)	4 (99.5)	4 (96.9)	4 (96.5)
Delivery Type C (%)	6 (54.3)	5 (52.1)	5 (47.8)	5 (59.0)	8 (51.3)
Dilatation C (%)	5 (99.5)	5 (99.7)	5 (99.5)	5 (96.9)	5 (96.5)
Effacement C (%)	5 (99.5)	5 (99.7)	5 (99.5)	5 (96.9)	5 (96.5)
Fetal Station C (%)	5 (99.5)	5 (99.7)	5 (99.5)	5 (96.9)	5 (96.6)
Followed physician C (%)	3 (96.9)	3 (99.4)	3 (97.8)	3 (99.2)	4 (97.8)
Followed physician hospital delivery C (%)	2 (62.1)	2 (63.2)	2 (69.4)	2 (83.1)	2 (69.0)
Followed physician primary care C (%)	2 (53.1)	2 (86.7)	2 (63.1)	2 (87.3)	2 (67.6)
Followed physician private clinic C (%)	2 (68.2)	2 (73.5)	2 (71.0)	2 (78.1)	2 (75.8)
Gestational Diabetes C (%)	2 (92.2)	2 (88.2)	2 (89.9)	2 (86.8)	2 (89.5)
Induced Delivery C (%)	2 (91.9)	2 (85.9)	2 (87.4)	2 (93.9)	2 (92.5)
Mother Age $\mu(\sigma)$	31.3 (5.2)	31.4 (5.4)	31.5 (5.6)	30.1 (5.6)	31.1 (5.6)
Nr Deliveries forceps C (%)	4 (82.0)	4 (86.0)	3 (94.0)	4 (89.3)	5 (82.6)
Nr Deliveries no assistance C (%)	8 (58.8)	9 (61.2)	10 (68.9)	9 (61.5)	12 (60.3)
Nr Deliveries vacuum C (%)	4 (78.9)	4 (81.6)	4 (88.0)	5 (82.3)	5 (77.4)
Nr of C-sections C (%)	6 (69.1)	6 (74.5)	5 (85.5)	6 (77.8)	6 (74.6)
Nr of Pregnancies C (%)	13 (44.2)	9 (42.9)	11 (42.0)	13 (40.2)	19 (42.1)
Nr of born babies C (%)	9 (38.4)	9 (42.6)	10 (41.2)	10 (43.2)	12 (38.8)
Nr of consultations $\mu(\sigma)$	6.8 (4.0)	7.7 (3.2)	9.3 (4.5)	8.9 (5.5)	8.4 (5.1)
Pelvis Adequacy C (%)	4 (89.6)	4 (52.9)	3 (93.1)	4 (81.5)	4 (82.6)
Position Admission C (%)	6 (84.5)	7 (61.3)	5 (89.2)	4 (74.2)	7 (73.1)
Position on Delivery C (%)	5 (93.0)	5 (93.6)	5 (94.8)	5 (94.2)	5 (93.9)
Pregnancy Type C (%)	7 (88.0)	7 (85.4)	7 (86.0)	7 (92.9)	7 (85.4)
Robson Group C (%)	11 (27.2)	11 (24.7)	11 (21.4)	11 (26.7)	11 (24.4)
Rupture amniotic pocket before delivery C (%)	2 (85.0)	2 (84.4)	2 (89.9)	2 (93.8)	2 (88.5)
Smoker C (%)	2 (91.0)	2 (90.7)	2 (85.5)	2 (89.9)	2 (88.1)
Spontaneous Delivery C (%)	2 (64.9)	2 (64.0)	2 (64.7)	2 (62.9)	2 (64.9)
Weeks on Admission $\mu(\sigma)$	38.7 (1.8)	39.0 (2.0)	38.6 (2.1)	38.8 (1.9)	38.7 (2.2)
Weeks on Delivery $\mu(\sigma)$	38.8 (1.8)	39.2 (1.7)	38.7 (2.0)	39.0 (1.6)	38.9 (2.0)
Weeks on Admission $\mu(\sigma)$	77.7 (13.4)	79.2 (14.7)	76.7 (13.0)	83.1 (15.2)	78.8 (14.5)
Weight start of pregnancy $\mu(\sigma)$	65.6 (13.5)	66.0 (13.7)	65.6 (14.1)	67.4 (14.6)	66.0 (14.1)

4.1.4 Methods

The section will cover the steps we took for evaluating the models. We first addressed the preprocessing of the data, then the training of the models and finally the evaluation of the models.

The evaluation was done by comparing the performance of the distributed model with the local and centralised models. The performance was measured by the AUROC, AUPRC, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The results were then compared using a 2-sample T-test.

4.1.4.1 Preprocessing

The initial dataset underwent preprocessing by eliminating attributes that were missing more than 90% of their data across all storage units (or silo). We standardized the representation of missing values, which varied widely, including representations such as "-1" "missing" or simply blank spaces. For imputation, we utilized the mean for continuous variables (calculated within site) and introduced a special category (NULLIMP) for categorical variables. We converted all categories into numerical values based on a predefined mapping that covered all potential categories across the datasets. Although this approach introduces an ordinal relationship and potential bias is created among features, we disregarded this concern because the methodology was uniformly applied across all datasets intended for training local, distributed and centralised. These preprocessing tasks were executed once for each dataset and silo.

However, in the context of training classification models, it is crucial that all classes of the target variable are known at the time of training and are represented in each split of the cross-validation process. To address this, we employed Synthetic Minority Oversampling Technique (SMOTE) [200] to up-sampled low-frequency target classes. We established a threshold of n<25 for low-frequency variables to ensure that each cross-validation split contained at least two instances of the class—although a minimum of 10 instances (10 splits) might suffice, we opted for 25 to mitigate potential distribution issues and have at least two examples of the class in each split. Additionally, we created dummy rows for missing target classes by imputing the mean for continuous variables and the mode for categorical variables (calculated within site). The necessity for up-sampling and missing variable creation was evaluated and applied as needed for each training session and for each target, considering that each session's split could result in a training set lacking instances of low-frequency classes.

All procedures were coded in python 3.9.7 with the usage of the scikit-learn library [107] and mlxtend library [108].

4.1.4.2 Model Training

To avoid pitfalls of inductive bias from a certain learning strategy, we learned six different models (i) Decision Trees, (ii) Bayesian methods, (iii) a logistic regression model with Stochastic Gradient Descent, (iv) KNN, (v) AdaBoost and (vi) Multi-layer Perceptron. The decision was to create diversity in the models used, in order to assess if the training methodology could have an impact on distributed model creation. The distributed model was an ensemble of models from each silo on a weighted soft-voting basis. The weights were defined by weighted averages of the scores each model obtained in the training set. Then the final result is obtained by creating a weighted

average of the class predictions for classification and a weighted average for regression. A model like this can be implemented with peer-to-peer or federated approaches. Nineteen features were used as target outcomes. These features were selected by filtering by the percentage of null values (below 50%). This choice was related to maintaining a equilibrium between having a wide range of variables to test how the target variables affects the outcome and having target variables that did go through an harsh imputation mechanism. For categorical outcomes, thirteen were selected (AA - Position Admission; ANP - Position on Delivery; AGESTA - Nr of Pregnancies; APARA - Nr of born babies; GS - Blood Group; GR - Robson Group; TG -Pregnancy Type; TP - Delivery Type; TPEE - Spontaneous Delivery; TPNP - Actual Type of Delivery; V - Followed physician; VCS - Followed physician primary care; VNH - Followed physician hospital delivery;). For continuous variables, six were selected (IA - Mother Age; IGA - Weeks on Admission; IMC - BMI; NRCPN - Nr of consultations; PI - Weight start of pregnancy; SGP - Weeks on Delivery;). Details can be seen in tables 4.1 and 4.2. Local models were built with each silo's data. The centralised model was trained with a training dataset from all the silos combined.

4.1.4.3 Model Performance Evaluation

All models were built for a certain outcome variable with a repeated cross-validation (2 times and 10 splits each) and then compared, over 10 stochastic runs, with evaluation being performed on a test set held out from each silo. By performing cross-validation twice, we aimed to generate a more robust estimation of the model's performance metrics by averaging the results over two separate runs, each partitioning the data differently. This approach is particularly useful in scenarios where data is limited or highly variable, as it provides a clearer insight into the model's expected performance in unseen data scenarios. The metrics used for classification models were Weighted AUROC computed as One-versus-Rest, Weighted AUPRC. The metrics for regression models were RMSE and MAE. The algorithm is shown in the algorithm 2. This rendered over 1000 different combinations. When a variable was used as outcome to predict, all others were used as predictors.

After all the data was collected, we used the standard independent 2-sample T-test to check if the differences were significant with a α of 0.05. First, we compared the overall performance of the distributed model vs their centralised and local counterpart. We also compared every distributed model per algorithm and sequentially the centralised and correspondent local model across all algorithms and repetitions and outcome variables with 2-sample T-test as well.

4.1.5 Results

Table 4.3 shows the aggregated metrics for AUROC, AUPRC, RMSE and MAE for distributed, centralised and local models predicting capabilities on each silo. The data refers to the mean of the metric values for all columns tested as targets for all methods and all silos. We also calculated the 95% confidence interval for each model (local and distributed per silo) in order to assess how

Pre-process all silos (null standardization, imputation, encoding);
for *target in target list* **do**
    for *n in 10 repetitions* **do**
        for *silo in imputed silos* **do**
            Train-Test Split (80:20);
            check for low frequency or nonexistent labels in train set ;
            train local model with hyper-parameter tuning with 2x10 repeated
              Cross-Validation ;
            define weights based on scores in the train set (weighted average for
              predicting the value) for the distributed model;
        end
        Create distributed (ensemble of all models) model with weights;
        predict local on the test set;
        predict distributed on the test set;

        Create a centralised model with all the data with a 2x10 repeated Cross-Validation
         ;
        Test the centralised model on the test set;
    end
end

Algorithm 2: Creation and evaluation of the 3 different models. We first preprocessed data. Then for each target, we created a distributed and centralised model. Then, over 10 repetitions per silo, we created a new train and test set and local model and tested the centralised, distributed and local on this test set.

well the distributed model would work as opposed to the local one per silo. We also calculated the *P* value for the means of the distributed vs centralised and distributed vs local.

Figure 4.1 shows the AUROC of each algorithm and silo on the Y axis and target variable and type of model on the X. The color bar refers to the value of the AUROC. Blue being lower values and red bigger values. The same type of graph was created for regression, where the Figure 4.2 shows the MAE for each silo and algorithm and target variable and type of model.

4.1.6   Discussion

A significant finding is that nearly 59% of distributed models demonstrated comparable, if not superior, performance relative to their centralised counterparts (table 4.4 last column, first two values for each algorithm). From these, 41.9% were also better or equal to the local model. If we take the best performing algorithm (SGD), we have 77.2% for distributed better than centralised and 66% better than centralised and local. This outcome underlines the potential of distributed models to offer reliable inference capabilities that match those of traditional centralised models, without sacrificing predictive accuracy. Furthermore, the adoption of distributed models enhances privacy for data owners, presenting a compelling case for their broader application in data-sensitive environments. Overall, our results suggest that it is possible to implement a distributed model without significantly losing information. Our analysis suggests that SGD, Adaboost and Naive Bayes approaches are suitable for such distributed approached with tabular data. However, MLPerceptron,

Table 4.3: Comparison of the distributed model with the centralised model and with the local model (Mean for all model and all columns). 2-sample T-test for the means was used as hypothesis test. Bold for *P* value below 0.05. AUPRC and AUROC for categorical target variable and RMSE and MAE for continuous target variable.

		M	SD	95% CI	*P*
AUPRC	distributed	0.691	0.216	(0.686, 0.696)	-
	centralised	0.706	0.225	(0.701, 0.711)	**1.10e-17**
	local	0.659	0.220	(0.654, 0.665)	**4.71e-05**
AUROC	distributed	0.723	0.182	(0.718, 0.727)	-
	centralised	0.729	0.180	(0.725, 0.734)	**2.98e-26**
	local	0.692	0.164	(0.688, 0.695)	**2.48e-02**
MAE	distributed	2.370	1.608	(2.315, 2.425)	-
	centralised	2.365	1.923	(2.298, 2.431)	**2.23e-04**
	local	2.527	1.799	(2.465, 2.589)	9.01e-01
RMSE	distributed	21.171	46.078	(19.584, 22.757)	-
	centralised	19.839	28.645	(18.853, 20.826)	**2.92e-02**
	local	23.771	49.776	(22.057, 25.485)	1.63e-01

Decision Trees and KNN do not seem to be a good approach for such use cases.

However, there are still issues to be addressed. This methodology presents hurdles regarding categorical class handling. Firstly, all classes should be known first-hand and should be given to each model even if that silo in particular has no cases of that class. Secondly, low-frequency classes are also an issue to be addressed, since training the model with cross-validation will raise problems because each split should have all classes present. Our approach relied on sample creation for low and non-existent target classes. However, this approach is adding information to the model that is not originally there. The way we chose for minimising this issue was by creating dummy variables with median and mode imputations based only on the information in the dataset. Nevertheless, non-existent classes are impossible to address without prior information. These class problems could be partially tackled in production by implementing data management and governance procedures, namely data dictionaries. Still on data preprocessing, we applied ordinal encoding to the variables which will create a natural hierarchy between variables. One solution for this is to create binary columns for each class in each column. This will remove the hierarchy between classes but increase variable numbers and training time considerably.

Another issue to consider is the path adopted to build the distributed model. In this case, it was decided to develop an ensemble of models with voting. However, other methods could have been employed, like parameter averaging, that should be tested as well. In particular, the usage of more robust neural networks could be assessed as well. We chose not to test state-of-the-art

Figure 4.1: Heatmap of classification algorithm and silo vs Target variable and model type. Value is the AUROC mean of all 10 experiments. Y axis is the algorithm and silo. X axis is Target variable and Method. AA - Position Admission; ANP - Position on Delivery; AGESTA - Nr of Pregnancies; APARA - Nr of born babies; GS - Blood Group; GR - Robson Group; TG -Pregnancy Type; TP - Delivery Type; TPEE - Spontaneous Delivery; TPNP - Actual Type of Delivery; V - Followed physician; VCS - Followed physician primary care; VNH - Followed physician hospital delivery;

neural networks since the data volume was low for that use case and several papers have already demonstrated that neural networks are not the most suitable tool for tabular data [201, 202]. We chose to add MLPerceptron as a baseline for comparison with the remaining algorithms. The results show us that the performance was below the other algorithms, but in this concrete case, the problem may reside in the architecture chosen and hyperparameters used in the Cross-validation which may have led to underfitting. Despite this, a precise and thorough demonstration of this use case would be important to consider such scenarios.

Furthermore, the algorithm underlying the distributed model is of importance as well for its performance versus the centralised model. Figures 4.1 and 4.2 and table 4.4 show us that Decision trees and KNN implemented in a centralised manner are consistently better than the distributed counterpart. This is especially notorious in the case of the decision trees. We believe this may be

Figure 4.2: Heatmap of regression algorithm and silo vs Target variable and model type. Value is the MAE mean of all 10 experiments. The y axis is the algorithm and silo. X axis is Target variable and Method. IA - Mother Age; IGA - Weeks on Admission; IMC - BMI; NRCPN - Nr of consultations; PI - Weight start of pregnancy; SGP - Weeks on Delivery;

Table 4.4: Model comparison: Distributed versus centralised and local for every test. Each cell is the total of distributed model when compared with centralised model (row) and local model (column) across different silos and outcome variable. ($>$ for better, $=$ for non significance and $<$ for worse). The first example is 72 which means that 72 iterations of the distributed SGD was better than the centralised and local. SGD: Stochastic Gradient Descent, NN: Neural Network, KNN: K-Nearest neighbours, ADA: AdaBoost, NB: Naive Bayes, DT: Decision Tree. Comparison was done with 2-sample T-test with a α of 0.05. (% in parentheses)

		Distributed $>$ Local	Distributed $=$ Local	Distributed $<$ Local	**Row Total**
	Distributed $>$ Centralised	72 (7.0)	14 (1.4)	9 (0.8)	**95 (9.3)**
SGD	Distributed $=$ Centralised	14 (1.4)	17 (1.7)	6 (0.6)	**37 (3.6)**
	Distributed $<$ Centralised	11 (1.1)	11 (1.1)	17 (1.7)	**39 (3.8)**
	Distributed $>$ Centralised	44 (4.3)	44 (4.3)	7 (0.7)	**95 (9.3)**
NN	Distributed $=$ Centralised	2 (0.2)	33 (3.2)	2 (0.2)	**37 (3.6)**
	Distributed $<$ Centralised	0 (0)	17 (1.7)	22 (2.1)	**39 (3.8)**
	Distributed $>$ Centralised	16 (1.6)	0 (0)	1 (0.1)	**17 (1.7)**
KNN	Distributed $=$ Centralised	10 (1)	2 (0.2)	1 (0.1)	**13 (1.3)**
	Distributed $<$ Centralised	72 (7)	28 (2.7)	41 (4)	**141 (13.7)**
	Distributed $>$ Centralised	64 (6.2)	25 (2.4)	22 (2.1)	**111 (10.8)**
ADA	Distributed $=$ Centralised	5 (0.5)	12 (1.2)	10 (1)	**27(2.6)**
	Distributed $<$ Centralised	10 (1)	6 (0.6)	17 (1.7)	**33 (3.2)**
	Distributed $>$ Centralised	51 (5)	19 (1.9)	34 (3.3)	**104 (10.1)**
NB	Distributed $=$ Centralised	5 (0.5)	19 (1.9)	12 (1.2)	**36 (3.5)**
	Distributed $<$ Centralised	3 (0.3)	4 (0.4)	24 (2.3)	**31 (3)**
	Distributed $>$ Centralised	27 (2.6)	0 (0)	1 (0.1)	**28 (2.7)**
DT	Distributed $=$ Centralised	8 (0.8)	0 (0)	0 (0)	**8 (0.8)**
	Distributed $<$ Centralised	97 (9.5)	12 (1.2)	26 (2.5)	**135 (13.2)**
Total		**511 (49.8)**	**263 (25.6)**	**252 (24.6)**	**1026 (100)**

related to way the algorithm is implemented. A centralised version may be able to create optimal splits in the data, while the distributed version may not be able to do so. This is a topic that should be further explored.

Even though this improvement may have a relationship to the target variable (i.e. figure 4.2 for IA and IGA variables), it is still an important fact to take into account when implementing such architectures. The performance of the models is also interesting to catch differences in silos. See silo 6 for TPNP (figure 4.1) where silo 6 consistently behaves differently than the rest. Checking performance data regarding regression tasks, we can see a drop in performance for PI and IA. While the explanation for the performance of IA can be explained by the average value of it which is 66. This is the highest average in the dataset. This means that the model will have a harder time predicting these values. This is also true for the distributed model. This is a topic that should be further explored.

As for implementation, such a mechanism could be implemented in at least two manners; with a central orchestrator or without. The first one would assume a central point that would make a

request to each silo for a prediction and then create the final prediction with the weighted averaging of each one. The second one would not require any additional platform and each silo would communicate with each of the others and receive the prediction and would create the final with their own. This implementation step would of course take into account variables that we were out of scope such as the communication between silos. Regarding the prediction capability as a whole, we found that this data is suitable to apply machine-learning models in order to predict several clinical outcomes, with very good results for several target variables.

4.1.7   Conclusion

This research demonstrated the efficacy of distributed models using real-world data by comparing their performance with that of local models, which are trained with data from individual silos, and centralised models, which utilize data from all silos. The findings reveal that an ensemble of models, essentially a distributed model as investigated in this study, can capture the nuances of the data, achieving performance comparable to a model constructed with comprehensive data. Even though The performance of these models is influenced by factors such as the inherent characteristics of the target variables and the data distribution across different silos, we are now fairly confident that distributed learning is a step forward regarding data privacy without loss of predictive performance when compared with centralised and local models. Considering the robust performance metrics observed, with AUROC/AUPRC scores exceeding 80% and MAE maintained below 1, further investigation into distributed models is warranted. Specifically, we aim to develop distributed models for predicting clinical outcomes, such as delivery type or Robson Group classifications, which hold significant potential for real-world clinical application like reducing unnecessary Caesarean Sections or accelerating diagnosis. These findings underscore that distributed learning not only advances data privacy but also maintains high prediction accuracy, promising substantial benefits for clinical practices.

4.2   Can Institutions Share Their Performance Metrics Without Hesitation of Retaliation?

This section is based on the paper entitled "Benchmarking institutions' health outcomes with clustering methods". This paper was focused on the fact that many healthcare institutions harbour reservations about openly sharing production metrics. One predominant concern is the potential for retaliatory actions, be it from regulatory bodies, competitors, or the public. In this paper, we propose the application of a clustering methodology that allows institutions to compare performance metrics without disclosing the actual values. The method is based on clustering, which involves grouping health institutions' outcomes into a known number of clusters, allowing institutions to position themselves in a range of clusters without sharing the true means of their target data. The proposed method uses the K-means and K-modes clustering algorithms and was tested on data from real Electronic health records and public datasets. This approach provides a valid

benchmark of hospital metrics and performances while protecting the privacy of participating institutions.

4.2.1 Introduction

Health institutions play a critical role in providing essential healthcare services to communities and ensuring that they operate efficiently and effectively is crucial. Benchmarking is a process that allows hospitals to compare their performance against that of other institutions, which can help identify areas of strength and weakness [203]. By analysing and evaluating performance metrics, such as patient outcomes, operational efficiency, and financial management, hospitals can identify best practices and make data-driven decisions to improve their overall performance. It can also help hospitals identify and implement innovative practices that can lead to better patient care and improved staff satisfaction [204].

However, despite the numerous benefits of benchmarking, some hospitals may be hesitant to participate due to concerns about revealing weaknesses or being perceived as inferior to their peers. The fear of being judged or penalized for poor performance can sometimes lead hospitals to avoid sharing data, making it difficult to accurately assess their performance and identify areas for improvement. Privacy issues and concerns turn this opportunity into an even less desirable path [204]. To address these concerns, benchmarking initiatives often ensure the confidentiality and anonymity of data to encourage participation and foster trust among participating institutions. However, this is usually not enough. In 2019, as stated in the work of Villanueva et al., [205], 26% of data-sharing initiatives are based on the aggregation of data and 24% are based on sharing data in closed consortia. Only 15% were based on open or controlled access.

To address concerns around privacy and confidentiality, we propose a new method of benchmarking based on clustering. This method involves grouping health institutions' outcomes into a known number of clusters, providing health institutions with the capability of positioning themselves in a range of clusters, without ever sharing the true means of their target data.

This approach to benchmarking not only addresses concerns around privacy and confidentiality. It has the potential to encourage greater participation in benchmarking initiatives, as hospitals can be assured of the anonymity and confidentiality of their data. By creating a more secure and private environment for benchmarking, hospitals can feel more comfortable sharing their data and participating in initiatives that can ultimately improve patient care and operational efficiency.

In conclusion, benchmarking is a crucial tool for hospitals to improve their performance and provide better care for their patients. While concerns around privacy and confidentiality may exist, the clustering approach to benchmarking provides a more accurate assessment of hospital performance while protecting the privacy of participating institutions. By embracing benchmarking initiatives and leveraging new approaches to benchmarking, hospitals can continuously improve their operations and ensure they provide the highest quality of care possible. In this paper we propose:

- study how to implement clustering mechanism for benchmark
- address preprocessing issues for the raw data

- highlight pain points to deployment in the real world.

4.2.2   Rationale and Related Work

This work was initially suggested as a follow-up to a previous work of Rodrigues et al., [206] where clustering is applied to streaming data sources. We then thought if a similar approach could be applied to healthcare in order to be able to compare data distributions without ever knowing their real values of them. Clustering in healthcare is often used to create clusters of patients, taking into account a given set of characteristics. This is used to find possible groups of phenotype and be able to characterise populations given the centroids [207, 208]. It is also used as a method of detecting regularities and patterns in multi-omics data that reveal different molecular subtypes [209, 210]. It can also be used to create unsupervised models for facilitating the annotation of data for supervised models [211].

K-means [212, 213, 214] is an unsupervised clustering algorithm used to group data points into K distinct clusters based on their similarity. It is widely used in ML, data mining, and image segmentation. The algorithm works by randomly initializing K centroids (or cluster centres) and assigning each data point to the nearest centroid. Then, the centroids are moved to the mean of the points assigned to each cluster. This process is repeated until convergence, where the clusters no longer change.

The objective of K-means is to minimize the sum of squared distances between each data point and its assigned centroid, which is also called the within-cluster sum of squares (WCSS). The algorithm attempts to find the best K clusters that minimize the WCSS. However, choosing the right value of K can be challenging, and the algorithm may converge to a suboptimal solution. Therefore, K-means is often run multiple times with different initializations to find the best clustering solution. Despite its simplicity, K-means can be computationally expensive when dealing with large datasets, and it may not work well with non-linearly separable data or when the clusters have different shapes and sizes.

K-modes is another clustering algorithm similar to K-means, but it is designed to work with categorical data. Unlike K-means, which computes the mean of continuous variables, K-modes computes the mode (or the most frequent value) of categorical variables within each cluster. The algorithm works by randomly initializing K centroids and assigning each data point to the nearest centroid based on the number of matching categories. Then, the centroids are moved to the mode of the categories within each cluster. This process is repeated until convergence, where the clusters no longer change.

The objective of K-modes is to minimize the dissimilarity between the data points within each cluster, which is often measured by the Hamming distance, Jaccard distance, or other similarity measures. Like K-means, choosing the right value of K is critical, and the algorithm may converge to a suboptimal solution. Therefore, K-modes is often run multiple times with different initializations to find the best clustering solution. K-modes is particularly useful when dealing with data

that have a large number of categorical variables or when the data contain missing values. However, like K-means, K-modes may not work well with non-linearly separable data or when the clusters have different shapes and sizes.

However, as far as we know, this is the first time clustering is tested for exchanging information privately.

4.2.3 Materials & Methods

4.2.3.1 Materials

We used two types of data in this paper. One is simpler and available online from the UCI dataset library, namely, the heart disease dataset [109]. We made fairly simple preprocessing on that dataset, namely removing the "?" by filling with null and then imputing missing values by imputing the mean on continuous variables and mode on categorical ones. We then separated the data into 3 distinct silos at random to mimic different health institutions.

In order to use real data and address problems found in the wild, we used clinical data gathered from nine different Portuguese hospitals regarding obstetric information, pertaining to admissions from 2019 to 2020. This originated from nine different files representing different sets of patients but with the same features associated with them. The software for collecting data was the same in every institution (although different versions existed across hospitals) - ObsCare. The data columns are the same in every hospital's database. Each hospital was considered a silo for comparison.

4.2.3.2 Method Overview

We used Python 3.9 to implement the mock example of such an use-case. The clustering was done with *scikit-learn* library [107]. The algorithm proposed is shown in algorithm 3.

for *variable in silo* **do**
 | initialize centroids;
end
while *No convergence* **do**

   • Send centroids to other silos

   • Receive other silo's information and add own centroids

   • Calculate new centroids

   • calculate score

end

Algorithm 3: Benchmarking with clustering

The method for assessing convergence is based on clustering metrics: the Rand Index (RI). This metric computes a similarity measure between two clusters by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusters [215]. The raw RI score is: $RI = (number\ of\ agreeing\ pairs)/(number\ of\ pairs)$. Furthermore, convergence must be obtained through several iterations to make sure it's stable, so a buffer period is also important. For the results section, we set the threshold as 0.9 and repetitions at 20.

In this paper, we propose to show how such an implementation could be done while addressing issues with data formats, types and preprocessing. So, we want to check if the encoding of categorical data affects the model and which method is better for encoding such variables. Additionally, we will try to understand if it is possible to create mechanisms for mixed data if categorical and continuous data must be used and evaluated separately and if so, through which mechanisms. We will test (1) continuous variables alone, and (2) encoded categorical variables as ordinal. We will also test (3) K-modes and (4) K-means with the proportion of each category for categorical data. K-means was used as implemented in *scikit-learn* [107] and K-modes, as implemented by J. de Vos [216].

4.2.4 Results

As for results, the data from heart disease rendered the figure 4.3. In this, we focused on continuous variables only. For easier reading, the data is as shown in the table 4.5. We used data from the real world to test if everything would work similarly, rendering the image 4.4. We added a binary category to show how meaningless the value turn in order to get any information out of it.

Figure 4.3: Clustering for 3 continuous variables with 3 silos and true centroids (S2) and true means (S2) for example purposes; The values were normalized for visualization purposes with MinMax
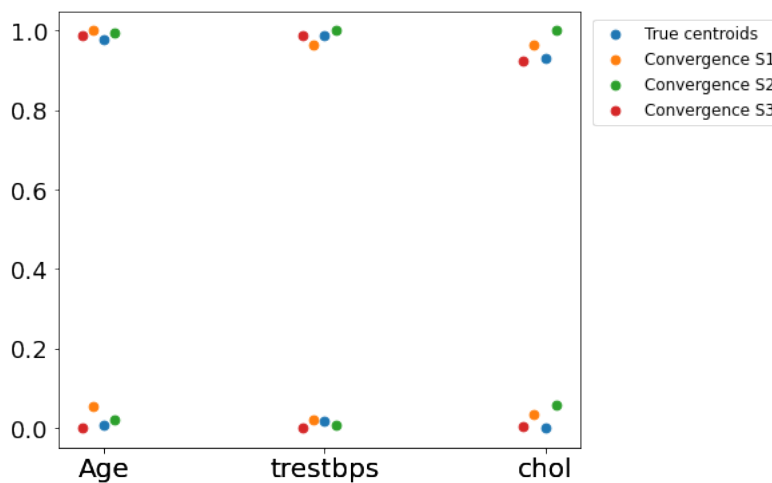
Table 4.5: Final Data points after convergence; S1, S2 and S3 are the centroids obtained in each silo (S) after convergence; True centroids are the centroids of the true means of all silos (TC)

	Age	trestbps	chol
S1	46.3 , 61.1	121.1 , 148.9	218.9 , 300.8
S2	45.8 , 61.0	120.7 , 149.9	220.9 , 304.0
S3	45.5 , 61.0	120.5 , 149.6	216.1 , 297.4
TC	45.6 , 60.8	121.0 , 149.6	215.8 , 297.9

Figure 4.4: Clustering for 3 variables with 9 silos and true centroids of the true means (TC); 2 continuous and 1 categorical one hot encoded, The values were normalised for visualisation purposes with MinMax
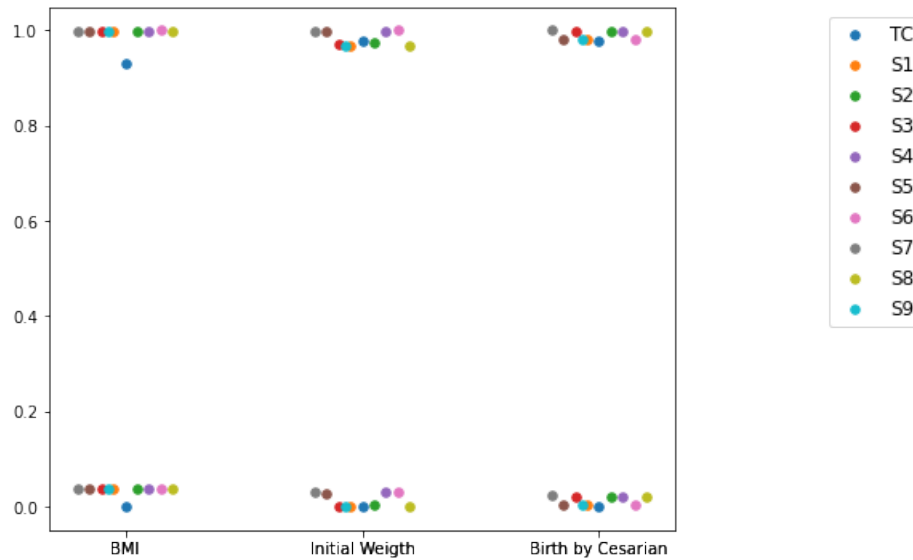
As before, the data is in table format in 4.6.

Then we experimented with categorical variables. Figure 4.5 shows the convergence of the silos with proportion data and K-means with that and with K-modes.

4.2.5 Discussion

As per the discussion, there are a few issues to be addressed. First as per data preprocessing. In order to cluster be obtained, the null data must be filled out. There are a few strategies to do so. One option is to eliminate records/rows with empty cells or impute data. Either is a possibility, with pros and cons but the capability of having a dataset where no null records are present across several features may be difficult to find in the wild, especially since there are often optional and conditional fields in most EHR. So imputation becomes more interesting, since it enables the usage of the whole dataset, even if biases are introduced. Mixed types of datasets are also an issue to be aware of. In this case, not only imputation but also encoding a categorical variable is a vital

Table 4.6: Final Data points after convergence and true centroids of the true means of each silo (TC)

	BMI	Initial Weight	Birth by Cesarian
TC	24.9 , 383.1	60.5 , 85.0	0 , 1
S1	40.1 , 409.4	60.4 , 85.0	0.96 , -0.04
S2	40.1 , 410.4	61.7 , 86.3	0.99 , -0.01
S3	40.0 , 410.4	61.9 , 86.5	0.96 , -0.04
S4	40.6 , 411.3	61.9 , 86.5	0.96 , -0.04
S5	40.0 , 410.4	60.5 , 85.1	1.0 , 0.0
S6	40.1 , 409.3	60.4 , 84.9	1.0 , 0.0
S7	40.7 , 411.3	60.5 , 85.0	0.96 , -0.04
S8	40.0 , 410.4	86.5 , 61.9	1.0 , 0.0
S9	41.0 , 410.4	85.0 , 60.4	1.0 , 0.0

step to take in the preprocessing phase. There are usually two main methods of data encoding, ordinal encoding and binary encoding. The first one keeps a unique column as the original data but maps every category to an increasing natural number. This creates an ordering in the data, often a misrepresentation of reality, not only due to this hierarchy but only because it assumes the differences between ranks of the hierarchy are always the same (1). The second is related to expanding the number of columns into the number of categories and creating 0s and 1s for the category. In machine-learning terms, binary seems more suited to be applied, but for benchmarking purposes, both are below par in terms of interpretability. For categorical data, we found out that K-modes seem to fulfil the requirements in a better way, providing better interpretability and reasoning about the results. However, it should be noted that we applied K-modes in a multivariate fashion and K-means in a univariate fashion. Given that no percentage is provided, only the mode of the data, we believe it is still hard to get any real insight from the centroids. However, K-modes provides less information, since it only shows the top two categories. Which, for example. binary targets, provide little to no information. However, for larger categorical sets, the information provided could be better. Moreover, the number of centroids pretended could be more important as well. Agreeing on only 1 centroid would render the mode of the data provided by all silos, which could be more interesting. As for continuous data, the use of real data was insightful, since Body Mass Index (BMI) had a few very big outliers around 300 and 400, which rendered centroids around that data. Even if not all silos had examples of these outliers, the ones that do have, pass that into the remaining. One possible workaround would be an addition of an extra cluster in order to catch possible outliers. However, this should be addressed in detail and assess how outliers could subvert the data from the silos and how to work around that.

As for the next steps, a few issues could be addressed in depth. Regarding imputation, it could

Figure 4.5: Clustering for 3 variables with 3 silos - (A) categorical variables with proportion with K-Means and (B) Categorical with K-modes
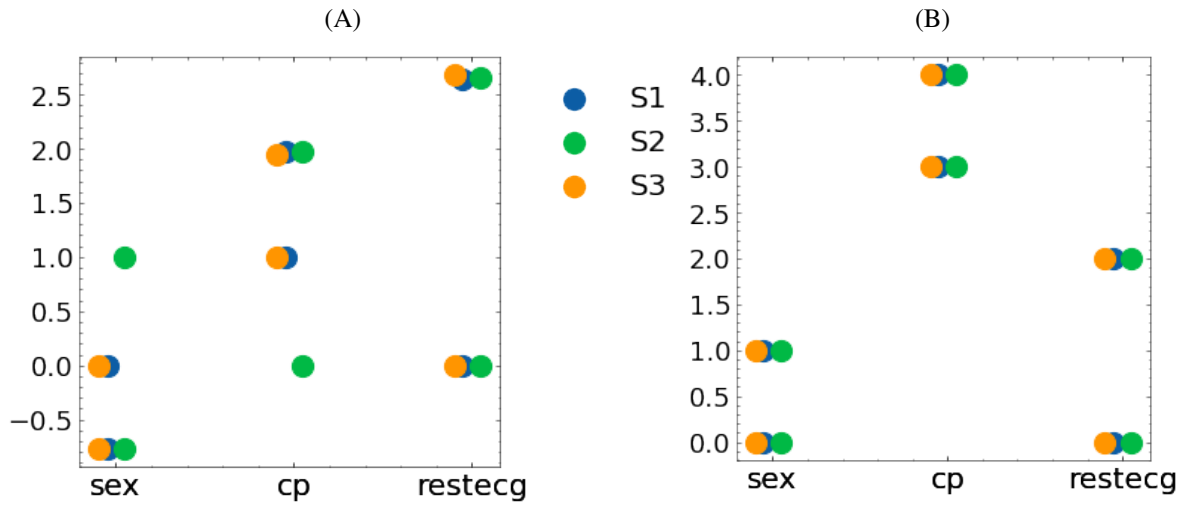
be interesting to understand how imputation, and which methods are more suitable to use for real-world scenarios. If the imputation of variables with a high null percentage influence significantly a centroid formation. Communication could be important as well. Which action is to be taken when a silo is "down" and does not send information to the remaining. Cluster information should be addressed as well. They need to be agreed upon beforehand in the scope of this paper. But if it could be selected by each silo? Would that be feasible or a convergence could be achieved? Finally, there is the question if there is the possibility of having leaks of true means across iterations by adversarial learning. At present time, we cannot be sure that the values are totally private, but then again, nothing is.

4.2.6 Conclusion

We believe that this work helps create the foundation for exchanging data across healthcare institutions without revealing the true data points. It could be useful for benchmarking and promoting a higher adoption rate. Even though there are still issues to be addressed, we think that the path is full of possibilities.

I don't want to insist on it, Dave, but I am incapable
of making an error.

HAL 9000

5

Explore Strategies to Transform Health Data Into Actionable Decisions and Policies

This chapter focuses on the practical application of health data in influencing decisions and policies, particularly in the realms of drug evaluation and obstetrics, as detailed in sections 5.1 and 5.2.

Section 5.1 represents an innovative application of causality principles and transparent ML models to assess the real-world effectiveness of two groups of breast cancer drugs. Beginning with traditional analysis methods, the study progressively adopts more complex techniques, including IPTW to enhance the comparative assessment of these treatments. This approach exemplifies how health data, analysed through advanced methodologies, can influence drug policy and treatment choices in clinical settings.

Section 5.2 is an extension of research in distributed data mechanisms (referenced in section 4.1), uses ML to develop a CDSS designed to assist in evaluating Caesarean Section (C-Section). The system's interoperability and its focus on supporting subpar evaluation of C-Section demonstrate the utilization of health data in creating tools that aid in decision-making processes in obstetrics. This section underscores the importance of applying health data analytics in real-time clinical environments to inform decisions and shape obstetrical policies.

5.1 How Can We Leverage Data to Assess Treatment Efficacy?

This section is based on the paper entitled "Comparative Analysis of Palbociclib and Ribociclib: A real world data and Propensity Score-Adjusted Evaluation with endocrine therapy". This was a method of applying the knowledge of causality and transparent ML models in order to assess

the real-world effect of two drugs for breast cancer. We started with traditional analysis and then moved to a more complex approach, using IPTW methods in order to further compare treatments.

5.1.1 Introduction

Currently, metastatic breast cancer is difficult to treat. Patients with Hormone Receptor positive (HR+) and Human Epidermal growth factor Receptor 2 negative (HER2-) breast cancer, the most common subtype, typically undergo Endocrine Therapy (ET). Therefore, new treatments can be very useful in improving quality of life, reducing toxicity, and decreasing scenarios of hormonal resistance. Medications from the group of Cyclin-dependent kinases 4 and 6 inhibitors (CDK4/6i) appear as a potential improvement in the therapeutic approach to advanced breast cancer. Within this group, there are palbociclib, ribociclib and abemaciclib. Cyclin-dependent kinases 4 and 6 (CDK4/6) are responsible for regulating the cell cycle at the transition between the G1 and S phases. In many neoplasms, this cycle is deregulated, and it promotes uncontrolled cell proliferation. It is then possible for these medications to have better effectiveness. These medications were approved by INFARMED, I.P. after an analysis of the therapeutic value they offer. This decision was made based on data provided by clinical trials done with these medications. The MONALEESA [217, 218, 219] studies were used for ribociclib, PALOMA [220, 221, 222] for palbociclib, and MONARCH [223, 224] for abemaciclib. These studies focused on testing the hypothesis of treating CDK4/6i in combination with an aromatase inhibitor or fulvestrant as an alternative to the gold standard. In these research findings, it was determined that there was a notable enhancement in effectiveness, supporting their application in clinical practice. However, this evaluation was based on clinical trials with very specific inclusion and exclusion criteria and in a highly controlled environment. It is then vital to study how these new molecules compare to current practice in terms of treatment effectiveness in a real-world setting. In the meticulously controlled setting of clinical trials, patient selection often skews towards relatively healthier individuals with fewer comorbidities. However, in real-world clinical practice, patients present a diverse range of health profiles, co-existing illnesses, and medication histories that may influence drug efficacy and safety. Real-world data, drawn from electronic health records, insurance claims databases, and patient registries, offers the advantage of reflecting a more heterogeneous patient population, thus potentially uncovering insights not readily apparent in clinical trial settings. Understanding the effectiveness and safety of CDK4/6i in real-world conditions is crucial for tailoring more individualized treatment regimens, optimizing outcomes, and enhancing the quality of life for patients with HR+, HER2- breast cancer [225]. Nevertheless, observational studies have inherent limitations, such as confounding by indication, which can lead to biased estimates of treatment effects. To tackle this, there are causality-based assessments that can be employed in order to better estimate the causal effects of treatments. Incorporating statistical techniques like IPTW can play an essential role in enhancing the quality of real-world evidence by accounting for treatment selection bias and balancing observed covariates between treatment groups. IPTW, grounded in the framework of causal inference, allows for the mimicking of a randomized control trial-like setting within observational studies. By assigning weights to individual patients based on their

propensity scores—the likelihood of receiving a particular treatment given a set of observed characteristics—analyses can achieve a balance between different treatment arms, thereby reducing bias and confounding factors. Establishing causality, rather than mere association, is vital for the robust interpretation of real-world data. As we strive to understand the long-term impact, efficacy, and safety of CDK4/6i in HR+, HER2- breast cancer, the rigorous application of IPTW and causal inference methods can substantially augment the validity of real-world findings, making them a more reliable basis for clinical decision-making [226, 227] So in this paper, we propose:

- To compare the effectiveness of the CDK4/6i drug class in terms of Progression Free Survival (PFS) and Overall Survival (OS);

- To assess the Hazard Ratio of using the CDK4/6i drug class in terms of PFS and OS.

- To compare the effectiveness of CDK4/6i in combination with letrozole or fulvestrant with the previous standard of care in terms of PFS and OS in patients with HR+, HER2- advanced breast cancer with bone only metastasis.

- To assess the differences in effectiveness between the three CDK4/6i in combination with letrozole or fulvestrant in terms of PFS and OS with causality principles in mind, especially the counterfactual theory and IPTW.

5.1.2 Materials & Methods

5.1.3 Study Design

This retrospective study was designed in 2022. The study aimed to evaluate the clinical benefit and long-term survival of patients with HR+/HER2- that started treatment with CDK4/6i plus ET in different lines of treatment between the 14th of March 2017 and the 31st of December 2021. The follow-up period was set until June 2022. Inclusion criteria: women and men, HR+ and HER2- in the primary tumour or metastatic site after biopsy. Exclusion criteria: Patients that had only one ambulatory medication, and patients involved in clinical trials, diagnosed with other neoplasms or with active treatment during the study period. The control group was defined by a population of patients, that were treated with hormone therapy as first-line (due to bone metastases only) between 2015 and 13 of match 2017. The evaluation of effectiveness will involve OS and progression-free analysis. We will compare the two different CDK4/6i in terms of effectiveness in real-world patients and will also compare the effectiveness of this class combined with ET against traditional ET.

5.1.4 Data collection

All data were collected from medical and administrative records from baseline to last visit or death. The data was collected from Instituto Português de Oncologia – Porto (IPO-P). Table 5.1 shows a comparison between the groups. Data included for population treated with CDK4/6i plus ET: demographic information, age at first diagnosis and age at the beginning of treatment, clinical

Table 5.1: Descriptive statistics of CDK4/6i group and ET group. The Drug/combination refers to the actual drug or the combination for CDK4/6

	ET	Palbociclib	Ribociclib
	(N=43)	(N=246)	(N=106)
Age at treatment start			
Mean (SD)	60.1 (12.4)	59.2 (11.7)	58.2 (10.7)
Median [Min, Max]	62.0 [34.0, 85.0]	60.0 [28.0, 84.0]	58.0 [32.0, 79.0]
Bone Only metastases			
No	NA	161 (65%)	74 (70%)
Yes	NA	85 (35%)	32 (30%)
Missing	43 (100%)	0 (0%)	0 (0%)
Visceral metastasis			
No	NA	121 (49%)	49 (46%)
Yes	NA	125 (51%)	57 (54%)
Missing	43 (100%)	0 (0%)	0 (0%)
Stage			
I	3 (7%)	22 (9%)	7 (7%)
II	20 (47%)	75 (30%)	22 (21%)
III	11 (26%)	74 (30%)	18 (17%)
IV	2 (5%)	65 (26%)	46 (43%)
Missing	7 (16.3%)	10 (4.1%)	13 (12.3%)
Drug/Combination			
Anastrozol	3 (7%)	NA	NA
Exemestane	4 (9%)	NA	NA
Fulvestrant	5 (12%)	180 (73%)	10 (9%)
Letrozol	31 (72%)	66 (27%)	96 (91%)

characteristics and performance status by Eastern Cooperative Oncology Group scale (ECOG), treatment line and treatment schema - CDK4/6 inhibitor and ET, stage of cancer, site of metastases (bone, soft tissue, visceral, central nervous system with or without another site). Data included for the population treated with ET as first-line: demographic information, age at first diagnosis and age at the beginning of treatment, clinical characteristics and performance status by ECOG, stage of the cancer. For comparison purposes, we used palbociclib and ribociclib since we had a small number of patients treated with abemaciclib (12).

5.1.5   Statistical Analysis

R was used for statistical analysis. Demographic, clinical characteristics and side effects were analysed using descriptive statistics (count, percentages and median/range). Kaplan–Meier test was used to determine the median PFS and OS in the entire population and subgroups. Log-rank test was used for comparisons of PFS and OS among different subgroups. Cox Regression was used to assess feature importance and impact. All statistical tests were two-sided, and the significance level was 0.05. The evaluation of the proportional hazards assumptions was done by *Schoen-*

feld residues analysis. We applied propensity score weights to achieve a more robust comparison between the two groups of CDK46i. We used the existence of visceral metastases, treatment line, age at treatment start, and stage. We used the WeightIt package for R [228]. We applied the weights to the Kaplan-Meier curves and to the Cox Regression. We applied the weights to get the ATE which is $E[Y_i(1) - Y_i(0)]$, the average effect of moving an entire population from untreated to treated, or from one drug to the other. Weights were used instead of matching since it is more suited for calculating ATE and the need to preserve the sample size since it is already small from the start. The formula for calculating the weights was through propensity score weighting with General Linear Model (GLM). Multiple comparisons were done with the *Benjamini-Hochberg* (BH) method.

5.1.6 Results

The median OS in the entire population treated with CDK4/6i was 46 months (95% CI 39.4–55.6). Median PFS was 20.1 months (95% CI 18.3–24.2). Following this, we compared Palbociclib and ribociclib only as first-line treatments. We found that regarding OS, there is no significant difference between the two, but ribociclib is significantly better in terms of PFS (P value ≤ 0.001) (Figure 5.1). Additionally, we compared the same CDK4/6i with letrozole as a combination only (PAL-LT and RIB-LT). Regarding this scenario, we found out that both were similar in terms of OS and PFS.

Figure 5.1: Survival curves for Palbociclib and Ribociclib (1st line) - PFS and OS
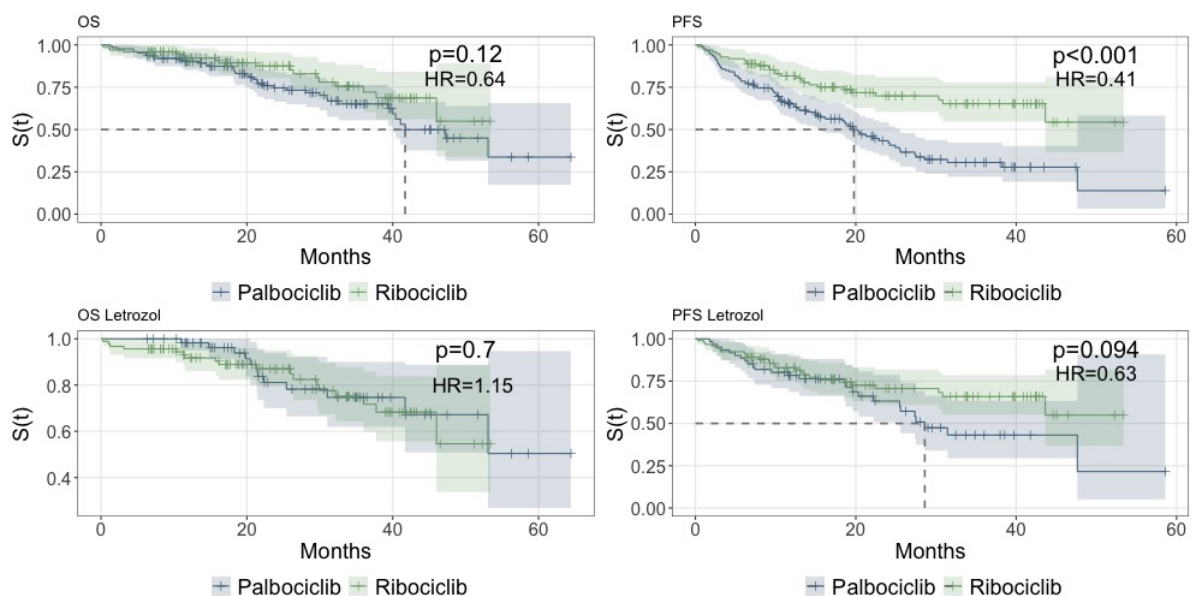
We then compared both with a cox regression, where OS shows no significant difference between palbociclib and ribociclib when adjusted to the stage, visceral metastases, age, treatment line, combination and ECOG. The proportional hazards' assumption was confirmed with P values all over 0.10.

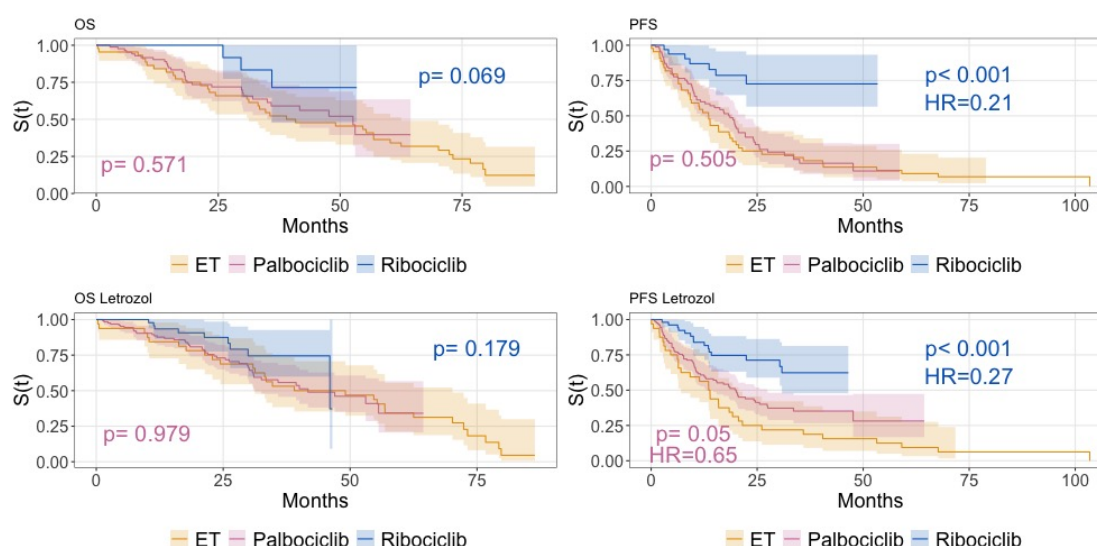Table 5.2: Cox Regression with palbociclib and Ribociclib - PFS and OS

Characteristic	OS			PFS		
	HR[1]	95% CI[1]	p-value	HR[1]	95% CI[1]	p-value
Drug						
Palbociclib	—	—		—	—	
Ribociclib	1.10	0.55, 2.19	0.8	0.67	0.41, 1.11	0.12
Menopausal Status						
Post-menopause	—	—		—	—	
Pre-menopause	1.02	0.57, 1.82	>0.9	1.12	0.72, 1.74	0.6
Combination						
Fulvestrant	—	—		—	—	
Letrozol	0.34	0.18, 0.67	**0.002**	0.38	0.24, 0.61	**<0.001**
Treatment Line						
1st Line	—	—		—	—	
2nd+ Lines	0.99	0.60, 1.63	>0.9	1.17	0.80, 1.73	0.4
Stage at Diagnosis						
I	—	—		—	—	
II	5.60	1.34, 23.4	**0.018**	1.87	0.97, 3.61	0.060
III	8.09	1.93, 33.9	**0.004**	3.04	1.58, 5.86	**<0.001**
IV	7.89	1.87, 33.4	**0.005**	2.24	1.15, 4.37	**0.018**
Visceral Metastasis						
No	—	—		—	—	
Yes	1.73	1.17, 2.55	**0.006**	1.34	0.99, 1.81	0.059
Age at treatment start	1.00	0.98, 1.02	0.9	0.99	0.97, 1.00	0.075
ECOG at treatment start						
0	—	—		—	—	
1	1.61	1.04, 2.49	**0.033**	1.23	0.88, 1.71	0.2
2	3.93	2.06, 7.51	**<0.001**	1.64	0.91, 2.97	0.10

[1] HR = Hazard Ratio, CI = Confidence Interval

When comparing ET with CDK4/6i as first-line treatment (figure 5.2). For this study we only compared patients with bone only metastasis. When comparing both CDK4/6i combined with Fulvestrant or letrozole, we see that Ribociclib (RIB+LT/FUL) is significantly better for PFS (*P* value $\leq$ 0.001 Hazard Ratio (HR)=0.21) but not OS. For Palbociclib as the first line with Fulvestrant or letrozole (PAL+LT/FUL), we see that there is no significant difference in terms of PFS and OS (*P*=0.57 and 0.51). We also applied the same analysis but comparing only the letrozole combination with letrozole alone (PAL-LT/RIB-LT vs LT). We found that both ribociclib and palbociclib are significantly better in terms of PFS (HR 0.65 for palbociclib and 0.27 for ribociclib) but not OS.

When comparing palbociclib and ribociclib adjusted for ATE weights, we found a different scenario from previous assessments. There is a significant difference between the two in terms of OS (figure 5.3). The weights were calculated as stated in the methods section.

Figure 5.2: Survival curves (OS and PFS) comparing ET to CDK4/6i combined with fulvestrant or letrozole as 1st line. First row is CDK4/6i combined fulvestrant or letrozole vs fulvestrant or letrozole. Second row is CDK4/6i combined with letrozole vs letrozole alone. *P* values shown as pairwise vs. ET.

The Cox regression adjusted for the variables and with the weights applied to render an HR=0.55 [95% CI 0.28-1.09;*P*=0.086] for OS. The HR for PFS is 0.56 [95% CI 0.32-1;*P*=0.05].

5.1.7 Discussion

The aim of this study was to evaluate the real-world use of palbociclib and ribociclib in combination with ET for HR+/HER2- and compare this drug class with traditional ET. Few real-world evidence studies of palbociclib and ribociclib used in daily clinical practice have been published identifying clinical benefit, patient profile, and sequencing of treatment, with even less evidence for the Portuguese population.

When comparing with clinical trials, regarding patient profile, in our study, 51% had visceral metastasis and 35% had bone-only metastases compared with 49% and 38% in PALOMA-2, and 60% and 25% in PALOMA-3, respectively [221, 229]. As for ribociclib and bone-only metastases, MONALEESA-7 [219] has 24% and MONALEESA-2 has 40% [217] and our study has 30%. Regarding menopausal status, our study has 20% premenopausal and 80% postmenopausal.

Of note, the range of median PFS for first-line palbociclib was 15.5–25.5 months, which is shorter than 27.6 months observed in a post hoc analysis of the PALOMA-2 clinical trial with extended follow-up [221], but in line with RWE studies (13.3–20.2 months) [225]. When assessed with only letrozole as a combination, the median PFS increased to 28.6 months [95% CI 25.5-not reached]. Additionally, analysing the postmenopausal women subgroup, palbociclib showed a median PFS of 16.3 months [95% CI 12.9 -20]. Furthering analysis of the postmenopausal and with letrozole, the median was 47.6 months [95% 25.6-2–not reached].

Figure 5.3: Comparison of palbociclib and ribociclib survival curves adjusted for propensity scores
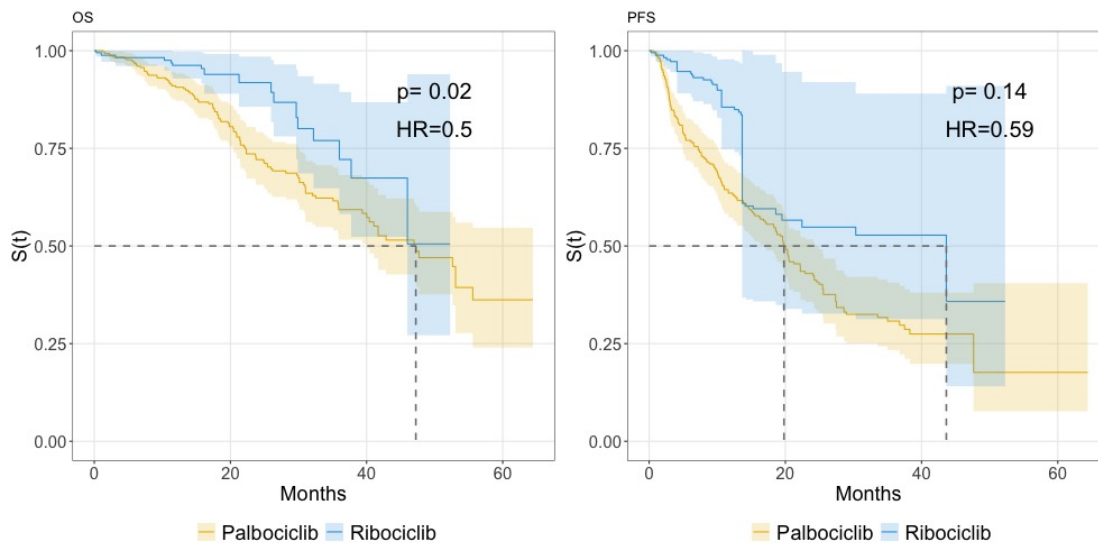
As for ribociclib, median survival time was not reached whether in OS and PFS. So we can at least say that the median PFS is longer than 50 months. This is longer than the median PFS of 23.8 months (95% CI 19.2–not reached) reported in the MONALEESA-7 trial [219] and longer than 25.3 months (95% CI 23.0–30.3) in the MONALEESA-2 trial [217]. Regarding the subgroup analysis of postmenopausal women, we noticed that the median was not reached for women treated with ribociclib and fulvestrant or letrozole (RIB-LT/FUL) and postmenopausal women treated with ribociclib in combination with letrozole (RIB-LT).

When directly comparing ribociclib and palbociclib without any adjustments, one might deduce that ribociclib is superior to palbociclib. However, after adjusting for confounding variables, there is no significant difference between the two inhibitors in terms of PFS or OS as indicated in table 5.2. This observation is further corroborated by the lower plots in figure 5.1, where even a subgroup analysis of CDK4/6i combined solely with letrozole reveals non-significant difference between the two.

In the first-line comparison, the analysis of OS outcomes reveals no substantial difference between ET alone and the combination of CDK4/6i with ET, irrespective of whether the CDK4/6i are administered with fulvestrant or letrozole (PAL-LT/FUL vs LT/FUL; P=0.57 | RIB-LT/FUL vs LT/FUL P=0.069) or exclusively with letrozole (PAL-LT vs LT; P=.979 | RIB-LT vs LT; P=0.179)(figure 5.2 left). With respect to PFS, ribociclib demonstrates superior efficacy when compared its combination with any of the adjuvants to these adjuvants alone (RIB-LT/FUL vs LT/FUL; HR=0.21) as well as when combined only with letrozole (RIB-LT vs LT; HR=0.27). Additionally, palbociclib exhibits significant improvement in PFS when combined with letrozole (PAL-LT vs LT; HR=0.65) (figure 5.2 right). When comparing with propensity scores weighting, we found out that ribociclib is significantly better than palbociclib for PFS and OS, providing a median OS of over 40 months and median PFS of around 42 months. Adjusted for the weighted

variables, Ribociclib is not significantly better for PFS, but has a *P* value of 0.013 for OS with an HR of 0.48. However, the Cox regression adjusted for variables and weights are not significant, even when the *P* value for PFS is 0.05. This suggests that a more in depth analysis may be necessary.

5.1.8 Conclusion

In conclusion, our findings underscore the efficacy of CDK4/6i in real-world settings. We can confidently affirm the impact of Ribociclib on PFS. This assertion aligns with clinical trial outcomes and real-world data further substantiates these findings. However, we cannot do the same for OS. Our results indicate that Ribociclib combined with letrozole or fulvestrant when compared to both is not superior to these alternatives used alone. The same happens when comparing ribociclib combined with letrozole with letrozole alone. However, we cannot do so for Palbociclib. Palbociclib combined with fulvestrant or letrozole was not significantly better than letrozole or fulvestrant alone for PFS nor OS. This is something interesting that we want to follow up with. Delving deeper into the characteristics of the patient population, including safety profiles, economic implications, and quality of life metrics, would be insightful. Additionally, a thorough examination of biomarkers within the population could offer invaluable insights. Finally, extending the follow-up period would be beneficial as well. We intend to explore these facets in subsequent publications. It's imperative to note that our data is sourced from a singular institution, limiting the capability of generalization of our results to a broader population. Nonetheless, we posit that this study lays a foundational groundwork for future research in this domain. While our evidence is rooted in observational data, and we've made adjustments for known confounders, the potential for residual confounding remains. Although the use of propensity score matching enhances the comparative robustness between the groups, the presence of unmeasured confounders cannot be entirely ruled out. Furthermore, the small sample size of our study limits the statistical power of our findings. For next steps we aim to further analyse the clinical variables that have an impact on the outcome of the combination of CDK4/6 with fulvestrant or letrozole and these drugs used alone in order to infer pharmaeconomic implications and possible profiles of patient that would not benefit from this combination which would be vital for economic reasons and to apply in countries with low access to these drugs.

5.2 How Can We Leverage Data to Create Clinical Decision Support Systems?

This section is based on the paper entitled "Machine-learning in Obstetrics: FHIR-based Support System for predicting delivery type". This work was in part a result of the work in section 4.1. While testing for distributed mechanisms, we kind of felt that some evaluation metrics were inspiring to pursue this further. We built a CDSS system that is interoperable and aims to provide support for subpar evaluation of a C-Section.

5.2.1 Introduction

The ability to provide care to both women and newborns during delivery is one of the most important aspects of healthcare and is often used as a metric to assess healthcare as a whole across different countries. C-Sections are one of the most important aspects of delivering babies since it has a considerable impact on the mother's health and well-being. Despite the increased prevalence of this procedure over the last few years, the reasons behind this trend still remain unclear. Reports suggest that this increment is a global phenomenon, with the rate of C-Sections almost tripling from 6.7% to 19.1% between 1990 and 2014 [230, 231]. Research on the impacts of C-Sections has focused on the risk of infection, haemorrhage, organ injury, and complications related to anaesthesia or blood transfusion [232, 233]. There is also a higher risk of complications in subsequent pregnancies, such as uterine rupture, abnormal placental implantation, and the need for hysterectomy [234, 235]. As for the infant, C-Sections can lead to respiratory problems, asthma, and childhood obesity [234]. In light of this, in 2015, World Health Organisation (WHO) stated that C-Sections rates higher than 10% were not associated with a reduction in maternal or newborn mortality, even though other complications could not be fully assessed [236]. In contrast, there is no evidence of the benefits of this procedure for women or babies when there is no clear medical need; therefore, it is paramount to focus on identifying and reducing such cases [231]. It was estimated that in 2018, there were 8.8 million unnecessary C-sections [237]. It was with this in mind that a committee was established in Portugal with the specific purpose of decreasing the percentage of C-Sections nationwide. One of the policies resulting from this committee's work was the reduction of government funding per inpatient C-Section episode for hospitals with rates of C-sections above 25%; as of 2020, the number of C-Sections in Portugal stands at approximately 36.3%, nearing the all-time high of 36.9% in 2009 [238]. Furthermore, studies have shown that several countries could benefit from similar policies [237]. A quantitative analysis estimated that a reduction in C-Sections could save millions of dollars [239] worldwide. Therefore, lowering the proportion of C-Sections can yield health and financial benefits for both institutions and patients alike. With these considerations in mind, we developed a machine-learning algorithm-based support system to assist clinical teams in identifying cases of potentially unnecessary C-Sections. As such, in this paper, we propose to:

- elaborate on how clinical decision support systems for C-Sections can be developed using interoperability standards;
- understand, based on the data collected, which features have the most significant impact on predicting delivery type;
- conduct a concise economic analysis to assess the potential financial impact of implementing the proposed clinical decision support tool;
- compare the system's output with clinicians' responses.

5.2.2   Rationale and Related Work

Regarding the related work, several teams already tackled the potential of predicting the delivery type before birth. We found studies related to predicting a successful vaginal birth after a previous C-section, such as the work of Lipschuetz et al., [240] where a gradient boosting method was used to predict such an event using prenatal data to do so. Grobman et al., [241] performed a similar study with a multivariable logistic regression model. Different modalities of data were also used to predict delivery type. Fergus et al. [242] introduces a method of predicting de- livery type using the fetal heart rate signals. Similarly, the work from Saleem et al. [243] proposed a method for predicting delivery type using interactions between the fetal heart rate and maternal uterine contraction. Finally, there are also studies that focus on predicting the delivery mode like the work of Ullah et al. [244] where a boosting algorithm was used in order to predict delivery mode with enriched datasets. In addition, Gimovsky et al. [245] introduced decision trees to predict C-Section by physician group with 0.73 AUROC. The works of [246] resulted in a seven-variable model with 0.78 AUROC and the works of [247] resulted in a model with 0.82 AUROC, reaching 0.93 with a first cervical examination. Finally, the works of Meyer et al. [248] focused around selecting suitable for a trial of labour after caesarean with AUPRC around 0.351. However, to the best of our knowledge, there was no model tested in clinical practice, with an interoperable format of communication like FHIR, which tried to not only predict delivery type but also provide support about possibly worn deliveries and none with simulation about financial implication, making our paper a potential novelty on different dimensions.

5.2.3   Methods

5.2.3.1   Materials

Data was retrospectively collected from nine different public Portuguese hospitals across the country, focusing on obstetric information, encompassing maternal data, various fetal data points, and the method of delivery in a retrospective manner. The inclusion criteria is all mothers with a registered outcome of the pregnancy from 2019 to 2020. There were no exclusion criteria. ach institution used identical EHR software, ensuring the columns remained consistent.

5.2.3.2   Clinical Comparison

The clinical comparison was performed by sending questionnaires to clinicians with a relationship with obstetrics in order to assess 10 patients, with only access to the variables used by the model and to answer three questions for each. The first was to give a score from 1-10 of how likely that patient would give birth through C-section, then to select the feature/variable that most influenced the decision and which feature they would require to make a better assessment. We sent the questionnaire to 20 people and obtained 6 answers, totaling 60 patient assessments. For these 10 patients, we also predicted the delivery type using our model in order to compare it with the

clinicians' answers. These patients were new and were not seen by the model during the training phase.

5.2.3.3 Analysis

All null representations were standardized. Data were prepossessed by removing features with high missing rates (>90% overall). The imputation process was performed using the KNN imputation method (for continuous variables) or a new category (NULLIMP) for categorical variables. Weight was categorized into percentiles defined specifically for Portuguese babies [249]. For the purpose of this study, the Birth Type was reduced to binary. All assisted birth were merged into vaginal birth and C-Section remained as the other class. Procedures and diagnoses were also used and were encoded as binary features, and we took the time to analyze each one of them in order to avoid leakage because there were procedures obviously related to C-sections and vaginal deliveries. Feature creation was performed through the free-text variable related to the prescribed medication. Medicine names were collected from it and converted into Anatomical Therapeutic Chemical (ATC) Classification Group level 4, which represents chemical subgroups. We also created some new features from data in the dataset, namely new categories related to the labour and condition of the baby. In addition, data quality issues were addressed, such as impossible values that were transformed into null values. The main variables affected by data quality were BMI/Weight and gestational age. The data were split into training and test sets in a 0.75:0.25 manner. From the overall datasets which comprised over 200 columns, only a few columns were selected (please see table 5.4 in the results section). We used a mixture of features selected by surveying the literature [250, 251, 252] and features with a high correlation with the outcome. The tested models were Logistic Regression, Decision Tree, Random Forest, three different Boosting methods (as implemented by eXtreme Gradient Boosting (XGBoost), Light Gradient-Boosting Machine (LightGBM) and *scikit-learn*) and a linear model based on Stochastic Gradient Descent. The evaluation was performed with repeated stratified cross-validation with 10 splits and 2 repetitions, with two full cycles of dividing the training set into 10 equal parts and using 9 as the training set and 1 as the validation set. This rendered table 5.5. The API for serving the prediction model was developed using FastAPI. We wrote all the code in Python 3.9.7.

5.2.3.4 Ethical Considerations

This study received Institutional Review Board approval from all hospitals included in this study with the following references: Centro Hospitalar São João; 08/2021, Centro Hospitalar Baixo Vouga; 12-03-2021, Unidade Local de Saúde de Matosinho; 39/CES/JAS, Hospital da Senhora da Oliveira; 85/2020, Centro Hosptilar Tamega Sousa; 43/2020, Centro Hospitalar Vila Nova de Gaia/Espinho; 192/2020, Centro Hospitalar entre Douro e Vouga; CA-371/2020-0t_MP/CC, Unidade Local de saúde do Alto Minho; 11/2021. All methods were carried out in accordance with relevant guidelines and regulations. The need for informed consent was waived by the ethics committee.

5.2.4 Results

5.2.4.1 Descriptive Statistics

The number of samples varied across the hospitals, ranging from 2364 to 18177. Distributions of the selected variables are presented in table 5.3. The sum of all samples totals 73351.

The outcome variable had the following distribution as stated in table 5.4.

5.2.4.2 The Model

The AUROC is presented in table 5.5 for the best hyper-parameters found for each algorithm in the training data. All models used the variables indicated in table 5.3. While XGBoost was the best-performing algorithm, we selected LightGBM [253] because of its speed and lower memory requirements, which we believe are better suited for deployment in a low-hardware environment. The threshold selected for deploying the model was 0.7457 which rendered the metrics in the test set, as shown in table 5.6.

5.2.4.3 Deployment

The purpose of this model is to serve as an API for usage within a healthcare institution and to act as a supplementary clinical decision support tool for obstetrics teams. For this to happen, a health information system must make the requests to the API. Even though a concrete, vendor-specific information model and input health information system were used, we hope to create a more interoperable clinical decision support system that can be used by every system that acts on birth and obstetrics departments. Therefore, we built it around the HL7 FHIR standard (R5 version) to simplify the method of interacting with the API. This decision, opposed as to using a proprietary model for the data, sits upon the usage of FHIR resources: Bundle and Observation for request and returning the result as a message through a custom operation called "$predict". It is intended to publish the profiles of these objects in order to facilitate access to the API using standardized mechanisms and data models. The current build of the profiles can be seen in the published FHIR Implementation Guide where the current specifications are described in detail `https://joofio.github.io/obs-cdss-fhir/`. The process is illustrated in figure 5.4. We deployed this model in production in a single hospital without a user interface, collecting only the data and predictions for later discussion and analysis. We collected 3231 requests. During this period, 123 (3.8%) alarms were triggered. From this, we tried to understand the level of certainty for the decision and check the difference from the threshold of these alarms. The distance to the threshold for 73 was lower than 0.1 and was bigger than 0.1 for 50 (1.55%) cases.

5.2.4.4 Clinical Evaluation

The median scores given by each clinician are presented in figure 5.5. We also predicted the result using our model as stated in figure 5.5. The model misclassified only one record (4). As for the analysis of missing features for the responders, they were divided into 3 categories: 1) Existent

Table 5.3: Distribution of features used for prediction, Mean and Standard Deviation (SD) for continuous variables. Mode and percentage for categorical variables. Number of samples is 73351.

Variable	M (SD)	Mode [%]
Mother Age	31.0 (5.6)	
Weight pre-pregnancy	65.8 (13.9)	
Weight on admission	78.6 (14.2)	
BMI	25.0 (5.4)	
Previous eutocic delivery	0.4 (0.7)	
Previous vacuum-assisted delivery	0.1 (0.3)	
Previous forceps	0.0 (0.1)	
Previous C-Section	0.1 (0.4)	
Fetal presentation on admission		cephalic [26.323%]
Bishop score	5.5 (3.0)	
Gestational age on admission	38.9 (1.9)	
Premature rupture of the membrane		No [87.991%]
Chronic hypertension		No [97.676%]
Gestational hypertension		No [97.749%]
Preeclampsia		No [98.299%]
Gestational diabetes		No [89.811%]
Gestational diabetes treated with diet		No [94.285%]
Gestational diabetes treated with insulin		No [98.083%]
Gestational diabetes treated with oral antidiabetic drugs		No [97.797%]
Maternal Diabetes		No [99.509%]
Type 1 Diabetes		No [99.816%]
Type 2 Diabetes		No [99.843%]
Presentation at birth		Vertex presentation [94.000%]
Delivery		Spontaneous [53.864%]
Gestational age on birth	39.0 (1.8)	
Smoking during pregnancy		No [88.442%]
Alcohol consumption during pregnancy		No [98.65%]
Consumed drugs during pregnancy		No [99.825%]
Nr of pregnancies (with current)	1.9 (1.1)	
Pregnancy type		Spontaneous [85.417%]
Surveillance		yes [97.699%]
Hospital surveillance		yes [67.807%]
Pelvis Adequacy		Adequate [17.512%]
Consistency of the cervix	1.6 (0.6)	
Fetal station	0.8 (0.8)	
Dilation of the cervix	1.3 (0.8)	
Effacement of the cervix	1.2 (1.2)	
Position of the cervix	0.6 (0.7)	
Haematologic disease		No [95.674%]
Respiratory disease		No [95.605%]
Cerebral disease		No [98.793%]
Cardiac disease		No [92.967%]
Neuroaxis techniques		1 [69.5%]
Number of children	0.6 (0.8)	

Table 5.4: Distribution of Delivery Methods

Type of delivery	Frequency (%)
C-Section	19 803 (27%)
Vaginal	38 189 (52%)
Instrumental delivery	15 359 (21%)

Table 5.5: Repeated Cross-validation (10x2) results in the training set with mean AUROC and 95% Confidence Interval (CI) for best hyper-parameters found for each algorithm. Wilcoxon Test for comparing with the best performing algorithm.

Metric	AUROC	CI 95%	*P*
XGBoost	0.8809	0.8799, 0.882	-
Decision Tree	0.8337	0.8324, 0.8349	≤ 0.001
Logistic Regression	0.8716	0.8706, 0.8726	≤ 0.001
AdaBoost	0.8753	0.874, 0.8766	≤ 0.001
LightGBM	0.8805	0.8793, 0.8817	0.003
Stochastic Gradient Descent	0.8704	0.8694, 0.8713	≤ 0.001
Random Forest	0.8752	0.8743, 0.8762	≤ 0.001

Table 5.6: Performance Metrics in the test set with chosen threshold

Metric	Value
Accuracy	0.8052
Sensitivity	0.8223
Precision	0.9023
F1 Score	0.8605

Figure 5.4: Deployment and decision mechanism of the model
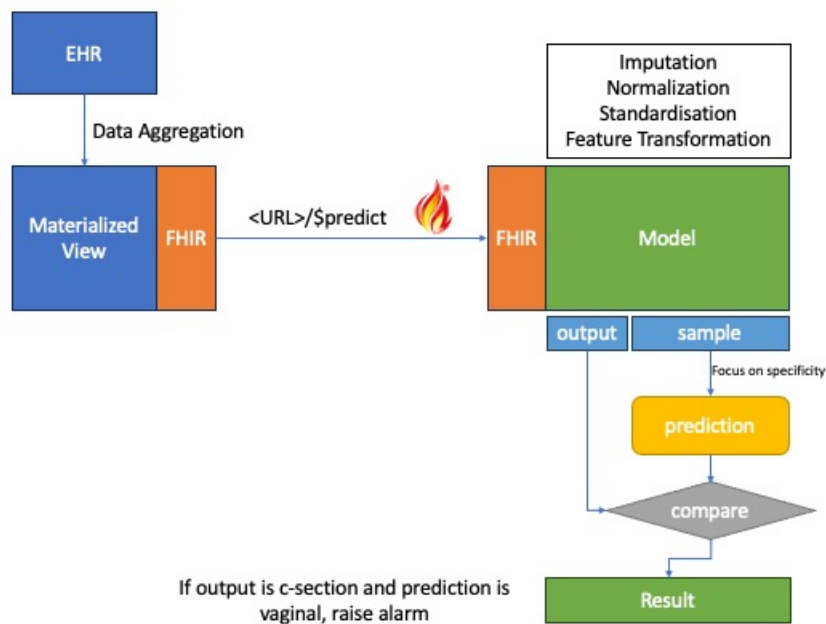
in the dataset but not included in the model, 2) Non-existent in the dataset and 3) existent in the dataset and included but that particular information was not filled for the patient assessed. This rendered a total of 62 % non-existent and 38 % existent but no information was provided at that moment. No feature mentioned existed in the dataset but had not been included in the model. From the non-existent, 38 % were new clinical assessments, 38% were linked to information from previous births, 15% connected in more in-depth information about provided information (i.e., motive for induction) and 11% were related to the mother's choice (if she wanted a C-section). As for feature importance, from the 60 answers, we got 55 % with labour being the most important factor. 15 % answered the number of previous vaginal births, 8 % the evolution of weight and another 8% the number of previous C-sections. The remaining 14% were various features, from BMI, neuroaxis techniques, gestational age and weight of the mother. Of all of these, 90 % were in the top 10 features of the model.

5.2.4.5 Potential Financial Impact

The financial support provided to public hospitals in Portugal is partially tied to the rate of C-sections. To assess the potential impact of this mechanism on Portuguese public hospitals, we conducted a simulation. We got data for every public hospital for the last 12 months and applied a 3.8% reduction (the rate of warnings triggered in the new dataset) and recalculated the rate of C-sections. The increase in support was calculated by the state-mandated rate as shown in table 5.7. With this new rate, we observed that implementing our tool would result in financial benefits for 30% (11 hospitals) of the public hospitals. Specifically, five hospitals would begin receiving support instead of no support at all. Three hospitals would experience a doubling of their financial

Figure 5.5: Validation data. The colour represents the actual birth type. The boxplot represents the median and IQR of the reviewers and the X represent each patient case. Contains 6 Vaginal births and 4 C-Sections. * represents wrong predictions of the model. (ID: 4)
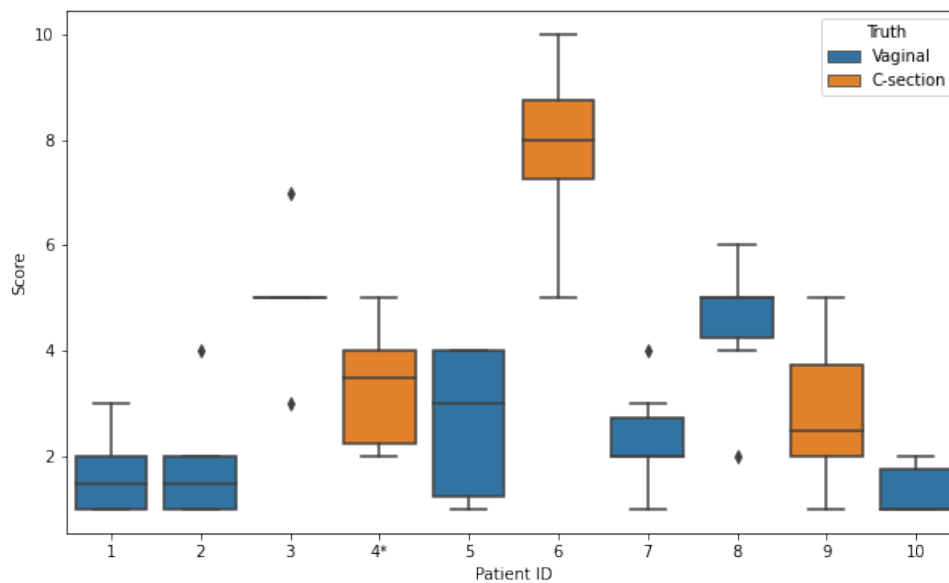
benefit, while two hospitals would see a 50% increase. Furthermore, one hospital would receive an additional one third of financial support. If we assumed that only half of the warnings found in the new data were actually true (1.9%) we found that only 6 hospitals would be benefited. 3 from 0 to 0.25, 2 from 0.25 to 0.50 and 1 from 0.50 to 0.75.

Table 5.7: Ruleset for state-provided financial support indexed to C-Sections. X is the current payment of a C-Section inpatient episode. Adapted from [254]

Rate of C-Sections	Support
<25%	x
[25%, 26.4%]	0.75 x
[26.5%, 27.9%]	0.5 x
[28%, 29.4%]	0.25 x
>29.5%	0

5.2.5 Discussion

The first thing to address about this model is the number of biases that we introduced in the model by choice. We joined all vaginal delivery types into a single category (assisted and non-assisted) which introduces a bias since these delivery modes are indeed different. Secondly, the fact that we want to predict if the delivery type was wrongly chosen, mainly for the case of a C-section

that did not need to be so, is also a bias. We used this approach because the initially collected data did not have the representation of such events. So the biases of possibly wrong delivery types were present in the training data. We attempted to minimize this issue by selecting a threshold that gave the model higher sensitivity than specificity so that only large probabilities would trigger an alarm for human consideration. Parallel to this, we are starting to gather labelled cases, with the help of clinicians in order to create a better training dataset. Furthermore, since the data was collected from different hospitals, differences in the data input can also occur. Even though the health information system is the same, the processes that originate the data and are being used for secondary purposes could introduce several biases in the data. This is an issue that was accepted from the start regarding the mechanism of data collection and model training. Despite this, we reached a model with a very high AUROC (88%, 95%CI [0.8795, 0.8815]), which is encouraging and versus the state of the art. Moreover, assuming that more data is provided and proper labelling is done regarding the outcome variable (like a clinical evaluation of needless C-sections) is added as well, a better model could be developed. Regarding the preliminary clinical evaluation, it was only possible to get an overview of the possible comparison due to the number of responders. Despite that, the results are encouraging, since the model seems to behave better than humans with the data provided. However, this is a biased vision, since clinicians in the real world have access to more data and information than the model has. It is encouraging, but caution is advised before more tests and evaluations are done. As for the deployment, future work could be the improvement of the API in order to map all variables to an ontology like SNOMED CT or similar, making it easier for every system and person to access it and get a suggestion of the delivery type. Finally, we believe the assessment can be improved. A more robust clinical assessment is necessary as well as a thorough analysis of the impact of the tool in the real world, since we need to create the bridge between the results of the model and how clinical decisions are affected by it. A full cost-effectiveness analysis is also necessary to understand the real world impact of the model. One interesting result is the fact that 38% of the answers regarding the most important data element missing from the patient record refers to data that is being collected but was missing for that specific patient, raising an important question about data input methodology, interoperability and quality. If we cannot have access to data when it matters most, it can become meaningless. Missing data is a problem of biomedical data as a whole. However, when specifically targeted at ML usage of this data for predicting something, we did not find any works comparing them with clinicians. However, we did find reportsof similar missing values in obstetrics data [255] and we also found works of similar nature using ML models with a robust handling of missing data such as XGBoost [256] to counter this problem. This indicates that our model has the potential to counter the missing data problem as well since LightGBM can also handle missing data natively.

5.2.6   Conclusion

We believe we have developed a robust system capable of detecting potentially incorrect C-section decisions, which could positively impact real-world medical practice. However, before implementation, several challenges must be addressed, particularly the need for further evaluation of

the system's impact on clinical decision-making and the reasons underlying sub-optimal delivery type decisions. C-sections may be performed for various reasons, from a mother's preference to a decision made by the obstetrics team. This system is not designed to impede medical practice or to highlight flawed decisions, potentially scrutinizing specific professionals. Such caution is necessary when implementing systems like these. While having a high AUROC is beneficial, the real-world impact is another consideration. The assumptions and biases associated with autonomous systems supporting clinical practice must be carefully considered. Nonetheless, the metrics and results we have achieved so far are promising for positively influencing health and economic outcomes.

We never are definitely right, we can only be sure we are wrong.

Richard P. Feynman

6

Discussion

Extracting knowledge from healthcare data is a complex and multifaceted challenge. This endeavour is contingent upon the availability and the interpretability of data. In this chapter, we delve into the key contributions of this thesis and the primary challenges encountered throughout the work, elucidating how these obstacles were navigated.

This thesis contributed to supporting the fact that generating synthetic data is not as easy as it may seem at first glance and it's not the silver bullet as advertised in some settings. The need for robust evaluation metrics, especially for tabular datasets, is still present. The fact that the current ones are similar to those used in the generators (i.e., GANs) can introduce a significant bias. The old saying that states that when a metric becomes a goal, it stops being the appropriate metric fits here perfectly. But not all are bad news. Synthetic data has its place in healthcare and can serve many purposes. Synthetic data, even with very low utility (as in the sense used throughout the thesis, like similarity to the original), can still be useful. Having a broad idea of how the variables are distributed, the possible categories, and a generic idea of null values can go a long way to accelerate data analysis. We would certainly enjoy that at least in the works stated in sections 4.1, 5.1 and 5.2.

Another contribution lies in the fact that we can implement several mechanisms to extract knowledge from data without taking the data from its original repositories. We have proved that distributed learning is not inferior to the gold standard (centralized) or to the local counterparts. We also showed that we can even compare performance across hospitals or health institutions, without sharing the actual values, which certainly will be appreciated by many boards. The technology is here, evidence exists, it's a matter of implementation, trust, and engineering.

Finally, we have shown (like others before us) that we can use data to support clinical decision-making. However, we have also seen that the path from data to decision-making is not as easy as it may seem. We need to have a robust infrastructure, a trust framework, and a legal and technical framework to support the tools developed. We need to have the clinicians involved in the process, and we need to have the tools be as transparent and explainable as possible. We need to have the tools be reliable and accountable. We need to have the tools be easy to use and to understand. We need to have the tools be useful. We need to have the tools be impactful. We need to have the tools be ethical. We need to have the tools be safe. We need to have the tools be innovative. We need to have the tools be the future of healthcare. We need to have the tools be the future of healthcare data science. This is all we need.

There could even be the argument that implementing a CDSS should be as hard as implementing a new drug. The impact of a CDSS could be as big as a new drug, and the risks could be as big as well. Should we implement RCT to prove the effectiveness of such systems? Or, like first discussed in chapter 1, we can use observational data to support the claims of efficacy. This leads to the final contribution where we tried to use novel approaches to data analysis and provide claims of causality using observational data based real-world data. Mechanisms like IPTW are useful and can provide insights into the causality framework of treatments. Personally, I do not believe we are **there** yet, but I would argue that we are close.
The following sections will go deeper on the limitation, hurdles and methods we employed to surpass them.

6.1   Accessing Data

The first problem is getting access to data. The data is not always available, and when it is, it is not always in the format we need. Ethics committees and Data Protection Officer (DPO) requirements are put in place in order to guarantee the patient's privacy and security, but a lot of times at the cost of timely access to data. We consider that synthetic data can have a good impact on this work. While we can leave the legal processes be, we may use synthetic data with a heavy focus on security to develop and test our algorithms. This is a very promising area of research, and we believe it will be a game-changer in the future. Parallel to this approach are distributed paradigms. Having a distributed approach to data analysis could be of great help. This would allow for the data to be analysed in its original location in a more secure way and timely manner. If metrics and models could be built by local teams and shared across regions and/or countries to leverage the power of the many for single institutions could be groundbreaking. However, underlying both these approaches are data dictionaries and data governance tools. Having the correct functional/clinical description of data could be of great impact on the usage of data. Having already the variables defined as categorical, numerical and so on could be of great help. This is a very important aspect of data science, and it is often overlooked. Simple statistics of datasets

could be useful as well. For example, the number of missing values, the number of unique values, the number of outliers, and so on. This would help the data scientist to understand the data better and to know what to expect from it.

6.2 Data Quality

This point relates to the second big hurdle of knowledge extraction from healthcare data - quality. As discussed in section 3.4, this is a very complex and sometimes elusive concept. In our case, this implied a lot of time spent with data preprocessing. We had to deal with missing values, outliers, and correctness in the context of the records, and data in different formats. We also had to link together different databases from different HISs which brought to light new problems like the new dimensions of correctness of data. There is a common saying that sums this pretty well *When we have one watch, we know the time, but when we have two, we may never know*. So if we had different information regarding the same variable in different systems, how to decide what is true? Another aspect that is often overlooked is the relationship with the clinicians. We need to understand that they are the ones who will use the tools we develop, and they need to be involved in the process. We need to understand their needs and their workflow. Furthermore, we need to understand what they need and how they need it. We need to understand that they are not data scientists, and they do not have the time to learn how to use our tools. We need to make it easy for them to use our tools. Now healthcare is often explained in terms of clinical teams of different backgrounds. A similar concept could be beneficial for harvesting knowledge from data.

6.3 Building robust software to support AI

Building software or tools based on this data is still an early subject that possibly requires a legal and technical framework. A legal is connected to the impact of such tools in healthcare. If drugs require such a long time to be approved in order to assess security, how can we approve a tool that can have a similar impact? A technical framework is connected to the fact that we are still in the early stages of a new HEADS paradigm. We are still trying to understand how to use data, and how to extract knowledge from it. We are still trying to understand how to evaluate the performance of our tools. We are still trying to understand how to evaluate the impact of our tools in healthcare in a timely manner in a way that is not biased and that is not too expensive. Imposing similar structures to drugs is ill-advised since it could possibly kill the innovation potential and the interest in providing such tools. And this is where a quality infrastructure could be of use. Seriously betting of biomedical informatics could render huge payoffs down the line. Having the human and material resources to build data infrastructures on local (healthcare institutions) and regional, or even country-wise or cross-country policies to use effective use healthcare data is essential. At the time of the writing of this thesis, examples like EHDS are very promising initiatives that could help to overcome the hurdles of data availability and quality. However, cross-country initiatives will always be as good as the weakest link, so it is important to have a common

framework and a common goal and to have the resources to achieve it. In concrete, having data pipelines, data governance and data interoperability tools, and data quality tools are essential. Having a common data dictionary and a common data format would also be of great help. This would allow for a more efficient use of data, and it would allow for the use of healthcare data to drive innovation. Tightly connected with this is the possibility of having Real World Evidence (RWE) support clinical decisions live. Having data like the one produced in 5.1 in real-time or with high update frequency could be leveraged in order to further support clinicians in making decisions based on data. However, we would require not only the premises already discussed, like data quality and cross-collaboration clinics, but a trust-framework would also be necessary. In order to make the automatic dashboard and metrics reliable, transparency is key. Having explainability and transparency in the process of evidence production will be key to building trust and accountability.

6.4   Evaluation of AI tools

The challenges of extracting knowledge from healthcare data are multi-faceted, as evident from the issues of data access, quality, and the complex relationship with clinicians. Another vital aspect is the integration of RWE into clinical decision-making processes. RWE, derived from data collected outside controlled clinical trials, offers immense potential for informing healthcare decisions. However, its integration requires meticulous attention to data quality, governance, and transparency. As healthcare data becomes increasingly digitized and voluminous, the opportunity to leverage RWE in real-time or with high-frequency updates grows. This could significantly enhance the ability of clinicians to make data-driven decisions. However, for RWE to be effectively integrated, it necessitates not only robust data infrastructure but also a trust framework. Clinicians and patients alike must have confidence in the accuracy, reliability, and transparency of the data and the algorithms used. Building this trust involves ensuring that data processing and decision-making algorithms are transparent and explainable, fostering a sense of accountability and reliability in the system.

Furthermore, the evolution of healthcare data science underscores the need for a comprehensive legal and technical framework. The comparison to drug approval processes highlights the importance of stringent evaluation for healthcare tools, balancing safety and innovation. The legal framework should address the ethical implications and societal impact of these tools, while the technical framework should focus on performance evaluation, data extraction techniques, and impact assessment. Establishing such frameworks is crucial for navigating the complexities of HEADS and for fostering an environment where innovation can thrive without compromising patient safety or data integrity. This approach also involves the creation of quality infrastructures, emphasizing biomedical informatics, and developing robust data infrastructures at various levels, from local healthcare institutions to regional and international collaborations.

6.5 Cross-disciplinary collaboration

The future trajectory of healthcare data science is heavily reliant on cross-disciplinary collaboration and shared frameworks. Initiatives like the EHDS signify positive strides towards enhanced data availability and quality through collaborative efforts. However, the efficacy of such initiatives is contingent upon the uniformity of standards, shared objectives, and adequate resourcing among all participating entities. Essential measures include establishing common data dictionaries, formats, and interoperability tools. These collaborative endeavours are pivotal in streamlining data usage, thereby catalysing innovation in healthcare. An integrative approach, melding technical expertise with legal and ethical considerations, is crucial for harnessing the full potential of healthcare data in improving patient outcomes and advancing medical science.

6.6 Summing up

In conclusion, this thesis presents a comprehensive examination of the multifarious aspects involved in harnessing healthcare data for knowledge extraction. The identified challenges and proposed solutions underscore the intricate interplay between data accessibility, quality, and interdisciplinary collaboration. The synthesis of this research contributes significantly to the field, providing a nuanced understanding of the complexities involved in leveraging healthcare data. This work not only advances academic knowledge but also holds the potential to inform and transform practical applications in healthcare, ultimately aiming to enhance patient care and outcomes.

I may not have gone where I intended to go, but I think I
have ended up where I needed to be.

Douglas Adams

7

Conclusion

On a more personal note, I want to conclude this thesis, taking a step back and contemplate the work done so far. What I could have been done better and what I have learned throughout this journey. But I also want to look ahead and point out possible directions in order to take the aim of this thesis further and create a substantial impact in the real world.

7.1 Looking Back

Reflecting on these last five years, my initial impression is one of shame regarding the earlier works in this thesis. Although it bothered me at first, I now believe it is a sign of growth and maturity. I want to believe that it shows how much I have learned and developed throughout these 5 years.

I have also discovered the value of collaboration and the importance of having a diverse team. The most successful projects are those that embrace interdisciplinary collaboration. Thus, the future trajectory of this field may well hinge on fostering such diverse, collaborative environments, enriching the scope and impact of healthcare data science.

Even though I've always liked the saying, 'If you want to go fast, go alone. If you want to go far, go together,' I never fully understood its importance until now. I've always tried to be a 'one-man show,' but there are limits to this approach, whether due to time or knowledge."

In contemplating the limitations of this thesis, it's pivotal to recognize that each project is inherently constrained by its unique focus and context. These projects were tailored to specific use cases and therefore may not offer broad generalizability. For instance, the work done in section 5.2 and 4.1 are oriented around specific data types and formats and ML models. Similarly, the project

111

in 3.4 is circumscribed to a certain data type within a specific clinical specialty. This specificity implies that the outcomes of these projects might not be directly extrapolatable to other diseases or data types. However, it is crucial to note that the methodologies employed are versatile. For example, the real-time prediction techniques used in 5.2 and the data analysis models in 5.1 and 4.1 can be adapted for other contexts. Additionally, the 3.3 method offers a versatile approach for analysing diverse datasets and can be seamlessly integrated into various data pipelines. But I cannot help to feel that the bigger limitation of this thesis is the lack of real-world deployment. While I have tried to simulate real-world scenarios, the lack of real-world deployment is a limitation. The deployment of real-world CDSS is, however, a complex undertaking requiring substantial investment in terms of time, finances, and perseverance. And of course, on top of this is the fact that we can bring problems to institutions and clinical workflow. Consequently, this aspect of tool deployment remains an avenue for future exploration. Nevertheless, extensive testing in real-world settings and the inclusion of clinicians in the developmental process imbue confidence in the readiness and potential impact of these tools upon deployment.

7.2   Looking Ahead

Looking towards future endeavours, the foundation established in this thesis paves the way for practical aid to healthcare teams. The deployment of real-world CDSS is, however, a complex undertaking requiring substantial investment in terms of time, finances, and perseverance. Consequently, this aspect of tool deployment remains an avenue for future exploration. Nevertheless, extensive testing in real-world settings and the inclusion of clinicians in the developmental process imbue confidence in the readiness and potential impact of these tools upon deployment.

The comprehensiveness of this work necessitated a confluence of knowledge from disparate domains. This included insights from biology and chemistry for understanding healthcare nuances, process design for process formalization, mathematics and statistics for ML and Exploratory Data Analysis, and interoperability standards for data amalgamation. Furthermore, ethical and privacy considerations were paramount for ensuring patient confidentiality and developing ethically sound models. Delving into healthcare terminologies, codifications, and semantics was essential for data interpretation, along with familiarity with clinical specialties like obstetrics and oncology. Bridging the gap between RCTs and observational or Real-World Data required adeptness in study design.

This multidisciplinary nature of the field underscores a key challenge: the necessity for a diverse skill set or, alternatively, a collaborative team with varied expertise. Our observations underscore that the most successful projects in this domain are those that embrace such interdisciplinary collaboration. Therefore, the future trajectory of this field may well hinge on fostering such diverse, collaborative environments, enriching the scope and impact of healthcare data science.

A

A.1 Data Dictionary

Acronym	Description
IA	Mother Age
GS	Blood Group
PI	Weight at the beginning of pregnancy
PAI	Weight on Admission
IMC	BMI
CIG	If Smoker During Pregnancy
APARA	Number of previously born babies
AGESTA	Number of Pregnancies
EA	Number of Previous Eutocic Deliveries with no assistance
VA	Number of Previous Eutocic Deliveries with help of vacuum extraction
FA	Number of Previous Eutocic Deliveries with help of forceps
CA	Number of Previous C-sections
TG	Pregnancy Type (spontaneous, In vitro fertilisation...)
V	If the pregnancy was accompanied by MD
NRCPN	Number of prenatal consultations
VH	If the pregnancy was followed by a MD in a hospital
VP	If the pregnancy was followed by a MD in a private clinic
VCS	If the pregnancy was followed by a MD in a primary care facility
VNH	If the pregnancy was followed by a MD in the same hospital the delivery was made
B	Pelvis Adequacy
AA	Baby's Position on Admission
BS	Bishop Score
BC	Bishop Score Cervical Consistency
BDE	Bishop Score Fetal Station
BDI	Bishop Score Dilatation
BE	Bishop Score Effacement
BP	Bishop Score Cervical Position
IGA	Number of Weeks on Admission
TPEE	If the delivery was spontaneous
TPEI	If the delivery was induced
RPM	If there was a rupture of the amniotic pocket before delivery began
DG	Gestational Diabetes
TP	Delivery Type
ANP	Baby's Position on Delivery
TPNP	Actual Type of Delivery
SGP	Pregnancy Weeks on Delivery
GR	Robson Group

B

B.1 C-section assessment questionnaire

Hospital X

Mãe

Idade da grávida no parto	32
Peso da grávida no início da gravidez	45
Peso da grávida quando é internada para ter o bébé.	S/Info.
IMC da grávida no início da gravidez	17.6
Nº partos eutócicos anter.	0
Nº partos distócicos anter. via vaginal com ventosas	0
Nº partos distócicos anter. via vaginal com forcéps	0
Nº de partos cesarianas anteriores	0
Hipertensão Crónica	S/Info.
Hipertensão Gestacional	S/Info.
Hipertensão Pré-eclâmpsia	S/Info.
Diabetes Gestacional	S/Info.
Diabetes Gestacional com Dieta	S/Info.
Diabetes Gestacional com Insulina	S/Info.
Diabetes Gestacional com antidiabéticos orais	S/Info.
Diabetes Materna	S/Info.
Diabetes Tipo 1	S/Info.
Diabetes Tipo 2	S/Info.
Fumou durante a gestação	S/Info.
Ingeriu álcool durante a gestação	S/Info.
Utilizou estupefacientes durante a gestação	S/Info.
Nº de gestações que teve (esta incluída)	2
Tipo de gravidez actual (se foi espontânea, FIV, etc)	Esp.
Se a gravidez foi vigiada (>= 5 consultas)	Sim
Se a gravidez foi vigiada no mesmo hospital do parto	Sim
Nº de filhos nascidos	0
Doença Hematológica	S/Info.
Doença Respiratória	S/Info.
Doença Cerebral	S/Info.
Doença Cardiaca	S/Info.

Evolução Peso

Percentil peso semana 39	S/Info.
Percentil peso semana 38	S/Info.
Percentil peso semana 37	25th-50th
Percentil peso semana 36	S/Info.
Percentil peso semana 35	10th-25th
Percentil peso semana 34	S/Info.
Percentil peso semana 33	S/Info.
Percentil peso semana 32	S/Info.
Percentil peso semana 31	10th-25th
Percentil peso semana 30	S/Info.
Percentil peso semana 29	S/Info.
Percentil peso semana 28	S/Info.
Percentil peso semana 27	S/Info.

Parto

Semanas de gestação na admissão	39.4
Rotura prematura de membranas	S/Info.
Apresentação no parto	Cefálica de vértice
Trabalho de parto	Espontâneo
Semanas de gestação no momento do Parto	39.4
Posição bebé na 1ª verificação no hospital	S/Info.
Avaliação da pelve óssea	S/Info.
BISHOP Consistência	S/Info.
BISHOP Descida	S/Info.
BISHOP Dilatação	S/Info.
BISHOP Extinção	S/Info.
BISHOP Posição	S/Info.
BISHOP Score	S/Info.
Técnicas do neuroeixo	Sim

Histórico Apresentação

Apresentação semana 39	S/Info.
Apresentação semana 38	S/Info.
Apresentação semana 37	cefálica
Apresentação semana 36	S/Info.
Apresentação semana 35	cefálica
Apresentação semana 34	S/Info.
Apresentação semana 33	S/Info.
Apresentação semana 32	S/Info.
Apresentação semana 31	cefálica
Apresentação semana 30	S/Info.
Apresentação semana 29	S/Info.
Apresentação semana 28	S/Info.
Apresentação semana 27	S/Info.

Com base nesta informação, diga de 1 a 10 quão provável seria originar uma cesariana.

Com base nesta informação, qual a característica/variável/elemento apresentado(a) acima que mais impactou a sua decisão?

Qual a característica/variável/elemento não existente que gostaria de ter para avaliar melhor?

B.2   Data quality questionnaire

Hospital X

Ficha nº 1

Mãe

Idade da grávida no parto	37.0
Grupo sanguíneo da grávida	0,RH_POSITIVO
Peso da grávida no início da gravidez	56.0
Peso da grávida quando é internada para ter o bebé.	S/ Info.
IMC da grávida no início da gravidez	21.9
Nº partos eutócitos anter. via vaginal sem nada	S/ Info.
Nº partos eutócitos anter. via vaginal com ventosas	S/ Info.
Nº partos eutócitos anteriores, via vaginal com fórceps	S/ Info.
Nº de partos cesarianas anteriores	S/ Info.
Posição bebé na 1ª verificação no hospital	S/info
BISHOP Score	S/ Info.
Semanas de gestação na admissão	34.0
Hipertensão Crónica	S/ Info.
Hipertensão Gestacional	S/ Info.
Hipertensão Pré-eclâmpsia	S/ Info.
Diabetes Gestacional	S/ Info.
Diabetes Gestacional com Dieta	S/ Info.
Diabetes Gestacional com Insulina	S/ Info.
Diabetes Gestacional com antidiabéticos orais	S/ Info.
Diabetes Materna	S/ Info.
Diabetes Tipo 1	S/ Info.
Diabetes Tipo 2	S/ Info.
Fumou durante a gestação	S/ Info.
Ingeriu álcool durante a gestação	S/ Info.
Utilizou estupefacientes durante a gestação	S/ Info.
Nº de gestações que teve (esta incluída)	2
Tipo de gravidez actual (se foi espontânea, FIV, etc)	ESPONTANEA
Se a gravidez foi vigiada (>= 5 consultas)	Sim
Se a gravidez foi vigiada no mesmo hospital do parto	S/ Info.
Avaliação da pelve óssea	S/ Info.
Doença Hematológica	S/ Info.
Doença Respiratória	S/ Info.
Doença Cerebral	S/ Info.
Doença Cardíaca	S/ Info.
Nº de filhos nascidos	S/ Info.

Parto

Tipo de gravidez actual (se foi espontânea, FIV, etc)	ESPONTANEA
Altura uterina. Medição da altura/tamanho da barriga em cm.	S/ Info.
Avaliação da pelve óssea	S/ Info.
Posição bebé na 1ª verificação no hospital	S/info
BISHOP Score	S/ Info.
BISHOP Consistência	S/ Info.
BISHOP Descida	S/ Info.
BISHOP Dilatação	S/ Info.
BISHOP Extinção	S/ Info.
BISHOP Posição	S/ Info.
Semanas de gestação na admissão	34.0
Indica se o trabalho de parto foi espontâneo	SIM
Indica se o trabalho de parto foi induzido	S/ Info.
Rotura prematura de membranas	S/ Info.
tipo de parto realizado da gravidez atual.	Parto eutócico cefálico
Apresentação no momento do parto	Cefálica de vértice
Trabalho de parto	Espontâneo
Semanas de gestação no momento do Parto	34.1
Classificação de Robson	10

Evolução Peso

Estimativa peso eco 24	S/ Info.
Estimativa peso eco 25	S/ Info.
Estimativa peso eco 26	S/ Info.
Estimativa peso eco 27	S/ Info.
Estimativa peso eco 28	S/ Info.
Estimativa peso eco 29	S/ Info.
Estimativa peso eco 30	S/ Info.
Estimativa peso eco 31	S/ Info.
Estimativa peso eco 32	S/ Info.
Estimativa peso eco 33	S/ Info.
Estimativa peso eco 34	2027.0
Estimativa peso eco 35	S/ Info.
Estimativa peso eco 36	S/ Info.
Estimativa peso eco 37	S/ Info.
Estimativa peso eco 38	S/ Info.
Estimativa peso eco 39	S/ Info.
Estimativa peso eco 40	S/ Info.
Estimativa peso eco 41	S/ Info.
Estimativa peso eco 42	S/ Info.

Histórico Apresentação

Apresentação na semana 42	S/ Info.
Apresentação na semana 41	S/ Info.
Apresentação na semana 40	S/ Info.
Apresentação na semana 39	S/ Info.
Apresentação na semana 38	S/ Info.
Apresentação na semana 37	S/ Info.
Apresentação na semana 36	S/ Info.
Apresentação na semana 35	S/ Info.
Apresentação na semana 34	cefálica
Apresentação na semana 33	S/ Info.
Apresentação na semana 32	S/ Info.
Apresentação na semana 31	S/ Info.
Apresentação na semana 30	S/ Info.
Apresentação na semana 29	S/ Info.
Apresentação na semana 28	S/ Info.
Apresentação na semana 27	S/ Info.
Apresentação na semana 26	S/ Info.
Apresentação na semana 25	S/ Info.
Apresentação na semana 24	S/ Info.

118

Bibliography

[1] Julia Adler-Milstein and Ashish K. Jha. "HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption". In: *Health Affairs* 36.8 (Aug. 2017), pp. 1416–1422. ISSN: 0278-2715, 1544-5208. DOI: 10.1377/hlthaff.2016.1651.

[2] Clemens Scott Kruse, Anna Stein, Heather Thomas, and Harmander Kaur. "The Use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature". In: *Journal of Medical Systems* 42.11 (Nov. 2018), p. 214. ISSN: 0148-5598, 1573-689X. DOI: 10.1007/s10916-018-1075-6.

[3] Venkataraman Palabindala, Amaleswari Pamarthy, and Nageshwar Reddy Jonnalagadda. "Adoption of Electronic Health Records and Barriers". In: *Journal of Community Hospital Internal Medicine Perspectives* 6.5 (Jan. 2016), p. 32643. ISSN: 2000-9666. DOI: 10.3402/jchimp.v6.32643.

[4] Barbara Di Camillo, Giuseppe Nicosia, Francesca Buffa, and Benny Lo. "Guest Editorial Data Science in Smart Healthcare: Challenges and Opportunities". In: *IEEE Journal of Biomedical and Health Informatics* 24.11 (Nov. 2020), pp. 3041–3043. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2020.3028398.

[5] Niels Peek and Pedro Pereira Rodrigues. "Three Controversies in Health Data Science". In: *International Journal of Data Science and Analytics* 6.3 (Nov. 2018). https://doi.org/10.1007/s41060-018-0109-y, pp. 261–269. ISSN: 2364-4168. DOI: 10.1007/s41060-018-0109-y.

[6] J. van der Lei. "Use and Abuse of Computer-Stored Medical Records". In: *Methods of Information in Medicine* 30.2 (Apr. 1991), pp. 79–80. ISSN: 0026-1270.

[7] Sertkaya, Aylin, Birkenbach, Anna, Berlind, Ayesha, and Eyraud, John. *Examination of Clinical Trial Costs and Barriers for Drug Development*. Tech. rep. Eastern Research Group, Inc., July 2014. URL: https://aspe.hhs.gov/sites/default/files/private/pdf/77166/rpt_erg.pdf (visited on 08/21/2023).

[8] Matthew Michelson and Katja Reuter. "The Significant Cost of Systematic Reviews and Meta-Analyses: A Call for Greater Involvement of Machine Learning to Assess the Promise of Clinical Trials". In: *Contemporary Clinical Trials Communications* 16 (Aug. 2019), p. 100443. ISSN: 2451-8654. DOI: 10.1016/j.conctc.2019.100443.

[9] Raghad Muhiyaddin, Alaa A. Abd-Alrazaq, Mowafa Househ, Tanvir Alam, and Zubair Shah. "The Impact of Clinical Decision Support Systems (CDSS) on Physicians: A Scoping Review". In: *The Importance of Health Informatics in Public Health during a Pandemic*. IOS Press, 2020, pp. 470–473. DOI: 10.3233/SHTI200597. URL: https://ebooks.iospress.nl/doi/10.3233/SHTI200597.

[10] E. Kilsdonk, L. W. Peute, and M. W. M. Jaspers. "Factors Influencing Implementation Success of Guideline-Based Clinical Decision Support Systems: A Systematic Review and Gaps Analysis". In: *International Journal of Medical Informatics* 98 (Feb. 2017), pp. 56–64. ISSN: 1872-8243. DOI: 10.1016/j.ijmedinf.2016.12.001.

[11] Eric J. Topol. "High-Performance Medicine: The Convergence of Human and Artificial Intelligence". In: *Nature Medicine* 25.1 (1 Jan. 2019), pp. 44–56. ISSN: 1546-170X. DOI: 10.1038/s41591-018-0300-7.

[12] *Why Do 87% of Data Science Projects Never Make It into Production?* July 2019. URL: https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/ (visited on 08/04/2023).

[13] Shannon C. Walker et al. "Model-Guided Decision-Making for Thromboprophylaxis and Hospital-Acquired Thromboembolic Events Among Hospitalized Children and Adolescents: The CLOT Randomized Clinical Trial". In: *JAMA Network Open* 6.10 (Oct. 13, 2023), e2337789. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2023.37789.

[14] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020. ISBN: 9780134610993. URL: http://aima.cs.berkeley.edu/.

[15] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. "Big Data in Healthcare: Management, Analysis and Future Prospects". In: *Journal of Big Data* 6.1 (June 19, 2019), p. 54. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0217-0.

[16] Tom M Mitchell. *Machine learning*. Vol. 1. 9. McGraw-hill New York, 1997.

[17] Casey Ross Swetlitz Ike. *IBM's Watson Supercomputer Recommended 'unsafe and Incorrect' Cancer Treatments, Internal Documents Show*. STAT, July 25, 2018. URL: https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ (visited on 10/18/2023).

[18] Jian Gao and Dashun Wang. *Quantifying the Benefit of Artificial Intelligence for Scientific Research*. Apr. 2023. DOI: 10.48550/arXiv.2304.10578. arXiv: 2304.10578 [physics]. URL: http://arxiv.org/abs/2304.10578 (visited on 05/09/2024).

[19] Stacey Tobin, Bamini Jayabalasingham, Sarah Huggett, and Maria de Kleijn. "A Brief Historical Overview of Artificial Intelligence Research". In: *Information Services & Use* 39.4 (Jan. 2019), pp. 291–296. ISSN: 0167-5265. DOI: 10.3233/ISU-190060. (Visited on 05/09/2024).

[20] *AI Index Report 2024 – Artificial Intelligence Index*. 2024. URL: https://aiindex.stanford.edu/report/ (visited on 05/09/2024).

[21] European Commission. *A Definition of AI: Main Capabilities and Disciplines*. Tech. rep. European Commission, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (visited on 08/03/2023).

[22] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson. "Evidence Based Medicine: What It Is and What It Isn't." In: *BMJ : British Medical Journal* 312.7023 (Jan. 13, 1996), pp. 71–72. ISSN: 0959-8138. pmid: 8555924.

[23] Achilleas Thoma and Felmont F. Eaves III. "A Brief History of Evidence-Based Medicine (EBM) and the Contributions of Dr David Sackett". In: *Aesthetic Surgery Journal* 35.8 (Nov. 1, 2015), NP261–NP263. ISSN: 1090-820X. DOI: 10.1093/asj/sjv130.

[24] Trisha Greenhalgh. *How to Read a Paper: The Basics of Evidence-based Medicine and Healthcare*. 6th ed. How To. Newark: Wiley, 2019. ISBN: 978-1-119-48472-1.

[25] John Graunt. "John Graunt on Causes of Death in the City of London". In: *Population and Development Review* 35.2 (2009), pp. 417–422. ISSN: 00987921, 17284457. URL: http://www.jstor.org/stable/25487673 (visited on 10/15/2023).

[26] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases". In: *AI Magazine* 17.3 (Mar. 1996), p. 37. DOI: 10.1609/aimag.v17i3.1230.

[27] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. *CRISP-DM 1.0 Step-by-step data mining guide*. Tech. rep. The CRISP-DM consortium, Aug. 2000. URL: https://maestria-datamining-2010.googlecode.com/svn-history/r282/trunk/dmct-teorica/tp1/CRISPWP-0800.pdf (visited on 08/21/2023).

[28] Seyyed Soroush Rohanizadeh and Mohammad Bameni Moghadam. "A Proposed Data Mining Methodology and Its Application to Industrial Procedures". In: *Journal of Industrial Engineering* (2009).

[29] Trishan Panch, Heather Mattie, and Leo Anthony Celi. "The "Inconvenient Truth" about AI in Healthcare". In: *npj Digital Medicine* 2.1 (Aug. 2019), p. 77. ISSN: 2398-6352. DOI: 10.1038/s41746-019-0155-4.

[30] IEEE. "IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries". In: *IEEE Std 610* (1991), pp. 1–217. DOI: 10.1109/IEEESTD.1991.106963.

[31] Jessica S Ancker, Lisa M Kern, Alison Edwards, Sarah Nosal, Daniel M Stein, Diane Hauser, and Rainu Kaushal. "How Is the Electronic Health Record Being Used? Use of EHR Data to Assess Physician-Level Variability in Technology Use". In: *Journal of the American Medical Informatics Association : JAMIA* 21.6 (Nov. 2014), pp. 1001–1008. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2013-002627. pmid: 24914013.

[32] Nicole Gray Weiskopf and Chunhua Weng. "Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research". In: *Journal of the American Medical Informatics Association : JAMIA* 20.1 (2013), pp. 144–151. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2011-000681. pmid: 22733976.

[33] Ricardo Cruz-Correia, Pedro Rodrigues, Alberto Freitas, Filipa Almeida, Rong Chen, and Altamiro Costa-Pereira. "Data Quality and Integration Issues in Electronic Health Records". In: *Information Discovery on Electronic Health Records*. Ed. by Vagelis Hristidis. Vol. 12. Chapman and Hall/CRC, Dec. 10, 2009. ISBN: 978-1-4200-9038-3 978-1-4200-9041-3. DOI: 10.1201/9781420090413-c4. URL: http://www.crcnetbase.com/doi/abs/10.1201/9781420090413-c4.

[34] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2870052.

[35] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1 (2020), p. 18. ISSN: 1099-4300. DOI: 10.3390/e23010018.

[36] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 15662535. DOI: 10.1016/j.inffus.2019.12.012.

[37] U. Kamath and J. Liu. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Springer International Publishing, 2021. ISBN: 978-3-030-83355-8.

[38] Cynthia Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature Machine Intelligence* 1.5 (5 2019), pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

[39] Paul R. Rosenbaum. *Causal Inference*. The MIT Press, Apr. 2023. ISBN: 978-0-262-37354-8. DOI: 10.7551/mitpress/14244.001.0001. URL: https://doi.org/10.7551/mitpress/14244.001.0001.

[40] M. A. Hernán. "A Definition of Causal Effect for Epidemiological Research". In: *Journal of Epidemiology and Community Health* 58.4 (Apr. 2004), pp. 265–271. ISSN: 0143-005X. DOI: 10.1136/jech.2002.006361. pmid: 15026432.

[41] Judea Pearl. *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution*. 2018. arXiv: 1801.04016 [cs.LG]. URL: https://arxiv.org/abs/1801.04016.

[42] Jingpu Shi and Beau Norgeot. "Learning Causal Effects From Observational Data in Healthcare: A Review and Summary". In: *Frontiers in Medicine* 9 (2022). ISSN: 2296-858X.

[43] Stephen Burgess and Simon G. Thompson. "Improving Bias and Coverage in Instrumental Variable Analysis with Weak Instruments for Continuous and Binary Outcomes". In: *Statistics in Medicine* 31.15 (July 10, 2012), pp. 1582–1600. ISSN: 1097-0258. DOI: 10.1002/sim.4498. pmid: 22374818.

[44] Neil M. Davies, George Davey Smith, Frank Windmeijer, and Richard M. Martin. "Issues in the Reporting and Conduct of Instrumental Variable Studies: A Systematic Review". In: *Epidemiology (Cambridge, Mass.)* 24.3 (May 2013), pp. 363–369. ISSN: 1531-5487. DOI: 10.1097/EDE.0b013e31828abafb. pmid: 23532055.

[45] Peter C. Austin. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". In: *Multivariate Behavioral Research* 46.3 (May 2011), pp. 399–424. ISSN: 0027-3171. DOI: 10.1080/00273171.2011.568786. pmid: 21818162.

[46] Peter C. Austin. "The Performance of Different Propensity-Score Methods for Estimating Differences in Proportions (Risk Differences or Absolute Risk Reductions) in Observational Studies". In: *Statistics in Medicine* 29.20 (Sept. 10, 2010), pp. 2137–2148. ISSN: 1097-0258. DOI: 10.1002/sim.3854. pmid: 20108233.

[47] Francis Bacon. "Of the Wisdom of the Ancients". In: *The Works of Francis Bacon*. Ed. by James Spedding, Robert Leslie Ellis, and Douglas DenonEditors Heath. Vol. 6. Cambridge Library Collection - Philosophy. Cambridge University Press, 2011, pp. 701–764. DOI: 10.1017/CBO9781139149594.028.

[48] European Commission. *Proposal for a Regulation of the european Parliament and of the council on the European Health Data Space*. 2022. URL: https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF (visited on 08/21/2023).

[49] European Commission. *Ethics Guidelines For Trustworthy AI*. Tech. rep. European Commission, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (visited on 08/03/2023).

[50] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. 2019. URL: http://www.fairmlbook.org (visited on 08/21/2023).

[51] 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: 68 (2017), pp. 1–20. arXiv: arXiv:1703.06490. URL: http://arxiv.org/abs/1703.06490.

[52]   JaWanna Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. *Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015*. Tech. rep. ONC, 2016. URL: https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php.

[53]   Mrinal Kanti Baowaly, Chao-Lin Liu, and Kuan-Ta Chen. "Realistic Data Synthesis Using Enhanced Generative Adversarial Networks". In: *2019 IEEE SECOND INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND KNOWLEDGE ENGINEERING (AIKE)*. IEEE; IEEE Comp Soc, 2019, pp. 289–292. ISBN: 978-1-72811-488-0. DOI: 10.1109/AIKE.2019.00057.

[54]   Comissão Nacional Proteção de dados. *Princípios aplicáveis aos tratamentos de dados efetuados no âmbito da investigação clínica*. 2015. URL: https://www.cnpd.pt/media/grhpa2y4/del_1704_2015_investclinica.pdf (visited on 01/26/2021).

[55]   Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. U.S. Department of Health and Human Services, 20 November 2013*. 2013. URL: %5Curl%7Bhttps://www.cnpd.pt/media/grhpa2y4/del_1704_2015_investclinica.pdf%7D (visited on 01/25/2021).

[56]   Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. "A systematic review of re-identification attacks on health data". In: *PLoS ONE* 6.12 (2011). ISSN: 19326203. DOI: 10.1371/journal.pone.0028071.

[57]   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). Publisher: BMC Medical Research Methodology, pp. 1–40. ISSN: 14712288. DOI: 10.1186/s12874-020-00977-1.

[58]   Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. "Synthetic data – A privacy mirage". In: *arXiv* (2020). arXiv: 2011.07018. ISSN: 23318422.

[59]   Ian J. Goodfellow et al. "Generative Adversarial Networks". In: *Communications of the ACM* 63.11 (June 2014), pp. 139–144. ISSN: 15577317. DOI: 10.1145/3422622. URL: http://arxiv.org/abs/1406.2661.

[60]   Matt J. Kusner and José Miguel Hernández-Lobato. "GANS for Sequences of Discrete Elements with the Gumbel-softmax Distribution". In: (2016), pp. 1–6. arXiv: arXiv:1611.04051. URL: http://arxiv.org/abs/1611.04051.

[61]   Brett K. Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P. Bhavnani, James Brian Byrd, and Casey S. Greene. "Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing". In: *Circulation: Cardiovascular Quality and Outcomes* 12.7 (July 2019), pp. 139–148. ISSN: 1941-7713. DOI: 10.1161/CIRCOUTC

OMES.118.005122. URL: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.

[62] Augustus Odena, Christopher Olah, and Jonathon Shlens. "Conditional image synthesis with auxiliary classifier gans". In: *34th International Conference on Machine Learning, ICML 2017* 6 (2017), pp. 4043–4055.

[63] Belen Vega-Marquez, Cristina Rubio-Escudero, Jose C Riquelme, and Isabel Nepomuceno-Chamorro. "Creation of Synthetic Data with Conditional Generative Adversarial Networks". In: *14TH INTERNATIONAL CONFERENCE ON SOFT COMPUTING MODELS IN INDUSTRIAL AND ENVIRONMENTAL APPLICATIONS (SOCO 2019)*. Vol. 950. Startup Ole; IEEE SMC Spanish Chapter, 2020, pp. 231–240. ISBN: 978-3-030-20055-8 978-3-030-20054-1. DOI: 10.1007/978-3-030-20055-8_22.

[64] Mehdi Mirza and Simon Osindero. "Conditional Generative Adversarial Nets". In: (2014), pp. 1–7. arXiv: arXiv:1411.1784. URL: http://arxiv.org/abs/1411.1784.

[65] Edward Choi. *medGAN Repository*. URL: https://github.com/mp2893/medgan (visited on 01/21/2022).

[66] Pei-Hsuan Lu and Chia-Mu Yu. "POSTER: A Unified Framework of Differentially Private Synthetic Data Release with Generative Adversarial Network". In: *CCS'17: PROCEEDINGS OF THE 2017 ACM SIGSAC CONFERENCE ON COMPUTER AND COMMUNICATIONS SECURITY*. ACM SIGSAC; Assoc Comp Machinery; AT & T Business; Baidu; NSF; CISCO; Internet Finance Authenticat Alliance; Samsung; Univ Texas Dallas; Google; IBM Res; Paloalto Networks; Visa Res; Army Res Off; Nasher Sculpture Ctr, 2017, pp. 2547–2549. ISBN: 978-1-4503-4946-8. DOI: 10.1145/3133956.3138823.

[67] POSTER. *POSTER Repository*. URL: https://goo.gl/94qyQz (visited on 01/21/2022).

[68] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. "Data synthesis based on generative adversarial networks". In: *Proceedings of the VLDB Endowment* 11.10 (2018), pp. 1071–1083. ISSN: 21508097. DOI: 10.14778/3231751.3231757.

[69] Mahmoud Mohammadi. *table-GAN Repository*. URL: https://github.com/mahmoodm2/tableGAN (visited on 01/21/2022).

[70] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. "Differentially private generative adversarial network". In: *arXiv* (2018).

[71] Illidan Lab. *dp-GAN Repository*. URL: https://github.com/illidanlab/dpgan (visited on 01/21/2022).

[72] Ramiro Camino, Christian Hammerschmidt, and Radu State. "Generating Multi-Categorical Samples with Generative Adversarial Networks". In: *ArXiv* (2018). URL: http://arxiv.org/abs/1807.01202.

[73] Ramiro Camino. *mc-medGAN Repository*. URL: https://github.com/rcamino/multi-categorical-gans (visited on 01/21/2022).

[74] Lei Xu and Kalyan Veeramachaneni. "Synthesizing tabular data using generative adversarial networks". In: *arXiv* (Nov. 2018). arXiv: 1811.11264. ISSN: 23318422. URL: http://arxiv.org/abs/1811.11264.

[75] The Synthetic Data Vault Project. *TGAN Repository*. URL: https://github.com/sdv-dev/TGAN (visited on 01/21/2022).

[76] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. "PATE-GaN: Generating synthetic data with differential privacy guarantees". In: *7th International Conference on Learning Representations, ICLR 2019* (2019), pp. 1–21.

[77] Brett Beaulieu-Jones, Casey Greene, and Steven Wu. *SPRINT-GAN Repository*. URL: https://github.com/greenelab/SPRINT_gan (visited on 01/21/2022).

[78] Pei-Hsuan Lu, Pang-Chieh Wang, and Chia-Mu Yu. "Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network". In: *PROCEEDINGS OF THE 9TH INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, MINING AND SEMANTICS (WIMS 2019)*. 2019. ISBN: 978-1-4503-6190-3. DOI: 10.1145/3326467.3326474.

[79] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling tabular data using conditional GAN". In: *arXiv* 32.NeurIPS (2019). ISSN: 23318422.

[80] The Synthetic Data Vault Project. *CTGAN Repository*. URL: https://github.com/sdv-dev/CTGAN (visited on 01/21/2022).

[81] Bauke Brenninkmeijer. "On the Generation and Evaluation of Tabular Data using GANs". PhD thesis. Radboud University, 2019.

[82] Bauke Brenninkmeijer. *WGAN-DP Repository*. URL: https://github.com/Baukebrenninkmeijer/On-the-Generation-and-Evaluation-of-Synthetic-Tabular-Data-using-GANs (visited on 01/21/2022).

[83] Yi Liu, Jialiang Peng, James J.Q. Yu, and Yi Wu. "Ppgan: Privacy-preserving generative adversarial network". In: *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS* 2019-Decem.201910212133 (2019), pp. 985–989. ISSN: 15219097. DOI: 10.1109/ICPADS47876.2019.00150.

[84] Yi Liu. *PPGAN Repository*. URL: https://github.com/niklausliu/PPGANs-Privacy-preserving-GANs (visited on 01/21/2022).

[85] 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. ISSN: 1527974X. DOI: 10.1093/jamia/ocy142.

[86] Mrinal Kanti Baowaly. *medWGAN Repository*. URL: https://github.com/baowaly/SynthEHR (visited on 01/21/2022).

[87] Jinsung Yoon, Lydia N Drumright, and Mihaela van der Schaar. "Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN)". In: *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS* 24.8 (2020), pp. 2378–2388. ISSN: 2168-2194. DOI: 10.1109/JBHI.2020.2980262.

[88] Amirsina Torfi and Edward A. Fox. "CorGAN: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records". In: *arXiv* (2020). ISSN: 23318422.

[89] Amirsina Torfi. *corGAN Repository*. URL: https://github.com/astorfi/cor-gan (visited on 01/21/2022).

[90] Uthaipon Tantipongpipat, Chris Waites, Digvijay Boob, Amaresh Ankit Siva, and Rachel Cummings. "Differentially Private Synthetic Mixed-Type Data Generation For Unsupervised Learning". In: (2020).

[91] DPautoGAN. *DPAutoGAN Repository*. URL: https://github.com/DPautoGAN/DPautoGAN (visited on 01/21/2022).

[92] Marcel Neunhoeffer, Zhiwei Steven Wu, and Cynthia Dwork. *Private Post-GAN Boosting*. 2021. arXiv: 2007.11934 [cs.LG]. URL: https://arxiv.org/abs/2007.11934.

[93] Marcel Neunhoeffer. *Post-GAN Boosting Repository*. URL: https://github.com/mneunhoe/post-gan-boosting (visited on 01/21/2022).

[94] Amirsina Torfi, Edward A. Fox, and Chandan K. Reddy. "Differentially Private Synthetic Medical Data Generation using Convolutional GANs". In: *arXiv:2012.11774 [cs]* (Dec. 2020). URL: http://arxiv.org/abs/2012.11774.

[95] Amirsina Torfi. *DRP-CGAN Repository*. URL: https://github.com/astorfi/differentially-private-cgan (visited on 01/21/2022).

[96] Manhar Walia, Brendan Tierney, and Susan Mckeever. "Synthesising Tabular Data using Wasserstein Conditional GANs with Gradient Penalty (WCGAN-GP)". In: *Irish Conference on Artificial Intelligence and Cognitive Science*. 2020.

[97] Sina Rashidian et al. "SMOOTH-GAN: Towards Sharp and Smooth Synthetic EHR Data Generation". en. In: *Artificial Intelligence in Medicine*. Ed. by Martin Michalowski and Robert Moskovitch. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 37–48. ISBN: 978-3-030-59137-3. DOI: 10/gjspkd.

[98] Anurag Dutt. *SMOOTH-GAN Repository*. URL: https://github.com/anuragdutt/synthehr%5C_medgan (visited on 01/21/2022).

[99] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3 (2016), p. 160035.

[100] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. "An empirical study on evaluation metrics of generative adversarial networks". In: *arXiv:1806.07755 [cs, stat]* (Aug. 2018). URL: http://arxiv.org/abs/1806.07755.

[101] Stefanie James, Chris Harbron, Janice Branson, and Mimmi Sundler. "Synthetic data use: exploring use cases to optimise data utility". In: *Discover Artificial Intelligence* 1.1 (2021), p. 15. ISSN: 2731-0809. DOI: 10.1007/s44163-021-00016-y. URL: https://doi.org/10.1007/s44163-021-00016-y.

[102] healthdatainsight.org.uk. *The Simulacrum*. healthdatainsight.org.uk. URL: https://healthdatainsight.org.uk/project/the-simulacrum/ (visited on 01/21/2022).

[103] integraal kankercentrum Nederland. *Synthetische dataset NKR beschikbaar voor onderzoekers*. URL: https://iknl.nl/nieuws/2021/synthetische-data-nkr-beschikbaar-voor-onderzoeker (visited on 01/21/2022).

[104] Clinical Practice Research Datalink. *CPRD cardiovascular disease synthetic dataset*. In collab. with Clinical Practice Research Datalink. 2020. DOI: 10.11581/YK6N-B652. URL: https://cprd.com/cprd-cardiovascular-disease-synthetic-dataset (visited on 01/21/2022).

[105] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[106] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. "The Synthetic Data Vault". In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Montreal, QC, Canada: IEEE, 2016, pp. 399–410. ISBN: 978-1-5090-5206-6. DOI: 10.1109/DSAA.2016.49. URL: http://ieeexplore.ieee.org/document/7796926/ (visited on 01/20/2022).

[107] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[108] Sebastian Raschka. "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack". In: *The Journal of Open Source Software* 3.24 (Apr. 2018). DOI: 10.21105/joss.00638. URL: http://joss.theoj.org/papers/10.21105/joss.00638.

[109] Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. *UCI Machine Learning Repository - Heart Disease*. DOI: 10.24432/C52P4X. 1988. URL: https://archive.ics.uci.edu/dataset/45/heart+disease.

[110] Beata Nowok, Gillian M. Raab, and Chris Dibben. "synthpop: Bespoke Creation of Synthetic Data in R". In: *Journal of Statistical Software* 74.11 (2016), pp. 1–26. DOI: `10.18637/jss.v074.i11`.

[111] Travers Ching et al. "Opportunities and Obstacles for Deep Learning in Biology and Medicine". In: *Journal of The Royal Society Interface* 15.141 (2018), p. 20170387. DOI: `10.1098/rsif.2017.0387`.

[112] Emily Muller, Xu Zheng, and Jer Hayes. "Evaluation of the Synthetic Electronic Health Records". In: (2022). arXiv: `2210.08655 [cs]`. URL: `http://arxiv.org/abs/2210.08655` (visited on 03/05/2023).

[113] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. "Synthetic Data - A Privacy Mirage". In: *arXiv* (2020). ISSN: 23318422. arXiv: `2011.07018`.

[114] 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), pp. 1–40. ISSN: 14712288. DOI: `10.1186/s12874-020-00977-1`. pmid: `32381039`.

[115] Khaled El Emam, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. "Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study". In: *JMIR Medical Informatics* 10.4 (Apr. 2022), e35734. DOI: `10.2196/35734`.

[116] Herkulaas MvE Combrink, Vukosi Marivate, and Benjamin Rosman. "Comparing Synthetic Tabular Data Generation Between a Probabilistic Model and a Deep Learning Model for Education Use Cases". In: (2022). arXiv: `2210.08528 [cs]`. URL: `http://arxiv.org/abs/2210.08528` (visited on 03/05/2023).

[117] 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: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens. Vol. 68. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 286–305. URL: `https://proceedings.mlr.press/v68/choi17a.html`.

[118] 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. ISSN: 1527974X. DOI: `10.1093/jamia/ocy142`. pmid: `30535151`.

[119] Paul R. Rosenbaum and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In: *Biometrika* 70.1 (1983), pp. 41–55. ISSN: 0006-3444. DOI: `10.1093/biomet/70.1.41`.

[120] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, Mukund Lahoti, and Pratik Narang. "A Universal Metric for Robust Evaluation of Synthetic Tabular Data". In: *IEEE Transactions on Artificial Intelligence* 5.1 (2024), pp. 300–309. DOI: `10.1109/TAI.2022.32 29289`.

[121] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling Tabular Data Using Conditional GAN". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: `https://papers.nip s.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstr act.html` (visited on 03/05/2023).

[122] Ross Quinlan. *UCI Machine Learning Repository - Thyroid Disease*. 1987. URL: `https ://archive.ics.uci.edu/dataset/102/thyroid+disease`.

[123] Richard S. Forsyth. *UCI Machine Learning Repository - Liver Disorders*. 1990. URL: `ht tps://archive.ics.uci.edu/dataset/60/liver+disorders`.

[124] William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. *UCI Machine Learning Repository - Breast Cancer Wisconsin (Diagnostic)*. 1995. URL: `https://archive.i cs.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic`.

[125] M. Zwitter and M. Soklic. *UCI Machine Learning Repository - Primary Tumor*. 1988. URL: `https://archive.ics.uci.edu/dataset/83/primary+tumor`.

[126] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. "A Theoretical Analysis of NDCG Type Ranking Measures". In: (2013). arXiv: `1304.6480 [cs.LG]`. URL: `https://arxiv.org/abs/1304.6480` (visited on 01/21/2023).

[127] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: `10.1177/0013164460020 00104`. eprint: `https://doi.org/10.1177/001316446002000104`.

[128] M. G. Kendall. "The Treatment of Ties in Ranking Problems". In: *Biometrika* 33 (1945), pp. 239–251. ISSN: 0006-3444. DOI: `10.1093/biomet/33.3.239`. pmid: `210068 41`.

[129] Sebastiano Vigna. "A Weighted Correlation Index for Rankings with Ties". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2015, pp. 1166–1176. ISBN: 978-1-4503-3469-3. DOI: `10.1145/2736277.2741088`. URL: `https://doi.org/10.1145/2736277.2741088` (visited on 02/15/2023).

[130] William Webber, Alistair Moffat, and Justin Zobel. "A Similarity Measure for Indefinite Rankings". In: *ACM Transactions on Information Systems* 28.4 (2010), 20:1–20:38. ISSN: 1046-8188. DOI: `10.1145/1852102.1852106`.

[131] Gonzalo Navarro. "A Guided Tour to Approximate String Matching". In: *ACM Computing Surveys* 33.1 (2001), pp. 31–88. ISSN: 0360-0300, 1557-7341. DOI: `10.1145/375360 .375365`.

[132] R. W. Hamming. "Error detecting and error correcting codes". In: *The Bell System Technical Journal* 29.2 (1950), pp. 147–160. DOI: `10.1002/j.1538-7305.1950.tb00463.x`.

[133] orsinium. *Textdistance: Compute Distance between the Two Texts.* Version 4.5.0. URL: `https://github.com/orsinium/textdistance` (visited on 03/05/2023).

[134] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17.3 (3 2020), pp. 261–272. ISSN: 1548-7105. DOI: `10.1038/s41592-019-0686-2`.

[135] Changyao Chen. *Rank-Biased Overlap (RBO)*. 2018. URL: `https://github.com/changyaochen/rbo` (visited on 03/16/2023).

[136] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. "The Synthetic data vault". In: *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Oct. 2016, pp. 399–410. DOI: `10.1109/DSAA.2016.49`.

[137] F. Martin-Sanchez and K. Verspoor. "Big Data in Medicine Is Driving Big Changes". In: *Yearbook of Medical Informatics* 9 (Aug. 2014), pp. 14–20. ISSN: 2364-0502. DOI: `10.15265/IY-2014-0020`.

[138] Muhammad F. Walji. "Electronic Health Records and Data Quality". In: *Journal of Dental Education* 83.3 (Mar. 2019), pp. 263–264. ISSN: 1930-7837. DOI: `10.21815/JDE.019.034`.

[139] Robert A. Verheij, Vasa Curcin, Brendan C. Delaney, and Mark M. McGilchrist. "Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse". In: *Journal of Medical Internet Research* 20.5 (May 2018), e9134. DOI: `10.2196/jmir.9134`. URL: `https://www.jmir.org/2018/5/e185`.

[140] Kristin M. Corey et al. "Assessing Quality of Surgical Real-World Data from an Automated Electronic Health Record Pipeline". In: *Journal of the American College of Surgeons* 230.3 (Mar. 2020), 295–305.e12. ISSN: 1879-1190. DOI: `10.1016/j.jamcollsurg.2019.12.005`.

[141] Chunhua Weng. "Clinical Data Quality: A Data Life Cycle Perspective". In: *Biostatistics & Epidemiology* 4.1 (Jan. 2020), pp. 6–14. ISSN: 2470-9360. DOI: `10.1080/24709360.2019.1572344`. URL: `https://doi.org/10.1080/24709360.2019.1572344` (visited on 08/18/2022).

[142] Andrew P. Reimer, Alex Milinovich, and Elizabeth A. Madigan. "Data Quality Assessment Framework to Assess Electronic Medical Record Data for Use in Research". In: *International Journal of Medical Informatics* 90 (June 2016), pp. 40–47. ISSN: 1872-8243. DOI: `10.1016/j.ijmedinf.2016.03.006`.

[143] Erik Joukes, Nicolette F. de Keizer, Martine C. de Bruijne, Ameen Abu-Hanna, and Ronald Cornet. "Impact of Electronic versus Paper-Based Recording before EHR Implementation on Health Care Professionals' Perceptions of EHR Use, Data Quality, and Data Reuse". In: *Applied Clinical Informatics* 10.2 (Mar. 2019), pp. 199–209. ISSN: 1869-0327. DOI: 10.1055/s-0039-1681054. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6426723/.

[144] Vojtech Huser et al. "Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets". In: *EGEMS (Washington, DC)* 4.1 (2016), p. 1239. ISSN: 2327-9214. DOI: 10.13063/2327-9214.1239.

[145] Yili Zhang and Güneş Koru. "Understanding and Detecting Defects in Healthcare Administration Data: Toward Higher Data Quality to Better Support Healthcare Operations and Decisions". In: *Journal of the American Medical Informatics Association: JAMIA* 27.3 (Mar. 2020), pp. 386–395. ISSN: 1527-974X. DOI: 10.1093/jamia/ocz201.

[146] Oren Kramer, Adir Even, Idit Matot, Yohai Steinberg, and Yuval Bitan. "The Impact of Data Quality Defects on Clinical Decision-Making in the Intensive Care Unit". In: *Computer Methods and Programs in Biomedicine* 209 (Sept. 2021), p. 106359. ISSN: 1872-7565. DOI: 10.1016/j.cmpb.2021.106359.

[147] Mark J. Giganti et al. "The Impact of Data Quality and Source Data Verification on Epidemiologic Inference: A Practical Application Using HIV Observational Data". In: *BMC public health* 19.1 (Dec. 2019), p. 1748. ISSN: 1471-2458. DOI: 10.1186/s12889-019-8105-2.

[148] Jiang Bian et al. "Assessing the Practice of Data Quality Evaluation in a National Clinical Data Research Network through a Systematic Scoping Review in the Era of Real-World Data". In: *Journal of the American Medical Informatics Association: JAMIA* 27.12 (Dec. 2020), pp. 1999–2010. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa245.

[149] Nicole Gray Weiskopf and Chunhua Weng. "Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research". In: *Journal of the American Medical Informatics Association* 20.1 (Jan. 2013), pp. 144–151. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2011-000681. URL: https://doi.org/10.1136/amiajnl-2011-000681 (visited on 02/22/2022).

[150] Carlos Sáez, Juan Martínez-Miranda, Montserrat Robles, and Juan Miguel García-Gómez. "Organizing Data Quality Assessment of Shifting Biomedical Data". In: *Studies in Health Technology and Informatics* 180 (2012), pp. 721–725. ISSN: 0926-9630. PMID: 22874286.

[151] Michael G. Kahn et al. "A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data". In: *eGEMs* 4.1 (Sept. 2016), p. 1244. ISSN: 2327-9214. DOI: 10.13063/2327-9214.1244. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051581/.

[152] Hang T. T. Phan, Florina Borca, David Cable, James Batchelor, Justin H. Davies, and Sarah Ennis. "Automated Data Cleaning of Paediatric Anthropometric Data from Longitudinal Electronic Health Records: Protocol and Application to a Large Patient Cohort". In: *Scientific Reports* 10.1 (June 2020), p. 10164. ISSN: 2045-2322. DOI: `10.1038/s41598-020-66925-7`. URL: `https://www.nature.com/articles/s41598-020-66925-7`.

[153] Siaw-Teng Liaw et al. "Quality Assessment of Real-World Data Repositories across the Data Life Cycle: A Literature Review". In: *Journal of the American Medical Informatics Association: JAMIA* 28.7 (July 2021), pp. 1591–1599. ISSN: 1527-974X. DOI: `10.1093/jamia/ocaa340`.

[154] George Hripcsak et al. "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers". In: *Studies in Health Technology and Informatics* 216 (2015), pp. 574–578. ISSN: 1879-8365.

[155] Roberto Álvarez Sánchez, Andoni Beristain Iraola, Gorka Epelde Unanue, and Paul Carlin. "TAQIH, a Tool for Tabular Data Quality Assessment and Improvement in the Context of Health Data". In: *Computer Methods and Programs in Biomedicine*. SI: Data Quality Assessment 181 (Nov. 2019), p. 104824. ISSN: 0169-2607. DOI: `10.1016/j.cmpb.2018.12.029`.

[156] Carsten Oliver Schmidt et al. "Facilitating Harmonized Data Quality Assessments. A Data Quality Framework for Observational Health Research Data Collections with Software Implementations in R". In: *BMC medical research methodology* 21.1 (Apr. 2021), p. 63. ISSN: 1471-2288. DOI: `10.1186/s12874-021-01252-7`.

[157] Hanieh Razzaghi, Jane Greenberg, and L. Charles Bailey. "Developing a Systematic Approach to Assessing Data Quality in Secondary Use of Clinical Data Based on Intended Use". In: *Learning Health Systems* 6.1 (2022), e10264. ISSN: 2379-6146. DOI: `10.1002/lrh2.10264`.

[158] Naresh Sundar Rajan, Ramkiran Gouripeddi, Peter Mo, Randy K. Madsen, and Julio C. Facelli. "Towards a Content Agnostic Computable Knowledge Repository for Data Quality Assessment". In: *Computer Methods and Programs in Biomedicine* 177 (Aug. 2019), pp. 193–201. ISSN: 1872-7565. DOI: `10.1016/j.cmpb.2019.05.017`.

[159] Lorenz A. Kapsner et al. "Linking a Consortium-Wide Data Quality Assessment Tool with the MIRACUM Metadata Repository". In: *Applied Clinical Informatics* 12.4 (Aug. 2021), pp. 826–835. ISSN: 1869-0327. DOI: `10.1055/s-0041-1733847`.

[160] Hossein Estiri and Shawn N Murphy. "Semi-Supervised Encoding for Outlier Detection in Clinical Observation Data". In: *Computer methods and programs in biomedicine* 181 (Nov. 2019), p. 104830. ISSN: 0169-2607. DOI: `10.1016/j.cmpb.2019.01.002`. pmid: `30658851`.

[161] Carlos Sáez, Alba Gutiérrez-Sacristán, Isaac Kohane, Juan M García-Gómez, and Paul Avillach. "EHRtemporalVariability: Delineating Temporal Data-Set Shifts in Electronic Health Records". In: *GigaScience* 9.8 (July 30, 2020), giaa079. ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa079. pmid: 32729900.

[162] Ricardo García-de-León-Chocano, Carlos Sáez, Verónica Muñoz-Soler, Ricardo García-de-León-González, and Juan M. García-Gómez. "Construction of Quality-Assured Infant Feeding Process of Care Data Repositories: Definition and Design (Part 1)". In: *Computers in Biology and Medicine* 67 (Dec. 1, 2015), pp. 95–103. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2015.09.024. (Visited on 11/19/2023).

[163] Ricardo García-de-León-Chocano, Verónica Muñoz-Soler, Carlos Sáez, Ricardo García-de-León-González, and Juan M García-Gómez. "Construction of Quality-Assured Infant Feeding Process of Care Data Repositories: Construction of the Perinatal Repository (Part 2)". In: *Computers in Biology and Medicine* 71 (Apr. 1, 2016), pp. 214–222. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2016.01.007.

[164] S&#225 et al. "A Standardized and Data Quality Assessed Maternal-Child Care Integrated Data Repository for Research and Monitoring of Best Practices: A Pilot Project in Spain". In: *Informatics for Health: Connected Citizen-Led Wellness and Population Health*. IOS Press, 2017, pp. 539–543. DOI: 10.3233/978-1-61499-753-5-539. (Visited on 11/19/2023).

[165] David A. Springate, Rosa Parisi, Ivan Olier, David Reeves, and Evangelos Kontopantelis. "rEHR: An R Package for Manipulating and Analysing Electronic Health Record Data". In: *PloS One* 12.2 (2017), e0171784. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0171784. pmid: 28231289.

[166] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann, 1988.

[167] Ankur Ankan and Abinash Panda. "pgmpy: Probabilistic graphical models using python". In: *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*. Citeseer. 2015.

[168] David Cortes. "Explainable Outlier Detection through Decision Tree Conditioning". In: (Jan. 2, 2020). DOI: 10.48550/arXiv.2001.00636. arXiv: 2001.00636 [cs, stat]. URL: http://arxiv.org/abs/2001.00636. preprint.

[169] *GX: A Proactive, Collaborative Data Quality Platform*. URL: https://www.greatexpectations.io/ (visited on 11/19/2023).

[170] João Almeida. *Obstetrics Clinical Decision Support System IG*. URL: https://joofio.github.io/obs-cdss-fhir/ (visited on 12/06/2023).

[171] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. "Deep Learning for Health Informatics". In: *IEEE Journal of Biomedical and Health Informatics* 21.1 (2017), pp. 4–21. ISSN: 2168-2208. DOI: 10.1109/JBHI.2016.2636665.

[172] Danton S. Char, Nigam H. Shah, and David Magnus. "Implementing Machine Learning in Health Care — Addressing Ethical Challenges". In: *The New England journal of medicine* 378.11 (Mar. 2018), pp. 981–983. ISSN: 0028-4793. DOI: 10/gddr8s.

[173] Jan Philipp Albrecht. "How the GDPR Will Change the World". en. In: *European Data Protection Law Review* 2.3 (2016), pp. 287–289. ISSN: 2364284X. DOI: 10/gc8z97. URL: https://edpl.lexxion.eu/article/EDPL/2016/3/4.

[174] Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. U.S. Department of Health and Human Services, 20 November 2013.* 2013. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-%20topics/de-identification/index.html (visited on 01/25/2021).

[175] Sawsan Abdulrahman, Hanine Tout, Hakima Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani. "A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond". In: *IEEE Internet of Things Journal* (2021). DOI: 10/gk35t4.

[176] Stefanie Warnat-Herresthal et al. "Swarm Learning for decentralized and confidential clinical machine learning". en. In: *Nature* 594.7862 (June 2021), pp. 265–270. ISSN: 1476-4687.

[177] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. "Machine Learning in Medicine". In: *New England Journal of Medicine* (2019).

[178] 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* (2020), pp. 1–19. ISSN: 2509-4971. DOI: 10/gjt8ds. pmid: 33204939.

[179] Fei Wang and Anita Preininger. "AI in Health: State of the Art, Challenges, and Future Directions". In: *Yearbook of Medical Informatics* 28.1 (2019), pp. 16–26. ISSN: 2364-0502. DOI: 10/gjt8d3.

[180] Divya Jatain, Vikram Singh, and Naveen Dahiya. "A Contemplative Perspective on Federated Machine Learning: Taxonomy, Threats & Vulnerability Assessment and Challenges". In: *Journal of King Saud University - Computer and Information Sciences* (2021). ISSN: 1319-1578. DOI: 10/gkd5f4.

[181] Anup Tuladhar, Sascha Gill, Zahinoor Ismail, and Nils D. Forkert. "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. ISSN: 1532-0464. DOI: 10/gktprm.

[182]   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. ISSN: 2509-4971. DOI: `10.1007/s41666-020-0008 2-4`. pmid: `33204939`.

[183]   Geun Hyeong Lee and Soo-Yong Shin. "Federated Learning on Clinical Benchmark Data: Performance Assessment". In: *Journal of Medical Internet Research* 22.10 (2020). ISSN: 1439-4456. DOI: `10/gjt8dr`.

[184]   Prayitno et al. "A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications". In: *Applied Sciences* 11.23 (23 2021), p. 11191. ISSN: 2076-3417. DOI: `10.3390/app112311191`.

[185]   Flavio Di Martino and Franca Delmastro. "Explainable AI for Clinical and Remote Health Applications: A Survey on Tabular and Time Series Data". In: *Artificial Intelligence Review* (2022), pp. 1–55. ISSN: 0269-2821. DOI: `10.1007/s10462-022-10304-3`. pmid: `36320613`.

[186]   Seyedeh Neelufar Payrovnaziri et al. "Explainable Artificial Intelligence Models Using Real-World Electronic Health Record Data: A Systematic Scoping Review". In: *Journal of the American Medical Informatics Association : JAMIA* 27.7 (2020), pp. 1173–1185. ISSN: 1067-5027. DOI: `10.1093/jamia/ocaa053`. pmid: `32417928`.

[187]   Duncan McElfresh et al. *When Do Neural Nets Outperform Boosted Trees on Tabular Data?* Oct. 30, 2023. arXiv: `2305.02997 [cs, stat]`. URL: `http://arxiv.or g/abs/2305.02997` (visited on 02/27/2024). preprint.

[188]   Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. *Self-Normalizing Neural Networks*. Sept. 7, 2017. arXiv: `1706.02515 [cs, stat]`. URL: `http://a rxiv.org/abs/1706.02515` (visited on 02/27/2024). preprint.

[189]   Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. "Deep Neural Networks and Tabular Data: A Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 35.6 (2024), pp. 7499–7519. DOI: `10 .1109/TNNLS.2022.3229161`.

[190]   Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. *Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?* July 18, 2022. DOI: `10.48550/arXiv .2207.08815`. arXiv: `2207.08815 [cs, stat]`. URL: `http://arxiv.org/a bs/2207.08815` (visited on 02/03/2023). preprint.

[191]   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 (May 2017), pp. 24–31. ISSN: 2405-6308. DOI: `10/gc882 7`.

[192] G. Price, M. van Herk, and C. Faivre-Finn. "Data Mining in Oncology: The ukCAT Project and the Practicalities of Working with Routine Patient Data". In: *Clinical Oncology (Royal College of Radiologists (Great Britain))* 29.12 (2017), pp. 814–817. ISSN: 1433-2981. DOI: `10.1016/j.clon.2017.07.011`.

[193] Dianbo Liu, Kathe Fox, Griffin Weber, and Tim Miller. "Confederated Learning in Healthcare: Training Machine Learning Models Using Disconnected Data Separated by Individual, Data Type and Identity for Large-Scale Health System Intelligence". In: *Journal of Biomedical Informatics* 134 (Oct. 1, 2022), p. 104151. ISSN: 1532-0464. DOI: `10.1016/j.jbi.2022.104151`. (Visited on 02/27/2024).

[194] Margarita Kirienko et al. "Distributed Learning: A Reliable Privacy-Preserving Strategy to Change Multicenter Collaborations Using AI". In: *European Journal of Nuclear Medicine and Molecular Imaging* 48.12 (2021), pp. 3791–3804. ISSN: 1619-7070. DOI: `10.1007/s00259-021-05339-7`. pmid: `33847779`.

[195] Yan Wang, Chuan Hong, Nathan Palmer, Qian Di, Joel Schwartz, Isaac Kohane, and Tianxi Cai. "A Fast Divide-and-Conquer Sparse Cox Regression". In: *Biostatistics (Oxford, England)* 22.2 (Sept. 23, 2019), pp. 381–401. ISSN: 1465-4644. DOI: `10.1093/biostatistics/kxz036`. pmid: `31545341`.

[196] Kunal Chandiramani, Dhruv Garg, and N Maheswari. "Performance Analysis of Distributed and Federated Learning Models on Private Data". In: *Procedia Computer Science* 165 (2019). 11, pp. 349–355. ISSN: 18770509. DOI: `10/gm2gbb`.

[197] Geun Hyeong Lee and Soo-Yong Shin. "Federated Learning on Clinical Benchmark Data: Performance Assessment". In: *Journal of Medical Internet Research* 22.10 (2020). 9. ISSN: 1439-4456. DOI: `10/gjt8dr`.

[198] Siqi Li et al. "Federated and Distributed Learning Applications for Electronic Health Records and Structured Medical Data: A Scoping Review". In: *Journal of the American Medical Informatics Association* 30.12 (Dec. 2023). `https://doi.org/10.1093/jamia/ocad170`, pp. 2041–2049. ISSN: 1527-974X. DOI: `10.1093/jamia/ocad170`.

[199] VirtualCare. *Obscare*. [Online; accessed 26/02/2024]. 2024. URL: `https://virtualcare.pt/portfolio/vc-obscare-2-2/` (visited on 02/26/2024).

[200] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of Artificial Intelligence Research* 16.1 (2002), pp. 321–357. ISSN: 1076-9757.

[201] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?" In: (July 2022). DOI: `10.48550/arXiv.2207.08815`. arXiv: `2207.08815 [cs, stat]`.

[202]   Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and
        Gjergji Kasneci. "Deep Neural Networks and Tabular Data: A Survey". In: (June 2022).
        DOI: 10.48550/arXiv.2110.01889. arXiv: 2110.01889 [cs].

[203]   Steven Suydam, Bryan A. Liang, Storm Anderson, and Matthew B. Weinger. "Patient
        Safety Data Sharing and Protection From Legal Discovery". In: *Journal of Medical Regu-
        lation* 93.2 (2007), pp. 19–25. ISSN: 2572-1852, 2572-1801. DOI: 10.30770/2572-1
        852-93.2.19.

[204]   Tim Hulsen. "Sharing Is Caring—Data Sharing Initiatives in Healthcare". In: *International
        Journal of Environmental Research and Public Health* 17.9 (9 2020), p. 3046. ISSN: 1660-
        4601. DOI: 10.3390/ijerph17093046.

[205]   Angela G. Villanueva, Robert Cook-Deegan, Barbara A. Koenig, Patricia A. Deverka,
        Erika Versalovic, Amy L. McGuire, and Mary A. Majumder. "Characterizing the Biomed-
        ical Data-Sharing Landscape". In: *The Journal of law, medicine & ethics : a journal of the
        American Society of Law, Medicine & Ethics* 47.1 (2019), pp. 21–30. ISSN: 1073-1105.
        DOI: 10.1177/1073110519840481. pmid: 30994069.

[206]   Pedro Pereira Rodrigues, João Araújo, João Gama, and Luís Lopes. "A Local Algo-
        rithm to Approximate the Global Clustering of Streams Generated in Ubiquitous Sen-
        sor Networks". In: *International Journal of Distributed Sensor Networks* 14.10 (2018),
        p. 155014771880823. ISSN: 1550-1477, 1550-1477. DOI: 10.1177/1550147718808
        239.

[207]   Abigail Walker and Pavol Surda. "Unsupervised Learning Techniques for the Investiga-
        tion of Chronic Rhinosinusitis". In: *The Annals of Otology, Rhinology, and Laryngology*
        128.12 (2019), pp. 1170–1176. ISSN: 1943-572X. DOI: 10.1177/00034894198638
        22. pmid: 31319675.

[208]   Anna Okula Basile and Marylyn DeRiggi Ritchie. "Informatics and Machine Learning to
        Define the Phenotype". In: *Expert Review of Molecular Diagnostics* 18.3 (2018), pp. 219–
        226. ISSN: 1473-7159. DOI: 10.1080/14737159.2018.1439380. pmid: 294315
        17.

[209]   Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman, and Riccardo Bel-
        lazzi. "Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning
        Methods and Tools". In: *Frontiers in Oncology* 10 (2020). ISSN: 2234-943X.

[210]   Nimrod Rappoport and Ron Shamir. "Multi-Omic and Multi-View Clustering Algorithms:
        Review and Cancer Benchmark". In: *Nucleic Acids Research* 46.20 (2018), pp. 10546–
        10562. ISSN: 1362-4962. DOI: 10.1093/nar/gky889. pmid: 30295871.

[211]   Ewen D. McAlpine, Pamela Michelow, and Turgay Celik. "The Utility of Unsupervised
        Machine Learning in Anatomic Pathology". In: *American Journal of Clinical Pathology*
        157.1 (2022), pp. 5–14. ISSN: 1943-7722. DOI: 10.1093/ajcp/aqab085. pmid:
        34302331.

[212] S. Lloyd. "Least Squares Quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489.

[213] Douglas Steinley and Michael J Brusco. "Initializing K-means batch clustering: A critical evaluation of several techniques". In: *Journal of Classification* 24.1 (2007), pp. 99–121.

[214] J MacQueen. "Classification and analysis of multivariate observations". In: *5th Berkeley Symp. Math. Statist. Probability*. University of California Los Angeles LA USA. 1967, pp. 281–297.

[215] Lawrence Hubert and Phipps Arabie. "Comparing Partitions". In: *Journal of Classification* 2.1 (1985), pp. 193–218. ISSN: 1432-1343. DOI: 10.1007/BF01908075.

[216] Nelis J. de Vos. *kmodes categorical clustering library*. https://github.com/nicodv/kmodes. 2015. URL: %5Curl%7Bhttps://github.com/nicodv/kmodes%7D.

[217] G. Hortobagyi et al. "Updated Results from MONALEESA-2, a Phase III Trial of First-Line Ribociclib plus Letrozole versus Placebo plus Letrozole in Hormone Receptor-Positive, HER2-negative Advanced Breast Cancer". In: *Annals of oncology : official journal of the European Society for Medical Oncology* (2018). DOI: 10.1093/annonc/mdy155.

[218] Dennis J. Slamon et al. "Phase III Randomized Study of Ribociclib and Fulvestrant in Hormone Receptor-Positive, Human Epidermal Growth Factor Receptor 2-Negative Advanced Breast Cancer: MONALEESA-3". In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 36.24 (Aug. 2018), pp. 2465–2472. ISSN: 1527-7755. DOI: 10.1200/JCO.2018.78.9909.

[219] Debu Tripathy et al. "Ribociclib plus Endocrine Therapy for Premenopausal Women with Hormone-Receptor-Positive, Advanced Breast Cancer (MONALEESA-7): A Randomised Phase 3 Trial". In: *The Lancet. Oncology* 19.7 (July 2018), pp. 904–915. ISSN: 1474-5488. DOI: 10.1016/S1470-2045(18)30292-4.

[220] Sunil Verma et al. "Palbociclib in Combination With Fulvestrant in Women With Hormone Receptor-Positive/HER2-Negative Advanced Metastatic Breast Cancer: Detailed Safety Analysis From a Multicenter, Randomized, Placebo-Controlled, Phase III Study (PALOMA-3)". In: *The Oncologist* 21.10 (Oct. 2016), pp. 1165–1175. ISSN: 1549-490X. DOI: 10.1634/theoncologist.2016-0097.

[221] H. S. Rugo et al. "Impact of Palbociclib plus Letrozole on Patient-Reported Health-Related Quality of Life: Results from the PALOMA-2 Trial". In: *Annals of Oncology: Official Journal of the European Society for Medical Oncology* 29.4 (Apr. 2018), pp. 888–894. ISSN: 1569-8041. DOI: 10.1093/annonc/mdy012.

[222] Richard S Finn et al. "The Cyclin-Dependent Kinase 4/6 Inhibitor Palbociclib in Combination with Letrozole versus Letrozole Alone as First-Line Treatment of Oestrogen Receptor-Positive, HER2-negative, Advanced Breast Cancer (PALOMA-1/TRIO-18): A Randomised Phase 2 Study". In: *The Lancet Oncology* 16.1 (Jan. 2015). https://linkinghub.elsevier.com/retrieve/pii/S1470204514711593, pp. 25–35. ISSN: 14702045. DOI: 10.1016/S1470-2045(14)71159-3.

[223] Matthew P. Goetz et al. "MONARCH 3: Abemaciclib As Initial Therapy for Advanced Breast Cancer". In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 35.32 (Nov. 2017), pp. 3638–3646. ISSN: 1527-7755. DOI: 10.1200/JCO.2017.75.6155.

[224] George W. Sledge et al. "MONARCH 2: Abemaciclib in Combination With Fulvestrant in Women With HR+/HER2- Advanced Breast Cancer Who Had Progressed While Receiving Endocrine Therapy". In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 35.25 (Sept. 2017), pp. 2875–2884. ISSN: 1527-7755. DOI: 10.1200/JCO.2017.73.7585.

[225] Nadia Harbeck et al. "CDK4/6 Inhibitors in HR+/HER2- Advanced/Metastatic Breast Cancer: A Systematic Literature Review of Real-World Evidence Studies". In: *Future Oncology* 17.16 (June 2021), pp. 2107–2122. ISSN: 1479-6694. DOI: 10.2217/fon-2020-1264.

[226] Peter C. Austin. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". In: *Multivariate Behavioral Research* 46.3 (May 2011), pp. 399–424. ISSN: 0027-3171. DOI: 10.1080/00273171.2011.568786. pmid: 21818162.

[227] Peter C Austin. "The Use of Propensity Score Methods with Survival or Time-to-Event Outcomes: Reporting Measures of Effect Similar to Those Used in Randomized Experiments". In: *Statistics in Medicine* 33.7 (Mar. 30, 2014), pp. 1242–1258. ISSN: 0277-6715. DOI: 10.1002/sim.5984. pmid: 24122911.

[228] Noah Greifer. *WeightIt: Weighting for Covariate Balance in Observational Studies*. 2023. URL: https://ngreifer.github.io/WeightIt/.

[229] Massimo Cristofanilli et al. "Fulvestrant plus Palbociclib versus Fulvestrant plus Placebo for Treatment of Hormone-Receptor-Positive, HER2-negative Metastatic Breast Cancer That Progressed on Previous Endocrine Therapy (PALOMA-3): Final Analysis of the Multicentre, Double-Blind, Phase 3 Randomised Controlled Trial". In: *The Lancet. Oncology* 17.4 (Apr. 2016), pp. 425–439. ISSN: 1474-5488. DOI: 10.1016/S1470-2045(15)00613-0. pmid: 26947331.

[230] Ana Pilar Betrán, Jianfeng Ye, Anne-Beth Moller, Jun Zhang, A. Metin Gülmezoglu, and Maria Regina Torloni. "The Increasing Trend in Caesarean Section Rates: Global, Regional and National Estimates: 1990-2014". In: *PLoS ONE* 11.2 (2016), e0148343. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0148343. pmid: 26849801.

[231] Innie Chen et al. "Non-clinical Interventions for Reducing Unnecessary Caesarean Section". In: *The Cochrane Database of Systematic Reviews* 2018.9 (2018), p. CD005528. ISSN: 1469-493X. DOI: 10.1002/14651858.CD005528.pub3. pmid: 30264405.

[232] J. R. Cook, S. Jarvis, M. Knight, and M. K. Dhanjal. "Multiple Repeat Caesarean Section in the UK: Incidence and Consequences to Mother and Child. A National, Prospective, Cohort Study". In: *BJOG: an international journal of obstetrics and gynaecology* 120.1 (2013), pp. 85–91. ISSN: 1471-0528. DOI: 10.1111/1471-0528.12010. pmid: 23095012.

[233] Nicole E. Marshall, Rongwei Fu, and Jeanne-Marie Guise. "Impact of Multiple Cesarean Deliveries on Maternal Morbidity: A Systematic Review". In: *American Journal of Obstetrics and Gynecology* 205.3 (2011), 262.e1–8. ISSN: 1097-6868. DOI: 10.1016/j.ajog.2011.06.035. pmid: 22071057.

[234] Oonagh E. Keag, Jane E. Norman, and Sarah J. Stock. "Long-Term Risks and Benefits Associated with Cesarean Delivery for Mother, Baby, and Subsequent Pregnancies: Systematic Review and Meta-Analysis". In: *PLoS medicine* 15.1 (2018), e1002494. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002494. pmid: 29360829.

[235] Ilan E. Timor-Tritsch and Ana Monteagudo. "Unforeseen Consequences of the Increasing Rate of Cesarean Deliveries: Early Placenta Accreta and Cesarean Scar Pregnancy. A Review". In: *American Journal of Obstetrics and Gynecology* 207.1 (2012), pp. 14–29. ISSN: 1097-6868. DOI: 10.1016/j.ajog.2012.03.007. pmid: 22516620.

[236] World Health Organization Human Reproduction Programme, 10 April 2015. "WHO Statement on caesarean section rates". In: *Reproductive Health Matters* 23.45 (2015), pp. 149–150. ISSN: 1460-9576. DOI: 10.1016/j.rhm.2015.07.007.

[237] Ilir Hoxha and Günther Fink. "Caesarean Sections and Health Financing: A Global Analysis". In: *BMJ Open* 11.5 (May 24, 2021), e044383. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2020-044383. pmid: 34031111.

[238] Pordata. *Cesarianas Nos Hospitais (%)*. URL: https://www.pordata.pt/Portugal/Cesarianas+nos+hospitais+(percentagem)-1985 (visited on 05/16/2022).

[239] Emily Callander, Antonia Shand, David Ellwood, Haylee Fox, and Natasha Nassar. "Financing Maternity and Early Childhood Healthcare in The Australian Healthcare System: Costs to Funders in Private and Public Hospitals Over the First 1000 Days". In: *International Journal of Health Policy and Management* 10.9 (May 12, 2020), pp. 554–563. ISSN: 2322-5939. DOI: 10.34172/ijhpm.2020.68. pmid: 32610760.

[240] Michal Lipschuetz et al. "Prediction of vaginal birth after cesarean deliveries using machine learning". In: *American Journal of Obstetrics and Gynecology* 222.6 (2020), 613.e1–613.e12. ISSN: 1097-6868. DOI: 10.1016/j.ajog.2019.12.267.

[241] William A. Grobman et al. "Development of a nomogram for prediction of vaginal birth after cesarean delivery". In: *Obstetrics and Gynecology* 109.4 (2007), pp. 806–812. ISSN: 0029-7844. DOI: 10.1097/01.AOG.0000259312.36053.02.

[242] Paul Fergus, Abir Hussain, Dhiya Al-Jumeily, De-Shuang Huang, and Nizar Bouguila. "Classification of caesarean section and normal vaginal deliveries using foetal heart rate signals and advanced machine learning algorithms". In: *BioMedical Engineering OnLine* 16.1 (2017), p. 89. ISSN: 1475-925X. DOI: 10.1186/s12938-017-0378-z.

[243] Saqib Saleem, Syed Saud Naqvi, Tareq Manzoor, Ahmed Saeed, Naveed ur Rehman, and Jawad Mirza. "A Strategy for Classification of "Vaginal vs. Cesarean Section" Delivery: Bivariate Empirical Mode Decomposition of Cardiotocographic Recordings". In: *Frontiers in Physiology* 10 (2019), p. 246. ISSN: 1664-042X. DOI: 10.3389/fphys.2019.00246.

[244] Zahid Ullah, Farrukh Saleem, Mona Jamjoom, and Bahjat Fakieh. "Reliable Prediction Models Based on Enriched Data for Identifying the Mode of Childbirth by Using Machine Learning Methods: Development Study". In: *Journal of Medical Internet Research* 23.6 (2021), e28856. ISSN: 1438-8871. DOI: 10.2196/28856.

[245] Alexis C. Gimovsky, Daisy Zhuo, Jordan T. Levine, Jack Dunn, Maxime Amarm, and Alan M. Peaceman. "Benchmarking Cesarean Delivery Rates Using Machine Learning-Derived Optimal Classification Trees". In: *Health Services Research* 57.4 (Aug. 2022), pp. 796–805. ISSN: 1475-6773. DOI: 10.1111/1475-6773.13921. pmid: 34862801.

[246] Robert M. Rossi, Erin Requarth, Carri R. Warshak, Kevin R. Dufendach, Eric S. Hall, and Emily A. DeFranco. "Risk Calculator to Predict Cesarean Delivery Among Women Undergoing Induction of Labor". In: *Obstetrics & Gynecology* 135.3 (Mar. 2020), p. 559. ISSN: 0029-7844. DOI: 10.1097/AOG.0000000000003696. URL: https://journals.lww.com/greenjournal/abstract/2020/03000/risk_calculator_to_predict_cesarean_delivery_among.10.aspx (visited on 10/23/2023).

[247] Joshua Guedalia et al. "Real-Time Data Analysis Using a Machine Learning Model Significantly Improves Prediction of Successful Vaginal Deliveries". In: *American Journal of Obstetrics & Gynecology* 223.3 (Sept. 1, 2020), 437.e1–437.e15. ISSN: 0002-9378, 1097-6868. DOI: 10.1016/j.ajog.2020.05.025. pmid: 32434000. URL: https://www.ajog.org/article/S0002-9378(20)30551-2/fulltext (visited on 10/23/2023).

[248] Raanan Meyer et al. "Implementation of Machine Learning Models for the Prediction of Vaginal Birth after Cesarean Delivery". In: *The Journal of Maternal-Fetal & Neonatal Medicine: The Official Journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians* (Oct. 25, 2020), pp. 1–7. ISSN: 1476-4954. DOI: 10.1080/14767058.2 020.1837769. pmid: 33103511.

[249] Ricardo F. Sousa-Santos, Rui F. Miguelote, Ricardo J. Cruz-Correia, Cristina C. Santos, and João F. M. A. L. Bernardes. "Development of a Birthweight Standard and Comparison with Currently Used Standards. What Is a 10th Centile?" In: *European Journal of Obstetrics and Gynecology and Reproductive Biology* 206 (Nov. 1, 2016), pp. 184–193. ISSN: 0301-2115, 1872-7654. DOI: 10.1016/j.ejogrb.2016.09.028. pmid: 27723549. URL: https://www.ejog.org/article/S0301-2115(16)3094 5-9/fulltext (visited on 02/16/2023).

[250] Rima Irwinda, Rabbania Hiksas, Angga Wiratama Lokeswara, and Noroyono Wibowo. "Maternal and Fetal Characteristics to Predict C-Section Delivery: A Scoring System for Pregnant Women". In: *Women's Health* 17 (Nov. 24, 2021), p. 17455065211061969. ISSN: 1745-5057. DOI: 10.1177/17455065211061969. pmid: 34818932. URL: https ://www.ncbi.nlm.nih.gov/pmc/articles/PMC8785277/ (visited on 09/18/2023).

[251] Alberto De Ramón Fernández, Daniel Ruiz Fernández, and María Teresa Prieto Sánchez. "Prediction of the Mode of Delivery Using Artificial Intelligence Algorithms". In: *Computer Methods and Programs in Biomedicine* 219 (June 1, 2022), p. 106740. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2022.106740. URL: https://www.scien cedirect.com/science/article/pii/S0169260722001262 (visited on 09/18/2023).

[252] Rashida Parveen, Mehnaz Khakwani, Anum Naz, and Rabia Bhatti. "Analysis of Cesarean Sections Using Robson's Ten Group Classification System". In: *Pakistan Journal of Medical Sciences* 37.2 (2021), pp. 567–571. ISSN: 1682-024X. DOI: 10.12669/pjms.37 .2.3823. pmid: 33679951. URL: https://www.ncbi.nlm.nih.gov/pmc/a rticles/PMC7931279/ (visited on 02/16/2023).

[253] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017), pp. 3146–3154.

[254] ACSS. *Termos Referência Para Contratualização de Cuidados de Saúde No SNS Para 2023*. 2023. URL: https://www.acss.min-saude.pt/wp-content/uplo ads/2016/10/Termos-Referencia-Contratualizacao%5C_2023.pdf (visited on 06/20/2023).

[255]   Kartik K. Venkatesh et al. "Machine Learning and Statistical Models to Predict Postpartum Hemorrhage". In: *Obstetrics and Gynecology* 135.4 (Apr. 2020), pp. 935–944. ISSN: 1873-233X. DOI: `10.1097/AOG.0000000000003759`. pmid: `32168227`.

[256]   Ghamar Bitar, Wei Liu, Jade Tunguhan, Kaveeta V. Kumar, and Matthew K. Hoffman. "A Machine Learning Algorithm Using Clinical and Demographic Data for All-Cause Preterm Birth Prediction". In: *American Journal of Perinatology* (Dec. 4, 2023). ISSN: 1098-8785. DOI: `10.1055/s-0043-1776917`. pmid: `38049100`.