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

DISSERTAÇÃO PROVISÓRIA

Programa Doutoral em Ciência de dados de saúde

Supervisor: Pedro Pereira Rodrigues

Second Supervisor: Ricardo Correia

January 4, 2024

Cover page

Integrity Declaration

Reproducibility

The code for all the experiments done in this thesis is stored online on GitHub. The link is as follows: https://github.com/joofio/heads-thesis From here, it should be possible to access the list of all the repositories involved in this thesis. Data is also available when possible. However, since most of the data used was directly retrieved from Electronic Health Record, it is blocked by ethical committees from sharing with third parties.

Professores Catedráticos

Patrício Manuel Vieira Araújo Soares Silva
Alberto Manuel Barros Da Silva
José Henrique Dias Pinto De Barros
Maria Fátima Machado Henriques Carneiro
Maria Dulce Cordeiro Madeira
Altamiro Manuel Rodrigues Costa Pereira
Manuel Jesus Falcão Pestana Vasconcelos
João Francisco Montenegro Andrade Lima Bernardes
Maria Leonor Martins Soares David
Rui Manuel Lopes Nunes
José Manuel Pereira Dias De Castro Lopes
Joaquim Adelino Correia Ferreira Leite Moreira
Raquel Ângela Silva Soares Lino
Fernando Manuel Mendes Falcão Dos Reis
Francisco José Miranda Rodrigues Cruz
José Paulo Alves Vieira De Andrade
Jorge Manuel Silva Junqueira Polónia
José Luís Dias Delgado
Isaura Ferreira Tavares
Fernando Carlos De Landér Schmitt
Acácio Agostinho Gonçalves Rodrigues
Maria De Fátima Moreira Martel
João Tiago De Sousa Pinto Guimarães
José Carlos Lemos Machado
José Carlos De Magalhães Silva Cardoso

Professores Catedráticos Jubilados e Aposentados

Alexandre Alberto Guerra Sousa Pinto
Álvaro Jerónimo Leal Machado De Aguiar
António Albino Coelho Marques Abrantes Teixeira
António Carlos De Freitas Ribeiro Saraiva
António José Pacheco Palha
António Manuel Sampaio De Araújo Teixeira
Belmiro Dos Santos Patrício
Cândido Alves Hipólito Reis
Carlos Rodrigo Magalhães Ramalhão
Cassiano Pena De Abreu E Lima
Deolinda Maria Valente Alves Lima Teixeira
Eduardo Jorge Cunha Rodrigues Pereira
Fernando Tavarela Veloso
Francisco Fernando Rocha Gonçalves
Isabel Maria Amorim Pereira Ramos
Jorge Manuel Mergulhão Castro Tavares
José Agostinho Marques Lopes
José Carlos Neves Da Cunha Areias
José Eduardo Torres Eckenroth Guimarães
José Fernando Barros Castro Correia
José Manuel Costa Mesquita Guimarães
José Manuel Lopes Teixeira Amarante
Levi Eugénio Ribeiro Guerra
Luís Alberto Martins Gomes De Almeida
Manuel Alberto Coimbra Sobrinho Simões
Manuel António Caldeira Pais Clemente
Manuel Augusto Cardoso De Oliveira
Manuel Machado Rodrigues Gomes
Manuel Maria Paula Barbosa
Maria Amélia Duarte Ferreira
Maria Da Conceição Fernandes Marques Magalhães
Maria Isabel Amorim De Azevedo
Rui Manuel Almeida Mota Cardoso
Rui Manuel Bento De Almeida Coelho
Serafim Correia Pinto Guimarães
Valdemar Miguel Botelho Dos Santos Cardoso
Walter Friedrich Alfred Osswald

To my and my

List of Publications

Core Research Papers
The 7 papers described below are the core structure of this thesis (4 were already published, and 3 are under review). The manuscripts are listed by order of appearance in the thesis.

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

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.

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).

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 our investigation is the acknowledgment, inspired by Richard P. Feynman, that absolute certainty is elusive in scientific inquiry; instead, our journey is marked by continual learning and improvement. We confront various obstacles, including data accessibility, quality concerns, and the integration of real-world evidence (RWE) into clinical practice, while also exploring innovative solutions like synthetic data and distributed data analysis paradigms.

A significant portion of our work is dedicated to addressing the dual challenges of data access and quality. The stringent requirements of ethics committees and Data Protection Officers (DPOs), designed to safeguard patient privacy, often impede timely data access. We propose synthetic data as a potential workaround, offering a secure and legal avenue for algorithm development and testing. Additionally, we underscore the importance of distributed data analysis, allowing for secure, location-based data analysis, thereby enhancing the timeliness and security of the process.

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 (HIS). The thesis highlights the criticality of clear functional and clinical data descriptions, advocating for comprehensive data dictionaries and governance tools to facilitate effective data utilization.

Moreover, we emphasize the necessity of a collaborative approach with clinicians, who are the end-users of the developed tools. Understanding their needs and workflows is paramount, necessitating user-friendly tools that clinicians can seamlessly integrate into their practice without the need for extensive data science training.

Another pivotal aspect of this thesis is the exploration of a legal and technical framework for healthcare data science, akin to the rigorous approval processes for pharmaceuticals. Such a framework should balance safety and innovation, ensuring that new tools and methodologies are both effective and ethical. This approach is complemented by a strong focus on biomedical informatics and the development of robust data infrastructures, both locally and internationally, to enhance data availability and quality.

The thesis also explores the potential of RWE to support clinical decisions in real-time, emphasizing the need for a trust framework that ensures transparency and explainability in evidence production. This is crucial for building clinician and patient trust in the data and decision-making processes.

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.

Keywords: real-world data, health data science, distributed-learning, machine-learning, data quality

Resumo

Esta tese mergulha no processo intrincado de extrair conhecimento dos dados de saúde, uma tarefa repleta de desafios, mas também de potencial. Central para nossa investigação é o reconhecimento, inspirado por Richard P. Feynman, de que a certeza absoluta é ilusória na pesquisa científica; em vez disso, nossa jornada é marcada por aprendizado contínuo e melhoria. Confrontamos vários obstáculos, incluindo acessibilidade dos dados, preocupações com a qualidade e a integração de evidências do mundo real (RWE) na prática clínica, explorando também soluções inovadoras como dados sintéticos e paradigmas de análise de dados distribuídos.

Uma parte significativa do nosso trabalho é dedicada a enfrentar os desafios duplos de acesso e qualidade dos dados. Os requisitos rigorosos dos comités de ética e dos Responsáveis pela Proteção de Dados (DPOs), projetados para salvaguardar a privacidade do paciente, muitas vezes impedem o acesso oportuno aos dados. Propomos dados sintéticos como uma solução potencial, oferecendo um caminho seguro e legal para o desenvolvimento e teste de algoritmos. Além disso, sublinhamos a importância da análise de dados distribuída, permitindo uma análise de dados segura e baseada na localização, aumentando assim a pontualidade e a segurança do processo.

A qualidade dos dados de saúde surge como um conceito complexo e elusivo, exigindo extenso pré-processamento de dados para gerir valores ausentes, outliers e inconsistências entre diferentes sistemas de informação de saúde (HIS). A tese destaca a criticidade de descrições claras de dados funcionais e clínicos, defendendo ferramentas abrangentes de dicionários de dados e governança para facilitar a utilização eficaz dos dados.

Além disso, enfatizamos a necessidade de uma abordagem colaborativa com os clínicos, que são os utilizadores finais das ferramentas desenvolvidas. Entender as suas necessidades e fluxos de trabalho é fundamental, necessitando de ferramentas de fácil utilização que os clínicos possam integrar sem problemas na sua prática, sem a necessidade de extensa formação em ciência de dados.

Outro aspecto fundamental desta tese é a exploração de um quadro legal e técnico para a ciência de dados de saúde, semelhante aos rigorosos processos de aprovação para produtos farmacêuticos. Tal quadro deve equilibrar segurança e inovação, garantindo que novas ferramentas e metodologias sejam eficazes e éticas. Esta abordagem é complementada por um forte enfoque na informática biomédica e no desenvolvimento de infraestruturas de dados robustas, tanto a nível local como internacional, para melhorar a disponibilidade e qualidade dos dados.

A tese também explora o potencial da RWE para apoiar decisões clínicas em tempo real, enfatizando a necessidade de um quadro de confiança que garanta transparência e explicabilidade na produção de evidências. Isso é crucial para construir confiança dos clínicos e pacientes na precisão, confiabilidade e transparência dos dados e dos algoritmos utilizados. A construção dessa confiança implica garantir que os processos de tratamento de dados e de tomada de decisão sejam transparentes e explicáveis, fomentando um sentido de responsabilidade e confiabilidade no sistema.

Em conclusão, esta tese contribui para o campo da ciência de dados de saúde, destacando os desafios multifacetados e propondo abordagens inovadoras para a extração eficaz de conhecimento dos dados de saúde. Sublinha a importância da colaboração interdisciplinar, infraestruturas de

dados robustas e um quadro legal e técnico equilibrado para aproveitar todo o potencial dos dados de saúde, impulsionando a inovação e melhorando os resultados dos pacientes.

Keywords: keyword1, Keyword2, keyword3

Acknowledgements

Queria aqui agradecer, fachabor, fachabor.

Author's Name

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

Jermaine Lamarr Cole

Outline

The idea for this thesis first formed in my mind during a mental process of understanding how clinical knowledge could be improved in terms of quality, quantity, and speed of generation. The feeling was that new technology, especially the ones related to digital and informatics domain took years to be fully implemented in practice and harness the potential benefits they provided. I felt that healthcare, like other domains, had a serious gap between academia and industry. So, the potential of all these discoveries was lost in "translation". So how could we leverage this?
This thesis is organized 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, like Knowledge Discovery in Databases (KDD), Evidence Based Medicine (EBM) or privacy and ethical concerns.
Chapter 3 corresponds to the papers published. The papers cover a wide range of the traditional KDD steps, so they are grouped around the phases they represent the most.
Chapter 4 represents the overall discussion of all of the papers and experiments done in the thesis.
Chapter 5 communicates the conclusion, limitations and future work.
Attachments include ethical permissions and supplementary data to some of the papers.

Contents

List of Figures

List of Tables

Abbreviations

AI Artificial Intelligence

API Application Programming Interface

ATE Average Treatment Effect

ATT Average Treatment Effect on the Treated

AUPRC Area Under the Precision Recall Curve

AUROC Area Under the Receiver Operating Characteristic Curve

BH *Benjamini-Hochberg*

BMI Body Mass Index

BN Bayesian Network

C-Section Cesarean Section

CausalML Causal Machine Learning

CDK4/6 Cyclin-dependent kinases 4 and 6

CDK4/6i Cyclin-dependent kinases 4 and 6 inhibitors

CDSS Clinical Decision Support System

CRISP-DM Cross-Industry Standard Process for Data Mining

DAG Directed Acyclic Graph

DPO Data Protection Officer

DQ Data Quality

EBM Evidence Based Medicine

ECOG Eastern Cooperative Oncology Group scale

EDA Exploratory Data Analysis

EHDS European Health Data Space

EHR Electronic Health Record

ET Endocrine Therapy

EU European Union

FHIR Fast Healthcare Interoperability Resources

GAN Generative Adversarial Network

GDPR General Data Protection Regulation

GLM General Linear Model

HER2- Human Epidermal growth factor Receptor 2 negative

HIPAA Health Insurance Portability and Accountability Act

HIS Health Information System

HL7 Health Level Seven

HR Hazard Ratio

HR+ Hormone Receptor positive

IKNL Integraal Kankercentrum Nederland

IPTW Inverse Probability of Treatment Weighting

IQR Inter-Quartile Range

IV Instrumental variable

JSD *Jensen-Shannon Divergence*

KDD Knowledge Discovery in Databases

KNN K-Nearest Neighbours

KS *Kolmogorov-Smirnov*

LIME Local Interpretable Model-Agnostic Explanation

MAE Mean Absolute Error

MHRA Healthcare Products Regulatory Agency

ML Machine Learning

MRE Mean Relative Error

OS Overall Survival

PCA Principal Component Analysis

PFS Progression Free Survival

POF Potential Outcome Framework

RCT Randomized Clinical Trial

RI Rand Index

RMSE Root Mean Squared Error

RWD Real World Data

RWE Real World Evidence

SCM Structural Causal Model

SEM structural equation model

SEMMA Sample, Explore, Modify, Model, and Assess

SHAP SHapley Additive exPlanations

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machines

UCI UC Irvine Machine Learning Repository

USA United States of America

VAE Variational Autoencoder

WHO World Health Organisation

XAI Explainable AI

Glossary

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

Albert Einstein

1 Introduction

1.1 Rationale

Healthcare practice revolves a lot around technology. Technology in the definitional sense of referring to *"methods, systems, and devices which are the result of scientific knowledge being used for practical purposes"*. Healthcare and medicine are in fact applied sciences where we use our knowledge of biology, physics, chemistry, and math and apply those concepts in order to create treatments, diagnoses, procedures, etc. However, in the last 20-30 years, computer science and informatics started gaining traction in the healthcare space [1]. A paper-based industry is now being digitalized and computerized. This has been leading to an increase in the amount of data generated by healthcare systems [2, 3]. This data has the potential to greatly improve the current methods and practices in healthcare. However, it is still not being used to its full potential [2, 4]. This is especially important when we note that the gold standard of evidence creation is Randomized Clinical Trials (RCTs) which can vary on 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 [5] can cost as much as 100 million USA dollars. This is indeed a very steep price to get 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 around 140 thousand dollars each [6]. Additionally, we have to take into account the time that 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 Electronic Health

Records (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 like a good approach for harvesting this potential. However, they are very different from the traditional statistical methods that are usually used in healthcare. So, in order to properly use these methods in healthcare and actually provide value to the 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 [7, 8], and how to proper evaluate the potential risks of AI in healthcare, just to mention a few [9]. This is a very complex problem that requires a lot of different approaches and solutions. It is a popular assumption that 87% of data science projects never get into production [10]. Even if numbers for the healthcare domain at not available at this time, it is safe to assume that the number is not much different, if not higher. And those that actually do, may never actually create any impact due to the lack of adoption by the healthcare practitioners or the lack of trust in the system [11].

So, with this introduction, we have there is still a long way to go to harvest all the potential healthcare data has to offer. And so my research objectives are focused on powering up this adoption. What can be done to improve these chances? What can we bring to the table to enhance the rate of success?

1.2 Research Objectives

This thesis has three main goals:

- Goal 1: Research methods for improving Data Quality. Whether through synthetic data generation to enlarge data volume and protect privacy (sections 3.1, 3.2 and 3.3) or by creating automatic data quality assessments (section 3.4)
- Goal 2: Assess alternative ways of usage of data without having access to all of it. This will be covered in sections 3.5 and 3.6.
- Goal 3: Difficulties and steps resulting from attempts to convert data into decisions and policies, whether through ML or traditional statistics. This will be covered in sections 3.7 and 3.8.

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

Isaac Asimov

2 State of the art

2.1 Artificial Intelligence

AI has already been under public focus for a few years now, but its concept is still elusive, mainly due to the fact that the definition has been changing rapidly as well. From the very beginning, the field of AI was about not only understanding but also building intelligent entities [12]. 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 could be credited to Warren McCulloch and Walter Pitts (1943) with the proposed model of artificial neurons. In the 50s, AI could be associated with the works of Christopher Strachey, of two chess-playing programs. In the 60s, the perceptrons could be indicated as state-of-the-art AI. In the 80s, expert systems were providing advanced reasoning that the so-called weak methods of previous iterations could not compete with. The 90s brought the probabilistic reasoning and ML which led to more robust systems that went further than the boolean logic used so far. In the 2000s, big data and ML got focused on. Big data was used as a matter symbolizing the increasing amounts of data in some industries [13], and ML as *the study of computer algorithms that improve automatically through experience* [14]. This last definition is especially important since it is currently used as a synonym of AI across several industries but have actual different meanings like discussed below. This era probably peaked around the IBM Watson victory in jeopardy, but with way fewer interesting results in healthcare [15], and 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, reports on AI show that papers regarding the subject have seen a 20-fold increase from 2010-2019. For defining AI, we could use the definition provided by a group of experts the European Commission asked to write

some guidelines on AI [16] From this document, it is understood, that first, we need to address the difference between intelligence and rationality. Since the first is more subjective and even philosophical, the second is more pragmatic and related to the capability of choosing the best action to take in a certain scenario towards a certain goal. This is a more concrete concept and although it is not the same as intelligence, it should be a part of it [16, 12]. 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 for 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 the one that 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 [16, 12]. So, with this is mind, we reached the definition of AI as:

> *Artificial intelligence (AI) 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).* [16]

Of course, since the developments of this area have been so vast, this concept may become outdated very quickly. However, it is a good starting point to understand the concept of AI and its implications.

2.2 Evidence Based Medicine

In 1996, David Sacket and colleagues defined EBM as *the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients* [17]. Despite having historical antecedents dating back to at least the 19th century, the first time the term "evidence-based medicine" was first coined by a team at McMaster University in Canada in the 1980s [18]. 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

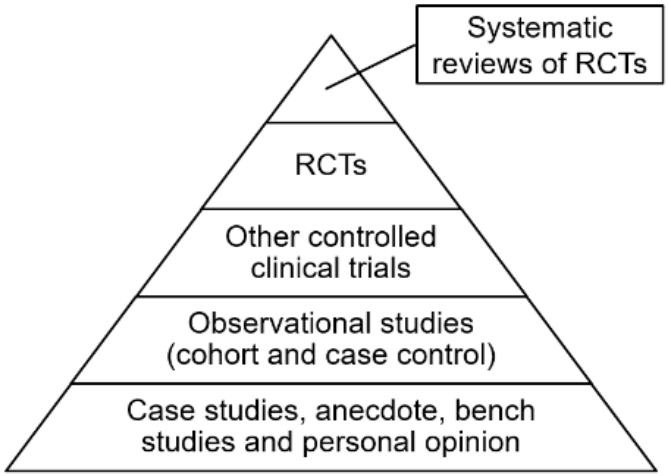


Figure 2.1: EBM adapted from [19]

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. With this, we can, as stated by Sacket, define EBM into 3 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 stagnated and unable to evolve [17].

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 opinions, Sackett frequently recounted the circumstances surrounding George Washington's un-

fortunate 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. However, EBM is not without critiques. The first one is that this is what medicine is all about and is already practiced all over. The data suggest something different [17]. The second refers to the virtually impossible task of keeping up with the literature. This argument, despite being refuted by examples of clinicians doing it, does raise the question of how we can deal with this, taking into account the increasing evidence overflow that the current times bring. How can we keep up with the literature and how can we make sure that the evidence is being applied in clinical practice? This is a very important question since the evidence is only useful if it is applied. This is where KDD and AI can play a role, as we will see in the next sections.

2.3 Extracting Knowledge of 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 [20]. This was the first time that data was 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. Additionally, the fact that data is being collected at an unprecedented rate, and the need to extract knowledge from it, 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 [21]. 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 [13] the digital data in the healthcare space has 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

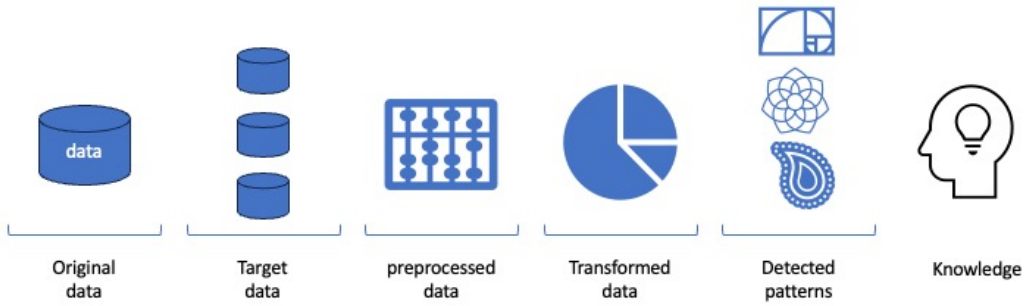


Figure 2.2: KDD Process, adapted from [21]

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 [22]. Sample, Explore, Modify, Model, and Assess (SEMMA) [23] 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. Like stated in [21], 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 [21].

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 subject and research areas, sits the most important of all - which is domain expertise.

2.4 Health Data Science

Health Data Science 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, Health Data Science involves the use of statistical and machine learning 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 Health Data Science is predictive modelling and hypothesis testing. Predictive modelling involves the creation and use of statistical models or machine learning 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 health data science as they allow us to make educated guesses about health trends and outcomes.

Importantly, Health Data Science 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 Health Data Science 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 health data science [24, 25]:

- 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* [26]. 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 health data science 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 [27, 28, 25]. 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 [29].

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 machine learning 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 health data science, since it can be the difference between a good and a bad model.

Privacy and ethics adds yet another layer to the problems of health data science. 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 health data science 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 health data science 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 health data science 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 Explainable AI (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 [30, 31, 32, 31, 33], a simplified approach based on [33] 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 machine learning 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 machine learning 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 machine learning 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 [34] 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 the healthy human person was a balance between 4 humours (blood, phlegm, yellow bile, and black bile). So, the treatment for George Washington was to rebalance those 4 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 4 senses of humour and illness 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. So that is why, nowadays, EBM, XAI can be brought together and expanded through Causal Machine Learning (CausalML). But what is causality? We could argue that 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 base of why RCTs are the gold-standard of experimentation, since they are the current best tool to achieve something similar to this [35].

CausalML is a branch of machine learning that focuses on understanding and quantifying causal relationships from data [36]. 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 [37]. There are currently two main frameworks for trying to unveil causality in data: the Structural Causal Model (SCM) and the Potential Outcome Framework (POF) [38]. **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 centers 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 causal machine learning, 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, machine learning 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 [39, 40]. 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 [41, 42]

   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 Health Data Science, KDD and AI in healthcare get 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."* [43]. 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 stringent 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 [44]:

- 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 don't 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 Clinical Decision Support Systems (CDSSs) that orient and support clinical practice, they can be biased by the type of data that originated said knowledge [45, 46].

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 machine learning 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.

Nothing great in the world was accomplished without passion.

Friedrich Hegel

3 Case Studies

This chapter will comprise the work done during this PhD. The works developed and corresponding papers were a search for improving data usage in several steps of the KDD process. We can see some work dedicated to leveraging data acquisition or alternatives to it, like the works depicted in section 3.5, 3.6, 3.3, 3.2 and 3.1. Others will focus more on how to use the data in order to make a difference in clinical practice like the section 3.7, 3.8 and 3.4.

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 Generative Adversarial Network (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 [47, 48]. 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 [49]. 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 [50], and internationally [51]. 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 [52]. 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 [53], 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 [53], but not without questions and doubts [54]. 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 [55] 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 $G_\theta(\mathrm{Z})$. Distribution of $G_\theta(\mathrm{Z})$ is denoted by P_θ. The goal of the generator is to choose θ such that the output $G_\theta(\mathrm{Z})$ has a distribution close to the real data. The discriminator is represented by D_ω, parametrized by weights 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

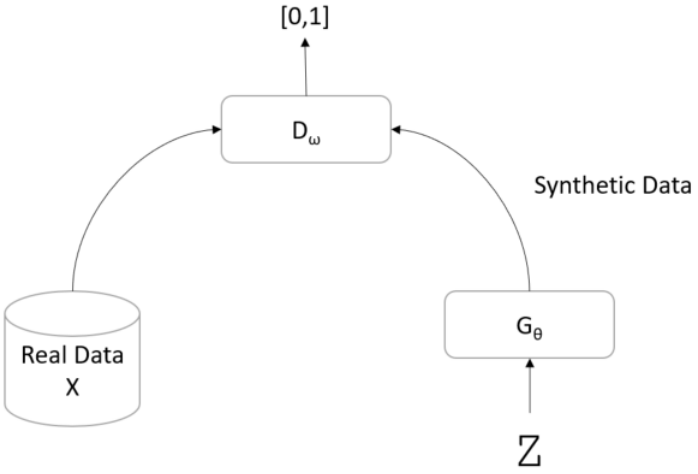


Figure 3.1: GAN framework

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) [55]:

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

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 [56], 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. [57] used a GAN architecture that was originally published with usage on image datasets [58]. Additionally, Vega-Marquez et al. [59] used an already known implementation of conditional GANs [60]. 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.

Table 3.1: Summary of the articles selected.

ID	year	Acronym	Article	Metric	Code
1	2017	medGAN	[47]	Utility, Privacy, Clinical	[61]
2	2017	POSTER	[62]	Utility, Privacy	[63]
3	2018	table-GAN	[64]	Utility, Privacy	[65]
4	2018	dp-GAN	[66]	Utility, Privacy	[67]
5	2018	mc-medGAN	[68]	Utility	[69]
6	2018	TGAN	[70]	Utility	[71]
7	2019	PATE-GAN	[72]	Utility, Privacy	–
8	2019	SPRINT-GAN	[57]	Utility, Privacy, Clinical	[73]
9	2019	GAN-based	[74]	Utility, Privacy	–
10	2019	CTGAN	[75]	Utility	[76]
11	2019	WGAN-DP	[77]	Utility, Privacy	[78]
12	2019	PPGAN	[79]	Utility, Privacy	[80]
13	2019	medBGAN	[49]	Utility	–
14	2019	medWGAN	[81]	Utility	[82]
15	2020	ADS-GAN	[83]	Utility, Privacy	–
16	2020	corGAN	[84]	Utility, Privacy	[85]
17	2020	CGAN	[59]	Utility	–
18	2020	DPAutoGAN	[86]	Utility, Privacy	[87]
19	2020	GAN Boosting	[88]	Utility, Privacy	[89]
20	2020	RDP-CGAN	[90]	Utility, Privacy	[91]
21	2020	WCGAN-GP	[92]	Utility, Privacy	–
22	2020	SMOOTH-GAN	[93]	Utility	[94]

Table 3.2: Metrics utilised for evaluation

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|R2 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	–

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 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 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 AUC, 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*, Wasserstein Distance (WD), Bhattacharyya Distance (BD) or Generate Scores (GS) that was a metric implemented by the authors of [79] 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). NMI (Normalised Mutual Information), 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 [95] (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.

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 [96], 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 cataloging, 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 Pulling the current metrics of assessing 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 [97]; 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 [97], 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 [98] is a nice example of how the 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 [99] and the Healthcare Products Regulatory Agency (MHRA) that uses synthetic data as well for its CPRD real-word evidence [100].

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 [54]. 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

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 [101], sdmetrics [102] and *scikit-learn* [103] and *mlxtend* [104]. 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 is 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 pair-wise 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.

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

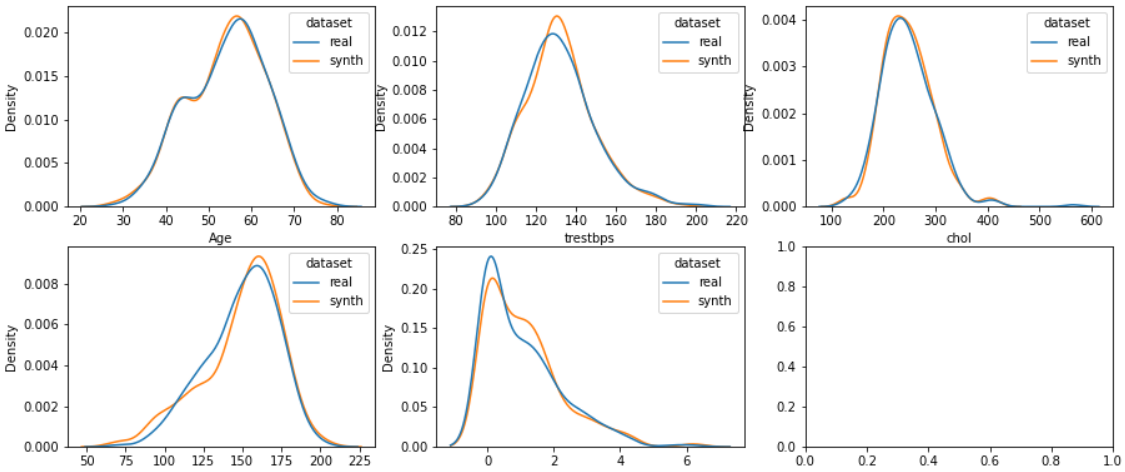


Figure 3.3: Continuous Variables plotted

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 [105]. The synthetic data was created by using the synthpop package [106]. 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).

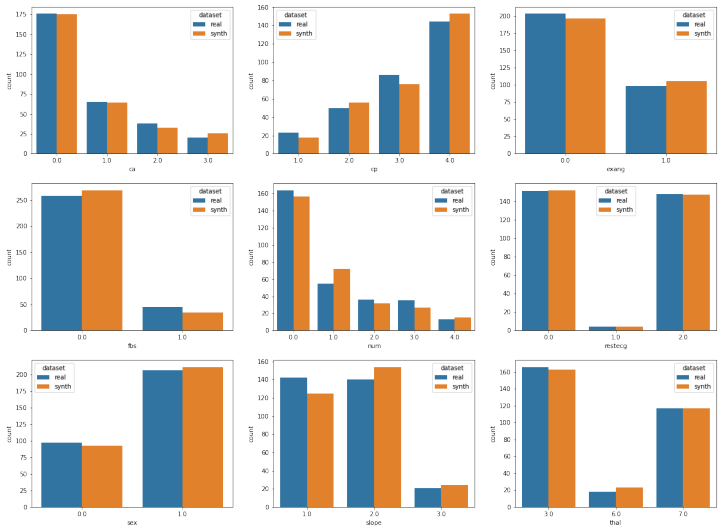


Figure 3.2: Categorical Variables plotted

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 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 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 [107]. 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 [108].

Synthetic data has the potential to overcome many of the limitations associated with real-world data, such as the lack of sufficient data volume, noise, and privacy concerns. Even though there are still doubts if the privacy part is the silver bullet sometimes referred to [109], the upsampling part is a standard use for years now. However, the quality of synthetic data generated by various techniques can vary significantly, and it is essential to assess the quality of synthetic data before

its usage. In healthcare, the assessment of synthetic data is crucial to ensure that it can provide valid insights and inform decision-making processes.

The assessment of synthetic data in healthcare is essential for its successful use in various applications, such as developing predictive models, testing algorithms, and conducting clinical trials. The use of synthetic data can significantly enhance the efficiency and effectiveness of healthcare research and practice. However, it is crucial to ensure that the synthetic data used in these applications are of high quality and validated to provide reliable and valid insights. The evaluation of synthetic data quality involves comparing its statistical properties and patterns with those of the original data. We can assess how similar columns are to each other through several statistical tests, and then we can infer some inter-column properties with methods like cross-validation, where two datasets are split into train tests and cross-tested and then the ratio between the evaluation result of both datasets is used as a metric [108, 110]. However, this methodology is a big proxy for such an inter-column relationship. Can we try to provide a better metric than this one to evaluate how similar are the inter-column relationship of two distinct datasets? In this paper, we suggest using feature importance values to create a more explainable and reasonable metric for inter-column relationships.

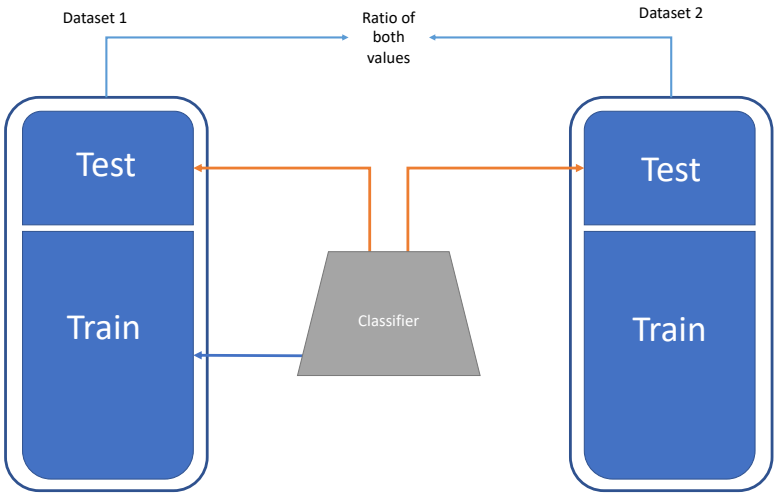
3.3.2 Rationale and Related Work

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. [111] 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 intracolumn 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 [112] among many others, acting more like sanity checks than anything else. Other known metrics are distance-based metrics like JSD, 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 [83, 113, 114].

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 [115, 108] 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., [116], known as TabSynDex. Other examples are likelihood of fitness like in the works of [117], coverage support [110] or very specific metrics implemented for evaluating specific data generators. However, the most used metric is cross-validation, 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, even if 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 image 3.4 represents this in detail. Several works used this metric as the comparing metric [108].

Figure 3.4: Cross-Validation of datasets



3.3.3 Materials & Methods

3.3.3.1 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) [118] 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

$$\text{NDGC} = \frac{\text{DCG}(P)}{\text{IDCG}(P)} \tag{3.3}$$

where $\text{DCG}(P)$ is the Discounted Cumulative Gain and $\text{IDCG}(P)$ is the Ideal Discounted Cumulative Gain.

Cohen's kappa coefficient [119] 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.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3.4}$$

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.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.5}$$

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 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 [120] 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.

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\text{total number of pairs}} \tag{3.6}$$

Weighted Kendall tau [121] 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.

$$\tau_w = \frac{\sum_{i<j} w_{ij} \cdot sgn(x_i - x_j)}{\sum_{i<j} w_{ij}} \tag{3.7}$$

where w_{ij} is the weight associated with the pair (x_i, x_j) and $sgn(\cdot)$ is the sign function. Rank-biased overlap (RBO) [122] is a measure of similarity between two ranked lists or rankings. It takes into account the order of items in the two lists, and it can be used to evaluate the quality of search results, recommendations, or any other kind of ranked list it has been shown to be more robust and accurate than other similarities measures such as Kendall tau or Spearman's rank correlation coefficient.

$$\text{RBO} = (1-\rho) \cdot \sum_{d=1}^{\infty} \left(\frac{g_d}{d}\right) \cdot \rho^d \tag{3.8}$$

where ρ is the weight, g_d is the gain at depth d, and $\sum_{d=1}^{\infty}$ indicates the summation over all depths.

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 be-

tween the two. Levenshtein distance [123] is the minimum number of single-character insertions, deletions, or substitutions required to transform one string into another. Damerau-Levenshtein distance [123] is similar to Levenshtein distance but also includes the transposition of two adjacent characters as an allowable operation. The hamming distance [124] 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 [123] 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 i in number of columns to test do
    for rep in 10 repetitions do
        permutate values in i columns
        for dataset in dataset pair do
            for target in dataset columns do
                Train-Test Split (95:5) model fit to train get feature importance per column
                Create an ordered rank of features
            end
        end
    end
end
```

Algorithm 1: Testing similarity scores in tabular datasets

The algorithms chosen were decision trees, random forests, SVM, KNN, and linear regression/logistic regression as implemented in the *scikit-learn* package [103]. The text distance metrics were implemented by the text-distance package [125]. Kendall tau, weighted Kendall tau were used as implemented by scipy [126] and RBO, as implemented in [127].

3.3.3.2 Data used

We used 5 datasets from the UCI repository. The ones chosen were related to healthcare and were heart disease [105], thyroid disease [128], liver disorders [129], breast cancer [130] and the primary tumour dataset [131]. 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 a synthetic dataset by applying the synthpop package to this data [106]. With this package, all variables were synthesised using the "cart" method, which is rpart implementation of a CART model.

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.

The number of repetitions and how that impacts the variance of the scores is shown in figure 3.7.

As for the test for the synthetic and real datasets, the results are displayed in figure 3.8.

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.

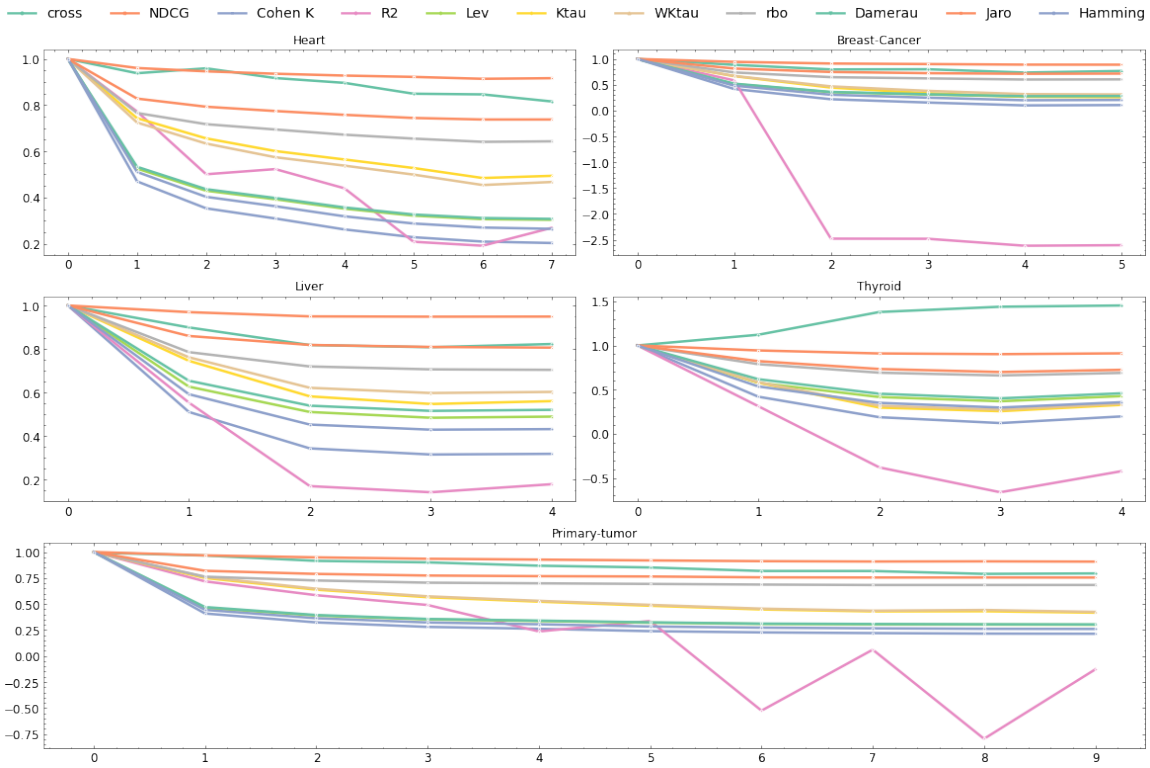


Figure 3.6: Plot showing the values at 50% columns mutated across all datasets and algorithms per metric type

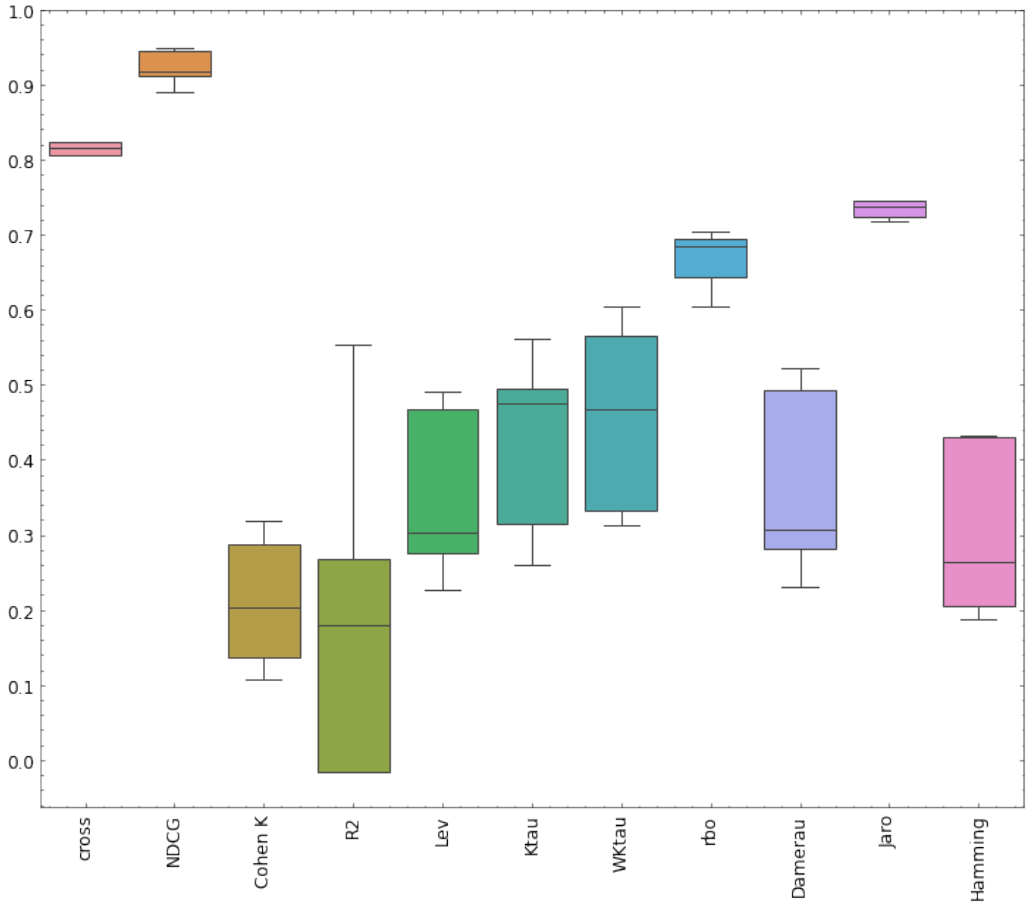


Figure 3.7: Plot the variance of different repetitions for every metric and the number of different columns changed. X is the number of columns mutated. Colour is the number of repetitions for each mutation and Y is the variance of the data.

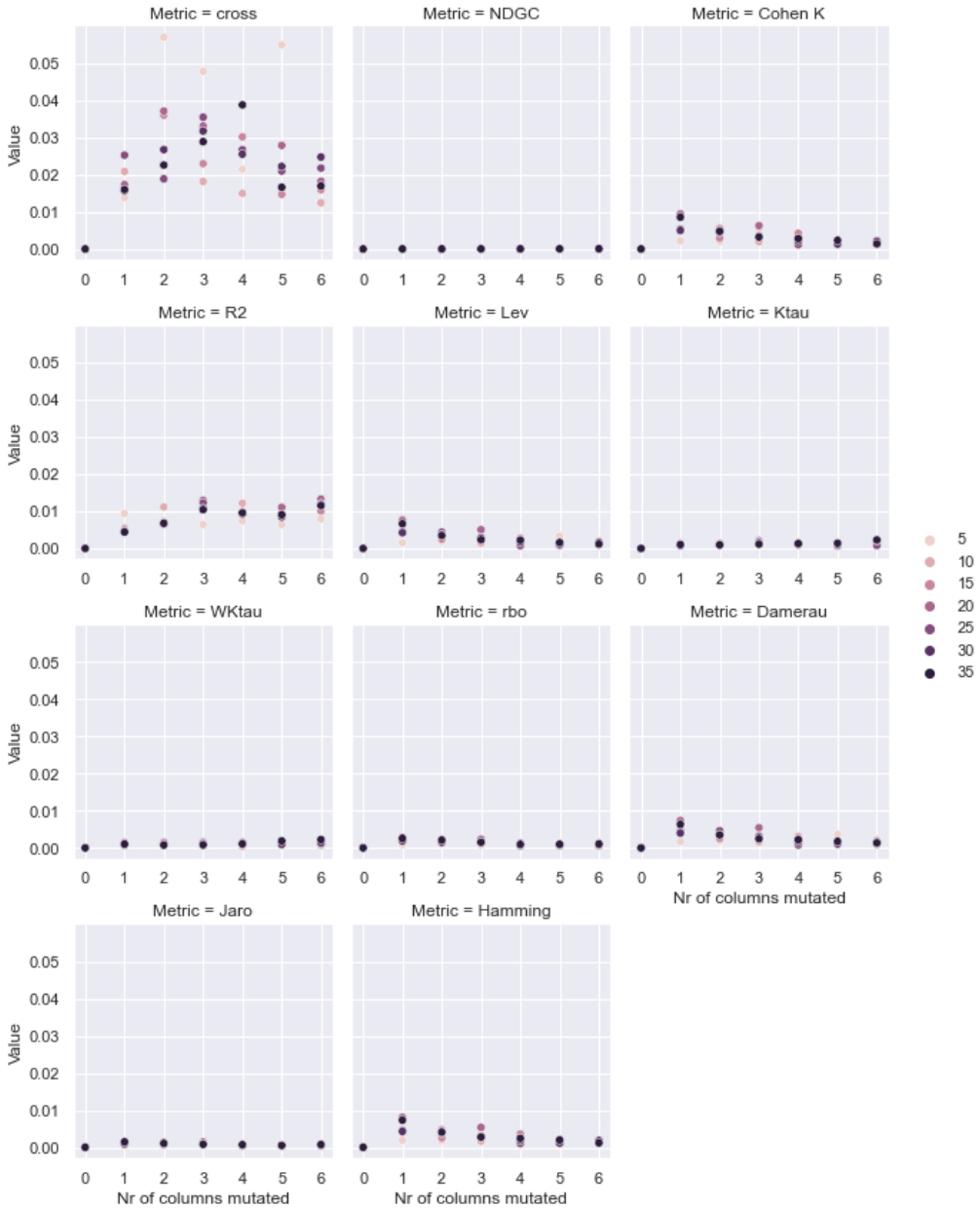
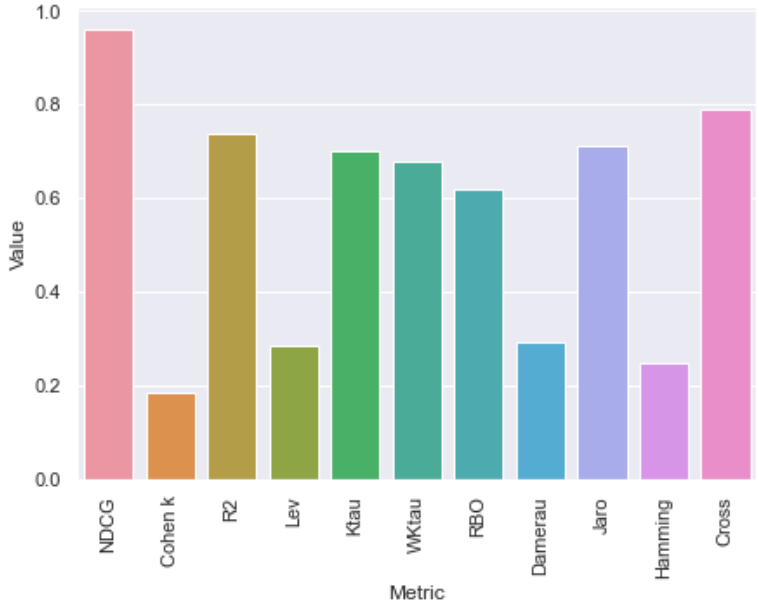


Figure 3.8: Result on a synthetic and real data



3.3.5 Discussion

With the results found, we feel that are better alternatives to cross-validation. At least Kendall tau, Weighted Kendall tau, and RBO seem like alternatives to cross-validation. Firstly, they seem to be directly connected to a difference in the dataset, secondly, they are a 0-1 metric and thirdly, variance across different iterations is also lower. From these, the metrics based on ranking metrics seem to work best, where Kendall tau, weighted Kendall tau and RBO have better performance than the rest. As for the variance with the number of repetitions, we also see that the ranking-based metrics have good stability, while the cross-validation and text-based metric have higher variability with a low number of repetitions (even if low) - figure 3.7. Text metrics also have a suitable performance, even though they have a drastic drop with only one column mutated (figure 3.5).

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. Ranking metrics (Kendall Tau, Weighted Kendall Tau, and RBO) and R^2 have shown the potential to capture inter-column relationships in a more consistent way than cross-validation. 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.

3.4 Data quality Metrics

This section is based on the paper entitled "Development and Validation of a Data Quality Evaluation Tool in Obstetrics Real-World Data through HL7-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 [132]. 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 construed as a linear concept; it is intrinsically 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 [133]. 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 [134]. 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 [135, 134, 136]. 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 [137, 138, 139, 140, 141, 142].

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 [136]. So, with all of 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. Consequently, in this paper, we propose:

- Create a tool for identifying data quality issues in obstetrics EHRs;
- Enlighten on the issues that can appear with a full deployment of such a tool
- Suggestion of a creation of a single score for data quality for comparison of high-quality and low-quality records in a database.
- Assess how such a tool can work in early-stage real-world scenarios and how to work with obstetricians to improve data quality.
- Identify data quality issues on obstetrics 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 [143]. 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., [144], 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 Data Quality (DQ) dimensions: completeness, consistency, duplicity, correctness, timeliness, spatial stability, contextualization, predictive value, and reliability [145]. Then a review of Bian et al. [143] 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 [146]. 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 [147] 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. [148] already reviewed some data quality assessment tools, like tools from OHDSI [149] or TAQIH [150]. Additionally, we found some others with similar purposes and characteristics like the work presented data dataquieR [151], 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 [152], taking into account the semantics of data and their meanings within their context. Furthermore, the work from Rajan et al. [153] 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 [154].

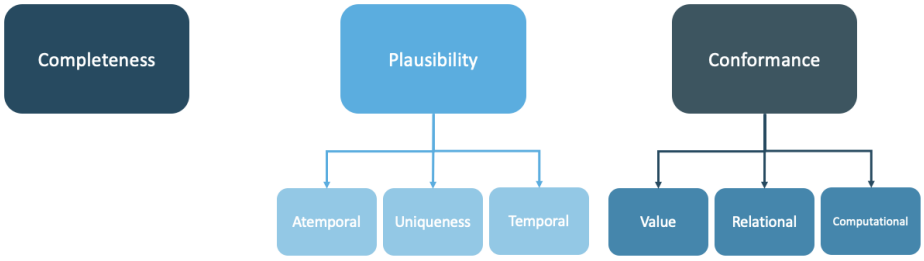
Regarding data quality assessment as a whole, the works of [155], focused on outlier detection in large-scale data repositories. The works of [156] focused on the exploration and identification of dataset shifts, contributing to the broad examination and repurposing of large, longitudinal data sets. The works of García-de-Léon-Chocano [157, 158, 159] 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 [160] 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 standard-based interoperability in mind. Finally, we have not seen, until the moment of this paper, any implementation that used machine learning 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 fetus 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.1. This study received Institutional Review Board approval from all hospitals included in this study with the following references: Centro Hospitalar São Jão; 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. 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.

Figure 3.9: 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 BN. BNs are probabilistic graphical models that represent a set of variables and their conditional dependencies via a directed acyclic graph. These networks are particularly adept at simultaneously predicting multiple variables, offering a cohesive framework for inferring several columns within a single network structure [161]. We used this model due to the possibility of using a single model for classifying the plausibility of all columns and due to its interpretable nature. The networks were created with the pgmpy package [162]. Secondly, we added the outlier-tree method [163] 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 identifiy 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 neighbors, identifying outliers without assuming a specific data distribution. It is versatile for different data structures but sensitive to parameter settings, like the number of neighbors. Interquartile Range (IQR) was also added as a supportive metric. It identifies outliers by marking data points that lie significantly outside the middle 50% (between the 1st and 3rd quartiles) of a dataset. Finally, a rule system was implemented to leverage domain knowledge in the overall scoring. The system is based on *great_expectations* package [164]. A set of rules was defined by the team, focusing on impossible numbers present in age, weight, or relationship between variables. The rules covered plausibility and conformance. The Conformance-based were related to technical issues like the format of dates, 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. We also added plausibility for the relationship between columns, namely weight across different weeks of gestation. 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 BNs 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 3.9 are described in the table 3.4.

As for Z-Scores, they were defined for all continuous variables based on the interquartile range. Then, rows were also assessed with distance analysis, with Local Outlier Factor and Elliptic Envelope from *scikit-learn* and the outlier-tree algorithm. We also added a rule engine, using the package. Rules were defined by the team, focusing on impossible numbers present in age, weight, or relationship between variables. As for missing information was created with all the data, creating the scoring based on the inverse of the missing percentage. Missing detection was based on primary key variables. For completeness, we used the inverse of the percentage of nulls in the training set. The API for serving the prediction models was developed with FastAPI. So, the methods applied in terms of the DQA framework shown in figure 3.9 are described in the table

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 3.12. To assess the tool's usefulness, we implemented it in a production environment and collect metrics regarding the data being produced. Then we presented some rows (or patient's records) to selected obstetrics clinicians for them to assess how likely the information is to be suitable for usage or rank it according to the perceived quality of the record. We also compared the results with the ones from the model to make sanity checks regarding the model's performance and

Table 3.4: 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

adequacy. We used Kendal Tau and Average Spearman's Rank Correlation Coefficient. Kendall Tau is a non-parametric statistic used to measure the strength and direction of the association between two ordinal variables. It calculates the difference between the number of concordant and discordant pairs of observations, normalized 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. It is based on the ranked values of the variables rather than their raw data, producing a value between -1 (perfect inverse relationship) and 1 (perfect direct relationship). We wrote all the code in Python 3.10.6 with the usage of the *scikit-learn* library for preprocessing, and evaluation [103].

3.4.5 Results

A BN with structure and parameters learned from the training dataset reached an average Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.857. The results are in the table 3.5.

The network is as represented in figure 3.10.

As for the rules created, they were conformance-based, like the format of dates, and conformance to the value set (i.e. Robson group, bishop scores, or delivery types). We also added plausibility rules, like expected values for BMI, weight, and gestational age. We also added plausibility for the relationship between columns, namely weight across different weeks of gestation. We have also added a relationship of greatness between ultrasound weights more than 5 weeks apart. The method of calculating the final score is stated in figure 3.11.

Table 3.5: Repeated Cross-Validation (10x2) Results: Column acronym with AUROC along with 95% Confidence Interval. Acronym description is available in appendix A.1

AP	0.944	[0.943, 0.945]	VNH	0.894	[0.893, 0.895]
AG	0.797	[0.778, 0.816]	TPEE	0.816	[0.815, 0.816]
EA	0.969	[0.968, 0.969]	AA	0.751	[0.743, 0.758]
CA	0.958	[0.958, 0.958]	GR	0.931	[0.93, 0.932]
IA	0.638	[0.637, 0.638]	V	0.983	[0.982, 0.983]
PI	0.881	[0.88, 0.881]	TP	0.866	[0.865, 0.868]
IMC	0.881	[0.881, 0.882]	VCS	0.79	[0.789, 0.791]
NRC	0.75	[0.75, 0.75]	ANP	0.942	[0.938, 0.946]
IGA	0.968	[0.968, 0.969]	GS	0.514	[0.507, 0.52]
SGP	0.974	[0.974, 0.974]	S	0.896	[0.896, 0.897]
VA	0.974	[0.974, 0.974]	VP	0.771	[0.77, 0.772]
TG	0.728	[0.726, 0.73]	TPNP	0.952	[0.951, 0.952]
Average 0.857 [0.846, 0.868]					

Figure 3.10: Network learned

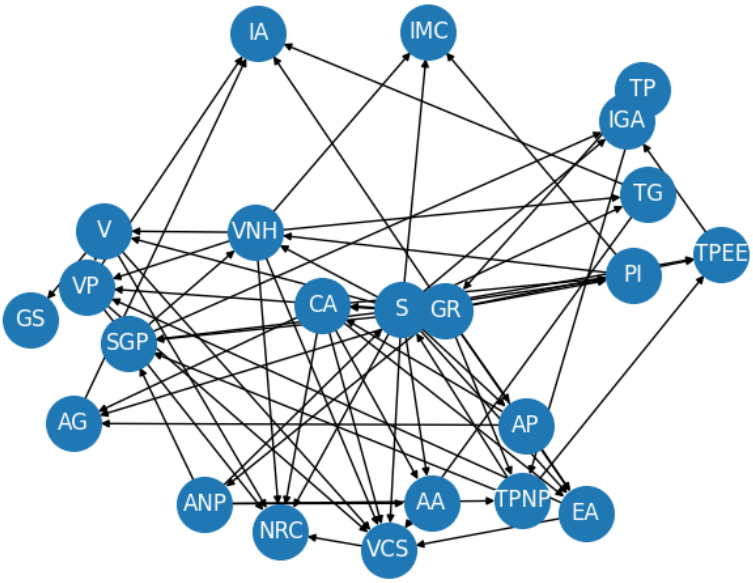
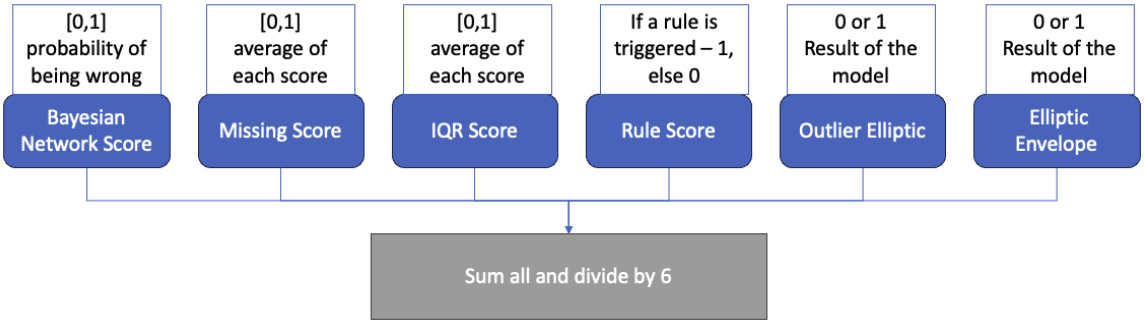


Figure 3.11: Workflow for creating the final score and which elements are used to do so



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 Health Level Seven (HL7) Fast Healthcare Interoperability Resources (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 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 [165].

For validation, we deployed the tool in docker format in a hospital to gather new data. We gathered 3231 new cases and returned a score for quality as exemplified in figure 3.12. Being that the score is from 0 to 1, the average score was 0.23 and Inter-Quartile Range (IQR) was 0.03. As for the clinicians' assessment, we got 4 answers. Figure 3.13 shows the aggregated rankings of the clinicians per record and is ordered by the rank provided by the model.

The Average Spearman's Rank Correlation Coefficient was 0.04 and the Kendall's Tau was 0.0 with a P value of 1. Defining the threshold for the model at 0.23 and the threshold for the physician at 8, we got an accuracy of 0.9, precision of 1, recall of 0.66 and f1-score at 0.8.

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 a automatic assessment tool to the human perception of quality and the

Figure 3.12: Model score for newly seen data

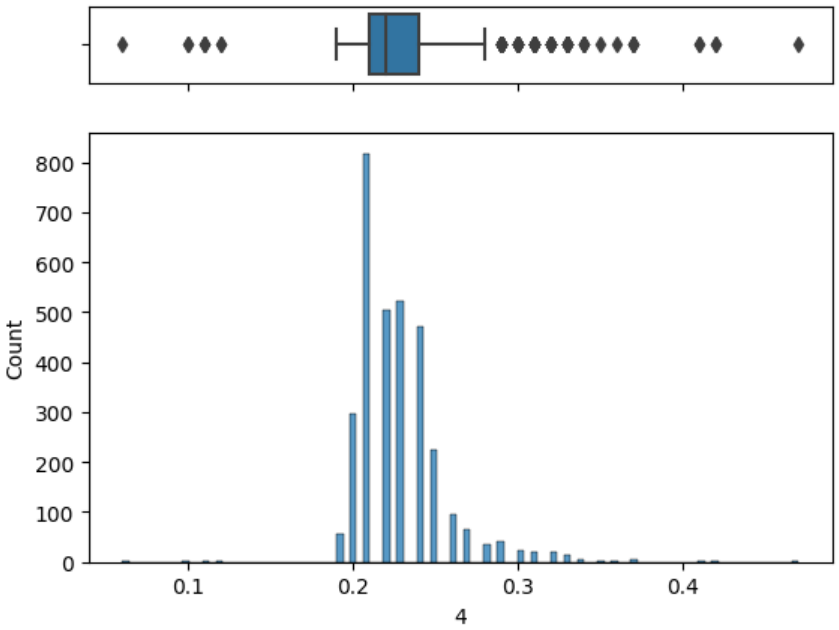
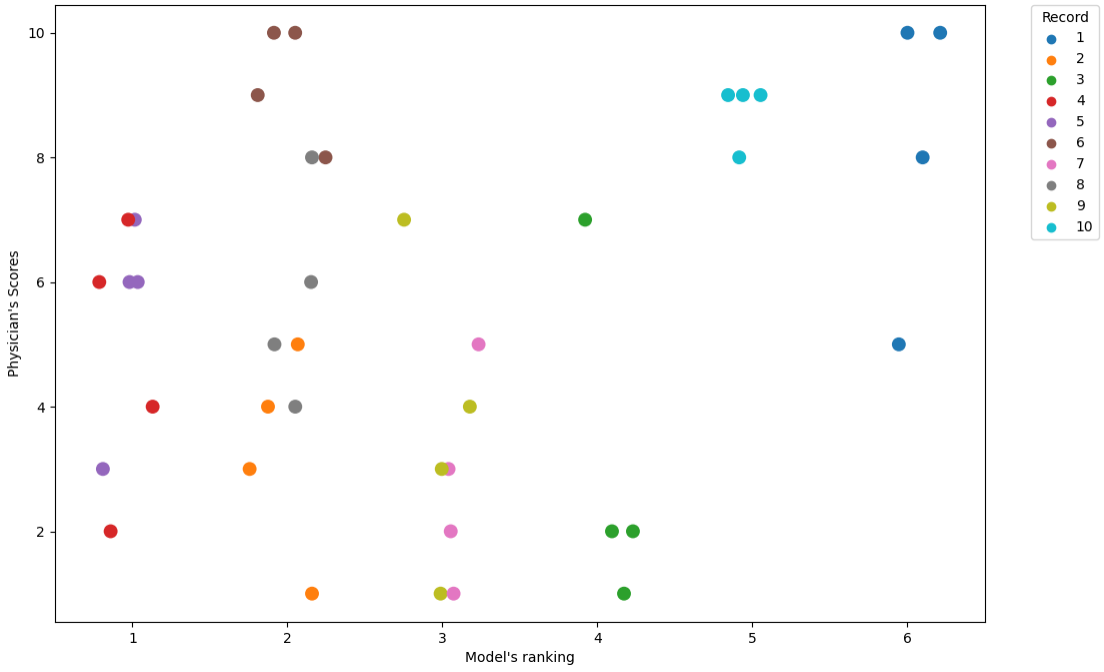


Figure 3.13: Comparison of clinical assessment of records with the model. Y is the clinicians' assessment, X is the ranking of the record per the model. Equal numbers on the X mean a tie per the model interpretation. Color is the record ID.



issues linked to doing so. Secondly, the fact that we applied explainable machine learning methods such as bayesian networks to leverage the potency of advanced data anlysis without compromising interperability 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. 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 out that asking clinicians to purely assess the quality of a record in an EHR is not an easy task. We found out that for a proper assessment, a context and objective must be defined in order to make the evaluation more objective. Moreover, the ranking methodology, even though is very useful for comparing with the model, it is not easy for clinicians to order 10 records when some of them have a perceived equal level of quality. This is a very important aspect to take into account when designing the evaluation of data quality. Probably a categorical evaluation of yes/no could be more useful and then compare it with the ordering of the model and define thresholds based on that. These reasons are probably the cause behind the great variability between clinicians and between clinicians and the model. However, we do see a tendency for a higher agreement of the worst quality results than the best ones. This result suggests that the system may not be suitable for ranking good-quality records and clinicians are also not able. However, it could be useful to alert for low-quality ones, which is also a very important task with a great impact on the quality of the data. These findings are supported not only by 3.13 but also by figure 3.12 where we can add some threshold for the need for a human review. From the preliminary data in

the questionnaires and looking at the graph, we believe that a threshold of around 0.23 could be a good starting point. As pointed out in the end of results section, we have tried to create binary outcomes with thresholds we chose. With a threshold of 8 for physician's ranking for defining a bad record and a threshold of 0.23 for the model evaluation, we achieved very good performance metrics. However, the thresholds defined were optimal and rely on very few samples. More data and research could be employed 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 to use the data for research, a higher threshold could be used. On the other hand, if the objective is to use the data for day-to-day clinical decisions, a lower threshold could be used. For the next steps, we believe a research path could be of identifying contexts for applying data quality checks like primary usage, research purposes, and aggregated analysis for decision-making among others could help better target those contexts and the importance of each variable for those use cases. This could be interesting to add to the tool in order to weigh the different variables according to the context.

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.

3.5 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 centralized 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 centralized paradigm, with the added benefit of not requiring heavy data sharing.

3.5.1 Introduction

As the use of AI is increasing in the healthcare space [166], increased demand for ethical usage of personal patient data is occurring as well [167]. 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 [168] and HIPAA in the USA [169]), and on the public side, with an increased concern with continuous data breaches across institutions [170]. 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 [171]. 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 [172], never quite raising into its full potential for promoting best healthcare practices [173, 174, 175] with data sharing between institutions. In order to overcome this issue, there are a few, more complex, systems that aggregate data from several institutions, so more robust algorithms could be trained. However, this globally centralised aggregation 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 [176].

There are already some implementations of distributed systems in the healthcare space, but we lack a robust understanding of how these models behave with real data, when compared with the classical models built with all the aggregated data. Additionally, the main issues regarding the development and implementation of such systems in healthcare are still elusive. So we aim to understand how distributed mechanisms behave compared to using all data in the healthcare space and if they are a suitable replacement for traditional machine-learning pipelines. The contributions of this paper are:

- Understand how to address the lack of data quality of real-world data regarding distributed model creation;
- Evaluate a distributed model against its local counterparts;
- Measure the prediction performance difference between a distributed model and a centralised one;

- Open a research path for using distributed models to predict several target variables in obstetrics clinical research.

3.5.2 Theoretical background and Related Work

Distributed learning [177] 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) [171] 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 [178, 179]. These methods use the strength of neural networks and several algorithms like federated averaging to create distributed models capable of handling complex data like text, sound, or image [180]. However, considering that there are great amounts of information, especially in healthcare, stored as tabular data [150, 181, 182] and that neural networks are often not the best tool for such data structures [183], there is a lack of knowledge in the traditional machine learning techniques in a distributed manner.

Nevertheless, there have been some health-related distributed machine-learning projects successfully implemented, such as euroCAT [184] 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 [185] did similar work, with an added centralised database in the middle, but the training being done with a decentralized system.

Finally, a few works have explored the evaluation of models in a distributed manner, for example, comparing centralised machine learning, distributed machine learning and federated learning on MNIST dataset [186]. Also, works that evaluate federated learning on MNIST, MIMIC-III and PhysioNet ECG datasets, but not in comparison with other methods [187]. The work by Tuladhar and colleagues [177] uses healthcare images and/or public and curated datasets. As far as we know, this is the first time a distributed machine learning evaluation is done with real-world clinical data from several different data sources.

3.5.3 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. 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 3.6 and 3.7. The data dictionary is in appendix A.1.

Table 3.6: Silos overview. categorical columns have a snippet of the most used category and a percentage. Continuous variables have a mean and standard deviation. Abbreviation meaning in the appendix A.1. The last row is the number of patients. * columns were used as target.

Column	Silo 1	Silo 2	Silo 3	Silo 4	Silo 5	Agrr.
IA*	31.1 **5.7**	30.7 **5.6**	31.1 **5.9**	31.1 **6.3**	31.3 **5.6**	31.1 **5.6**
GS*	a,rh.. **40%**	a,rh.. **40%**	a,rh.. **39%**	o,rh.. **38%**	a,rh.. **41%**	a,rh.. **40%**
PI	66.4 **14.4**	66.1 **13.5**	65.5 **14.1**	65.5 **14.1**	65.5 **14.4**	66.0 **14.1**
PAI	81.4 **14.9**	79.5 **14.5**	78.0 **15.2**	79.6 **16.3**	78.3 **14.2**	78.8 **14.5**
IMC*	25.2 **8.6**	25.2 **6.2**	25.0 **5.3**	25.0 **8.9**	24.9 **7.8**	25.1 **7.0**
CIG	Null **84%**	Null **85%**	Null **87%**	Null **90%**	Null **88%**	Null **88%**
APARA	Null **45%**	Null **41%**	1.0 **37%**	Null **42%**	1.0 **35%**	Null **39%**
AGESTA*	1 **41%**	1.0 **43%**	1.0 **39%**	1 **39%**	1 **43%**	1.0 **42%**
EA*	Null **75%**	Null **60%**	Null **75%**	Null **67%**	Null **45%**	Null **60%**
VA	Null **90%**	Null **80%**	Null **89%**	Null **93%**	Null **55%**	Null **77%**
FA	Null **99%**	Null **83%**	Null **94%**	Null **96%**	Null **60%**	Null **83%**
CA*	Null **88%**	Null **73%**	Null **86%**	Null **90%**	Null **62%**	Null **75%**
TG*	espo.. **62%**	espo.. **90%**	espo.. **85%**	espo.. **63%**	espo.. **89%**	espo.. **85%**
V	s **99%**	s **92%**	s **99%**	s **94%**	s **99%**	s **98%**
NRCPN*	7.3 **4.7**	7.0 **6.4**	6.4 **3.9**	5.5 **3.6**	10.5 **5.1**	8.4 **5.1**
VP	Null **82%**	Null **85%**	Null **81%**	Null **79%**	Null **73%**	Null **76%**
VCS	s **61%**	s **53%**	s **78%**	s **50%**	s **70%**	s **68%**
VNH	s **88%**	s **76%**	s **81%**	Null **52%**	s **71%**	s **69%**
B	Null **95%**	Null **78%**	Null **90%**	Null **97%**	Null **81%**	Null **83%**
AA	Null **89%**	Null **78%**	apr... **52%**	Null **96%**	Null **71%**	Null **73%**
BS	Null **98%**	Null **79%**	Null **97%**	Null **86%**	Null **97%**	Null **95%**
BC	Null **99%**	Null **83%**	Null **99%**	Null **87%**	Null **97%**	Null **97%**
BDE	Null **99%**	Null **83%**	Null **99%**	Null **88%**	Null **97%**	Null **97%**
BDI	Null **99%**	Null **83%**	Null **99%**	Null **87%**	Null **97%**	Null **96%**
BE	Null **99%**	Null **83%**	Null **99%**	Null **87%**	Null **97%**	Null **96%**
BP	Null **99%**	Null **83%**	Null **99%**	Null **87%**	Null **98%**	Null **97%**
IGA*	38.1 **3.5**	38.8 **2.2**	38.9 **1.6**	38.8 **2.4**	38.6 **2.1**	38.7 **2.2**
TPEE	Null **70%**	Null **75%**	Null **65%**	Null **64%**	Null **60%**	Null **65%**
TPEI	Null **98%**	Null **84%**	Null **93%**	Null **92%**	Null **99%**	Null **93%**
RPM	Null **91%**	Null **94%**	Null **89%**	Null **92%**	Null **85%**	Null **88%**
DG*	Null **88%**	Null **90%**	Null **90%**	Null **91%**	Null **90%**	Null **89%**
TP*	part.. **43%**	part.. **53%**	part.. **44%**	part.. **52%**	part.. **49%**	part.. **51%**
ANP	cefá.. **92%**	cefá.. **94%**	cefá.. **95%**	cefá.. **95%**	cefá.. **94%**	cefá.. **94%**
TPNP*	espo.. **53%**	espo.. **52%**	espo.. **58%**	espo.. **62%**	espo.. **62%**	espo.. **53%**
SGP*	38.5 **2.8**	38.9 **2.0**	39.1 **1.7**	39.0 **2.3**	38.9 **2.0**	38.9 **2.0**
GR*	1 **22%**	1 **20%**	1 **24%**	Null **81%**	1 **28%**	1 **24%**
N (total)	8039	8566	4989	2364	18177	80874

3.5.4 Methods

Data was prepossessed with the removal of features with high missing rates ($> 90\%$ in all silos). All missing value representations were standardized. The imputation process was done using the

Table 3.7: Silos overview part 2. categorical columns have a snippet of the most used category and a percentage. Continuous variables have a mean and standard deviation. Abbreviation meaning in the appendix A.1. The last row is the number of patients. * columns were used as target.

Column	Silo 6	Silo 7	Silo 8	Silo 9	Agrr.
IA	31.3 **5.2**	31.4 **5.4**	31.5 **5.6**	30.1 **5.6**	31.1 **5.6**
GS	a,rh.. **42%**	a,rh.. **39%**	a,rh.. **40%**	a,rh.. **42%**	a,rh.. **40%**
PI	65.6 **13.5**	66.0 **13.7**	65.6 **14.1**	67.4 **14.6**	66.0 **14.1**
PAI	77.7 **13.4**	79.2 **14.7**	76.7 **13.0**	83.1 **15.2**	78.8 **14.5**
IMC	24.9 **5.1**	24.9 **7.0**	24.8 **8.0**	25.7 **5.6**	25.1 **7.0**
CIG	Null **91%**	Null **91%**	Null **86%**	Null **90%**	Null **88%**
APARA	1.0 **38%**	Null **43%**	Null **41%**	Null **43%**	Null **39%**
AGESTA	1.0 **44%**	1 **43%**	1.0 **42%**	1.0 **40%**	1.0 **42%**
EA	Null **59%**	Null **61%**	Null **69%**	Null **61%**	Null **60%**
VA	Null **79%**	Null **82%**	Null **88%**	Null **82%**	Null **77%**
FA	Null **82%**	Null **86%**	Null **94%**	Null **89%**	Null **83%**
CA	Null **69%**	Null **75%**	Null **85%**	Null **78%**	Null **75%**
TG	espo.. **88%**	espo.. **85%**	espo.. **86%**	espo.. **93%**	espo.. **85%**
V	s **97%**	s **99%**	s **98%**	s **99%**	s **98%**
NRCPN	6.8 **4.0**	7.7 **3.2**	9.3 **4.5**	8.9 **5.5**	8.4 **5.1**
VP	Null **68%**	Null **74%**	Null **71%**	Null **78%**	Null **76%**
VCS	Null **53%**	s **87%**	s **63%**	s **87%**	s **68%**
VNH	Null **62%**	s **63%**	s **69%**	s **83%**	s **69%**
B	Null **90%**	Null **53%**	Null **93%**	Null **82%**	Null **83%**
AA	Null **84%**	apr... **61%**	Null **89%**	Null **74%**	Null **73%**
BS	Null **99%**	Null **98%**	Null **99%**	Null **95%**	Null **95%**
BC	Null **100%**	Null **100%**	Null **100%**	Null **97%**	Null **97%**
BDE	Null **100%**	Null **100%**	Null **100%**	Null **97%**	Null **97%**
BDI	Null **100%**	Null **100%**	Null **99%**	Null **97%**	Null **96%**
BE	Null **100%**	Null **100%**	Null **100%**	Null **97%**	Null **96%**
BP	Null **100%**	Null **100%**	Null **100%**	Null **97%**	Null **97%**
IGA	38.7 **1.8**	39.0 **2.0**	38.6 **2.1**	38.8 **1.9**	38.7 **2.2**
TPEE	Null **65%**	Null **64%**	Null **65%**	Null **63%**	Null **65%**
TPEI	Null **92%**	Null **86%**	Null **87%**	Null **94%**	Null **93%**
RPM	Null **85%**	Null **84%**	Null **90%**	Null **94%**	Null **88%**
DG	Null **92%**	Null **88%**	Null **90%**	Null **87%**	Null **89%**
TP	part.. **54%**	part.. **52%**	part.. **48%**	part.. **59%**	part.. **51%**
ANP	cefá.. **93%**	cefá.. **94%**	cefá.. **95%**	cefá.. **94%**	cefá.. **94%**
TPNP	espo.. **64%**	Null **100%**	espo.. **50%**	espo.. **65%**	espo.. **53%**
SGP	38.8 **1.8**	39.2 **1.7**	38.7 **2.0**	39.0 **1.6**	38.9 **2.0**
GR	1 **27%**	1 **25%**	1 **21%**	3 **27%**	1 **24%**
N (total)	12002	8258	6693	11786	80874

mean value (for continuous variables) or a new category (NULLIMP) for categorical variables. All categories were encoded as numbers using a previous mapping created based on all possible categories in all silos. Even though an ordinal relationship is created among features, we believe that

since we are applying this methodology to all datasets, which will be the source for all tests (local, distributed and centralised), that fact may be ignored. When training classification models, all of the target variable classes must be known at that moment and should be present in each split of the cross-validation. So, when assessing the training dataset, low-frequency target classes ($n < 25$) were up-sampled with Synthetic Minority Oversampling Technique (SMOTE) [188] and missing target classes were addressed with dummy rows creation by the imputation of the mean for continuous variables and mode for categorical variables (per silo). These preprocessing mechanisms were applied in each run and for each target. The distributed model was an ensemble of models from each silo on a weighted soft-voting basis, defining weights and thresholds based on the training set scores. All procedures were coded in Python 3.9.7 with the usage of the *scikit-learn* library [103] and *mlxtend* library [104]. This study received Institutional Review Board approval from all hospitals included in this study with the following references: CHUSJ; 08/2021, CHBV; 12-03-2021, ULSM; 39/CES/JAS, HSOG; 85/2020, CHTS; 43/2020, CHVNGE; 192/2020, CHEDV; CA-371/2020-0t_MP/CC, ULSAM; 11/2021.

3.5.4.1 Model Performance Evaluation

Local models were built with each silo's data. The distributed model was built as an ensemble of all the local models with weighted averaging. The centralised model was trained with a training dataset from all the silos combined. All models were built for a certain outcome variable with cross-validation and then compared, over 10 stochastic runs, with evaluation being performed on a test set held out from each silo. The metrics used for classification models were Weighted AUROC computed as One-versus-Rest, Weighted AUPRC. The metrics for regression models were Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The algorithm is shown in the algorithm 2. This rendered over 1000 different combinations.

3.5.4.2 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. Nineteen features were used as target outcomes. These features were selected by filtering by the percentage of null (below 50%). For categorical outcomes, thirteen were selected. For continuous variables, six were selected. Details can be seen in tables 3.6 and 3.7.

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. We did the comparison between the distributed model and sequentially the centralised and correspondent local model across all algorithms.

Pre-process all silos (null standardization, imputation, encoding);
for *target in target list* **do**
  Create a centralised model with all the data with a 2x10 Cross-Validation
  Create distributed (ensemble of all models) model with:
  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 CV
    - define weights based on scores in the train set
  end
  for *n in 10 repetitions* **do**
    for *silo in imputed silos* **do**
      - Train-Test Split (80:20)
      - train local model with hyper-parameter tuning with 2x10 CV
      - predict local on the test set
      - predict distributed on the test set
      - predict centralised on the test set
    end
  end
end

Algorithm 2: Creation and evaluation of the 3 different models

3.5.5 Results

Table 3.8 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 well the distributed model would work as opposed to the local one per silo. We also calculated the *P* value for the means.

Table 3.8: Comparison for the centralised model, distributed model and local model (Mean for all model and all columns). Bold for *P* value below 0.05.

		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

3.5.6 Discussion

The imputation process was done using the mean value (for continuous variables) or a new category (NULLIMP) for categorical variables. All categories were encoded as numbers using a previous mapping created based on all possible categories in all silos. Even though an ordinal relationship is created among features, we believe that since we are applying this methodology to all datasets, which will be the source for all tests (local, distributed and centralised), that fact may be ignored. When training classification models, all of the target variable classes must be known at that moment and should be present in each split of the cross-validation. So, when assessing the training dataset, low-frequency target classes ($n < 25$) were up-sampled with SMOTE [188] and missing target classes were addressed with dummy rows creation by the imputation of the mean for continuous variables and mode for categorical variables (per silo). These preprocessing mechanisms were applied in each run and for each target. The distributed model was an ensemble of models from each silo on a weighted soft-voting basis, defining weights and thresholds based on

Figure 3.14: 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.

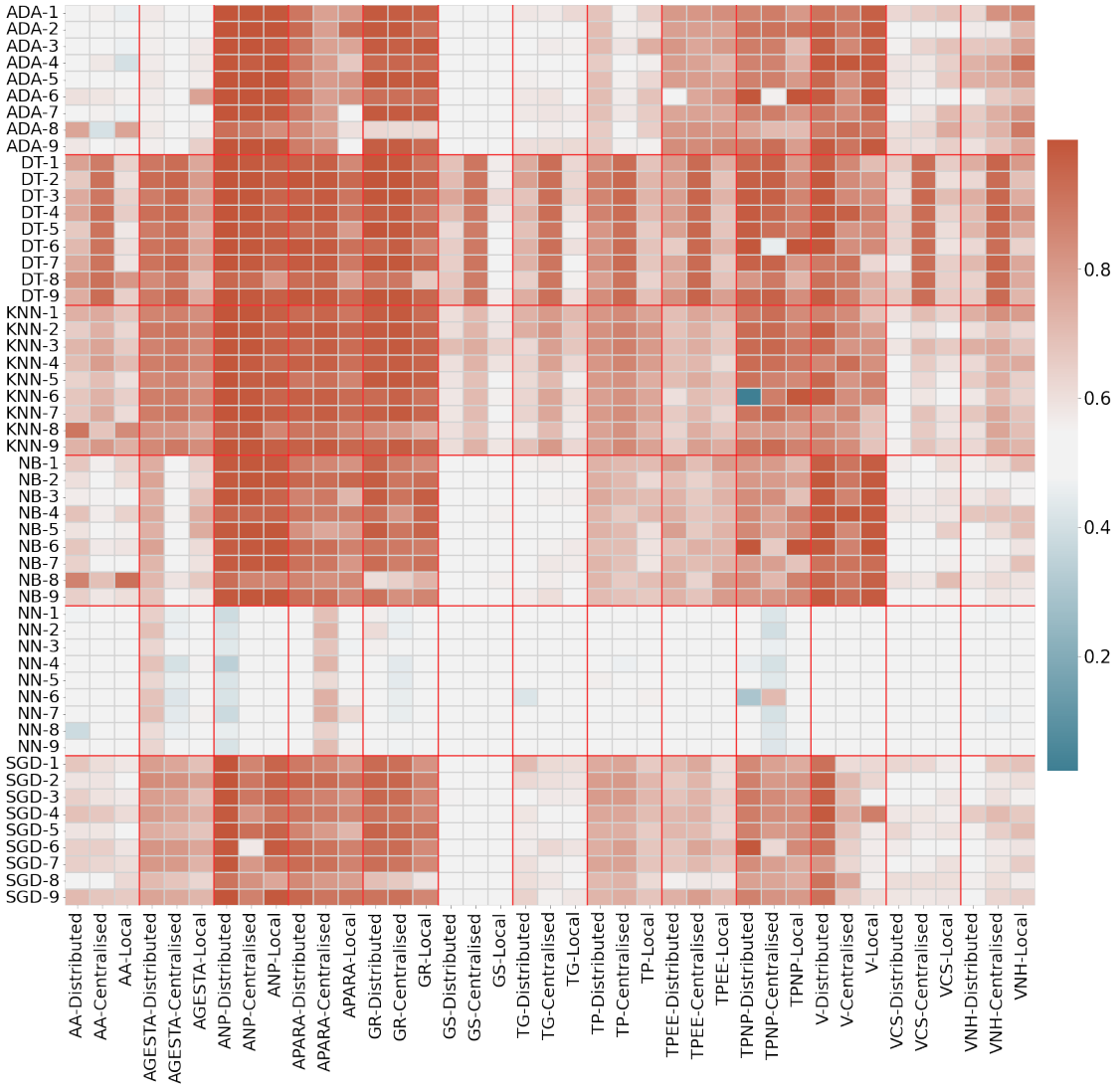


Figure 3.15: Heatmap of regression algorithm and silo vs Target variable and model type. Value is the MAE mean of all 10 experiments. The y axis is the algorithm and silo. X axis is Target variable and Method.

Table 3.9: Hypothesis testing of 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). ($>$ for better, = for non significance and $<$ for worse)

		$>$ Local	= Local	$<$ Local	Total
SGD	$>$ Centralised	72	14	9	95
	= Centralised	14	17	6	37
	$<$ Centralised	11	11	17	39
NN	$>$ Centralised	44	44	7	95
	= Centralised	2	33	2	37
	$<$ Centralised	0	17	22	39
KNN	$>$ Centralised	16	0	1	17
	= Centralised	10	2	1	13
	$<$ Centralised	72	28	41	141
ADA	$>$ Centralised	64	25	22	111
	= Centralised	5	12	10	27
	$<$ Centralised	10	6	17	33
NB	$>$ Centralised	51	19	34	104
	= Centralised	5	19	12	36
	$<$ Centralised	3	4	24	31
DT	$>$ Centralised	27	0	1	28
	= Centralised	8	0	0	8
	$<$ Centralised	97	12	26	135
Total		511	263	252	1026

the training set scores. The first thing that is noticeable is the high scores achieved in our analysis which show that all algorithms in all forms (local, distributed and centralised) have a good grasp on ranking data (negative on the bottom and positives on the top of a scale) for classification or predicting the value for regression. We notice that distributed models have performance similar to their centralised counterparts. $\sim$59% of all of the distributed models had similar or better performance than the centralised models. This suggests that a distributed model can be used to reliably infer information and does not compromise prediction performance when compared with the gold standard (centralised) while increasing privacy for the data owners.

Overall, our results suggest that it is possible to implement a distributed model without significantly losing information. However, there are still issues to be addressed. This methodology presents hurdles regarding categorical class handling. Firstly, all classes should be known firsthand 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.

Moreover, like in most secondary usage of data, other issues are important to keep in mind, even in such a controlled environment as this one. Even though the software is the same in every hospital, the clinical service is the same and the underlying data models are the same, the version of the software is not the same across all hospitals. This difference alone can alter the way each column is populated, mainly through front-end changes or label modification, among other aspects. Additionally, each hospital has its own workflows in practice that can also alter the way data is collected; changing timings or steps in a certain workflow can dramatically change the data acquisition and the reality it represents.

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 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 [189, 190]. 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. 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 3.14 and 3.15 and table 3.9 show us that decision trees and K-nearest neighbours implemented in a centralised manner are consistently better than the distributed counterpart. Even though this improvement may have a relationship to the target variable (i.e. figure 3.15 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 3.14) where silo 6 consistently behaves differently than the rest. 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 ML models in order to predict several clinical outcomes, with very good results for several target variables.

3.5.7 Conclusion

With this paper, it was possible to assess how well-distributed models can perform with real data, when compared to local models (trained with data from each silo) and global centralised models (trained with all data). These results show that an ensemble of models is able to fully grasp the specificity of the data, with performance similar to that of a model built with all the data. Even though the nature of the target and the silos can impact the performance, and several issues should be considered during the implementation phase, we are now fairly confident that distributed learning is a step forward regarding data privacy without loss of prediction performance. Finally, taken into account that the scores for several target variables are AUROC/AUPRC above 80% and MAE below 1, we will explore this further in a different work. We hope to be able to develop distributed models for predicting clinical outcomes like delivery type or Robson group, that could turn out useful in real-world clinical practice.

3.6 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 harbor 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.

3.6.1 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 [191]. 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 [192].

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 [192]. 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., [193], 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.

3.6.2 Rationale and Related Work

This work was initially suggested as a follow-up to a previous work of Rodrigues et al., [194] 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 [195, 196]. It is also used as a method of detecting regularities and patterns in multi-omics data that reveal different molecular subtypes [197, 198]. It can also be used to create unsupervised models for facilitating the annotation of data for supervised models [199].

K-means [200, 201, 202] is an unsupervised clustering algorithm used to group data points into K distinct clusters based on their similarity. It is widely used in machine learning, 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.

3.6.3 Materials & Methods

3.6.3.1 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 [103]. 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 [203]. 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* [103] and K-modes, as implemented by J. de Vos [204].

3.6.3.2 Data used

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 [105]. 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.

3.6.4 Results

As for results, the data from heart disease rendered the figure 3.16. In this, we focused on continuous variables only. For easier reading, the data is as shown in the table 3.10. We used data from the real world to test if everything would work similarly, rendering the image 3.17. We added a binary category to show how meaningless the value turn in order to get any information out of it.

Figure 3.16: 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

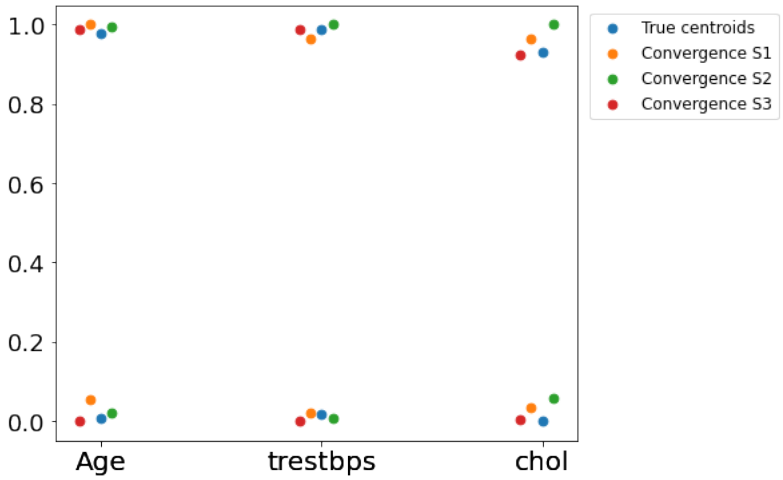
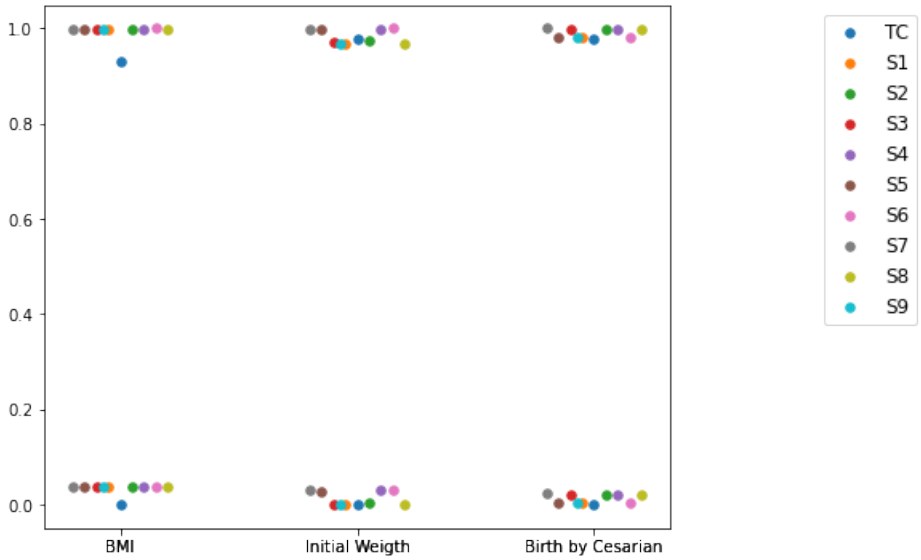


Table 3.10: Final Data points after convergence; S1, S2 and S3 are the centroids obtained in each silo (S) after convergence; True centroids are the centroids of the true means of all silos (TC)

	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 3.17: 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 3.11.

Then we experimented with categorical variables. Figure 3.18 shows the convergence of the silos with proportion data and K-means with that and with K-modes.

3.6.5 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

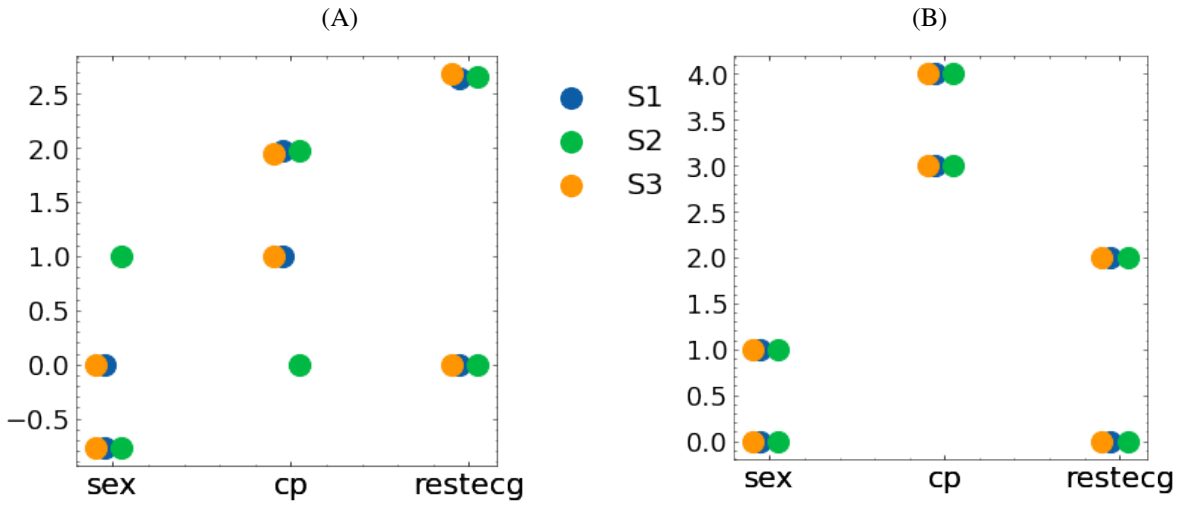
Table 3.11: Final Data points after convergence and true centroids of the true means of each silo (TC)

	Body Mass Index (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 3.18: 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.

3.6.6 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.

3.7 Leveraging 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.

3.7.1 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 [205, 206, 207] studies were used for ribociclib, PALOMA [208, 209, 210] for palbociclib, and MONARCH [211, 212] 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 [213]. 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 [214, 215] 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.

3.7.2 Materials & Methods

3.7.3 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 tumor 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.

3.7.4 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 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 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

Table 3.12: 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%)

(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).

3.7.5 Statistical Analysis

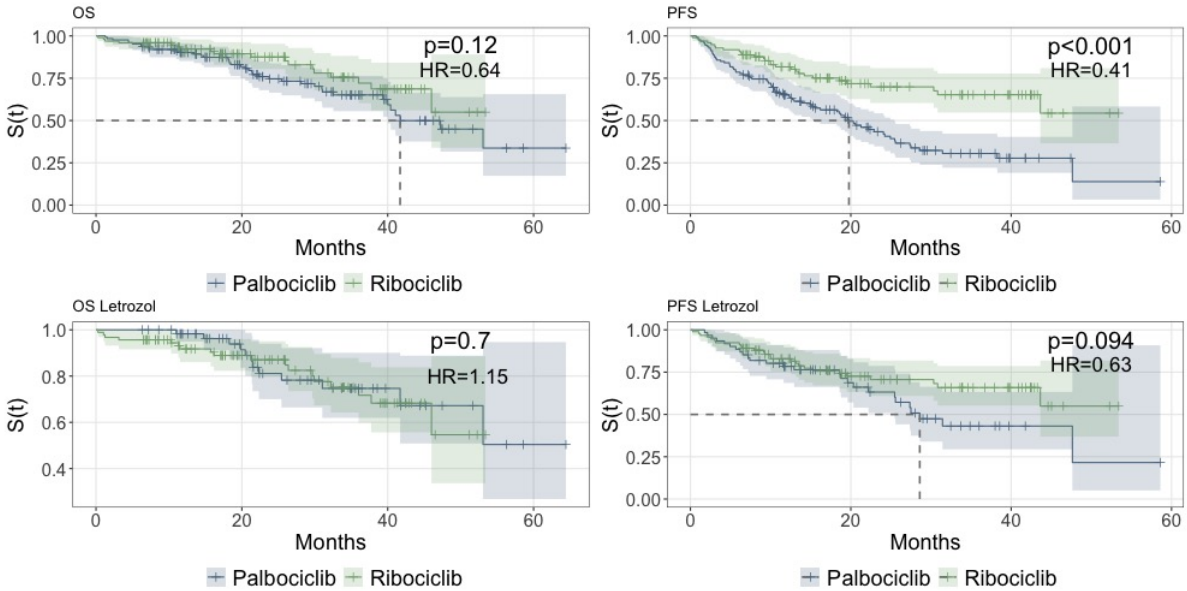
R was used for statistical analysis. Demographic, clinical characteristics and side effects were analysed using descriptive statistics (count, percentages and median/range). Kaplan–Meier test was used to determine the median PFS and OS in the entire population and subgroups. Log-rank test was used for 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 *Schoenfeld* 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 [216]. 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.

3.7.6 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 3.19). 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 3.19: 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.

When comparing ET with CDK4/6i as first-line treatment (figure 3.20). For this study we only compared patients with bone only metastasis. When comparing both CDK4/6i combined

Table 3.13: Cox Regression with palbociclib and Ribociclib - PFS and OS

	OS			PFS		
Characteristic	**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

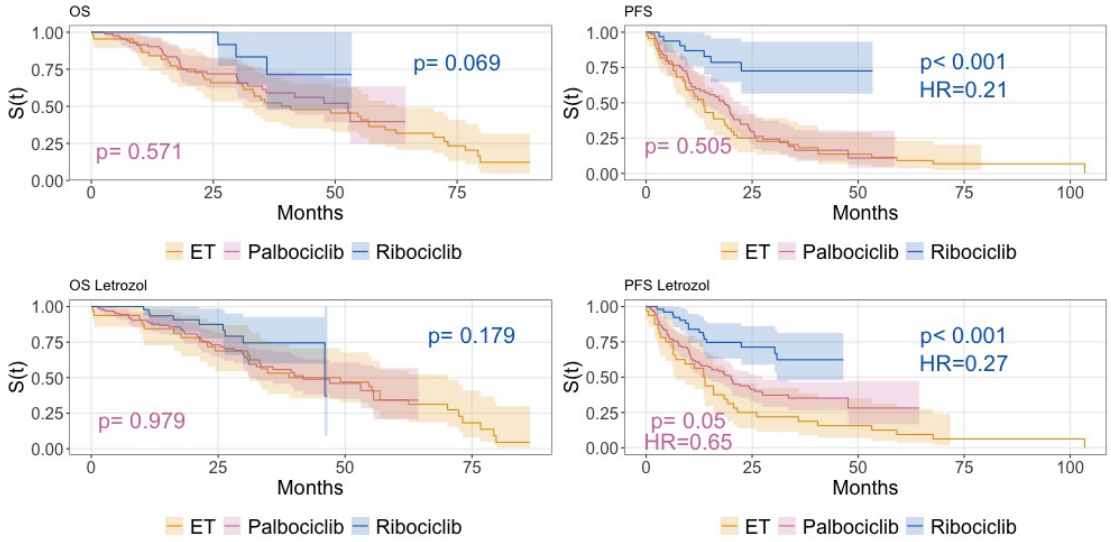
[1] HR = Hazard Ratio, CI = Confidence Interval

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 3.21). The weights were calculated as stated in the methods section.

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 HRfor PFS is 0.56 [95% CI 0.32-1;P=0.05].

Figure 3.20: 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.



3.7.7 Discussion

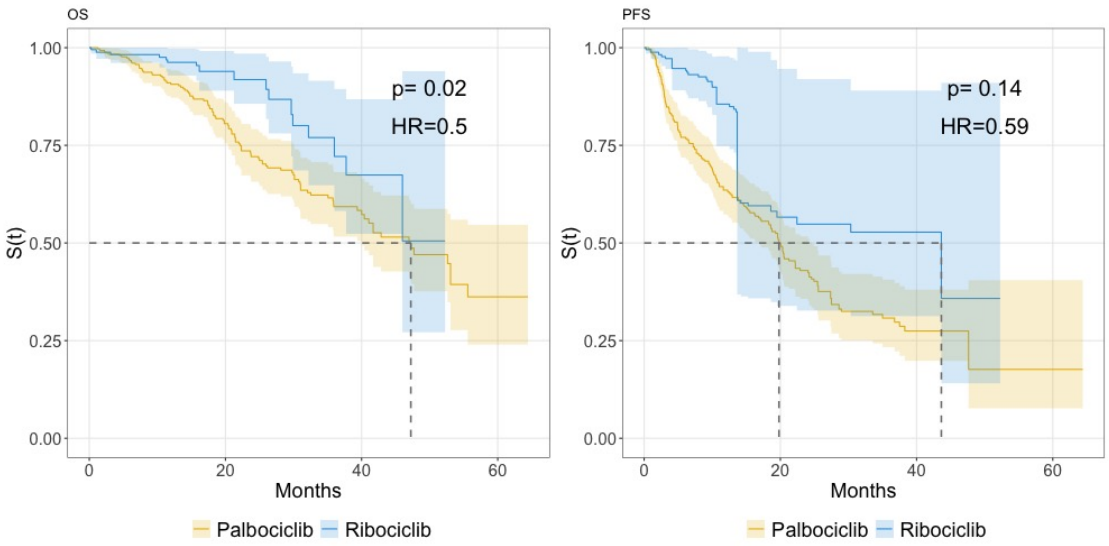
The aim of this study was to evaluate the real-world use of palbociclib and ribociclib in combination with ETfor HR+/HER2- and compare this drug class with traditional ET. Few real-world evidence studies of palbociclib and ribociclib used in daily clinical practice have been published identifying clinical 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 [209, 217]. As for ribociclib and bone-only metastases, MONALEESA-7 [207] has 24% and MONALEESA-2 has 40% [205] 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 [209], but in line with RWE studies (13.3–20.2 months) [213]. 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].

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 [207] and longer than 25.3 months (95% CI 23.0–30.3) in the MONALEESA-2 trial [205]. Regarding the subgroup

Figure 3.21: Comparison of palbociclib and ribociclib survival curves adjusted for propensity scores



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 2. This observation is further corroborated by the lower plots in figure 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 = 0.979 | RIB-LT vs LT; *P*=0.179)(figure 3.20 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 3.20 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.

3.7.8 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 ribocilib 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.

3.8 Leveraging 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 3.5. 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 Cesarean Section (C-Section).

3.8.1 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 this type of procedure

increasing over the last few years, it is still illusive the reasons behind such events. Reports from 2016 suggest that this increment is a global phenomenon, being that from 1990 to 2014, this type of delivery almost increases by 3-fold from 6.7% to 19.1% [218, 219]. Some of these impacts, being more prone to investigation in the last years, including the risk of infection, haemorrhage, organ injury and complications related to the use of anaesthesia or blood transfusion [220, 221]. There is also a higher risk of complications in subsequent pregnancies like uterine rupture, abnormal placental implantation and the need for hysterectomy [222, 223]. As for the infant, C-Sections include the risk of respiratory problems, asthma and obesity in childhood [222]. Facing this, in 2015, World Health Organisation

(WHO) released a statement regarding C-Sections rates. Even when other complications could not be totally assessed, it was concluded that C-Section rates higher than 10% were not associated with a reduction in maternal or newborn mortality [224].

Since there is no evidence that this type of procedure is beneficial for women or babies when there is no clear need for it, the focus on filtering such cases is important [219]. Moreover, particularly in Portugal, C-Sections are used as a way of financing healthcare institutions. This was implemented as a strategy of decreasing C-Sections across the country. A committee was created especially with the purpose of reducing the percentage of C-Sections nationally. One of the actions taken along this creation was the reduction of government funding for hospitals with rates of C-Sections above 25%. In 2020, the number of C-Sections in Portugal is about 36.3%. Almost at the all-time high of 36.9% in 2009 [225]. So, lowering the proportion of C-Section can provide health and financial benefits to institutions and populations alike. With this in mind, we developed a machine-learning algorithm-based support system to assist clinical teams to detect cases of potentially unnecessary C-Sections for analysis. So in this paper, we propose:

- help to provide a method of bringing to the discussion of clinical staff possible less than optimal care regarding deliveries;
- elaborates on how clinical decision support systems can be developed using interoperability standards;
- understand, based on the gathered data, which are the more impacting features for predicting delivery type outcome;
- open a research path regarding the evaluation of this type of clinical decision support system prior to the delivery;
- Perform a concise economic analysis to assess the potential financial impact of implementing the proposed clinical decision support tool.

3.8.2 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., [226] where a gradient boosting method was used to predict such an event using prenatal data to do so. Grobman et al., [227] performed a similar study with a multivariable logistic regression model. Different modalities of data were

also used to predict delivery type. Fergus et al. [228] introduces a method of predicting de- livery type using the fetal heart rate signals. Similarly, the work from Saleem et al. [229] 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. [230] where a boosting algorithm was used in order to predict delivery mode with enriched datasets. In addition, Gimovsky et al. [231] introduced decision trees to predict C-sections by physician group with 0.73 AUROC. The works of [232] resulted in a seven-variable model with 0.78 AUROC and the works of [233] resulted in a model with 0.82 AUROC, reaching 0.93 with a first cervical examination. Finally, the works of Meyer et al. [234] focused around selecting suitable for a trial of labor after cesarean 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 Fast Health- care Interoperability Resources (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.

3.8.3 Methods

3.8.3.1 materials

Data was 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 data is from all patients that had information registered in the obstetrics EHR and had a registered outcome of the pregnancy from 2019 to 2020. Despite differing software versions across hospitals, each institution used identical EHR software, ensuring the columns remained consistent.

3.8.3.2 Clinical Comparison Analysis

3.8.3.3 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.

3.8.3.4 Analysis

We obtained full approval from the ethics committee before commencing the study, as detailed in the 'Ethics Approval and Consent to Participate' subsection. The need for informed consent was waived by the ethics committee. 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 [235]. 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 labor 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 1 in the results section). We used a mixture of features selected by surveying the literature [236, 237, 238] 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 XGBoost, 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 3. The API for serving the prediction model was developed using FastAPI. We wrote all the code in Python 3.9.7.

3.8.4 Results

3.8.4.1 Descriptive Statistics

The number of samples varied across the hospitals, ranging from 2364 to 18177. Distributions of the selected variables are presented in table 3.14. The sum of all samples totals 73351.

The outcome variable had the following distribution as stated in table 3.15

3.8.4.2 The model

The AUROC is presented in table 3.16 for the best hyper-parameters found for each algorithm in the training data. All models used the variables indicated in table 3.14. While XGBoost was the

Table 3.14: 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 3.15: Distribution of Delivery Methods

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

best-performing algorithm, we selected LightGBM [239] 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 3.17.

3.8.4.3 Deployment

The purpose of this model is to be served as a API for usage within a healthcare institution and act as a supplementary management decision support tool for obstetrics teams. And for that 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 which can be used by every system that acts upon births and obstetrics departments. That is why we built it around the HL7 FHIR standard (R5 version) in order 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 - current build of the profiles `https://joofio.github.io/obs-cdss-fhir/`. The current spec is detailed in the following link. The process is identified in figure 3.22. We deployed this model in production in a single hospital and without a user interface, only collecting the data and prediction for later discussion and analysis. We collected 3231 requests. During this time, the number of alarms that were triggered was 123 (3.8%). 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.

3.8.4.4 Clinical Evaluation

The median scores given by each clinician are presented in figure 3.23. We also predicted the result using our model as stated in figure 2. 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 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

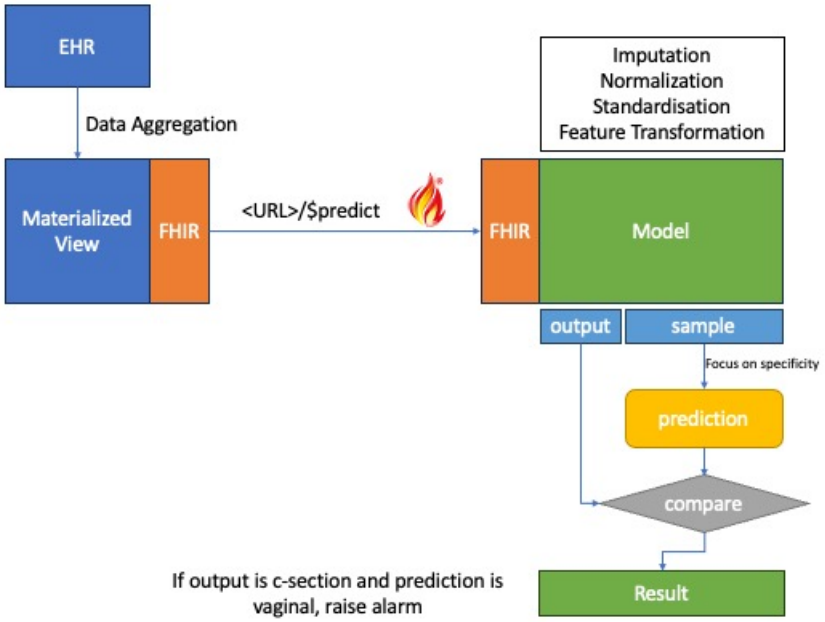
Table 3.16: Performance Metrics in the training set with mean AUROC and 95% Confidence Interval

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

Table 3.17: Performance Metrics in the test set with chosen threshold

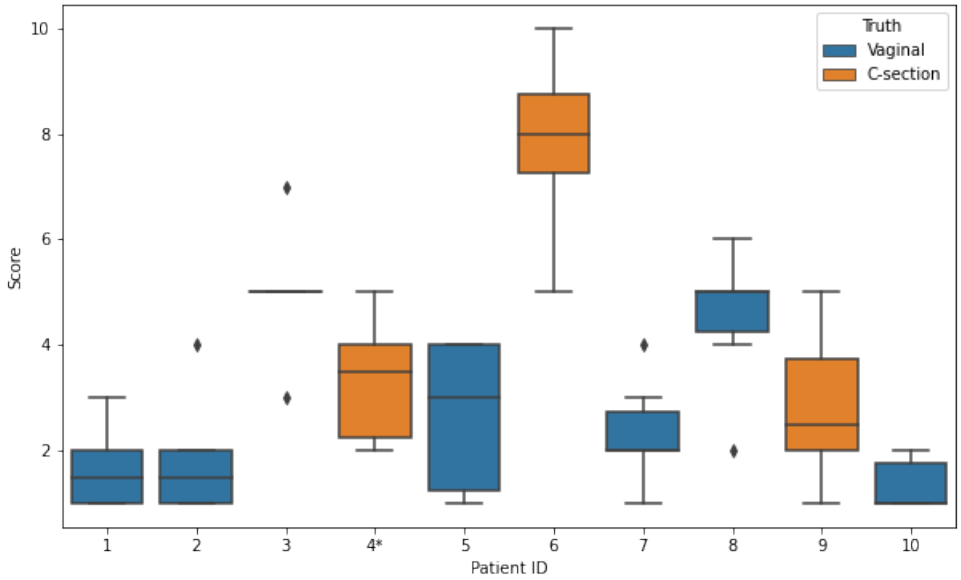
Metric	Value
Accuracy	0.8052
Sensitivity	0.8223
Precision	0.9023
F1 Score	0.8605

Figure 3.22: Deployment and decision mechanism of the model



that moment. No feature mentioned existed 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 labor 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.

Figure 3.23: 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.



3.8.4.5 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. 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 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 3.18: Ruleset for state-provided financial support indexed to C-Sections. X is the current payment of a C-Section inpatient episode. Adapted from [240]

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

3.8.5 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 labeled 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 labeling 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 issue 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.

3.8.6 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

4 Discussion

Extracting knowledge from healthcare data is not easy. It relies on the availability of data, which is not always the case, and on the ability to extract knowledge from it. In this chapter, we discuss the main challenges we faced during the development of this thesis, and how we overcame them. We also discuss the limitations of our work, and how it can be improved in the future. Finally, we discuss the main contributions of this thesis, and how they can be used to improve the quality of healthcare. 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. I 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 I 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.

This issue also 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. 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. Thirdly, 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 health data science 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 3.7 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 premisses

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.

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 real-world evidence (RWE) into clinical decision-making processes. RWE, derived from data collected outside of 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 health data science 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.

Lastly, the future of healthcare data science depends heavily on cross-disciplinary collaboration and common frameworks. Initiatives like the European Health Data Space (EHDS) are steps in the right direction, promoting data availability and quality through collaborative efforts. However, the success of such initiatives relies on the strength of their weakest links, necessitating uniform standards, shared goals, and adequate resources across all participating entities. Concrete measures like establishing common data dictionaries, data formats, and interoperability tools are essential. These efforts will pave the way for more efficient data utilization, driving innovation in healthcare. Such an integrated approach, combining technical prowess with legal and ethical considerations, is vital for realizing the full potential of healthcare data in improving patient outcomes and advancing medical science.

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

Niels Bohr

5 Limitations, future work and conclusions

Regarding limitations, all the projects done in this thesis focus on different aspects of the process of extracting knowledge from healthcare data. Also, they are heavily reliant on specific use cases that are not necessarily generalizable. For example, the 3.7 project is focused on a specific disease, and the 3.8 and 3.5 projects are focused on a specific type of data, and the 3.4 project is focused on a specific type of data and a specific clinical specialty. This means that the results of these projects are not necessarily generalizable to other diseases or other types of data. However, the methods used in these projects are generalizable, and they can be used in other projects. For example, the 3.7 methods can be used to predict in real-time, and the 3.8 and 3.5 models can be used to analyse other types of data. The 3.3 method can be used to analyse any type of dataset and incorporated into data pipelines.

In future work, I think the groundwork is laid for actually providing assistance to healthcare teams. However, actually deploying real-world CDSSs is seldom an easy task and requires time, money and patience. This is why that part, the actual deployment of the tools, is left for future work. However, we did many tests in the real world and included clinicians in most of our work, so we are confident that the tools are ready to be deployed and create an impact.

For this work to be complete, I had to gather knowledge from different areas. From biology and chemistry, for the healthcare part of it, process design to understand and formalize processes, and math and statistics for machine learning and EDA, interoperability and standards for getting data together, ethics and privacy to gather data with guarantees to the patient's privacy and for creating ethical-aware models. Had to dwell into terminologies and healthcare codification and semantics to interpret data and also get acquainted with some clinical specialties like obstetrics and oncology. Had to collect evidence and make the bridge between RCTs and observational data

or Real World Data (RWD) so study design was also needed to bridge the gap. Maybe this is one of the main issues with this domain, where a different set of skills is required to do everything. The alternative would be a team of different people and honestly, the most successful projects I have seen are the ones that have a team of different people. Finally, the

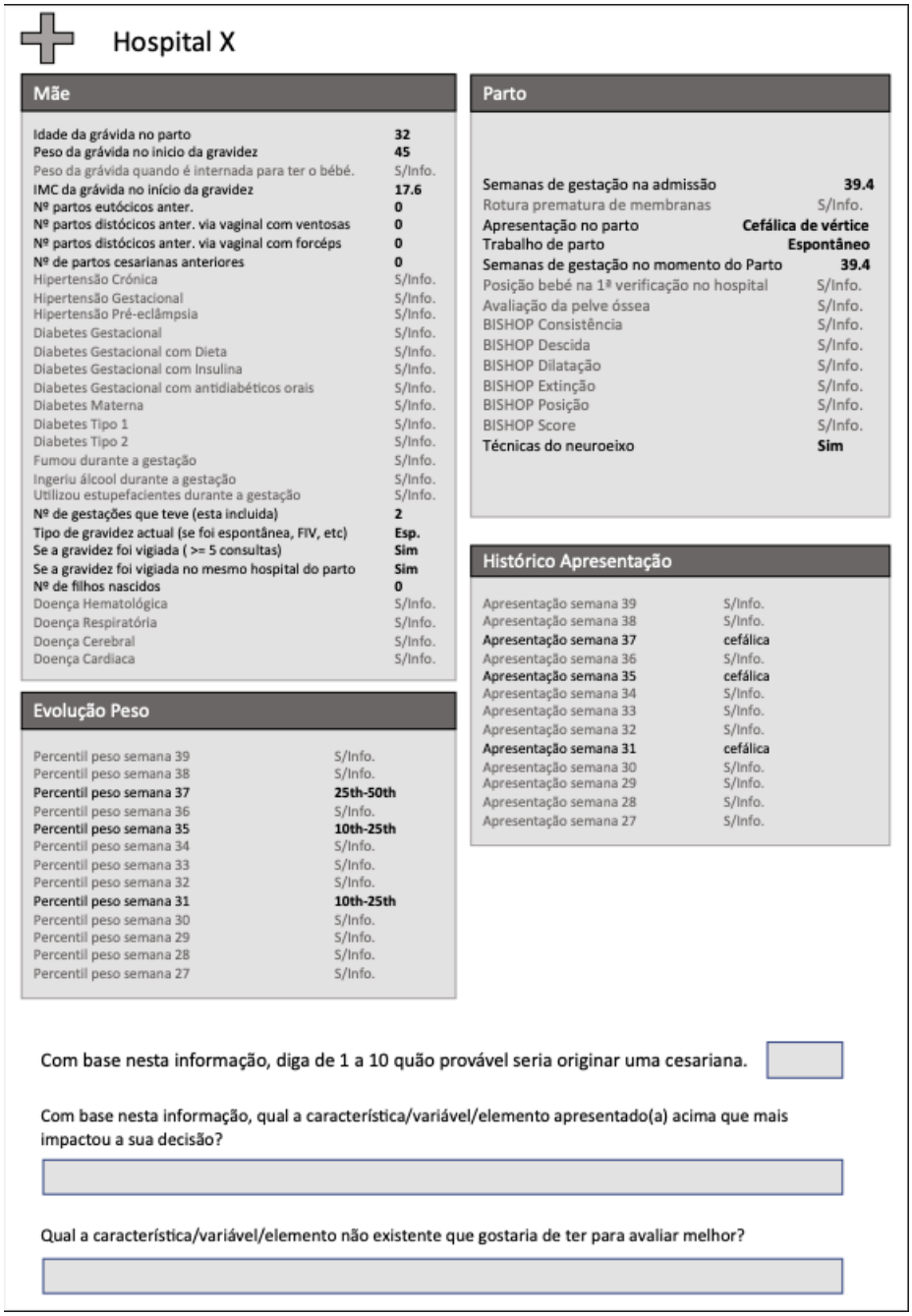
A

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 inicio 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 incluida)	**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.

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**

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.

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 nrº 1

Mãe

Idade da grávida no parto	37.0
Grupo sanguíneo da grávida	0,RH_POSITIVO
Peso da grávida no inicio 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.

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). http://www.healthaffairs.org/doi/10.1377/hlthaff.2016.1651, 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). http://link.springer.com/10.1007/s10916-018-1075-6, 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). https://www.tandfonline.com/doi/full/10.3402/jchimp.v6.32643, 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. URL: https://ieeexplore.ieee.org/document/9248660/.

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

[6] 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). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6722281/, p. 100443. ISSN: 2451-8654. DOI: 10.1016/j.conctc.2019.100443.

[7] 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*. https://ebooks.iospress.nl/doi/10.3233/SHTI200597. IOS Press, 2020, pp. 470–473. DOI: 10.3233/SHTI200597. (Visited on 08/04/2023).

[8] 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.

[9] 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. URL: https://www.nature.com/articles/s41591-018-0300-7.

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

[11] 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. URL: https://doi.org/10.1001/jamanetworkopen.2023.37789.

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

[13] 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. URL: https://doi.org/10.1186/s40537-019-0217-0.

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

[15] 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).

[16] European Commission. *A Definition of AI: Main Capabilities and Disciplines*. Tech. rep. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. European Commission, 2019.

[17] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2349778/.

[18] 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. URL: https://doi.org/10.1093/asj/sjv130.

[19] 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.

[20] "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).

[21] 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.

[22] 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.

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

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

[25] 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.

[26] 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.

[27] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215048/ (visited on 10/14/2023).

[28] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3555312/.

[29] 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.

[30] 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.

[31] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1 (2020-12-25), p. 18. ISSN: 1099-4300. DOI: 10.3390/e23010018. URL: https://www.mdpi.com/1099-4300/23/1/18.

[32] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI". In: *Information Fusion* 58 (2020-06), pp. 82–115. ISSN: 15662535. DOI: 10.1016/j.inffus.2019.12.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S1566253519308103.

[33] U. Kamath and J. Liu. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Springer International Publishing, 2021. ISBN: 978-3-030-83355-8. URL: https://books.google.pt/books?id=_O2LzgEACAAJ.

[34] 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 2019-05), pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x. URL: https://www.nature.com/articles/s42256-019-0048-x.

[35] 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.

[36] 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.

[37] Judea Pearl. "Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution". In: (Jan. 11, 2018). arXiv: 1801.04016 [cs, stat]. URL: http://arxiv.org/abs/1801.04016. preprint.

[38] 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. URL: https://www.frontiersin.org/articles/10.3389/fmed.2022.864882.

[39] 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.

[40] 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.

[41] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.

[42] 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.

[43] 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.

[44] European Commission. *Proposal for a Regulation of the european Parliament and of the council on the European Health Data Space*. https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd897

4-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF. 2022. (Visited on 08/21/2023).

[45] European Commission. *Ethics Guidelines For Trustworthy AI*. Tech. rep. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. European Commission, 2019.

[46] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. http://www.fairmlbook.org. 2019.

[47] 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.

[48] 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: %5Curl%7Bhttps://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php%7D.

[49] 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.

[50] Comissão Nacional Proteção de dados. *Princípios aplicáveis aos tratamentos de dados efetuados no âmbito da investigação clínica*. 2015. (Visited on 01/26/2021).

[51] 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).

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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/CIRCOUTCOMES.118.005122. URL: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.

[58] 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.

[59] 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.

[60] 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.

[61] Edward Choi. *medGAN Repository*. URL: %5Curl%7Bhttps://github.com/mp2893/medgan%7D (visited on 01/21/2022).

[62] 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.

[63] POSTER. *POSTER Repository*. URL: %5Curl%7Bhttps://goo.gl/94qyQz%7D (visited on 01/21/2022).

[64] 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.

[65] Mahmoud Mohammadi. *table-GAN Repository*. URL: %5Curl%7Bhttps://github.com/mahmoodm2/tableGAN%7D (visited on 01/21/2022).

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

[67] Illidan Lab. *dp-GAN Repository*. URL: %5Curl%7Bhttps://github.com/illidanlab/dpgan%7D (visited on 01/21/2022).

[68] 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.

[69] Ramiro Camino. *mc-medGAN Repository*. URL: %5Curl%7Bhttps://github.com/rcamino/multi-categorical-gans%7D (visited on 01/21/2022).

[70] 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.

[71] The Synthetic Data Vault Project. *TGAN Repository*. URL: %5Curl%7Bhttps://github.com/sdv-dev/TGAN%7D (visited on 01/21/2022).

[72] 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.

[73] Brett Beaulieu-Jones, Casey Greene, and Steven Wu. *SPRINT-GAN Repository*. URL: %5Curl%7Bhttps://github.com/greenelab/SPRINT_gan%7D (visited on 01/21/2022).

[74] 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.

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

[76] The Synthetic Data Vault Project. *CTGAN Repository*. URL: %5Curl%7Bhttps://github.com/sdv-dev/CTGAN%7D (visited on 01/21/2022).

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

[78] Bauke Brenninkmeijer. *WGAN-DP Repository*. URL: %5Curl%7Bhttps://github.com/Baukebrenninkmeijer/On-the-Generation-and-Evaluation-of-Synthetic-Tabular-Data-using-GANs%7D (visited on 01/21/2022).

[79] 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.

[80] Yi Liu. *PPGAN Repository*. URL: %5Curl%7Bhttps://github.com/niklausliu/PPGANs-Privacy-preserving-GANs%7D (visited on 01/21/2022).

[81] 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.

[82] Mrinal Kanti Baowaly. *medWGAN Repository*. URL: %5Curl%7Bhttps://github.com/baowaly/SynthEHR%7D (visited on 01/21/2022).

[83] 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.

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

[85] Amirsina Torfi. *corGAN Repository*. URL: %5Curl%7Bhttps://github.com/astorfi/cor-gan%7D (visited on 01/21/2022).

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

[87] DPautoGAN. *DPAutoGAN Repository*. URL: %5Curl%7Bhttps://github.com/DPautoGAN/DPautoGAN%7D (visited on 01/21/2022).

[88] Marcel Neunhoeffer, Zhiwei Steven Wu, and Cynthia Dwork. *Private Post-GAN Boosting*. 2020.

[89] Marcel Neunhoeffer. *Post-GAN Boosting Repository*. URL: %5Curl%7Bhttps://github.com/mneunhoe/post-gan-boosting%7D (visited on 01/21/2022).

[90] 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.

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

[92] 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.

[93] 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.

[94] Anurag Dutt. *SMOOTH-GAN Repository*. URL: %5Curl%7Bhttps://github.com/anuragdutt/synthehr%5C_medgan%7D (visited on 01/21/2022).

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

[96] 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.

[97] 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.

[98] healthdatainsight.org.uk. *The Simulacrum*. healthdatainsight.org.uk. URL: %5Curl%7Bhttps://healthdatainsight.org.uk/project/the-simulacrum/%7D (visited on 01/21/2022).

[99] integraal kankercentrum Nederland. *Synthetische dataset NKR beschikbaar voor onderzoekers*. URL: %5Curl%7Bhttps://iknl.nl/nieuws/2021/synthetische-data-nkr-beschikbaar-voor-onderzoeker%7D (visited on 01/21/2022).

[100] 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).

[101] 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.

[102] 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).

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

[104] 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.

[105] Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. *Heart Disease*. DOI: 10.24432/C52P4X. 1988. URL: UCI%20Machine%20Learning%20Repository.

[106] 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.

[107] Travers Ching et al. "Opportunities and Obstacles for Deep Learning in Biology and Medicine". In: *Journal of The Royal Society Interface* 15.141 (2018-04-04), p. 20170387. DOI: 10.1098/rsif.2017.0387. URL: https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0387.

[108] Emily Muller, Xu Zheng, and Jer Hayes. "Evaluation of the Synthetic Electronic Health Records". In: (2022-10-16). arXiv: 2210.08655 [cs]. URL: http://arxiv.org/abs/2210.08655 (visited on 03/05/2023).

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

[110] 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.

[111] 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. URL: https://medinform.jmir.org/2022/4/e35734.

[112] 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: (2022-10-16). arXiv: 2210.08528 [cs]. URL: http://arxiv.org/abs/2210.08528 (visited on 03/05/2023).

[113] 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: 1703.06490. URL: http://arxiv.org/abs/1703.06490.

[114] 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.

[115] Paul R. Rosenbaum and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In: *Biometrika* 70.1 (1983-04-01), pp. 41–55. ISSN: 0006-3444. DOI: 10.1093/biomet/70.1.41. URL: https://doi.org/10.1093/biomet/70.1.41.

[116] Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, Mukund Lahoti, and Pratik Narang. "TabSynDex: A Universal Metric for Robust Evaluation of Synthetic Tabular Data". In: (2022-07-12). DOI: 10.48550/arXiv.2207.05295. arXiv: 2207.05295 [cs]. URL: http://arxiv.org/abs/2207.05295 (visited on 02/03/2023).

[117] 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.nips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html (visited on 03/05/2023).

[118] 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]. (Visited on 01/21/2023).

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

[120] 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: 21006841.

[121] 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, 2015-05-18, 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).

[122] William Webber, Alistair Moffat, and Justin Zobel. "A Similarity Measure for Indefinite Rankings". In: *ACM Transactions on Information Systems* 28.4 (2010-11-23), 20:1–20:38. ISSN: 1046-8188. DOI: 10.1145/1852102.1852106. URL: https://doi.org/10.1145/1852102.1852106.

[123] Gonzalo Navarro. "A Guided Tour to Approximate String Matching". In: *ACM Computing Surveys* 33.1 (2001-03), pp. 31–88. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/375360.375365. URL: https://dl.acm.org/doi/10.1145/375360.375365.

[124] 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.

[125] orsinium. *Textdistance: Compute Distance between the Two Texts.* Version 4.5.0. URL: %5Curl%7Bhttps://github.com/orsinium/textdistance%7D (visited on 03/05/2023).

[126] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17.3 (3 2020-03), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2. URL: https://www.nature.com/articles/s41592-019-0686-2.

[127] Changyao Chen. *Rank-Biased Overlap (RBO)*. 2018-02-11. URL: %5Curl%7Bhttps://github.com/changyaochen/rbo%7D (visited on 03/16/2023).

[128] Ross Quinlan. *Thyroid Disease*. 1987. URL: UCI%20Machine%20Learning%20Repository.

[129] Richard S. Forsyth. *Liver Disorders*. 1990. URL: UCI%20Machine%20Learning%20Repository.

[130] William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. *Breast Cancer Wisconsin (Diagnostic)*. 1995. URL: UCI%20Machine%20Learning%20Repository.

[131] M. Zwitter and M. Soklic. *Primary Tumor*. 1988. URL: UCI%20Machine%20Learning%20Repository.

[132] 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.

[133] 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.

[134] 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.

[135] 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.

[136] 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).

[137] 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.

[138] 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/.

[139] 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.

[140] 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.

[141] 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.

[142] 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.

[143] 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.

[144] 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).

[145] 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.

[146] 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/.

[147] 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.

[148] 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.

[149] 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.

[150] 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.2

018.12.029. URL: https://www.sciencedirect.com/science/article/pii/S0169260718304188.

[151] 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.

[152] 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. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/lrh2.10264 (visited on 02/21/2022).

[153] 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.

[154] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8387126/.

[155] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10111961/.

[156] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7391413/.

[157] 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. URL: https://ebooks.iospress.nl/doi/10.3233/978-1-61499-753-5-539 (visited on 11/19/2023).

[158] 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. URL: https://www.sciencedirect.com/science/article/pii/S0010482516000111.

[159] 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. URL: https://www.sciencedirect.com/science/article/pii/S0010482515003340 (visited on 11/19/2023).

[160] 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.

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

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

[163] 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.

[164] *GX: A Proactive, Collaborative Data Quality Platform*. URL: %5Curl%7Bhttps://www.greatexpectations.io/%7D (visited on 11/19/2023).

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

[166] 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 (2017-01), pp. 4–21. ISSN: 2168-2208. DOI: 10.1109/JBHI.2016.2636665.

[167] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5962261/.

[168] 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: %5Curl%7Bhttps://edpl.lexxion.eu/article/EDPL/2016/3/4%7D.

[169] 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).

[170] 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.

[171] 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.

[172] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. "Machine Learning in Medicine". In: *New England Journal of Medicine* (2019-04-03). URL: https://www.nejm.org/doi/10.1056/NEJMra1814259.

[173] 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* (2020-11-12), pp. 1–19. ISSN: 2509-4971. DOI: 10/gjt8ds. pmid: 33204939. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7659898/.

[174] Yan Cheng Yang, Saad Ul Islam, Asra Noor, Sadia Khan, Waseem Afsar, and Shah Nazir. "Influential Usage of Big Data and Artificial Intelligence in Healthcare". In: *Computational and Mathematical Methods in Medicine* 2021 (2021-09-06), p. 5812499. ISSN: 1748-670X. DOI: 10/gm2dpf. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8437645/.

[175] Fei Wang and Anita Preininger. "AI in Health: State of the Art, Challenges, and Future Directions". In: *Yearbook of Medical Informatics* 28.1 (2019-08), pp. 16–26. ISSN: 2364-0502. DOI: 10/gjt8d3.

[176] 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* (2021-06-05). ISSN: 1319-1578. DOI: 10/gkd5f4. URL: https://www.sciencedirect.com/science/article/pii/S1319157821001312.

[177] 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. URL: https://www.sciencedirect.com/science/article/pii/S1532046420300526.

[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* 5.1 (2021), pp. 1–19. ISSN: 2509-4971. DOI: 10.1007/s41666-020-00082-4. pmid: 33204939.

[179] Geun Hyeong Lee and Soo-Yong Shin. "Federated Learning on Clinical Benchmark Data: Performance Assessment". In: *Journal of Medical Internet Research* 22.10 (2020-10-26). ISSN: 1439-4456. DOI: 10/gjt8dr. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7652692/.

[180] 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 2021-01), p. 11191. ISSN: 2076-3417. DOI: 10.3390/app112311191. URL: https://www.mdpi.com/2076-3417/11/23/11191.

[181] 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* (2022-10-26), pp. 1–55. ISSN: 0269-2821. DOI: 10.1007/s10462-022-10304-3. pmid: 36320613. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9607788/.

[182] 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 (2020-05-17), pp. 1173–1185. ISSN: 1067-5027. DOI: 10.1093/jamia/ocaa053. pmid: 32417928. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647281/.

[183] 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* (2022), pp. 1–21. ISSN: 2162-237X, 2162-2388. DOI: 10.1109/TNNLS.2022.3229161. arXiv: 2110.01889 [cs]. URL: http://arxiv.org/abs/2110.01889.

[184] 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/gc8827. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5833935/.

[185] 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.

[186] 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. URL: https://linkinghub.elsevier.com/retrieve/pii/S1877050920300478.

[187] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7652692/.

[188] 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.

[189] 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].

[190] 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].

[191] Steven Suydam, Bryan A. Liang, Storm Anderson, and Matthew B. Weinger. "Patient Safety Data Sharing and Protection From Legal Discovery". In: *Journal of Medical Regulation* 93.2 (2007-06-01), pp. 19–25. ISSN: 2572-1852, 2572-1801. DOI: 10.30770/2572-1852-93.2.19. URL: https://meridian.allenpress.com/jmr/article/93/2/19/453914/Patient-Safety-Data-Sharing-and-Protection-From.

[192] Tim Hulsen. "Sharing Is Caring—Data Sharing Initiatives in Healthcare". In: *International Journal of Environmental Research and Public Health* 17.9 (9 2020-01), p. 3046. ISSN: 1660-4601. DOI: 10.3390/ijerph17093046. URL: https://www.mdpi.com/1660-4601/17/9/3046.

[193] Angela G. Villanueva, Robert Cook-Deegan, Barbara A. Koenig, Patricia A. Deverka, Erika Versalovic, Amy L. McGuire, and Mary A. Majumder. "Characterizing the Biomedical Data-Sharing Landscape". In: *The Journal of law, medicine & ethics : a journal of the American Society of Law, Medicine & Ethics* 47.1 (2019-03), pp. 21–30. ISSN: 1073-1105. DOI: 10.1177/1073110519840481. pmid: 30994069. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6730647/.

[194] Pedro Pereira Rodrigues, João Araújo, João Gama, and Luís Lopes. "A Local Algorithm to Approximate the Global Clustering of Streams Generated in Ubiquitous Sensor Networks". In: *International Journal of Distributed Sensor Networks* 14.10 (2018-10), p. 155014771880823. ISSN: 1550-1477, 1550-1477. DOI: 10.1177/1550147718808239. URL: http://journals.sagepub.com/doi/10.1177/1550147718808239.

[195] Abigail Walker and Pavol Surda. "Unsupervised Learning Techniques for the Investigation of Chronic Rhinosinusitis". In: *The Annals of Otology, Rhinology, and Laryngology* 128.12 (2019-12), pp. 1170–1176. ISSN: 1943-572X. DOI: 10.1177/0003489419863822. pmid: 31319675.

[196] Anna Okula Basile and Marylyn DeRiggi Ritchie. "Informatics and Machine Learning to Define the Phenotype". In: *Expert Review of Molecular Diagnostics* 18.3 (2018-03-04), pp. 219–226. ISSN: 1473-7159. DOI: 10.1080/14737159.2018.1439380. pmid: 29431517. URL: https://doi.org/10.1080/14737159.2018.1439380.

[197] Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman, and Riccardo Bellazzi. "Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools". In: *Frontiers in Oncology* 10 (2020). ISSN: 2234-943X. URL: https://www.frontiersin.org/articles/10.3389/fonc.2020.01030.

[198] Nimrod Rappoport and Ron Shamir. "Multi-Omic and Multi-View Clustering Algorithms: Review and Cancer Benchmark". In: *Nucleic Acids Research* 46.20 (2018-11-16), pp. 10546–10562. ISSN: 1362-4962. DOI: 10.1093/nar/gky889. pmid: 30295871.

[199] 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 (2022-01-06), pp. 5–14. ISSN: 1943-7722. DOI: 10.1093/ajcp/aqab085. pmid: 34302331.

[200] S. Lloyd. "Least Squares Quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982-03), pp. 129–137. ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489. URL: http://ieeexplore.ieee.org/document/1056489/.

[201] 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.

[202] 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.

[203] Lawrence Hubert and Phipps Arabie. "Comparing Partitions". In: *Journal of Classification* 2.1 (1985-12-01), pp. 193–218. ISSN: 1432-1343. DOI: 10.1007/BF01908075. URL: https://doi.org/10.1007/BF01908075.

[204] Nelis J. de Vos. *kmodes categorical clustering library*. https://github.com/nicodv/kmodes. 2015–2021. URL: %5Curl%7Bhttps://github.com/nicodv/kmodes%7D.

[205] 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.

[206] 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.

[207] 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.

[208] 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.

[209] 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.

[210] 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.

[211] 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.

[212] 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.

[213] 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. URL: https://www.futuremedicine.com/doi/full/10.2217/fon-2020-1264.

[214] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.

[215] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4285179/.

[216] Noah Greifer. *WeightIt: Weighting for Covariate Balance in Observational Studies*. 2023. URL: %5Curl%7Bhttps://ngreifer.github.io/WeightIt/%7D.

[217] 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.

[218] 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 (2016-02-05), e0148343. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0148343. pmid: 26849801. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4743929/.

[219] Innie Chen et al. "Non-clinical Interventions for Reducing Unnecessary Caesarean Section". In: *The Cochrane Database of Systematic Reviews* 2018.9 (2018-09-28), p. CD005528. ISSN: 1469-493X. DOI: 10.1002/14651858.CD005528.pub3. pmid: 30264405. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6513634/.

[220] 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 (2013-01), pp. 85–91. ISSN: 1471-0528. DOI: 10.1111/1471-0528.12010. pmid: 23095012.

[221] 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.

[222] 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.

[223] 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 (2012-07), pp. 14–29. ISSN: 1097-6868. DOI: 10.1016/j.ajog.2012.03.007. pmid: 22516620.

[224] 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.

[225] Pordata. *Cesarianas Nos Hospitais (%)*. https://www.pordata.pt/Portugal/Cesarianas+nos+hospitais+(percentagem)-1985. URL: %5Curl%7Bhttps://www.pordata.pt/Portugal/Cesarianas+nos+hospitais+(percentagem)-1985%7D (visited on 05/16/2022).

[226] 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.

[227] 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.

[228] 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. URL: http://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-017-0378-z.

[229] 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. URL: https://www.frontiersin.org/article/10.3389/fphys.2019.00246/full.

[230] 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.

[231] 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* n/a (n/a 2021). _eprint: https://onlinelibrary.wiley.c 6773.13921. ISSN: 1475-6773. DOI: 10.1111/1475-6773.13921. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.13921.

[232] 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).

[233] 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).

[234] 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: 10.1080/14767058.2020.1837769. pmid: 33103511.

[235] 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)30945-9/fulltext (visited on 02/16/2023).

[236] 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).

[237] 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.sciencedirect.com/science/article/pii/S0169260722001262 (visited on 09/18/2023).

[238] 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/articles/PMC7931279/ (visited on 02/16/2023).

[239] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017), pp. 3146–3154.

[240] ACSS. *Termos Referência Para Contratualização de Cuidados de Saúde No SNS Para 2023*. 2023. URL: %5Curl%7Bhttps://www.acss.min-saude.pt/wp-content/uploads/2016/10/Termos-Referencia-Contratualizacao%5C_2023.pdf%7D (visited on 06/20/2023).