Synthetic data generation for research – enabler for privacy-enhancing health data sharing?

A multidisciplinary research on synthetic data generation in healthcare

Master Thesis

Delft University of Technology

MSc Complex Systems Engineering and Management

April 8, 2024

Author
Iris van der Wel 4876881

Graduation committee

Chairperson and first supervisor:	Prof.dr.ir. G.A. (Mark) de Reuver
Second supervisor:	Dr. S. (Saba) Hinrichs-Krapels
Advisor:	Dr. J.H.F. (Jacobien) Oosterhoff

Executive Summary

Sharing personal health data can boost scientific research, in terms of increased comprehension of diseases, diagnostics, treatments, and predictive capabilities. But data-driven health research, specifically the development of AI models, is hampered by poor health data availability. Sharing health data for research purposes can therefore help, but the complex and fragmented data protection regulations in the European Union (EU) have resulted in lengthy procedures for researchers wishing to access health data.

The use of synthetic health data is seen as a promising technical solution to enable the benefits of using high-quality health data, while reducing privacy risks for patients. Synthetic data generation is the process of applying a mathematical model or algorithm to a dataset with personal data to generate a dataset with synthetic data. This output represents the statistical characteristics of the original dataset, while disclosing minimal information about patients, thus safeguarding data protection. Synthetic data generation is presented as the better alternative for data anonymisation, as it limits identifying the original patients while still capturing the high dimensionality of health datasets, increasing its utility for research.

Current research primarily focuses on a technical development and implementation of data protection, leaving substantive matters on the application of synthetic data within its context open. By combining an institutional and technical approach, this thesis aims to fill this gap, providing insight in how synthetic data generation can enable health data sharing for research purposes. It uses a design science research approach to combine academic knowledge on synthetic data and institutional analysis, with the (largely unknown) information about health data sharing practices, to design a framework that organises the factors that inhibit or enable synthetic data sharing in a privacy-enhancing manner. The research question is formulated as follows:

> *How could synthetic data generation enable health data sharing for research in a privacy-enhancing manner?*

To answer the research question, the institutional environment of health data sharing practice was mapped; the concepts of synthetic data generation, privacy risks, and evaluation thereof were analysed; technical and practical knowledge were studied to identify how synthetic data relates to the institutional environment which is structured into a framework with different phases of synthetic data sharing. To concretise the research, a use case of health data sharing between a healthcare provider and research institute was defined.

The institutional analysis of health data sharing showed that the data protection environment is complex and multifaceted, involving various actors at multiple levels of governance. Relevant actors include the healthcare provider as data collectors, research institute as data users, and the Ministry of Healthcare, Welfare, and Sports (HWS) as policymakers. The institutional landscape is composed of formal regulations on EU and Dutch level and is characterised by uncertainties regarding the legal definition of anonymisation, a contradiction between the EU and Dutch interpretation of data protection principles, and time- and labour-intensive interactions to conclude data sharing agreements. The challenges occurred at various levels. For example, the interaction patterns regarding use of consent and definition of anonymisation are decided at the national and EU level, while data sharing procedures emerge at an interorganisational level.

To understand how synthetic data may (not) solve these issues, the technicalities of synthetic data generation and data protection were examined next. Even though the great potential of synthetic data generation, it is important to acknowledge that some privacy-preserving models are vulnerable to re-identification in practice. For instance, researchers have identified instances where synthetic datasets unintentionally disclosed patients' identities or specific attributes when compared to the original dataset, demonstrating the need for formal privacy evaluations. This result conflicts with the absence of such evaluations, which could lead to unintentional information disclosure. Moreover, researchers use various metrics due to the lack of agreement on how privacy can best be evaluated. Lastly, researchers rarely

interpret the quantitative values of their privacy evaluations. Concerted efforts aimed at setting standards for the privacy-preserving generation of synthetic data are necessary to ensure data protection within research environments.

To answer the research question, I propose a framework with data protection-related barriers, drivers and solution directions to conclude what enables or inhibits synthetic health data sharing. Synthetic data can enable health data sharing by embodying the principles of data protection law, easing the current health data sharing process, and increasing research opportunities.

An important legal barrier concerns the very broad definition of personal data in EU data protection law makes it difficult to determine whether synthetic health data qualifies as personal or anonymous data. On top of that, the definition of personal data makes that the anonymisation is difficult to operationalise; whether certain data can be considered anonymous depends on the context. This makes it difficult for researchers, healthcare providers and technologies providers to know which standards must be met for synthetic health data to classify data as anonymous. In practice, this results in uncertainties regarding the applicable legal rules in health data sharing, which significantly inhibits synthetic health data sharing. Combined with issues related to interpretability of quantitative privacy metrics of synthetic health data, healthcare providers must resort to a case-by-case assessment of privacy risks that their data sharing activities cause, in order to share health data in a privacy-enhancing manner. This stresses the need for multidisciplinary collaborations for such risk assessments.

A second barrier arises regarding the Dutch approach towards the need for consent to share health data for research. There is a friction between the Dutch 'consent-by-default' approach, the presumption of the GDPR's that processing personal data for scientific research is in line with the original purpose for which the data was collected; and the presumption of such compatibility when data is anonymised. Due to this friction, it is uncertain whether the law allows for the generation of synthetic health data for research or requires a separate legal base in the GDPR.

To this end, one of the policy opportunities I specifically deem important, is for Dutch governmental bodies and EU bodies to clarify concepts on the intersection of data protection law, synthetic data, and anonymisation. The current approach towards the definition of personal data is not fit to the technological advances of our society.

These fundamental discussions are political in nature, so how can actors involved in health data sharing work around these uncertainties, caused by legal ambiguities and a lack of interpretability? This thesis proposes the following process:

The application of synthetic data should start with an exploration of the intended synthetic data use, as this affects the legal requirements and privacy risks of the context, executed by healthcare providers and research institutes. Second, because the real-world dataset that is synthesised needs to suit the intended use, healthcare providers should analyse privacy risks of the original dataset and this data should be curated to fit the generation method. These two steps are important to analyse the requirements that arise from legislation, privacy risks, and analytical intentions. Third and similarly, healthcare providers should select a generation model that suits the intended use and sufficiently preserve privacy. Some models allow for an explicit utility-privacy trade-off in the model parameters. The model parameters should then be tweaked to the intended use in terms of analytical requirements and privacy requirements. In the fourth phase, synthetic data should be evaluated for remaining privacy risks, to prevent unintentional information disclosure. It is vital that technology providers aid synthetic healthcare providers in this evaluation, by providing them with appropriate and understandable privacy metrics to quantitatively assess privacy risks. This is necessary because data users might not be familiar with the intricacies that the synthetic data generation technologies imply. Healthcare providers are responsible for performing an institutional privacy evaluation, consisting of an interpretation of the privacy metrics, the intended use and their earlier analysis of real-world data. The technology provider, however, thus has an important role to support them in this task.

List of abbreviations

Abbreviation	Meaning
AD	Attribute disclosure
AI	Artificial intelligence
CFREU	Charter of fundamental rights of the European Union
CJEU	Court of Justice of the European Union
DP	Differential privacy
DPA	Data Protection Authority
DPIA	Data Protection Impact Assessment
DSR	Design Science Research
ECHR	European Convention of Human Rights
ECtHR	European Court of Human Rights
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EHDS	European Health Data Space
EHR	Electronic health record
EMA	European Medicines Agency
EU	European Union
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act (HIPAA)
IAD framework	Institutional analysis and development framework
IC	Information and communication
IEEE	Institute of Electrical and Electronics Engineering
LLM	Large Language Model
MI	Membership Inference
Ministry of HWS	Dutch Ministry of Health, Welfare and Sports ('Ministerie van Volksgezondheid, Welzijn en Sport')
ML	Machine learning
NN	Nearest Neighbour
RQ	Research question
UAVG	Uitvoervoeringswet Algemene Verordening Gegevensbescherming (Dutch GDPR implementation act)
USA	United States of America
VAE	Variational autoencoders
WGBO	Wet voor geneeskundige behandelingsovereenkomst (Dutch act on medical treatment contracts)
WP29	Working Party on the Protection of Individuals with regard to the Processing of Personal Data

Table of contents

1 Introduction

1.1 Problem statement

The secondary use of health data for research offers opportunities to improve the healthcare sector (Safran et al., 2007). Secondary use refers to the use of health data for other purposes than its initial collection, such as delivering care to patients (Safran et al., 2007); this thesis specifically focuses on secondary use by external researchers, referred to as 'health data sharing'. Sharing health data can catalyse scientific research by advancing our comprehension of diseases, diagnostics, treatments, and predictive capabilities (Hendolin, 2022; Iacob & Simonelli, 2020). For example, artificial intelligence (AI) based research in healthcare presents opportunities to improve prevention of diseases, accuracy of diagnoses, effectiveness of treatments, personalised medicines, and efficient healthcare management (Davenport & Kalakota, 2019).

However, data-driven research such as the development of AI models requires large and representative datasets to develop, train and test models (Chen et al., 2021). In medical research, this amounts to the use of patients' personal data (Commission, 2020; Kokosi & Harron, 2022), such as the patient's physical and mental health status, medical assessments, and the patient's use of healthcare services (Hordern, 2022).[1] Exposure of such sensitive information can have severe ethical and social consequences for individuals, such as discrimination and stigmatisation (Marks, 2019), and may raise legal issues for the responsible organisation (Article 29 Data protection working party, 2011; F. Li et al., 2011). Hence, health data cannot be shared unconditionally.

Data-driven health research is hampered by the poor data availability and associated administrative burdens; problems that are rooted in the EU and national laws for data protection and a lack of technical infrastructures or privacy-preserving technologies (European Commission, 2020; Hendolin, 2022).[2] For example, the development of complex models often requires large and multidimensional health datasets that can be realised through multi-institutional collaborations (Rajendran et al., 2021). However, to share data across institutions, personal data cannot always be 'simply' anonymised: intensive anonymisation, such as data aggregation, can result in a deterioration of the data quality to the point where it becomes unsuitable. (Abay et al., 2019; Kokosi & Harron, 2022).

Hence, we want to harvest the benefits of sharing and using high-quality health data, while safeguarding data protection of patients (Kokosi & Harron, 2022). To this end, health research needs solutions to cope with the legal and technical complexities, of which the use of synthetic data is deemed a promising method (Kokosi & Harron, 2022; Murtaza et al., 2023). Synthetic data is generally referred to artificially or algorithmically generated data that can

[1] This thesis uses the terms 'personal health data', 'health data' and 'patient data' interchangeably. These terms refer to *personal* data, unless mentioned otherwise.

[2] This thesis uses the terms 'privacy' and 'data protection' interchangeably. On the one hand, these terms convey different notions. In international treaties, the right to data protection focuses on measures to hide or safeguard personal data. The right to privacy has a broader scope than data protection, encompassing one's private life, home and communications, determining who can have access to which data. Therefore, data protection can be considered a subset of privacy (Charter of Fundamental Rights of the European Union, 2000, arts. 7–8; Hildebrandt, 2019). On the other hand, researchers in computer science community and also the overwhelming majority of literature cited in this thesis mainly refer to privacy as disclosure of personal information, aligning with the legal concept of data protection.

represent the statistical characteristics of the real-world dataset, while preserving patients' privacy (Rajotte et al., 2022; Tsao et al., 2023). Synthetic data has the potential to increase health data sharing between organisations in a way that reduces privacy risks, by reducing the need to share personal data (Murtaza et al., 2023). The next section will further clarify the concept 'synthetic data'.

However, this area of research is still developing; in its current forms, synthetic data generation is still associated with risks of re-identification, or issues regarding its representativeness of real patient data, which undermines utility of synthetic data (Murtaza et al., 2023). Synthetic data literature primarily focuses on the technical features of the technology, without considering the institutional environment, i.e. the rules and resources that enable or disable actors to interact. As such, how the technology fits within organisational data policies, connects with health data sharing processes, and complies with (inter)national regulations, remains out of view. Understanding this institutional context helps to identify the factors that enable synthetic health data sharing in a privacy-enhancing manner.

1.2 Scientific problem

To further substantiate this problem statement, this section explores current literature on synthetic data generation. This is done by defining synthetic data (§1.2.1) and a specification to electronic health records (§1.2.2). A search in PubMed and Scopus on terms related to synthetic health data, data protection and its context showed how literature addresses data protection from a technical (§1.2.3) and institutional (§1.2.4) perspective. The findings are concluded in the definition of a scientific knowledge gap (§1.2.5).

1.2.1 Definition and types of synthetic data

In an attempt to provide a widely acceptable definition of synthetic data, the Alan Turing Institute proposes the following definition: "synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)" (Giuffrè & Shung, 2023, p. 1; Jordon et al., 2022). This thesis adopts this definition, as it supports the various generation methods that exists for synthetic data generation, and does not presume privacy; this is important, as will be shown in §1.2.3. Numerous methods to generate synthetic data in the context of healthcare exist (Hernandez et al., 2022). For instance, there are classical approaches, that use 'simpler' anonymisation techniques, such as replacing real values with fake values, adding noise to data, or using statistical models based on correlation (Hernandez et al., 2022). These methods are primarily relevant when the amount of available data is scarce, for example in case of rare diseases or experimental treatments (Wang et al., 2021). Further, synthetic data generation is increasingly studied in the advancing research in the field of deep learning, and AI in general (Murtaza et al., 2023; Tsao et al., 2023). These methods have a variety of underlying generation models and use numerous quality evaluation metrics to both determine the synthetic data's resemblance to the real-world data, and privacy preservation (Hernandez et al., 2022; Murtaza et al., 2023. Consequently, synthetic data can be generated with technologies that vary in complexity, comprehensibility, and representativeness of the real data.

1.2.2 Specification of scope: synthetic electronic health records

As there are different methods to generate data, there are also different types of data that can be generated, such as medical images (Han et al., 2018), biomedical signals, and electronic health records (EHRs), both as free-text and tabular data (Dahmen & Cook, 2019; Guan et al., 2018; Venugopal et al., 2022). This thesis focuses on the generation of EHRs for time and scope reasons, as well as to make the concept of (synthetic) health data more tangible. The argumentation for scoping to EHRs is twofold.

First, EHRs are currently often unavailable to researchers, even though they have great value for health data sharing for research, by giving insight into medical history and outcomes of treatments (Scholte et al., 2019). Today, almost all patient data is recorded in EHRs provided by each healthcare provider (Ministry of Internal Affairs, 2018). A record contains *inter alia* a patient's medical test and treatments, their progress, diagnoses, lab test results, referral and discharge letters, medical imaging scans, and nursing reports (Ministry of Internal Affairs, 2018). Hence, EHRs are a fruitful source for research of all kinds; in recent years specifically to develop ML-based health solutions for better healthcare delivery (J. Li et al., 2023). Despite their great value for e.g. aetiology of diseases, EHRs are yet to be shared beyond the border of healthcare providers (Yan et al., 2022).

Second, EHRs contain a broad range of data types, often recorded in longitudinal fashion. For example, biomedical signals and lab test results are processed as continuous timeseries data, and medication use and diagnoses are typically processed as discrete values (e.g. Booleans and categories) mapped over time (J. Li et al., 2023). The high-dimensionality and causality in relationships between variables raise unique challenges for generating synthetic data based on health data (Lee et al., 2017; Murtaza et al., 2023). This is property is specifically reflected in EHRs, as they record mixed-type data from one patient over a certain period of time (Abedi et al., 2022; Lee et al., 2017). This complicates the synthetic data generation process (Abedi et al., 2022). Therefore, EHRs are representative for synthetic data generation in the healthcare context.

1.2.3 Gaps in technical literature

The search in PubMed and Scopus showed that most researchers view synthetic data as a solution to challenges imposed by data protection law, because of its potential to anonymise personal data (see e.g. Abay et al., 2019; Abedi et al., 2022; Rajotte et al., 2022). However, other than mentioning the relevant laws, they do not further assess whether their models preserve privacy. Synthetic data literature primarily focuses on performance and reliability of generation methods, rather than devoting attention to data protection in this context (see e.g. Abedi et al., 2022; Rajotte et al., 2022). This is worrisome, considering the fact that synthetic data can still impose privacy risks, for example regarding re-identification of patients (Chauhan et al., 2023; Chen et al., 2021; Giuffrè & Shung, 2023; Stadler et al., 2022). Assuming that synthetic data generation *per se* results in anonymous datasets may lead to re-identification risks that are unaccounted for, infringing patients' data protection rights (Giuffrè & Shung, 2023).

To address unintentional re-identification risks, as well as information leakage as referred to by Chen et al. (2021), privacy risks are often quantitatively measured (Appenzeller et al., 2022). For example, regarding Generative Adversarial Networks (GANs), there are various researchers that acknowledge the need for additional data protection assessments in the form of quantitative metrics (Dikici et al., 2021; Diller et al., 2020; Goncalves et al., 2020; Hernandez et al., 2022; Nik et al., 2023; C. Sun et al., 2023; Venugopal et al., 2022; Wang et al., 2021; Yale et al., 2019b; Yan et al., 2022). As these data generation models have similar structures and sometimes build on each other, they allow for a privacy benchmark of different models (Appenzeller et al., 2022; Yan et al., 2022). Similarly, Mosquera et al. (2023) and Zhou et al. (2022) perform a quantitative privacy risk assessment of their neural network-based model and textual model, respectively. Coutinho-Almeida et al. (2021) conducted one of the first literature reviews that evaluate what privacy metrics are available for generation methods in a healthcare setting. They provide a general overview of the available metrics. Murtaza et al. (2023) and Hernandez et al. (2022) performed a similar analysis, but more extensively. What characterises these articles, is that they only present the metrics, without proper explanations or normative interpretations. For example, although Coutinho-Almeida et al (2021) aim to help organisations understand the available privacy metrics, the interpretation of their results requires an understanding of the (computer science) concepts that underly these privacy metrics, such as privacy loss, differential privacy and distance-based metrics. This applies to the analyses of the other mentioned articles as well.

Contrary to most researchers, Appenzeller et al. (2022) start their quantitative assessment with an analysis of legal requirements. Considering the risk of re-identifying patients, they consider synthetic data generation as a technique to mitigate privacy risks, rather than an anonymisation technique. The authors take a positive step towards an institutional analysis. However, the technical study interprets the General Data Protection Regulation (GDPR) in a unnuanced and, in my view, incorrect manner. They consider the GDPR's view regarding anonymisation as similar to the interpretation of US health data protection law (as defined in the Health Insurance Portability and Accountability Act (HIPAA). However, anonymisation under HIPAA concerns the removal of direct identifiers, such as names or social security numbers. This understanding does not apply to the GDPR, as shown in §1.2.4. The incorrect interpretation shows the importance of a multidisciplinary research that bridges the knowledge of institutional rules and technical knowledge of synthetic data generation.

In summary, the referenced articles provide a starting point for identifying what privacy metrics are available. However, due to a lack of explanation and assumed prerequisite knowledge of mathematical and computer science concepts, they are hard to interpret for readers without a technical background. Thus, they require further explanation and translation to the context of users of synthetic data in healthcare.

1.2.4 Gaps in institutional literature

Next to the technical articles that propose or discuss synthetic data generation models, there are articles that study (parts of) the institutional context of synthetic data generation.

Within the area of open data science, Haendel et al. (2021) analysed how COVID-19 datasets may be shared to stimulate health data sharing for research in the United States of America (US). Differentiating between synthetic data and de-identified data, they assessed the applicability of data protection mechanisms from US law to the process of obtaining health data for research. This thesis has a similar focus, but instead focuses on EU data protection law, particularly the Dutch implementation. The HIPAA and GDPR both regulate the processing of personal health data. However, they maintain different definitions of personal data and provide different guidelines for anonymisation. Therefore, the institutional environments differ significantly. Lessons can be drawn from relevant process components, such as data sharing resources and actors involved. A limitation of their research, as is more often the case with synthetic open data, is that the granularity of the analysed synthetic data is low. In their research, the data is highly aggregated to a small set of dimensions to mitigate re-identification risks. This suffices for their specific study purpose of COVID-19 testing on a national level, but such data does not fulfil the quality requirements for patient-level medical research. Consequently, the privacy risks that inherently arise in highly granular synthetic datasets that aptly capture the high dimensionality of EHRs necessary for research purposes, are not considered or analysed. What we can learn from this field of research, however, is that the required level of data protection depends on the types of synthetic data and its intended use for research. Moreover, it confirms the importance of analysing the context in which synthetic data is applied.

Kamel Boulos et al. (2022) emphasised the need for a wider socio-technical framework for privacy-preserving technologies, including rules for disclosure, use, and dissemination of personal health data. They acknowledge that privacy-preserving methods themselves do not necessarily lead to secure and ethical use. For example, next to the technological evaluation, such a framework should consider harmonisation of data protection regulations, establishment of collaborations between different stakeholders, bureaucratic simplification, and guidelines for using and reporting on synthetic data. However, the authors do not discuss what rules should be harmonised or how actors can collaborate. This thesis gives more substance to this question by delving into the (legal) challenges of current data sharing practice and the extent to which synthetic data may (not) solve them.

Furthermore, Alloza et al. (2023) identified barriers to the adoption of synthetic data generation in healthcare, such as the training of professionals to analyse synthetic datasets, computational costs, and the need for a clarification of how synthetic data can be evaluated. The authors stress the importance of a data protection evaluation and introduce some methods to assess data protection, such as attribution or membership disclosure. However, these concepts are only referred to, without proper explanations or interpretations.

The benefits and drawbacks of the application of synthetic data in health research as analysed by Giuffrè and Shung (2023) come closest to an institutional analysis of synthetic data, by analysing regulatory concerns. Regarding privacy, they argue that synthetic data poses risks to GDPR compliance, as it is unclear how synthetic data relates to the GDPR's core concept of 'personal data'. Therefore, they call on regulators to update the data protection rules. Similarly, according to Tsao et al. (2023), the use of synthetic data for research is governed and studied on a case-by-case basis due to its specific use cases, leaving important overarching questions unanswered. For example, despite synthetic data being artificial, there remains a certain risk to identification of outliers, that depends on the specific context wherein the synthetic data is used and wherein the real data is collected. These authors too, raise the question whether synthetic data should be regarded as personal data, but leave the question open for further research.

Also from a legal perspective, Bellovin et al. (2019) analysed how synthetic data relates to health data protection rules in the US and identified conflicts with the understanding of personal health data in the HIPAA (Appenzeller et al., 2022; Bellovin et al., 2019). They identify the misclassification of synthetic data as anonymous data as important challenge that developers and users of synthetic data should be aware of to ensure compliance with data protection regulations. The authors provide a good example of a multi-disciplinary research into synthetic data, introducing anonymisation as important theme in the institutional context. Hence, this thesis has a similar scope, but instead, focuses on the EU jurisdiction.

Chauhan et al. (2023) articulate that synthetic health data may give an illusion of privacy, even though personal information can be inferred. Moreover, there are no methods to determine whether synthetic data is “truly anonymous”. The authors try to map broader, ethical concerns around the application of synthetic data generation, but fall to identify the causes of these concerns.

In summary, many unanswered questions remain about synthetic health data in its institutional context, especially for the EU.

1.2.5 Scientific problem

On the one hand, synthetic data literature gives an overview of the developments of synthetic data in healthcare as well as technical reviews of data protection. This provides a solid foundation for understanding what techniques are available, its maturity level and how data protection is assessed. Nevertheless, researchers stop at assessing their models with privacy metrics, without taking the necessary step to evaluate whether this level of data protection conforms to the institutional context.

Moreover, the institution-oriented articles hardly refer to the technical privacy evaluations of the researchers, resulting in a literature gap between the technicalities of synthetic data generation and the institutional environment. Thus, while synthetic data generation presents promising features, there remain unanswered questions surrounding its application. To relieve the administrative burdens of current data exchange procedures, it is essential for the actors in the health data sharing process to have a clear understanding of whether the proposed synthetic data generation applications comply with the institutions and how this can be demonstrated. Inherently, this requires a clarification of institutions for the use of synthetic data. The objective of this thesis is to identify the factors that enable and hamper synthetic health data sharing in a privacy-enhancing manner.

Filling the identified knowledge gaps, requires understanding of how health data is currently shared. However, institutional literature on health data sharing primarily studies the formal rules (see e.g. Boyd et al., 2021a, 2021a; Hansen et al., 2021; Molnár-Gábor et al., 2022; Slokenberga, 2022), leaving institutional arrangements on lower governance levels underexposed. For example, these authors mention health data sharing for research is associated with high administrative burdens, but fail to explain how such problems emerge within organisations, such as the coordination issues during the health data sharing process or the interpretation of legal concepts at the organisation level. Therefore, this thesis will look beyond the formal institutions of health data sharing.

1.3 Research objective and question

The research objective of this thesis is to design a framework that structures the data protection-related factors that influence the extent to which synthetic health data enables health data sharing for research. The framework builds on the current institutional data protection environment of health data sharing, making up for the lack of knowledge on current health data sharing practices. This thesis will first expand the literature on formal institutions of secondary use of health data for research, providing insights into how challenges emerge in health data sharing practice. Second, this thesis provides insights into the technical data protection evaluation of synthetic EHRs, paying specific attention to how researchers interpret their privacy metrics as starting point for understanding how the evaluations can be translated to the institutional context. To bridge the gap between institutional and technical literature, the findings of the synthetic data generation analysis are studied in relation to the institutional environment of health data sharing. This provides insights into whether synthetic data generation can address the challenges of health data sharing for research. The identified factors are structured in a framework that helps researchers and users of synthetic data generation to understand to what extent synthetic data can enable health data sharing for research, while protecting patients' data. This thesis defines 'factors' as data protection-related barriers and drivers that determine the potential impact of synthetic data generation on health data sharing. Based on the institutional and technical analysis, solution directions are presented for the phases of synthetic data sharing.

To address the scientific problem identified in §1.2.5, this thesis formulates the following main research question:

> ***RQ: How could synthetic data generation enable health data sharing for research in a privacy-enhancing manner?***

To answer this research question, the following sub-research questions are formulated. These sub questions are further clarified in the methodology (Chapter 2).

> *RQ1 What is the institutional data protection environment of health data sharing for research?*
>
> *RQ2 How could synthetic data generation and evaluation contribute to health data protection?*
>
> *RQ3 Which data protection-related factors influence how synthetic data generation enables health data sharing for research?*

1.4 Relevance for master programme

This thesis aligns with the objectives of the Complex Systems Engineering and Management master programme, because an interdisciplinary approach is adopted to design a framework within a socio-technical system that is characterised by complex interactions. The technical complexity stems from the development and evaluation of synthetic data generation, whereas the social complexity arises from data sharing interactions between healthcare providers, research institutes, technology providers and patients. Additionally, the legal data protection framework causes uncertainties for the actors within this complex systems, troubling health data sharing for research. Furthermore, data sharing in healthcare necessitates a delicate balance between various, sometimes conflicting, values. While healthcare providers stand to gain from enhanced research outcomes, society as a whole benefits from advancements in the healthcare system. Nonetheless, these healthcare advantages must be carefully weighed against the imperative of protecting individuals' data. The insights furnished in this study attempt to bridge the technical and institutional knowledge on synthetic data generation and data sharing practice, serving as a basis for sharing synthetic data in a manner that upholds the protection of personal health data.

1.5 Thesis outline

Chapter 2 outlines the methodology, clarifying the research approach and methods used to answer each research question. Chapter 3 delves into the institutional data protection environment to answer RQ1, constituting an analysis of the patterns of interaction of actors involved in health data sharing practice. This further comprises mapping the applicable (in)formal rules, characteristics of data sharing systems, and the actors' roles and responsibilities in the health data sharing process. Chapter 4 answers RQ2, by looking at synthetic health data generation – specifically at synthetic EHRs, how different generation models could contribute to data protection, and what barriers remain for synthetic data users to assess this protection. RQ3 is answered in chapter 5, that connects the characteristics of synthetic health data generation (chapter 4) with the institutional framework for health data sharing (chapter 3). This includes exploring the implications for data access, anonymisation, consent, and implementation within the Dutch data protection environment. Lastly, this chapter synthesises the findings into a practical framework that structures the factors that drive and inhibit synthetic data sharing in a way that safeguards patient data, as well as possible solutions for the barriers. Chapter 6 concludes the thesis by answering the main research question, and discussing limitations and possibilities for future research.

2 Methodology

This methodology chapter outlines the research approach (§2.1), clarifies the research questions (§2.2), the framework that underlies this thesis' analysis (§2.3), data requirements and corresponding methods and limitations (§2.4).

2.1 Design-based approach

Generally, this research employs qualitative methods to design an artefact, with the artefact being a framework that structures the barriers and solution directions for synthetic health data sharing in consideration of data protection institutions. Theories of institutionalism start from the idea that actors' behaviour is attributed to institutions (Schneiberg & Clemens, 2006). Institutions can be defined as systems of rules, including legal rules, social norms and culture that structure actions of actors (Kim & Stanton, 2013; Klijn & Koppenjan, 2006; Scharpf, 1997). The institutional environment can be seen as a social infrastructure that provides a certain and stable basis, enabling actors to interact (Kim & Stanton, 2013; Klijn & Koppenjan, 2006; Powell, 1991). To analyse institutions and design for these institutions, this thesis combines theories from institutional analysis with established design science strategies.

Specifically, this thesis follows a Design Science Research (DSR) approach, which concerns the process of designing any artefact with a solution to an understood research problem, such as constructs, models, methods and instantiations (Hevner et al., 2004), and other properties of technical, social or informational resources (Peffers et al., 2007). Methods direct the performance of purpose-driven activities, by providing protocols, practices or algorithms to accomplish a task (Hevner & Wickramasinghe, 2018). By identifying factors that enable synthetic health data sharing, this research starts the debate of a method to evaluate data protection of synthetic data generation conform its institutional environment. Thus, a framework that structures these factors fit the description of an artefact.

Numerous researchers have proposed frameworks for performing DSR (Hevner et al., 2004; Johannesson & Perjons, 2014; Peffers et al., 2007). Within the discipline of information systems (IS), Hevner et al. (2004) have presented an influential framework for understanding, executing and evaluating SDR (Figure 1). It combines behavioural science and design science, fulfilling the demand for theoretical verification of artefacts (Hevner et al., 2004). Hevner et al.

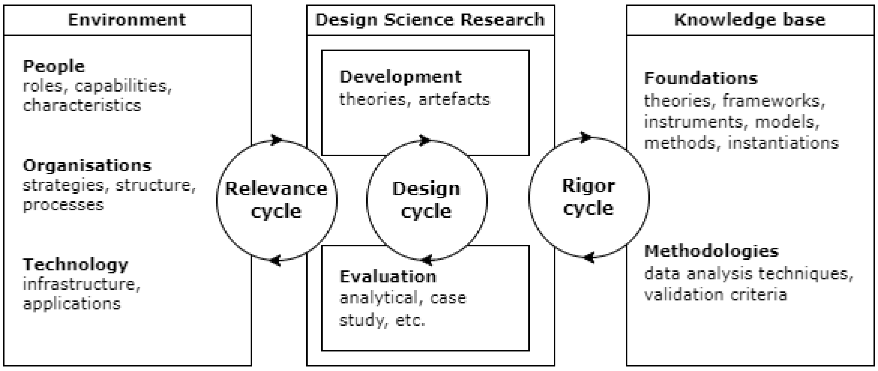

Environment
People
roles, capabilities, characteristics
Organisations
strategies, structure, processes
Technology
infrastructure, applications
Relevance cycle
Design Science Research
Development
theories, artefacts
Design cycle
Evaluation
analytical, case study, etc.
Rigor cycle
Knowledge base
Foundations
theories, frameworks, instruments, models, methods, instantiations
Methodologies
data analysis techniques, validation criteria

Figure 1. Three cycle DSR approach (Hevner et al., 2004)

(2004) propose a three-cycle view on DSR: the relevance cycle assures research activities address its contextual needs; the rigor cycle applies existing scientific foundations, expertise and experiences to design an innovative artefact, and the design cycle iterates between developing and evaluating the artefact. These cycles together determine the extent to which the artefact achieves its purpose of improving the application context (Hevner & Wickramasinghe, 2018). Results of this approach include the design of artefacts, as well as understanding of why the artefacts contribute to solve problems in its application context (Hevner & Wickramasinghe, 2018). The approach by Hevner et al. (2004) suits this thesis, because it values input from the contextual environment of the artefact, an underrepresented discussion in literature on synthetic data generation, and existing technical knowledge, providing a theoretical basis to come to a first design. The cycles are executed iteratively, leading to some overlap exists between the cycles. Literature on institutional analysis methods, synthetic data generation and evaluation methods provide the scientific base. For instance, literature about the institutional analysis (knowledge base) informs its application to map actors' and organisations' roles, responsibilities, and rules and data infrastructures (environment) (framework for this analysis is specified in §2.3). Similarly, the rigor cycle connects the methods to generate synthetic data, phases of synthetic data life cycle, and privacy risks evaluation methods taken from scientific literature (knowledge base) to inform the design of a framework (design science research). Insight in the current knowledge base is important to assure the design provides new knowledge to the organisation (Hevner & Wickramasinghe, 2018). In short, the findings from the knowledge base inform the design of a framework that structures the factors that enable health data sharing, in line with the environmental requirements to assure purposefulness.

Given the scope of this thesis and its accompanying time and resource limitations, this research focuses on unravelling the institutional environment and its relation to the knowledge base. This thesis provides a first iteration of a framework that structures the (non-exhaustive) factors that enable synthetic health data sharing, which can be built on by other researchers.

2.2 Research questions

Following the DSR approach, this thesis formulates three sub-research questions to determine how synthetic data generation can enable health data sharing: gain insight in the institutional environment of health data sharing (RQ1), understanding the methods to generate and evaluate synthetic data generation that preserves patients' data protection (RQ2, and designing an artefact that combines the practical and technical knowledge into a purpose-bound artefact (RQ3).

RQ1: Understanding the institutional environment

The first research phase comprises the analysis of the institutional data protection environment based on Ostrom's institutional analysis and development (IAD) framework (see §1.3.3). This framework allows us to identify the factors that enable or constrain certain behaviour; in this thesis the objective is to identify how institutions result in interaction patterns in health data sharing practice. Formal institutions are important mechanisms to enforce desired data sharing behaviour, yet, research practice is also important to encourage data sharing (Kim & Stanton, 2013). Therefore, this thesis focuses on formal rules, as well as informal data sharing institutions. An analysis of the institutional data protection environment provides insights into the challenges of current health data sharing practice and their cause. This serves as a basis to assess how synthetic data generation can address current challenges. The first research question is formulated as follows:

> *RQ1 What is the institutional data protection environment of health data sharing for research?*

RQ2: Understanding the methods to generate and evaluate synthetic health data

Before the interaction between synthetic data generation and the institutional environment can be analysed, the concept of synthetic data generation, its possibilities, and limitations should be clarified. Therefore, the second research phase explores features of the state-of-the-art technologies and how these contribute to protection of personal data, specifically focusing on EHRs to build a knowledge base. To say something about the extent to which synthetic data generation contributes to personal data protection, it is important to examine how developers argue that their model meets this requirement. Therefore, this research question requires an analysis of how developers evaluate data protection. In terms of the three cycle DSR approach, this thesis explores the knowledge base on data protection by synthetic EHRs and contributes to the knowledge gap by providing a structured overview of privacy evaluations and their interpretations by researchers. The second research question is formulated as follows:

> *RQ2 How could synthetic data generation and evaluation contribute to health data protection?*

RQ3: Designing a framework

The third research phase consists of executing the relevance and rigor cycle, connecting the knowledge base and the environment. With understanding of the key actors, corresponding legal concepts, procedures and institutional challenges of health data sharing, as well as scientific knowledge on synthetic data generation and evaluation, we can assess how the institutional environment interacts with synthetic data. This relation should be defined in terms of how current challenges may be solved and how synthetic data generation may impose new challenges for the institutional environment. These interactions identify the data-protection related factors that enable and inhibit synthetic health data sharing and solution directions. These are structured based on the phases of synthetic data sharing in a framework to conclude how synthetic data health can influence data sharing and what (or who) is needed to address data protection barriers. The third research question is formulated as follows:

> *RQ3 Which data protection-related factors influence how synthetic data generation enables health data sharing for research?*

2.3 IAD framework

By recognising the importance of the artefact's environment for design, Hevner et al.'s DSR approach allows for a more thorough analysis of the institutions of the artefact; this aligns with the thesis' focus on data sharing institutions. The Institutional Framework for Policy Analysis and Design serves as scientific foundation of the institutional analysis (Polski & Ostrom, 2017). Polski and Ostrom apply Ostrom's IAD framework for common pool resources for policy analysis (Ostrom, 1990; Polski & Ostrom, 2017); it can be applied as a general institutional analysis framework for a broad range of goods and services (Ostrom, 2011). It combines contextual concepts in which the good/service with the positions of participants and their competence regarding the analysed action situation, interaction patterns, and the possible outcomes (Figure 2) (Ostrom, 1990; Polski & Ostrom, 2017). This IAD framework shows similarity to the concepts of Hevner et al. (2004), regarding the people of the technology's environment, the way the system is organised, and the physical conditions of its application. This section clarifies the application scope of the framework and the purposes it serves.

The strength of this framework lays in studying resources not as stand-alone goods or services but as part of a greater, complex system (Filgueiras & Silva, 2021; Polski & Ostrom, 2017; Purtova & Van Maanen, 2023). In its core, the IAD identifies the following concepts (see Figure 2). First, the characteristics of the resource, categorised in physical and material characteristics, community attributes, and rules-in-use. Second, the action arena defines the boundaries for interaction patterns and consists of the actors and the action situation. The action situation is a social space where actors interact and undertake actions, such as exchanging goods and services, solving issues, or exercising control over the others (Ostrom, 2011). Third, the characteristics of the resource along with the action arena determines the behaviour of actors and how this results in certain interaction patterns. The outcomes of these interaction patterns are evaluated based on the policy objectives. These concepts are further defined in the sections where they are applied (§3.1-3.5).

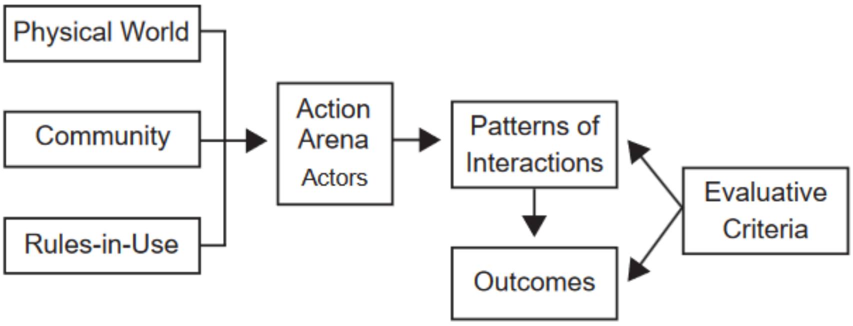


Figure 2. Institutional Analysis and Development Framework (Polski & Ostrom, 2017)

To determine in what order components should be studied, Polski and Ostrom (2017) suggest to first define the policy analysis objective. This thesis uses the IAD framework as an analytical tool, with the objective to structure knowledge from an empirical use case to an overview of institutions for sharing health data with third party researchers. The focus of the analysis will be describing the behaviour of actors, based on the characteristics of the resource (Polski & Ostrom, 2017). The objectives of this analysis are to identify factors in the characteristics of the resource that result in certain behaviour, evaluate how the actors' behaviour in the action arena leads to certain patterns of interactions, and eventually what outcomes these interaction patterns result in. Specifically, the outcomes of this institutional analysis give insight the challenges of health data sharing for research and its causes. With understanding of the institutional environment, the IAD framework can be deployed for a second objective: prediction. The factors influencing synthetic data sharing can be identified by exploring how synthetic data generation relates to the institutional environment and its challenges. The framework components should be analysed in the order that suits the study (Polski & Ostrom, 2017). An analysis of the action situation and interaction patterns and a more thorough analysis of the outcomes requires understanding of the characteristics of the resource in this thesis. As the rules-in-use form the foundation of health data sharing practice and define the relevant concepts, therefore, these are discussed first. Then, the physical and material conditions are discussed, followed by the community attributes.

2.4 Methods per research question

Figure 3 summarises how the research questions are studied within the three-cycle design approach (Hevner et al., 2004). This section discusses the methods used to answer each research question; the methods are summarised in Figure 3.

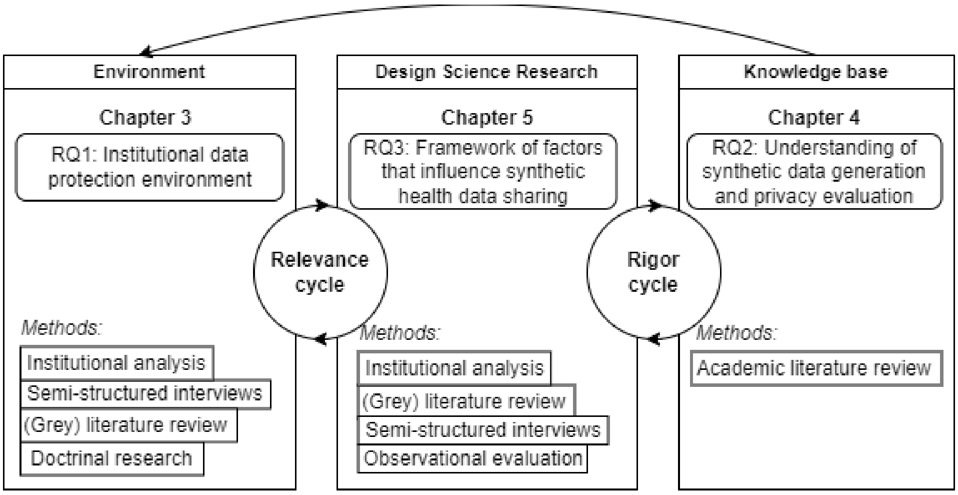

Environment
Chapter 3
RQ1: Institutional data protection environment
Methods:
Institutional analysis
Semi-structured interviews
(Grey) literature review
Doctrinal research
Relevance cycle
Design Science Research
Chapter 5
RQ3: Framework of factors that influence synthetic health data sharing
Methods:
Institutional analysis
(Grey) literature review
Semi-structured interviews
Observational evaluation
Rigor cycle
Knowledge base
Chapter 4
RQ2: Understanding of synthetic data generation and privacy evaluation
Methods:
Academic literature review

Figure 3. Adapted three cycle design approach with methods (based on Hevner et al., 2004)

2.4.1 RQ1: (grey) literature review, doctrinal research & semi-structured interviews

Following the DSR approach, the first research question aims to map the institutional environment of health data sharing practice, structured by the IAD framework.

Doctrinal research

The rules-in-use component of the IAD framework requires understanding of the formal data protection rules, with a focus on secondary use of health data. Therefore, a doctrinal research method suits best (Hutchinson, 2016), meaning that the current legislation and corresponding recitals, available case law, additional guidelines of authoritative institutions are analysed; this includes documents by the Commission and data protection authorities (DPA). Such information is publicly online available via EU and national government sources.

(Grey) literature review

Regarding the other components of the IAD framework, academic literature is consulted where possible. However, current literature on secondary use of health data is often limited to regulatory hurdles; the positions of actors and data sharing processes are not defined in literature (§1.2.5). This literature gap is complemented with grey literature, existing of reports commissioned by governmental bodies or healthcare actors and policy documents.

Semi-structured interviews

Academic and grey literature lack insight into the health data sharing procedures of individual organisations, the relevant actors within organisations, and opinions of such actors regarding this process. As this qualitative information is not publicly available, interviews with actors involved in health data sharing are conducted. Qualitive research interviews can contribute to a conceptual and theoretical body of knowledge based on the experiences of interviewees (Qu & Dumay, 2011). Individual in-depth interviews are often used in healthcare research to reconstruct (perception) events and experience related to health(care) (Qu & Dumay, 2011). The objective of the individual in-depth interviews in this thesis is to understand the interviewees' personal experience with and opinion on (synthetic) data sharing. A key question was asking their expectation on how synthetic data could facilitate sharing health data for research. To allow for some generalisation of the individual interviews, all interviews are based on a predefined set of topics queried by a consistent set of questions with a certain prioritisation (Adams, 2015; Qu & Dumay, 2011). Therefore, this thesis conducts semi-structured interviews based on the predefined concepts of the IAD framework, supplemented with questions that emerge during the interview (interview protocol in Appendix A). The general set of concepts remains the same for each interview; however, as the roles of interviewees differ, the concepts that relate to the interviewee's expertise are expanded for individual interviews (Qu & Dumay, 2011). Moreover, semi-structured interviews leave room for elaborating on parts of questions and discussions that emerge during the interview (Adams, 2015; Qu & Dumay, 2011).

Interviewees within the use case are selected. This thesis studies one use case of (personal) health data exchange between a healthcare provider and a research institute, as such health data only becomes available after lengthy ethical and legal procedures. Specifically, drafting data user agreements is seen as a bottleneck. This use case is selected because of its multi-organisational focus, the relevance of health data sharing for research, encompassing the relevant aspects of the research question. Moreover, both the healthcare provider and research institute have expressed their interest in synthetic data generation. The interviewees selected for this thesis are described in Table 1; each identifier represents one interview. Only interviewees from the same healthcare provider and research institute are selected. These interviewees are approached via the professional network of the author and referrals from other interviewees.

The goal of the interviews is to gather insights into the process of synthetic data generation and its problems. This thesis performs a theoretical thematic analysis (Friese et al., 2018). By applying a theoretical framework (i.e. the IAD framework) to explain a process, the interviews for the institutional environment (RQ1) are deductive in nature; the interview questions map the themes of theory, and the analysis interprets the results in this predefined framework (Friese et al., 2018). Contrary to analysing an entire dataset to induce themes, coding in theoretical thematic analysis is based on the specific research interests of the researcher (Friese et al., 2018). This analysis method thus sacrifices richness of results for a more thorough analysis of the most important themes in the research data. In addition, the rules-in-use analysis already surfaced important themes prior to the interview analysis, justifying a more focused analysis of the interviews.

Table 1. Overview of interviewees and their relevance

Actor	Relevance for thesis	Identifier
Data steward at healthcare provider	Data stewards are the first contact person when it comes to data management for researchers and are concerned with data protection compliance. These interviewees provide insight into health data sharing process and hurdles (*RQ1*), and their ideas of how synthetic data may change their role (*RQ3*).	DS1-H, DS2-H
Privacy officer at healthcare provider	Privacy officers are consulted with all kinds of data protection issues during the health data sharing process. This interviewee provides insight in health data sharing process and hurdles, specifically focusing on data protection issue (*RQ1*), and their ideas of how synthetic data may change their role (*RQ3*).	PO-H
Legal counsellor at healthcare provider	Legal counsellors support the arrangements of health data sharing agreements. The interviewee provides insight in the conclusion of legal agreements and provide their opinion on legal discussions (*RQ1*), and their ideas of how synthetic data may change their role (*RQ3*).	LC-H
Legal counsellor at research institute	Legal counsellors support the arrangements of health data sharing agreements. The interviewees provide insight in the conclusion of legal agreements and their opinion on legal discussions, from the perspective of data recipients (*RQ1*), and their ideas regarding how synthetic data may change their role (*RQ3*).	LC1-R, LC2-R
Researcher at research institute	Researchers are subject to heath data sharing policy and directly experience the hurdles of the process to evaluate the outcomes of the patterns of interactions. This interviewee specifically needs health data for developing AI models (*RQ1*).	RS-R
Policymaker secondary use of health data at Ministry of HWS	The Ministry of HWS is concerned with interpreting the EU data protection rules and providing guidance to organisations on this. The interviewee can confirm legal discussion and offer insight in progress on guidance of data policy or legal concepts (*RQ1*)	PM-HWS
Technology provider for synthetic data generation	Technology providers can bridge issues identified in synthetic data literature to its application context (*RQ3*).	TP1, TP2
Data steward at healthcare provider	Data steward can evaluate the identified barriers and suggestions from an organisational perspective, offering insight into how the framework suits the institutional environment (*RQ3*).	DS1-H-VAL
Technology provider for synthetic data	Technology provider can evaluate the identified challenges and suggestions from a technical perspective (*RQ3*).	TP1-VAL

At the start of the analysis, a general coding frame should be developed (Braun & Clarke, 2006; Friese et al., 2018). In this thesis, this comprises the detailed components of the IAD framework, along with challenges related to the legal base and anonymisation (Appendix A.2). The general coding frame was supplemented during the analysis with relevant inductively generated codes (Braun & Clarke, 2006). These are seen as the overlapping themes between interviewees, such as how they fill in the IAD components. The results are reported following the structure of the IAD framework and validated by analysing overlap with other interviews and (grey) literature (Yin, 1984).

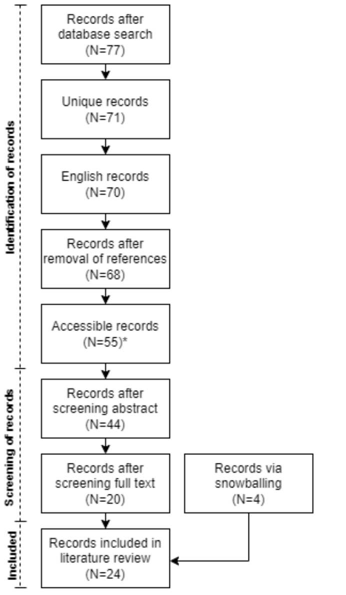
One practical limitation of semi-structured interviews as a research method is that it requires time and resources to collect, conduct, and analyse the interviews (Hove & Anda, 2005). A limitation of analysing one use case concerns that research findings may not be generalised to other settings (Yin, 1984). For instance, this research involves only one healthcare provider, whereas others presumably have different health data exchange procedures. Time limitations make these shortcomings acceptable. Rather than interviewing actors from different organisations, this thesis explores one use case in-depth, to describe the institutional environment in sufficient detail. To account for this limitation, the findings should be generalised with care: only findings that can be validated via other sources may be generalised.

2.4.2 RQ2: Academic literature review

The second research question aims to analyse the privacy-related features of synthetic data generation. Synthetic data generation models are primarily discussed in academic literature. Therefore, an academic literature review is performed to understand what metrics are proposed to evaluate privacy of proposed synthetic EHRs models, and especially, how these metrics are interpreted by researchers.

Article selection process

The literature study was performed and reported in line with Wee and Banister (2016) and Kable et al. (2012). To scope the literature study, the search terms, used in Scopus (which covered all results of ScienceDirect) and PubMed, are limited to the concepts of synthetic data generation of EHRs, and evaluate privacy (Table 2). To explain the search string: for the concept of synthetic data, other terms such as fake or artificial data identified records with synthetic data generation as primary focus and are therefore not included in search string. For EHRs, other terms such as personal health record identified records with a different scope, namely records in control of individuals instead of healthcare providers, combining data from multiple sources. These are therefore not included in search string. A more detailed overview of the article selection process is presented in Figure 4. The primary selection criterion is that the researchers explicitly address privacy or perform a privacy evaluation. Via forward snowballing, 4 additional articles regarding evaluation of privacy metrics are added to the literature review.


Identification of records
Records after database search (N=77)
Unique records (N=71)
English records (N=70)
Records after removal of references (N=68)
Accessible records (N=55)*
Screening of records
Records after screening abstract (N=44)
Records after screening full text (N=20)
Records via snowballing (N=4)
Included
Records included in literature review (N=24)

* The author had not access to the IEEE library. Where possible, these articles were retreived from other sources

Figure 4. Article selection process (based on Kable et al.,

Table 2. Search strings for and results of literature study

Database	Search string	Records retrieved	Unique records	Records in selection
Scopus	TITLE-ABS-KEY(("data generation" OR "synthetic data") AND ("electronic health record" OR "electronic patient record" OR "electronic medical record") AND ("privacy" OR "data protection" OR "confidentiality" OR "anonymity"))	67	61	15
PubMed	("data generation" [Title/Abstract] OR "synthetic data" [Title/Abstract]) AND ("electronic health record" [Title/Abstract] OR "electronic patient record" [Title/Abstract] OR "electronic medical record" [Title/Abstract]) AND ("privacy" [Title/Abstract] OR "data protection" [Title/Abstract] OR "confidentiality" [Title/Abstract] OR "anonymity" [Title/Abstract])	10	10	5
Forward snowballing				4
Total				24

Analysis method of literature review

To structure the information in the various articles, the following questions were answered:

- What type of article? (review article or model proposal)
- Which data types are discussed and for what use case?
- What generation methods are discussed?
- What are limitations of the generation method?
- How is privacy defined in the study?
- What privacy metrics are proposed?
- How are these privacy metrics interpreted?

2.4.3 RQ3: (grey) literature review and semi-structured interviews and observational evaluation

(Grey) literature review and semi-structured interviews

The interaction of synthetic data generation and its institutional environment builds on the identified challenges, both technical and institutional, and are grounded in literature, as well as policy documents and guidelines. Similar to RQ1, the results are structured via the IAD framework. However, as shown in the scientific problem (§1.2.4), there is limited data available about the institutions of synthetic data; presumably due to the novelty of the topic, it has not been subject to public debate.

One way to gather this knowledge is to conduct exploratory interviews with experts in the field of synthetic data generation in healthcare (Yin, 2011). Therefore, this thesis conducts two interviews with technology providers that deliver tools for synthetic data generation to organisations in the healthcare sector (Table 1). The interview questions were similar to the questions presented in Appendix A, hence, building on the IAD framework. Yet, they focused more on the role of the technology provider, and discussed how synthetic data generation could change the current environment. A thematic analysis was performed to analyse the results of these interviews. However, the general coding framework was less restrictive in comparison to the interviews for RQ1, as the answers from technology providers were more diverse and lied apart. The themes are presented in Appendix C. Prior to the analysis, these themes were identified based on the conclusions of the institutional analysis and the literature review of synthetic data generation and evaluation. The results are again structured similarly, as RQ2, following the components of the IAD framework.

Framework formulation and observational evaluation

The formulation of a framework builds on the previous research methods. The insights gathered via the aforementioned research methods are synthesised in a structured manner. To avoid subjectivity, the framework requires a foundation in literature, serving as theoretical background for the recommendations (See e.g. Wicks & St. Clair, 2007).

To validate the proposed framework, this thesis initiates an observational evaluation in line with the design evaluation methods mentioned by Hevner et al. (2004). To validate the proposed recommendations, this thesis performs an expert validation via interviews, as defined in the previous paragraph, as well as an analysis of documents are performed. The emphasis is not on the evaluation phase (§2.1), but to contribute to the knowledge base of applying synthetic data generation in practice. A case study is considered as observational way to evaluate an artefact design (Hevner et al., 2004). A case study evaluation concerns an in-depth analysis of the artefact in its business environment, or institutional environment as referred to in this thesis. A descriptive way to evaluate an artefact design is by informed argumentation based on information from the knowledge base. These evaluation methods should build an argument for quality metrics, such as the utility and efficacy of the artefact (Hevner et al., 2004). A design is finished when it meets the requirements of the problem it aimed to solve (Hevner et al., 2004).

In this thesis, the 'case study' comprises the use case analysed in the institutional environment of health data sharing and synthetic data sharing. Where interviewees were first questioned to map current institutions, they can now be questioned about the validity of the proposed framework, and eventually, but not in this thesis, bringing the solution directions that apply to the corresponding actors in practice. The case study is performed by interviewing actors that understand synthetic data generation and health data sharing practice (Yin, 2011). Semi-structured interviews, as explained above, are conducted with a healthcare provider's data stewards and a technology provider of synthetic data generation. The evaluation criteria are the artefact's fit to institutional environment and, certainly, preservation of privacy. The evaluation criteria are applied to the following framework components: phases in synthetic data sharing process, identified data-protection related drivers and barriers, involved actors and proposed solution directions. Therefore, this interview could be seen as a starting point of a case study that integrates the framework and lessons learned in health data sharing practice;

the evaluation indicates how the framework is received by relevant actors. Additional considerations from the interviewees regarding these problems and measures are reiterated in the framework of Chapter 5.

A limitation regarding this validation method is that it rests on expert knowledge. However, given the variety of disciplines involved, it is unlikely that a single expert would possess knowledge of all necessary aspects, such as regulation, technical features and data exchange practices. This may impose a challenge for the reliability of expert statements. It is incumbent upon the interviewer to pose appropriate queries to suitable experts. To mitigate this risk, the recommendations are categorised in technical, organisational, and procedural requirements. In doing so, experts can validate requirements relevant to their expertise.

3
Institutional environment of sharing health data

This chapter aims to map the institutional environment of sharing health data for research purposes across different organisations, aiming to answer the following research question:

> *RQ1 What is the institutional data protection environment of health data sharing for research?*

Based on the IAD framework explained in §2.3, this chapter begins with describing the contextual factors that influence health data sharing behaviour. First, applicable formal data protection rules are analysed (§3.1). Second, the physical and material attributes explain the current data sharing infrastructures as well as legal and institutional means to facilitate data sharing (§3.2). Third, the characteristics of community are analysed to identify how actors feel towards changes to health data sharing practices (§3.3). Prior to the actor analysis, the acceptable outcomes are defined based on policy objectives (§3.4). The interactions between actors and their environment are analysed in three action arenas (§3.5); to look ahead, these cover 1) day-to-day health data sharing practices, 2) the implications of the definition of personal data for research, and 3) legal grounds to share health data for research. Outcomes of these arenas explain how the institutional environment fails to meet the policy objectives. These IAD concepts are further defined in the corresponding sections.

3.1 Rules in use

This section analyses the rules in use, i.e. the set of formal and informal rules necessary to explain the policy-related actions, interactions, and challenges (Polski & Ostrom, 2017). The focus should be on commonly used rules that influence the operations of most participants (Polski & Ostrom, 2017). Figure 5 presents an overview of the relations between the IAD framework and the rules-in-use, visualising the importance of the rules-in-use on various aspects of the action arena.

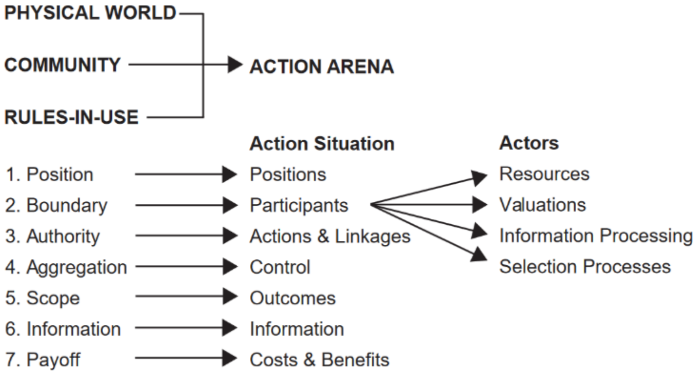


Figure 5. Relation between rules-in-use and components of IAD framework (Polski & Ostrom, 2017)

To give substance to Figure 5, Table 3 defines these aspects (after Polski and Ostrom (2017)) and illustrates how the types rules apply in the context of health data sharing.

Table 3. Overview of types of rules-in-use in health data sharing

Type	Definition	Example application to secondary use of health data for research use of health data
Position rules	The roles of participants in the action arena	Mainly determined by the roles defined in the GDPR and the actors' positions in the health data sharing process.
Authority rules	Action capacity that participants can have in their role	The actions participants can undertake are determined by data protection principles and responsibilities of roles, and delegated powers to EU member states, all defined in the GDPR
Boundary rules	The exit and entry conditions to the system	The reach of data protection regulations, determined by their definition of processing personal health data; the data protection rules that determine who may access health data also determine what actors may participate in health data sharing
Aggregation rules	Rules that determine how interactions lead to a certain outcome, determined by the level of control participants may exercise in their positions	Identify what participants can change the health data sharing process, and the competence of participants to engage in policy formulation of health data sharing
Scope rules	Criteria for acceptable outcomes	This thesis identifies the scope rules based on policy objective for secondary use of health data. At the actor level, desired outcomes may differ
Information rules	Rules that determine what information is accessible to participants	Determine what information about the process of obtaining health data for research is known to which participants
Payoff rules	Distribution of the costs and benefits of the interactions and outcomes of the action arena	The distribution of data-driven research results (or lack thereof)

Table 3 shows that the types of rules occur at multiple decision-making levels, ranging from operational to policy and regulatory levels. This thesis distinguishes formal from informal rules. Formal rules are codified in laws, regulations, and case law. Informal rules are norms and rules that actors adopt without formally documenting them, for instance creating traditions, codes of conducts, and rules or informal ways to monitor or sanction other actors (Ostrom, 2005). Generally, the formal rules apply to all three action arenas discussed later in this chapter (§3.5), across decision-making levels, as many issues related to secondary use of health data for research stem from formal data protection regulations (Hansen et al., 2021; Slokenberga, 2022).[3] Therefore, this section primarily discusses the formal rules (at the Dutch and EU level). Where possible, the sections of the corresponding action arenas are complemented with informal rules that apply to that specific action arena, primarily referring to rules regarding scope, information and pay-off.

This section focuses on three types of rules. First, an introduction to the EU data protection rules clarifies the roles and associated responsibilities and obligations that the GDPR defines regarding personal data processing in healthcare (position & authority rules) (§3.1.1). Second, this section clarifies the legal definition of personal (health) data and implications thereof to determine to scope of data protection rules (boundary rules) (§3.1.2). Lastly, the specific rules for secondary use of health data for research help to understand what actors are entitled to share data under what conditions (authority rules) (§3.1.3).

3.1.1 Introduction to GDPR

This section introduces the GDPR. The GDPR enshrines the right to data protection, by codifying requirements for processing personal data (General Data Protection Regulation, 2016, rec 6). As the right to data protection is not absolute, the GDPR seeks a balance between

[3] Please note this is a *general* distinction: some types of rules-in-use can be informal and formal, such as scope rules can follow from the GDPR, public documents and norms of specific actors. However, the distinction clarifies the analysis in this section and ensures that rules that strongly relate to interaction patterns between actors are described in those sections.

protecting personal data and enabling a free flow of personal information (General Data Protection Regulation, 2016, art. 1(1) jo. rec. 6). Appendix B defines relevant legal concepts.

The GDPR adopts a risk-based approach. Actors that determine the means and purposes of personal data processing ('data controllers') must assess their processing activities considering its risks to the rights of the individuals they process data of ('data subjects') (Gonçalves, 2020). Therefore, contrary to organisations' typical interpretation, the GDPR does not comprise a set of requirements that data controllers must each comply with, to comply with the GDPR (DS1-H). Instead, processing personal data requires a continuous risk assessment that considers the nature of the data, scope, context, and purposes, and implementing measures appropriate to protecting data subjects (General Data Protection Regulation, 2016, art. 24).

To apply this to our use case, this thesis considers healthcare providers as data controllers: they determine the means of data collection (when treating patients) and subsequently for data sharing. Patients are 'data subjects', as they are the natural persons described by the data. Research institutes can be viewed as data recipients. The GDPR mentions but does not define this term. This thesis interprets data recipients as external parties that obtain access to data. Notably, data recipients simultaneously qualify as controllers for their own processing activities, such as analysing data for their research. To avoid this ambiguity, healthcare providers are henceforth referred to as data collectors, because they collect data via healthcare delivery to patients (qualifying as data controllers in the GDPR); and research institutes are referred to as data recipients, because they receive data for further research (qualifying as both data controllers and recipients).

3.1.2 Scope of data protection law

Central to the (material) scope of data protection rules is that they apply to the processing of personal data (General Data Protection Regulation, 2016, art. 2(1)). In literature, the question has been raised whether synthetic data can be classified as anonymous data, exempting it from data protection rules (§1.2.4). This section explores the concepts of personal, pseudonymous, and anonymous data to clarify the scope of the GDPR in the research context. In terms of the IAD framework, this scope determines what actors are subject to the rules for health data sharing, and therefore, constitutes boundary rules.

Definition of personal (health) data

Personal data means "*any information relating to an identified or identifiable natural person (*'data subject'*)*". An '*identifiable natural person'* is a person who can be identified, directly or indirectly; particularly via, but not limited to, identifiers (e.g. a name or number) or factors that are specific to the natural person, such as their physical or genetic identity (General Data Protection Regulation, 2016, art. 4(1)). Health data specifically is defined as "*personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status*" (General Data Protection Regulation, 2016, art. 4(15)). Information about someone's health can expose intimate facts about a person's life, such as their medical history, treatments, personality, and well-being (Hordern, 2022; General Data Protection Regulation, 2016, rec 35). Because of its sensitive nature, the GDPR recognises health data as a special type of personal data that merits a higher level of data protection than non-sensitive personal data (General Data Protection Regulation, 2016, art. 9(1)). In general, processing special categories of personal data is prohibited, unless it is in the interest of society and subject to suitable safeguards (General Data Protection Regulation, 2016, rec 52). The GDPR defines an exhaustive list of purposes that justify processing of health data, that must apply jointly with the general legal bases for processing personal data (General Data Protection Regulation, 2016, art. 9(2) jo. 6(1)). The acknowledgment of the value of processing data is reflected in these purposes, including health security, prevention, management of healthcare services, and scientific research.

Definition of pseudonymisation and anonymisation of personal data

Medical research often involves sharing pseudonymised data (Mostert et al., 2016). The GDPR defines pseudonymised data as data that can identify a person with the use of additional information (General Data Protection Regulation, 2016, rec 26). Pseudonymisation should be seen as a security metric that reduces privacy risks – not as a measure to avoid the applicability of the data protection laws, as it still qualifies as personal data (General Data Protection Regulation, 2016, rec 28).

Although the GDPR does not define anonymisation, its meaning can be derived from the scope of personal data; data is considered anonymous when it no longer identifies a natural person. To assess whether a person is identifiable, all means that can reasonably be expected to be used should be taken into account, considering factors such as available technology, and effort and costs required for identification (General Data Protection Regulation, 2016, rec 26). Because this data is no longer identifiable, anonymous data sharing is not subject to data protection laws (Mostert et al., 2016).

The notion of personal data is not merely defined by legislation – case law plays an important role in scoping this definition, and thereby the reach of data protection law. In *Breyer*, the CJEU held that it should be ascertained whether data held by one entity could identify people when combined with data held by another entity (*Breyer*, 2016, para. 43); and when the former reasonably likely has means to obtain this data (*Breyer*, 2016, para. 45). The CJEU's interpretation is so strict that even the possibility of legal proceedings counts as such a means (*Breyer*, 2016, para. 49). If this is the case, the data qualifies as personal; showing that the EU data protection framework does not easily assume data to be anonymous.

The General Court of the EU applied *Breyer* in ruling on the difference between pseudonymised and anonymised data (*Single Resolution Board v European Data Protection Supervisor*, 2023). The question was what perspective to take when a data controller provides pseudonymised data to another party, where the data is personal data to the provider, but may not be personal data to the recipient (Kroes, 2023). The General Court held that the mere potential for reidentification does not mean that information is personal data *per se* – the possibility that persons are identified in practice should be considered in view of the circumstances of the case (*Single Resolution Board v European Data Protection Supervisor*, 2023, para. 97). Because the third party did not have access to information that identifies the data subjects, the pseudonymised data they received should be considered as anonymous data (*Single Resolution Board v European Data Protection Supervisor*, 2023, paras 94–106). It should be further investigated whether the third party has legal means to identify the data subjects, depending on the type of institution and processing purposes (Kroes, 2023). The General Court uses a more pragmatic approach, in comparison to the highly hypothetical approach of the CJEU, prescribing consideration of concrete possibilities rather than legal means that might be possible but remain unutilised. Nota bene, EU data protection authorities (DPAs) have appealed against the ruling, so the final word is up to the CJEU.

Relevance of personal data, pseudonymisation, and anonymisation for research

Following the reasoning of the GDPR, health data shared for research should preferably be anonymous (van Bon-Martens & van Veen, 2019). However, pseudonymisation or anonymisation, specifically when it concerns health or genomic data, is not always desirable or possible for medical research (Kroes, 2023; Mostert et al., 2016).

First, depending on research purposes, datasets with more individual information generally allow researcher to identify new correlations and causations; pseudonymising or anonymising data might hamper this and thus reduce the utility (Determann, 2020).

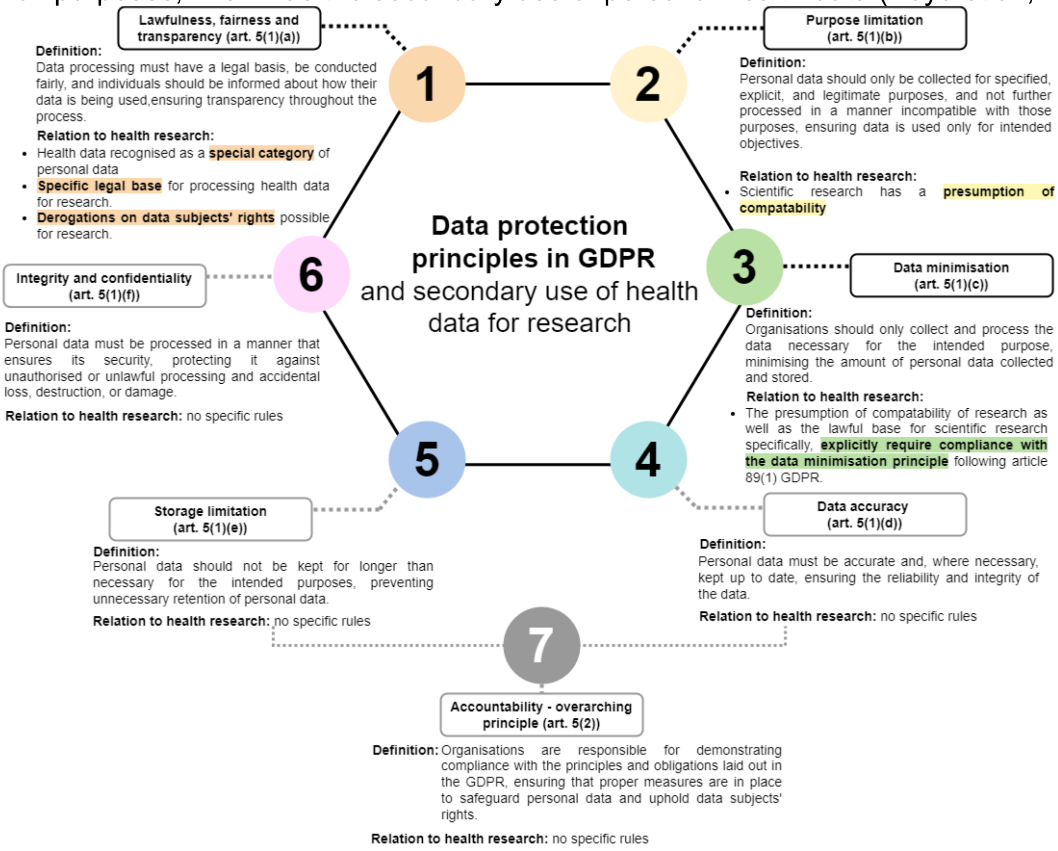
Relatedly, full anonymisation is often unattainable due to the high standard of anonymising data in the Netherlands (van Bon-Martens & van Veen, 2019); the comprehensiveness of datasets and the computational means to link data make it difficult for organisations to fully anonymise data (Determann, 2020). There is a high risk of re-identification in health data, due to the rich nature of health datasets (Kroes, 2023; Mostert et al., 2016). According to Determann (2020), anonymisation is only possible when health data is aggregated to group level, so that data points cannot be connected to individuals (Determann, 2020). Coding or encryption measures, for example, do not suffice, as an organisation still has the keys to identify data subjects (Determann, 2020). For healthcare providers sharing patients' data for research, the *Single Resolution Board v European Data Protection Supervisor* case implies they require a lawful basis for processing health data and informing data subjects about disclosure of their data to third parties (Kroes, 2023). As a result, pseudonymous data may be considered anonymous for external researchers when they lack means to reidentify data subjects, reducing administrative burdens.

3.1.3 Scientific research in data protection law

The balance between protecting personal data and promoting a free flow of data for public interests can be seen in the GDPR's scientific regime. Figure 6 outlines how the GDPR's data protection principles apply for health research, showing how the GDPR balances data protection and research interests. The principle of data accuracy, storage limitation, integrity and confidentiality and accountability do not specify rules for secondary use of health data for research. Also, they play a less prevalent role in the challenges around the regulatory framework. The position of scientific health research is reflected in the principles of purpose limitation, lawfulness, fairness and transparency, and data minimisation, discussed hereafter.

Purpose limitation

The purpose limitation principle prohibits processing data for a goal that is incompatible with the purpose for which the data was originally processed, which must be explicitly specified to the data subject (General Data Protection Regulation, 2016, art. 5(1)(b)). An exception is processing data for the goals of statistics, scientific, and historical research, and to serve the public interest: these ends are deemed compatible with the original purpose ('presumption of compatibility'), to stimulate data-driven research (Becker et al., 2022). Health data that are initially processed for healthcare delivery purposes, may therefore be used for scientific research purposes, known as the secondary use of personal health data (Boyd et al., 2021b).


Data protection principles in GDPR
and secondary use of health data for research
1
2
3
4
5
6
7

Lawfulness, fairness and transparency (art. 5(1)(a))

Definition: Data processing must have a legal basis, be conducted fairly, and individuals should be informed about how their data is being used,ensuring transparency throughout the process.

Relation to health research:

- Health data recognised as a **special category** of personal data
- **Specific legal base** for processing health data for research.
- **Derogations on data subjects' rights** possible for research.

Purpose limitation (art. 5(1)(b))

Definition: Personal data should only be collected for specified, explicit, and legitimate purposes, and not further processed in a manner incompatible with those purposes, ensuring data is used only for intended objectives.

Relation to health research:

- Scientific research has a **presumption of compatability**

Data minimisation (art. 5(1)(c))

Definition: Organisations should only collect and process the data necessary for the intended purpose, minimising the amount of personal data collected and stored.

Relation to health research:

- The presumption of compatability of research as well as the lawful base for scientific research specifically, **explicitly require compliance with the data minimisation principle** following article 89(1) GDPR.

Data accuracy (art. 5(1)(d))

Definition: Personal data must be accurate and, where necessary, kept up to date, ensuring the reliability and integrity of the data.

Relation to health research: no specific rules

Storage limitation (art. 5(1)(e))

Definition: Personal data should not be kept for longer than necessary for the intended purposes, preventing unnecessary retention of personal data.

Relation to health research: no specific rules

Integrity and confidentiality (art. 5(1)(f))

Definition: Personal data must be processed in a manner that ensures its security, protecting it against unauthorised or unlawful processing and accidental loss, destruction, or damage.

Relation to health research: no specific rules

Accountability - overarching principle (art. 5(2))

Definition: Organisations are responsible for demonstrating compliance with the principles and obligations laid out in the GDPR, ensuring that proper measures are in place to safeguard personal data and uphold data subjects' rights.

Relation to health research: no specific rules

Figure 6. Data protection principles in view of secondary use of health data for research

Lawfulness, fairness and transparency

The principle of lawfulness, fairness and transparency consists of three connected concepts. Lawfulness requires a legal base for processing personal data (General Data Protection Regulation, 2016, art. 6(1) jo. 9(2)). Processing health data requires the application of one of the legal bases in Article 6(1) as well as one of the derogations in Article 9(2) GDPR.

Fairness regulates the relation between data controllers and data subjects, imposing duties on data controllers to inform data subjects regarding the lawfulness and transparency of the data processing, including what data are processed and potential risks (General Data Protection Regulation, 2016, rec 39). Establishing rights for data subjects in order to decrease information asymmetries between data controllers and data subjects is an example of fair and transparent processing. For research purposes, the GDPR permits certain derogations on the rights for data subjects (Figure 6). Where data controllers usually have to inform data subjects about secondary use of personal data (General Data Protection Regulation, 2016, art. 14(1)-(4)), the obligation does not extend to cases where providing such information would be impossible or would impose a disproportionate effort, particularly for purposes of scientific research (art. 14(5)(b)). However, there are no clear rules that indicate when efforts can be considered disproportionate (Mostert et al., 2018). Similar restrictions to the rights of data subjects apply to the right to erase personal data (art. 17(3)(d)), or the right to object to processing of personal data (art. 21(6)). National legislators may further restrict the rights of data subjects when necessary to achieve scientific research purposes (art. 89(2)). In the Netherlands, the right to access, rectification, and restriction of processing can be derogated by the data controller for scientific purposes (Uitvoeringswet Algemene Verordening Gegevensbescherming, 2018, art. 44).

To process data lawfully, researchers can process health data for research for the purposes described in art. 9(2). Processing that is necessary for purposes in the public interest, scientific, historical or statistical research are allowed when this is provided for by EU or national law, and in accordance with art. 89(1) (art. 9(1)(j)). The requirement of there being a legal base, means this exception cannot be relied upon in every member state (Scheibner et al., 2020), complicating international data sharing. Art. 89(1) GDPR requires data controllers to adopt appropriate safeguards to protect data subjects' rights. Such safeguards comprise technical and organisational measures, particularly to aid data minimisation. Besides this legal base that specifically refers to research, there are other grounds on which researchers can rely for secondary processing of health data, e.g. by obtaining informed consent from data subjects (a), or when processing data for public sector bodies' duties under social employment and security laws (b), protecting vital interests of the individual (c), preventive medicine, medical diagnoses and healthcare system management (h), and public health (i).

Moreover, to further regulate processing of health data, there is room for national legislators to implement additional conditions for processing health data for research (Hansen et al., 2021). Based on Art. 9(4) GDPR, member states may adopt additional conditions to process health data. For example, member states may decide what kind of research, such as publicly or privately funded, may rely on what legal bases (Verhoeven et al., 2021). In their GDPR implementation act (UAVG), the Netherlands has adopted additional conditions for research. For secondary use of health data, consent is the default, from which can be deviated where consent is (almost) impossible (van Bon-Martens & van Veen, 2019; Verhoeven et al., 2021). Following the line of reasoning in the Dutch Civil Code regarding medical treatment contracts (art. 7:458 BW, 'Wet op de geneeskundige Behandelingsovereenkomst' or WGBO), which includes research as an exception for medical confidentiality, personal data may be processed for research without consent if it serves the public interest, requesting informed consent is impossible or would take disproportionate effort, and safeguards are put in place such that the privacy of data subjects is not disproportionately affected, such as pseudonymisation (Verhoeven et al., 2021; Uitvoeringswet Algemene Verordening Gegevensbescherming, 2018, art. 24). Consent can reasonably be skipped for example when patients have passed away or are non-traceable; the research involves a very large number of participants; or there being a substantial risk of response bias in sampling activities, leading to inaccurate datasets and subsequentially research (Scholte et al., 2019). Moreover, the

Dutch Civil Code requires that patients have not expressly objected to the disclosure of their data ('opt-out') for use other than healthcare delivery (art. 7:458(2)BW).

Data minimisation

The data minimisation principle is closely related to the aforementioned art. 89(1) GDPR. It holds that processing personal data should be limited to what is necessary considering its purposes (art. 5(1)(c). Processing personal data for scientific research purposes is subject to appropriate safeguards following art. 89(1), meaning that they must adhere to data minimisation via pseudonymisation, for example. However, other than requiring such safeguards, the content of these measures and when they are deemed appropriate remains unspecified (Mostert et al., 2018; Slokenberga, 2022). If such measures would restrict research purposes, and that purpose cannot be achieved by other reasonable means, an exemption is provided (Mostert et al., 2016).

Scientific research thus has a privileged position in the GDPR. The following sections will clarify how these formal rules impact data sharing behaviour of actors, and led to discussions on core concepts of data protection law.

3.2 Physical and material conditions

To analyse the action situation, the IAD framework requires understanding of the real world and the attributes of the goods or service in question (Ostrom, 2011). The resources and capabilities related to providing goods and services are referred to as physical and material conditions, e.g. available technology, labour, and funding (Polski & Ostrom, 2017). These conditions are important to consider as they affect actions situations and constrain the design space of institutional arrangements (Polski & Ostrom, 2017). Following Polski and Ostrom's approach, saying something about how data is shared, requires understanding of the concept of data and its economic nature (§3.2.1), and how data is produced and provided via technical (§3.2.2) and institutional (§3.2.3) means.

3.2.1 Ambiguous economic nature of data

Historically, the object of an IAD analysis concerns physical assets. Data however is different in the sense that it concerns intangible content (Purtova & Van Maanen, 2023), with deviating opinions regarding its economic value (Filgueiras & Silva, 2021). Personal health data is a broad concept, covering all information concerning one's health (§3.1.2). Consistent with data governance and economic literature, this thesis understands data as a digital representation of information (Purtova & Van Maanen, 2023).

In assessing the physical and material conditions of a good as part of an IAD framework analysis, it is common to assess the economic nature hereof. This entails examining the level of control regarding access hereto (excludability) and the extent to which one party's consumption limits availability to others (subtractability) (Polski & Ostrom, 2017). Accordingly, different kinds of goods are accompanied with different kinds of governance challenges, thus providing grips for how to govern the good or service. For example, common pool resources are inherently related to collective action problems (Ostrom, 1990). Appendix D performs such an analysis, but shows that personal health data cannot be clearly classified as one of the good archetypes; it has characteristics of multiple types. Therefore, these results are put aside. What we can learn from the analysis, however, is that data should not be studied as a standalone good, but as part of a larger system that comprises actors that need to collaborate via (technical) means to extract value from personal health data.

3.2.2 Data sharing infrastructures

According to Frischmann et al. (2014), the technical infrastructure is a resource necessary to extract value from data. Therefore, the infrastructure is part of the action situations and collective action dilemmas. To support data-driven policies, national and EU governments should govern infrastructures that facilitate health data collection, storage, and sharing (Filgueiras & Silva, 2021).

Health data information systems are characterised by heterogeneity because of the differences in local, regional and national data infrastructures (Pavlenko et al., 2020). For instance, hospital data infrastructures comprise various IT solutions, such as laboratory information systems, electronic health records, and other portals, databases and registries (Pavlenko et al., 2020; Writers collective Nictiz, VWS, VZVZ, ZN, 2023). Health data is mostly stored locally, under responsibility of healthcare providers (Writers collective Nictiz, VWS, VZVZ, ZN, 2023).

The Dutch government is increasingly investing in easing health data exchange for primary health care delivery, with the implementation of the Electronic Data Sharing in Health Care Act ('Wet elektronische gegevensuitwisseling in de zorg' or 'Wegiz') in July 2023 as prime example (Zorginstituut Nederland, 2023). This act stimulates and obliges data flows via standards for various types of exchanges (Zorginstituut Nederland, 2023). Contrary to some other EU Member States, the Netherlands have not centrally developed such a system (Verhoeven et al., 2021). Healthcare providers decide on the use of information systems for secondary use of health data (de Mul et al., 2012). As a result, health data are often siloed within the infrastructures of healthcare providers (DS1-H). Like primary use, secondary use of health data requires appropriate data infrastructures (Doutreligne et al., 2023; Pavlenko et al., 2020). Data warehouses, deployed by healthcare providers, for example, pool data from different information systems to more homogeneous formats for research, management and healthcare delivery purposes (Doutreligne et al., 2023). Such systems, provided by private or hybrid organisations, are increasingly used in the Dutch healthcare sector (de Mul et al., 2012; Health RI, 2022; Pavlenko et al., 2020).

Sharing health data within and between organisations for secondary use requires forms of standardisations (DS1-H). For example, electronic health record systems are not designed for research purposes, but to serve the needs of healthcare providers for delivering care (Lima et al., 2019). To enable data exchange, architectures, such as data warehouses, protocols and other methods are designed (Lima et al., 2019). First, to extract value from data by the parties involved there needs to be common understanding of what a data point means and how it is structured (DS1-H). Second, the underlying semantic standards should be clear. For instance, there are different systems for registering medical diagnoses, requiring harmonisation of definitions for the same diagnoses or lab values (DS1-H). A common data model standardises data structures of disparate datasets, after which standardised, analyses can be performed across datasets (Kent et al., 2021). As example, for secondary use of health data for scientific research, the Observational and Medical Outcomes Partnerships (OMOP) Common Data Model has been developed (DS1-H; Lima et al., 2019). The OMOP common data model structures all clinical and health system data and standardises its semantic representation to improve data interoperability between information systems (Kent et al., 2021). However, the existence of such standards does not mean they are adopted broadly: there are still challenges regarding understanding of data structure and semantic in research (DS1-H).

Human resources and capabilities are necessary to design data sharing policies and bring it to practice, by mobilising people to work with health data (Filgueiras & Silva, 2021). The relevant roles are further described in the §3.4.

3.2.3 Institutional means

In addition to technical resources, there are institutional means to share health data: data protection impact assessments (DPIA), different kinds of data-related agreements, and data management plans (DMP).

A DPIA is a risk analysis performed by organisations to analyse whether (new) types of processing of personal data imposes high risks for data subjects, especially when a new technology is used (General Data Protection Regulation, 2016, art. 35), such as synthetic data. The analysis is specifically important when the processing of personal data includes the large-scale processing of sensitive data (General Data Protection Regulation, 2016, art. 35). The analysis covers topics such as processing purposes, the necessity and proportionality of processing personal data, and mitigating risks for data subjects.

Data agreements ensure that the shared data is used only for its intended purpose under predefined conditions, in line with the obligations of the GDPR (van Bon-Martens & van Veen, 2019). Three types are data transfer agreements, where the researcher receives (pseudonymised) health data; data access agreements, where the researcher obtains access to health data while it remains on the healthcare provider's site; and material transfer agreements, regulating the tangible transfer of materials – and is therefore beyond the scope of this thesis (LC1-RI, LC-H; Nuffield Council on Bioethics, 2015; Smit et al., 2024)). Data agreements usually differ for different data sharing use cases (PO-H). In general, these define roles and responsibilities of involved parties; describe the data and processing purposes, methods of data sharing, and storage; and other measures to protect pseudonymised data (LC1-RI). Where Smit et al. (2024) identify such agreements as additional governance arrangements for processing health data for research purposes without consent, the interviews showed that such formal data agreements form an integral part of research collaborations, regardless of whether consent is obtained (DS1-H, DS2-H, LC1-RI, LC-H).

Lastly, to assure responsible data management, researchers are to record their risk assessments and data protection safeguards in a DMP (DS1-H, DS2-H, LC-H, PO-H, R-RI). Formulating DMPs requires researchers to think about how they want to process personal and non-personal data (Van Gend & Zuiderwijk, 2023), capturing how they comply with data protection principles (PO-H).

3.3 Community attributes

The community attributes refer to the community's shared understanding of policy activities and the extent to which their beliefs, values, and desired outcomes are homogenous (Polski & Ostrom, 2017).

There is a general consensus among actors that health data sharing is important for medical research, as data-driven research can improve healthcare delivery for patients – as long as it is shared in accordance with the data protection rules (DS1-H, DS2-H, PO-H, LC-H, LC1-RI, LC2-RI, R-RI, PM-HWS). Also, all interviewees believe that the process can be improved to increase this benefit. There are individual differences in the motivations for such improvements, focused on easing data sharing for researchers (DS1-H, LC-H), or improving the data protection of patients (PO-H, LC1-RI, LC2-RI). Regarding rules, both the healthcare provider and research institute confirmed that the Dutch approach regarding the definition of pseudonymised data lacks guidance and is strict (DS1-H, LC-H, PO-H, LC2-H). Considering this common understanding, stakeholders are willing to rethink current data sharing practices to address deficiencies (DS1-H, DS2-H, PO-H, LC-H, LC1-RI, LC2-RI, R-RI).

To understand the position of researchers of healthcare providers, it is pointed out that storing data is a very labour-intensive process, requiring big investments that researchers want to protect and get the most out of themselves (DS1-H, PO-H). Therefore, a natural reaction from these researchers is to maximise value of these data before publishing or sharing it (DS1-H). However, from the organisation's perspective, this reduces the use of health data: health organisations can benefit from the outcomes of the data-driven research. There is insufficient expertise within healthcare providers to make maximum use of data (DS1-H, DS2-H).

3.4 Evaluative criteria

To evaluate the interaction patterns in the action arenas (§3.5), evaluative criteria determine what aspects of the outcomes are (un)satisfactory (Ostrom, 2011). Overall, similar criteria apply to the actors involved in the action arenas, driven by other rationales as shown for the community attributes (§3.3). This section formulates these criteria based on the scope rules.

From a policy perspective, ensuring that the richness of data in healthcare systems is put to optimal use is considered important by the Dutch Ministry of Health, Welfare and Sports (HWS), that underlines its utility for improving healthcare (Kamerbrief Visie En Strategie

Secundair Datagebruik, 2023). One of the policy objectives is enlarging the knowledge base for researchers by making health data available in a digital and standardised manner (Kamerbrief Visie En Strategie Secundair Datagebruik, 2023). This policy builds on three pillars: increasing data availability, increasing trust via high-quality health data and privacy-enhancing use of health data, and facilitating central control over health data for patients (Kamerbrief Visie En Strategie Secundair Datagebruik, 2023). The policy objectives are based on an EU level; national level; and network level, i.e. in consultation with the playing field, which includes healthcare providers.Accordingly, the overall policy objective relevant to this thesis is the stimulation of secondary use of health data while safeguarding privacy of patients, as presented by the Ministry of HWS (Kamerbrief Visie En Strategie Secundair Datagebruik, 2023; Writers collective Nictiz, VWS, VZVZ, ZN, 2023). Subsequently, informed by the interviews, criteria used to evaluate the outcomes are availability of data (DS1-H, DS2-H, PO-H, LC-H, LC1-RI, R-RI)via efficient use of resources; safeguarding protection of patients' personal data (DS1-H, PO-H, LC1-RI, LC2-RI); and data findability (DS1-H, DS2-H, R-RI).

3.5 Interaction patterns and outcomes

This section integrates the contextual analysis of health data sharing information into explanations of the behaviour of actors. Central components of the IAD framework are the 'action arenas', also referred to as 'action situations' in other works of Ostrom. The objective of this section is to understand how actors' behaviour generates certain interaction patterns. The action arena can be explained with knowledge of the set of actors, their roles and level of participation, their possible actions, the level of control they can exercise in comparison to other actors, and eventually, how this results in certain outcomes (Polski & Ostrom, 2017). Decisions made in the action arena may concern operational decisions that affect day-to-day decision-making (Polski & Ostrom, 2017), or decisions at higher abstraction levels, for example, the definition of policies (Ostrom, 2011).

Informed by interviews and the previous analysis, decisions in three action arenas are highlighted (Table 4). The identified action arenas are operational decisions in health data sharing (§3.5.1), decisions in defining personal data (§3.5.2), and decisions in the legal base of health data sharing for research (§3.5.3). The interaction patterns naturally result in certain outcomes that indicate the performance of policy systems (Polski & Ostrom, 2017). Table 4 summarises the characteristics of the action arenas that are discussed hereafter in turn.

Table 4. Overview of action arenas

Overview of action arenas
Action arena 1: Operational decisions in health data sharing
• ***Decision-making level***: between organisations • ***Decisions***: operational decisions are made regarding health data sharing activities. This comprises initiating data sharing, the interactions to obtain access to health data, and the factual sharing of health data. • ***Actors***: healthcare providers and research institutes • ***Impact***: decisions impact whether external individuals can access health data for research.
Action arena 2: Decisions in definition of personal data
• ***Decision-making level***: EU and Dutch implementation • ***Decisions***: discussions in case law and guidance regarding the material scope of data protection law: definitions of personal data, pseudonymisation, and anonymisation. • ***Actors***: the discussion lies mainly with policymakers, such as the Ministry of HWS, who give substance to the data protection rules. • ***Impact***: the extent to which healthcare providers and research institutes are subject to data protection laws.
Action arena 3: Decisions in legal base for health data sharing
• ***Decision-making level***: EU and Dutch implementation • ***Decisions***: the rules in use showed there are various legal bases applicable for secondary use of health data for research. There are discussions regarding the appropriate legal base. • ***Actors***: the Ministry of HWS, who supports in the implementation of (EU) data protection regulations and provides guidance for its interpretation. Also, healthcare providers and collectives thereof, make their own decisions in choosing a legal base, based on the available information provided by policymakers. • ***Outcomes***: impact how healthcare providers and research institutes must set up their data sharing process.

3.5.1 Action arena 1: Operational decisions in health data sharing

This action arena describes decisions made in health data sharing practice: 1) the process starts with the initiation of health data sharing, is followed by 2) an internal data protection assessment, 3) the conclusion of data agreements, and 4) the factual sharing of data.

1. Interactions in health data sharing request

Sharing health data across organisations is often an ad-hoc process: opportunities of secondary use of health data arise within networks of physicians and researchers that are connected to healthcare providers and research institutes (DS1-H, DS2-H, R-RI1). There is no single entry point where third-party researchers can view what data is available and find how to access it (DS1-H, DS2-H, R-RI; Veen & Verheij, 2023). Data is often shared as part of research collaborations between researchers, or between physicians within healthcare providers with external researchers (DS1-H, DS2-H, R-RI). Therefore, data sharing agreements are part of broader research agreements, driven by mutual benefit (LC-H), as will be discussed under 4.

2. Internal handling of a request

Before concluding data sharing agreements, researchers from healthcare providers as well as the research institute have to account for privacy risks (DS1-H, PO-H, R-RI). This process is followed for each data sharing activity (DS1-H, DS2-H, LC2-H, PO-H, LC1-RI), and particularly focused on compliance with the GDPR and, if necessary, the Dutch law on medical research with human participants ('Wet medisch-wetenschappelijk onderzoek met mensen' or 'WMO') (DS1-H, LC-H, PO-H). For this process, researchers are ought to perform a DPIA when there are high privacy risks; show compliance with the data protection principles, such as compatibility with data processing purposes, selecting a legal base, securing data, and minimising data; and define to whom it is disclosed (PO-H, LC-H, LC1-RI, LC2-RI; van Bon-Martens & van Veen, 2019). Moreover, researchers must document the informed consent procedure, if applicable, and create a DMP (DS1-H, DS2-H, R-RI).

3. Interactions in health data sharing process

When researchers decide to share data, the data collector (healthcare provider) and data recipient (research institute) must agree on data sharing conditions in a data sharing agreement (Smit et al., 2024) (§3.2.3). The actors involved, the decisions they can take, and their control are summarised in Figure 7.[4] The arrows represent the actions taken or decisions made by the actors, and how these contribute to health data sharing. The combinations of these decisions result in the interaction patterns.

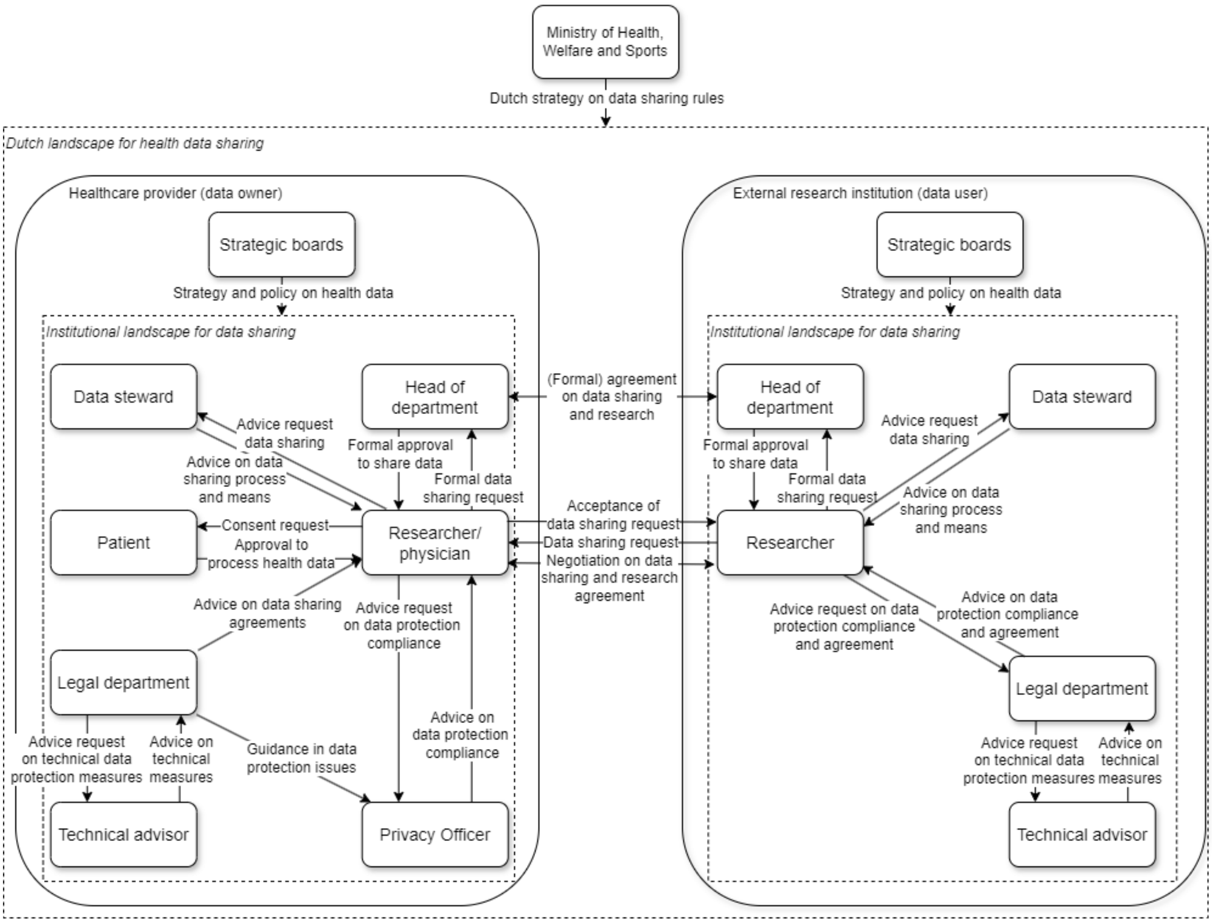
To support researchers in health data sharing, both in the internal assessment and the data sharing agreement, healthcare providers and research institutes appointed data stewards. Their primary goal is to facilitate data to flow in a way that complies with laws, regulations and organisational policies. Data stewards support researchers in compliant re-use of health data, partly by calling in the right experts at the right time and advising on use of technical infrastructures to store, share and analyse data (DS1-H, DS2-H).

The privacy teams of healthcare providers and research institutes advise on all kinds of privacy-related issues. In the use case. the healthcare provider offered two lines of privacy support. The first point of contact for researchers are privacy officers, who support *inter alia* in conducting DPIAs, dealing with data breaches, and establishing simple data sharing agreements (DS1-H, LC-H, PO-H). Legal counsellors, part of the legal affairs department, form the second line of support (LC-H). They evaluate the content of data sharing agreements (§3.2.3), especially when they deviate from standard models. Legal counsellors support first-line privacy officers by providing them with guidance and standards (LC-H). Legal counsellors also support the board of director with formulating its privacy strategy and organisational data

[4] Please note that the Data Protection Officer is not included in this figure. This independent actor oversees internal GDPR compliance and ensures that the organisation's privacy maturity level is adequate. Although this actor is mentioned by LC2-RI, and indeed, has an important role in data protection within organisations, the position of this actor in the health data sharing process is less prevalent.

policies (DS1-H, DS2-H, PO-H, LC2-RI). The document passes by the actors upon request, sometimes more often after iterations. If those involved cannot agree on health data sharing terms, the process is aborted (LC-H).

Figure 7. Overview of institutional environment of health data sharing


Ministry of Health, Welfare and Sports
Dutch strategy on data sharing rules
Dutch landscape for health data sharing
Healthcare provider (data owner)
Strategic boards
Strategy and policy on health data
Institutional landscape for data sharing
Data steward
Head of department
Advice request data sharing
Advice on data sharing process and means
Formal approval to share data
Formal data sharing request
Patient
Consent request
Approval to process health data
Researcher/ physician
Advice on data sharing agreements
Advice request on data protection compliance
Legal department
Advice request on technical data protection measures
Advice on technical measures
Guidance in data protection issues
Advice on data protection compliance
Technical advisor
Privacy Officer
(Formal) agreement on data sharing and research
Acceptance of data sharing request
Data sharing request
Negotiation on data sharing and research agreement
External research institution (data user)
Strategic boards
Strategy and policy on health data
Institutional landscape for data sharing
Head of department
Data steward
Advice request data sharing
Formal approval to share data
Formal data sharing request
Advice on data sharing process and means
Researcher
Advice request on data protection compliance and agreement
Advice on data protection compliance and agreement
Legal department
Advice request on technical data protection measures
Advice on technical measures
Technical advisor

Upon request of legal counsellors, the security team can advise on the technical measures in data sharing agreements (DS2-H, LC-H, LC2-RI). One of the technical measures required by the GDPR (§3.1.2-3.1.3) is to pseudonymise data (General Data Protection Regulation, 2016, art. 89(1)). The researchers issuing health data are responsible for pseudonymising data (DS2-H). In practice, this usually means removing personal identifiers (DS1-H, LC1-RI).

Regarding formal control, however, these experts only have an advisory role: researchers, and ultimately department heads, remain responsible. Researchers are not authorised to sign legal agreements, including data sharing agreements (as shown in Figure 7); this competence is granted exclusively to heads of faculties (for universities) or principal investigators (for research teams within healthcare providers) (DS1-H, LC-H, LC1-RI, LC2-RI). In practice, the researcher conducting the research is responsible for preparing the relevant documentation and when having obtained approval for it, facilitating data sharing. For larger contracts, the department head is authorised to sign; they have most control (LC-RI).

4. Interactions in sharing health data

Data may be shared via approved IT platforms, for example shared cloud folders by industry parties or secure initiatives from public parties from the healthcare sector (DS2-H). Digital research environment myDRE, for example, focuses on enabling secure data sharing and analysis specifically for research ('myDRE platform', n.d.). This platform is developed collaboratively by some Dutch university medical centres (Radboud UMC, 2019). This example illustrates that the methods for data sharing are diverse, provided by different types of actors.

In current health data sharing processes, there are no formal compensation models for secondary use of health data (DS1-H); this has yet to be discussed by actors in health research. For example, there are not necessarily monetary pay-off rules. Actors are paid-off in terms of (use of) research outputs and distribute publication rights and authorship attribution of the research outputs (DS1-H, DS2-H, PO-H, R-RI).

Evaluation of outcomes

These interaction patterns show that the process of obtaining access to health data for research is both time-consuming and labour intensive, as articulated by the actors themselves (DS1-H, DS2-H, PO-H, LC-H, LC1-RI, LC2-RI, R-RI).

First, there is inefficient use of technical and human resources. There are many kinds of actors involved, on the side of the data collector and the data recipients, all with their own expertise domain. To reach an agreement, the document has to pass back and forth between the data collector and data recipient several times; a process that takes months when the provisions deviate from the templates (LC-H, R-RI). A healthcare provider's data steward explains this outcome clearly: "*Currently, a lot of data is not shared because concluding agreements is too big of a hassle. Moreover, too much data is shared in a way that is not secure enough*" (DS2-H). This quote reflects most of the interviewees' opinions regarding the current status of the health data sharing process; the current process does not sufficiently facilitate health data sharing, and in addition, the data is insufficiently protected.

Second, the process does not fully support the objective of protecting personal data of patients (LC1-RI). Or, how a healthcare provider's privacy officer put it: "*Sharing health data is crucial: you cannot conduct research without it. Yet, it has to be done safely and responsibly, and that is where the process sometimes falls short*" (PO-H). The clauses on data protection move towards issues to prevent liability; whereas the patient perspective should be most important (PO-H, LC1-RI, LC2-RI). After signing the agreements, the involved parties do not re-evaluate them despite actual changes in the collaboration, nor do they monitor each other's compliance (LC-H, LC1-RI). The focus on legal compliance, rather than patient privacy, may be explained by the trend of increasing data availability: making data more available also implies that researchers are more cautious in sharing data out of fear for non-compliance; the consequences then affect more patients and extend to more organisations (DS2-H). What they forget, however, is that adherence to the GDPR is not an issue of compliance or non-compliance (§3.1.1). Conform the GDPR, data sharing requires a balance of the research interests and privacy risks, and appropriate mitigating measures (DS1-H, LC2-H).

Third, the findability of health data is limited for research institutes. A reason is that collecting data is a very labour-intensive process that requires large investments from researchers. Therefore, researchers may be hesitant in sharing data with third parties, before they have maximised their private gains (DS1-H), for example in terms of scientific publications. Another reason is the lack of support from national governments in providing an appropriate technical infrastructure (Verhoeven et al., 2021), essentially trapping valuable health data inside researchers' social professional networks, e.g. hampering discovery of data held by someone that a researcher is not acquainted with. The lack of a central technical infrastructure also resulted in different data governance models and repositories by healthcare providers (Verhoeven et al., 2021). The question arises who should be responsible for creating such an infrastructure and how to compensate for the data collection costs (DS1-H).

3.5.2 Action arena 2: Decisions in definition of personal data

This section discusses the implications of the ongoing debate regarding the scope of data protection law, by analysing the decisions regarding the definition of personal data, pseudonymisation, and anonymisation and concludes with an evaluation of outcomes regarding the implications of the definitions for data availability and data protection.

Interactions regarding interpretation of data protection law

As discussed in action arena 1, primarily pseudonymised health data is shared for research. The analysis of the rules-in-use showed that the concepts of personal data, pseudonymisation, and anonymisation are in development at EU level (§3.1.2). In short, data is considered as anonymous when data subjects cannot reasonably and likely be identified (referred to as the 'reasonableness standard'), considering the state of the art of re-identification technologies (General Data Protection Regulation, 2016, rec 26).

Regarding anonymisation, there is discussion on how the concepts of "reasonably" and "likely" should be operationalised. In the EU's attempt to account for new developments in re-identification tools, it has deliberately chosen to leave those concepts open in the GDPR (Groos & Van Veen, 2020). However this has led to different interpretations by EU member states of what constitutes such a tool; for example, some DPAs presume that anonymisation as referred to in the GDPR is not possible while maintaining value of health data for research, whereas others do (Hansen et al., 2021). Furthermore, some member states uphold a more pragmatic approach (Groos & Van Veen, 2020), investigating what means an actor can actually use to re-identify data subjects in consideration of contextual factors. Other member states, among which the Netherlands, follow a stricter, more hypothetical approach, taking into account all means that could possibly be used to re-identify data subjects. The various operationalisations of anonymisation and pseudonymisation among EU member states (Vukovic et al., 2022) spurred a debate in EU case law that remains ongoing (*Breyer*, 2016; *Single Resolution Board v European Data Protection Supervisor*, 2023).

This approach is important for the question whether data that has undergone pseudonymisation, can be considered as anonymous for the recipient. The Dutch government strictly interpret pseudonymisation, which in practice means that as long as long as there exists an identification key "somewhere in the world" that can identify the data subjects, pseudonymised data remains personal (DS1-H, LC2-RI; Hansen et al., 2021). Unlike the Netherlands, the United Kingdom, for example, applies a less strict interpretation, which amounts to data being anonymous if the recipient does not and cannot reasonably access the key (DS1-H). Translating this to the context of secondary use of health data, means that when healthcare providers have pseudonymised health data in a way that does not enable the research institute to reasonably and likely re-identify data subjects, the research institute may receive health data without data protection regulations applying.

What complicates the application of anonymisation and pseudonymisation rules, is that for healthcare data specifically, it is very difficult to reduce re-identification risks to the GDPR standards. Health datasets can be very comprehensive which means that without direct identifiers, the possibility remains that data subjects can be reidentified (Kroes, 2023).

Evaluation of outcomes

Action arena 2 showed how interactions between regulators, policymakers, and courts at EU and national level bring about a dynamic landscape regarding the scope of data protection law.

Purely from a data protection point of view, a broad definition of pseudonymisation and thus a broad material scope of the GDPR is preferable. However, from the policy objective of stimulating research, a stricter definition would increase data availability for research, by easing the health data sharing process. Implications of adopting a narrower reach of pseudonymous data (or a wider scope for anonymisation) in practice should be further studied, as it could significantly change areas where pseudonymous data is often shared, such as medical research. The extent to which rules are imposed on pseudonymised data influences the effort required for healthcare data exchange. Consequently, so does the demand for techniques that enable this, such as synthetic data generation (DS1-H).

This thesis takes the stance that the current hypothetical approach regarding anonymisation is undesired, as it disproportionately restricts secondary use of health data. The means reasonably and likely to be used by the recipient of anonymous data should be considered, and not all hypothetical scenarios. To compensate data subjects, data collectors and recipients can deliberately choose to incorporate data protection safeguards for anonymised data in health research as well. An organisational safeguard could be assessing whether the intended research purpose is compatible with the patient's initial consent or evaluating the research's contribution to societal interests (Groos & Van Veen, 2020). Also, technical safeguards, such as state-of-the-art anonymisation techniques, should be implemented. Moreover, an institutional safeguard such as research agreements could legally prohibit attempts by data recipients to re-identify data subjects from anonymous datasets.

Regardless of the outcomes of this discussion, what is worse about these interactions in current health data sharing practice, is that the ambiguous legal definitions and interpretations thereof, both at an EU and national level, cause uncertainties for healthcare providers and research institutes. For example, the lack of a harmonised and clear approach results in the fact that researchers are often unaware of what is considered anonymous or pseudonymous data, leading to personal data being wrongly classified as anonymous (LC1-RI, LC2-RI). This is problematic, as this also prevents the necessary application of data protection safeguards. The dissatisfaction is reflected, for example, in universities' initiative to discuss how they define identifiability (LC2-RI).

The Ministry of HWS may resolve these uncertainties by providing guidance on pseudonymisation and anonymisation in health research. Neither this Ministry nor the Dutch DPA have undertaken efforts to further conceptualise the identifiability criteria in a health research setting, which could result organisations using the strictest possible interpretation to make sure they operate compliantly (LC2-RI, PO-H) – in leading to (unnecessarily) restricting data sharing. A new or clearer approach is hence desired, and, going off of the judgement in (*Single Resolution Board v European Data Protection Supervisor*, 2023)*,* even required, as the hypothetical identifiability criterium has had its day.

3.5.3 Action arena 3: Decisions in legal base for health data sharing

The analysis of the rules in use showed that various legal bases exist for secondary use of health data for research. To determine the applicable legal base, healthcare providers and research institutes in the Netherlands have to look into three acts that simultaneously apply: the GDPR, the Dutch GDPR implementation act (UAVG), and the Dutch Civil Code regarding medical treatment (WGBO). This action arena describes the challenges and uncertainties that arise from the Dutch consent-by-default approach.

Interactions of rules with actors: a consent-by-default approach

As *lex specialis,* the WGBO should be consulted first (Hansen et al., 2021). To ensure medical confidentiality, the WGBO regulates under what conditions the physician may disclose health data. The WGBO applies to the relationship of physicians and patients, but extends to sharing electronic health records, as these records are a product of primary healthcare delivery. According to the WGBO, researchers should first obtain consent from patients for secondary use of health data, unless this is unfeasible or impossible (§3.1.3). The WGBO only mentions medical confidentiality; it does not provide the legal base for secondary use of health data (consent in WGBO does not equal consent in GDPR) (Hansen et al., 2021).

For secondary use of health data by researchers within one data controlling organisation (i.e. the healthcare provider), researchers may presumably rely on the presumption of compatibility (Hansen et al., 2021) (General Data Protection Regulation, 2016, art. 5(1)(b)). However, when health data is shared with external research institutes, these require their own legal base as data controller. The UAVG links the WGBO's exception to medical confidentiality for research to the GDPR, stating that explicit consent should be obtained, unless one of the exceptions apply – which are similar to the exceptions in the WGBO. If consent cannot be asked, researchers can rely on the legal base of scientific research, Art. 9(2)(j) GDPR. Here, the Netherlands implemented an additional condition that research with health data should serve the public interest (Uitvoeringswet Algemene Verordening Gegevensbescherming, 2018, art. 24(b)). This thesis refers to this as the 'consent-by-default' approach.

With this consent-by-default approach, the Netherlands restricts health data sharing (DS1-H, DS2-H, PO-H, LC-H, LC2-RI), deviating from other EU member states, such as France, Finland, or Denmark (Veen & Verheij, 2023). Advocating for this approach is that it directly empowers data subjects to decide whether their health data can be processed for research. Consent is even the only legal base that allows data subjects to exercise control (Lynskey, 2016). From a legal perspective, the consent-by-default approach is problematic when viewed in relation to the GDPR. The Dutch government seems to ignore the other research exceptions of the GDPR, and therefore does not seize all the possibilities offered by the EU framework to promote research (Mostert et al., 2018). Outcomes are (unfairly) high administrative burdens for researchers (Mostert et al., 2016). This strict interpretation of consent is accompanied by a lack of guidance regarding the application of other legal bases (Veen & Verheij, 2023). Regarding the principle of purpose limitation, for example, it is unclear how the presumption of compatibility applies when health data is shared with external research organisations and whether they need an additional legal base (Hansen et al., 2021).

Reaction to consent-by-default approach

To further challenge the consent-by-default approach, the healthcare sector has expressed their concerns regarding this strict approach. They confirm that many of the problems related to health data access are caused by misunderstandings about the application of data protection laws, advocating for a reorientation of the WGBO and UAVG principles (Writers collective Nictiz, VWS, VZVZ, ZN, 2023). The Ministry of HWS will deliver policies to reinterpret the legal bases for health data exchange in the healthcare sector, focusing on when consent can be invoked lawfully and when other grounds may apply (Writers collective Nictiz, VWS, VZVZ, ZN, 2023). It remains to be seen when the Ministry of HWS actually delivers the policy, considering they are often long overdue (LC-H).

As an answer to this uncertainty, important actors in the healthcare sector, including (academic) hospitals, have joined forces to create uniformity about the application of regulatory concepts (Writers collective Nictiz, VWS, VZVZ, ZN, 2023). In such self-governance initiatives, the most common adaptations of consent are models that shift away from specific consent, such as 'broad consent' covering a broad range of future data uses (Mostert et al., 2018; General Data Protection Regulation, 2016; Veen & Verheij, 2023; Writers collective Nictiz, VWS, VZVZ, ZN, 2023, rec 33). One example is the non-binding Code of Conduct developed by COREON, that participants of health data sharing usually adhere to – at least to some extent (DS2-H).[5] However, as Appendix E discusses, this Code conflicts with the GDPR for too liberally interpreting patients' consent, as the Dutch DPA has expressed (Veen & Verheij, 2023). This highlights that compliance with industrial self governance initiatives does not guarantee legal compliance.

Yet, who is to blame? Ostrom's design principles for self-governance prescribe the need for clear rules about system boundaries and who is entitled to which actions (Ostrom, 1990). The Dutch consent-by-default approach fails to achieve this due to the ambiguous legal concepts. The lack of clear rules constrains the sector to device a workable and appropriate solution through self-governance.

Evaluation of outcomes

The multitude of applicable legal bases coupled with the possibility for national legislators to introduce additional rules for processing health data and for processing personal data for research purposes, resulted in a fragmented regulatory framework for the secondary use of health data (Hansen et al., 2021) (§3.1.3).

At the EU-level, the fragmentation poses challenges for cross-border data sharing between organisations: the administrative burdens associated with locating and adhering to applicable regulations are high (Hansen et al., 2021). At the Dutch level, the discretion for national legislators to implement further conditions has led to a consent-by-default approach – in practice raising consent fatigue among patients and researchers (Veen & Verheij, 2023). This consent fatigue among researchers can best be illustrated by this statement from a healthcare provider's privacy officer: "*You notice that people sometimes use exceptions to consent, even when you think it doesn't feel right*" (PO-H). It shows how researchers try to find

[5] COREON is a professional network of parties involved with medical research, such as academic hospitals, universities and other research institutes ('About Coreon: purpose and mission', n.d.).

ways to avoid requesting consent, when in fact the Dutch consent-by-default approach prohibits this. Hence, whilst the GDPR tries to stimulate the data flow for scientific research, the mere acknowledgement of research does not guarantee desired outcomes (Vukovic et al., 2022).

3.6 Conclusion

This chapter analysed the institutional data protection environment of secondary use of health data for research using Ostrom's IAD framework. The objective was to identify the interaction patterns that lead to challenges in data protection, data availability, and data findability in health data sharing.

The actors involved in defining the institutional environment were mapped. The health data sharing for research is governed by multiple actors (healthcare providers, governmental organisations, and research institutes) at multiple levels (organisational, national, and EU). Rules are found in both legislation and organisational policies, affecting both the health institution where the data is held, and the research institute willing to reuse the data. In both organisations, there is a strong role for the researcher or physician (to coordinate which data will be shared, to share the actual data, and to obtain approval for the sharing with their head of department). But organisational support staff and superiors also play a large role. These are in both organisations the heads of department (to conclude a data sharing agreement); data steward (to provide advice); and privacy officer, legal counsel and security advisor (to provide legal and technical advice, e.g. to what extent a technical measure can contribute to legal compliance). On the data collector's side, the patient plays an important role in giving consent. All these actors are embedded in the Dutch landscape for health data sharing, covered by formal rules of the GDPR, UAVG and WGBO.

Three action arenas surfaced in the research. First, at a network level, a challenge emerges from interaction patterns in the data sharing process. The conclusion of data agreements, the main instrument to facilitate data access agreements concerns a time-consuming process and the level of data protection leaves something to be desired. Moreover, data findability is an issue due to a lack of (central) infrastructures that can disclose available data to (external) parties. Second, at the national level, challenges emerge from the rules regarding the broad definition of personal data, specifically regarding pseudonymisation and anonymisation. Health institutions and data recipients, or the Netherlands in general, use a very strict definition of anonymous data and provide little guidance on when data can be considered anonymous, in consideration of the identification risks. Third, also at the national level, challenges emerge from the Dutch consent approach. EU rules that should ease secondary use of health data are 'ignored' by the Dutch rules on medical confidentiality, requiring consent by patients where possible. This resulted in unanimous disagreement among actors involved in health data sharing for research.

The institutional data protection analysis in chapter 3 showed the main challenges in health data sharing for research. Overall, the available mechanisms to share health data, along with uncertainties regarding the EU and Dutch approach towards anonymisation and consent, hamper health data sharing for research.

4
Synthetic data generation and evaluation

The institutional analysis in chapter 4 identified the challenges that follow from interactions between actors in health data sharing for research. The premise of this thesis is that technical solutions should be further exploited to solve health data sharing issues, enabling data to flow while protecting patients' data. Synthetic data generation is analysed as such a privacy-enhancing technology. Before we can say anything about how synthetic data may enable health data sharing for research, the concept of synthetic data generation and how it contributes to data protection should be further explored. This chapter aims to answer the following research question:

> *RQ2 How could synthetic data generation and evaluation contribute to personal data protection?*

We have adopted the following definition of synthetic data: "synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)" (Jordon et al., 2022). Via an academic literature review (methodology in §2.4.2 and overview in Appendix F), this chapter first studies what data-related tasks synthetic data generation is up for by introducing its application and main characteristics (§4.1). This thesis focuses specifically on synthetic EHRs as challenging use case that represents one of the complex (mixed) data types (§1.2.2); second, it explains the various synthetic data generation models for electronic health records (§4.2) as well as their limitations (§4.3); lastly, the various privacy measures are presented and how technology developers interpret these (§4.4).

4.1 Introduction to synthetic data generation

This section explores the concept of synthetic data generation, by discussing how synthetic data can be applied and placing it in the context of (traditional) anonymisation techniques (§4.1.1). Also, core characteristics of synthetic data are discussed, focusing on the data types used to train generation models and produced by the generation models (§4.1.2).

4.1.1 Application of synthetic data generation in healthcare

Any new solution should be validated extensively on effectives and representation of reality, especially in healthcare, as application can highly impact patients, for example with regards to treatments or diagnostic tooling (Murtaza et al., 2023). This requires vast amounts of health data, which is not always available to researchers (Dove & Phillips, 2015; Murtaza et al., 2023). As shown in Chapter 1, privacy issues arise in data sharing scenarios with external research institutes (Hernandez et al., 2022), and are therefore strictly regulated (Dove & Phillips, 2015). The need for personal data sharing steered the development of privacy-enhancing technologies (Alloza et al., 2023). Earlier efforts to prevent re-identification while upholding statistical characteristics of a dataset focused on data masking and anonymisation techniques (Murtaza et al., 2023). In attempts to reduce identification risks and preserve data utility after data anonymisation, these earlier techniques did not seem to find the right balance (Hernandez et al., 2022) – transformed datasets compromise on truthfulness and fail to capture the complexity of the original datasets, thus providing limited value for more complex analyses (Murtaza et al., 2023; Pawar et al., 2018). In addition, the transformed data remains vulnerable for privacy risks, failing to meet its privacy objectives (Pawar et al., 2018).

In recent years, synthetic data generation has been proposed as a new approach to anonymisation (Murtaza et al., 2023).[6] In synthetic data generation, a dataset is generated that fits to the original dataset (Hernandez et al., 2022). Synthetic data generation models are trained to 'resemble' real-world data, showing a similar structure in terms of distribution shape, variance, and correlations between variables (Hernandez et al., 2022; Murtaza et al., 2023). Yet, the approach demonstrates improved resilience to privacy attacks (Bellovin et al., 2019).

The process of generating synthetic data generally consists of four main steps (Figure 8). Based on real-world input data, synthetic data can be generated. The generation process can be subdivided as follows: first, the real-world is stripped of (direct) identifiers; second, a model is trained based on this real-world data; third, real-world input data can be transformed into synthetic data (Yoon et al., 2023). Then, the synthetic data is evaluated based on utility, resemblance and privacy (Jadon & Kumar, 2023). When the evaluation shows acceptable privacy risks, the synthetic data can be applied for the desired purposes (Yoon et al., 2023).

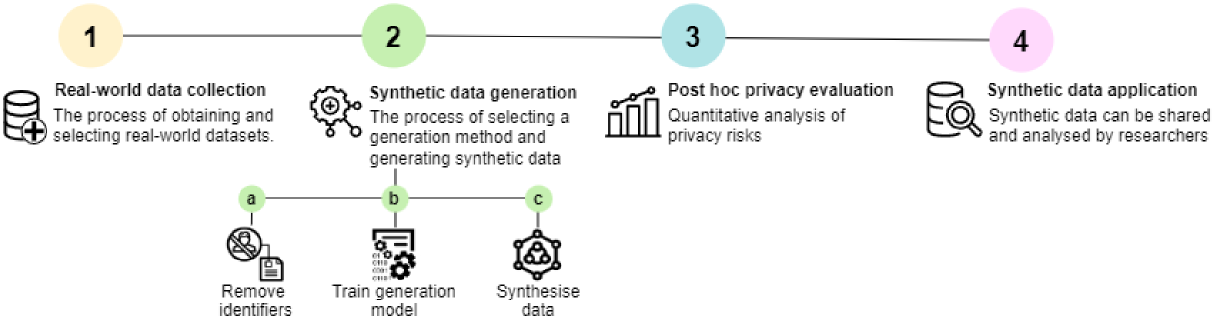

1
Real-world data collection
The process of obtaining and selecting real-world datasets.
2
Synthetic data generation
The process of selecting a generation method and generating synthetic data
3
Post hoc privacy evaluation
Quantitative analysis of privacy risks
4
Synthetic data application
Synthetic data can be shared and analysed by researchers
a
b
c
Remove identifiers
Train generation model
Synthesise data

Figure 8. Process of applying synthetic health data (based on Yoon et al., 2023)

Looking ahead to phase four, synthetic data can be applied for various purposes in a health care (research) context Figure 8. It can be used for data augmentation, i.e. to complement imbalance or scarce datasets (Jadon & Kumar, 2023). This way, AI models can be trained for all desired (yet unavailable) scenario's (Hernandez et al., 2022), increasing its generalisability (Jadon & Kumar, 2023). Moreover, synthetic data can be used to test a model that has been trained on real-world data (Hernandez et al., 2022). In a cross-organisational context, synthetic data generation can be used as a privacy preserving technology to avoid sharing sensitive data (Hernandez et al., 2022). From a data protection point of view, the advantage of synthetic data generation over other anonymisation methods, is that it does not contain data from the original dataset (Hernandez et al., 2022). Therefore, synthetic data can be used to share health data with research institutions, for example in epidemiological research which is concerned with analysing the distribution and determinants of diseases, or clinical trials, to simulate populations and outcomes to improve trial designs (Jadon & Kumar, 2023). As privacy-enhancing technology, synthetic data can also be applied to medical training: students that interact with health data, such as medical or computer science students, can use synthetic patient cases instead of using real patient data, which often happens in practice (Jadon & Kumar, 2023; Wiedekopf et al., 2021).

This thesis studies synthetic data as a privacy-enhancing technology to enable health data sharing for research between organisations.

[6] Next to synthetic data generation, there is another AI-based approach that gathers specific attention in research: federated learning (Kamel Boulos et al., 2022; P. Zhang & Kamel Boulos, 2022). Whereas with synthetic data generation manipulated datasets are shared, federated learning allows to share aggregated data based on real-world data analysis of local models (P. Zhang & Kamel Boulos, 2022). An argument to study synthetic data generation over federated learning, is that practice shows that synthetic data generation can be applied more efficiently in terms of implementation time, effectively, regarding privacy metrics, and on a more detailed level (Azizi et al., 2023).

4.1.2 Characteristics of synthetic data generation models

This section further explores synthetic data generation models by characterising them on the information used to generate data, and the detail-level of data produced by generation models.

Comparable to Wiedekopf et al. (2021), Murtaza et al. (2023) distinguishes three types of synthetic data generation models: models developed through real-world data, models developed through expert knowledge, and models that combine both approaches. This thesis focuses on data-driven models only, as these are associated with data protection issues (Wiedekopf et al., 2021). Knowledge-driven approaches are based on general theories and frameworks, and therefore, are not necessarily based on patient data (Murtaza et al., 2023; Wiedekopf et al., 2021).

Regardless of the generation method, the quality of synthetic data is determined by the appropriateness and quality of input data (Jadon & Kumar, 2023; Murtaza et al., 2023). The level of precision vary strongly per dataset depending on the data source (referred to as ‘granularity’). On the one hand, the value for research can be maximised with “fine-grained" datasets, yet, these datasets are also more vulnerable to privacy attacks as they contain more details about individuals’ health status (Murtaza et al., 2023). In line with the problem central to this thesis, Murtaza et al. (2023) found that most of the proposed synthetic data generation models are based on public datasets, due to limited access to private datasets and fewer privacy risks of generated data. Public datasets are, with some exceptions, anonymised. In these public datasets, full anonymisation is obtained by taking snapshots of health data, for example, by including only certain parts of EHRs. Another means for anonymisation is aggregation, for example, the object of observation can be aggregated from individuals to groups of individuals, or the observation can be aggregated from detailed descriptions of observations to the presence of certain general observations in binary format (Murtaza et al., 2023). Aggregated data can cover a wider range of a patient’s health status in comparison to snapshots, however, compromises sequential relations between variables. Longitudinal data include most details of a patient’s health status over time, such as EHRs, but, also comes with most privacy risks. Therefore, access to such datasets is often restricted for developing synthetic data generation models.

Understanding of these granularity levels will help to evaluate the utility and privacy levels of synthetic data generation models in the upcoming paragraphs. Depending on the purposes of application, a different level of granularity suffices to develop a high-quality synthetic data generation model. (Murtaza et al., 2023).

4.2 Generation of synthetic electronic health records

On the one hand, synthetic EHRs should resemble real-world data. On the other, they need to differ sufficiently to avoid an unacceptable level of information disclosure (Yale et al., 2020). Researchers have proposed classical approaches using, relatively, simpler statistical methods, such as replacing values based on correlation, noise injunction, Bayesian networks, or a collection of data distributions to estimate the probability density function (Hernandez et al., 2022; Murtaza et al., 2023; Theodorou et al., 2023; Thomas et al., 2022). Other, simpler versions of ML models to generate synthetic data, are decision trees. For example, Braddon et al. (2023) apply classification and regression trees to generate synthetic data that captures the prenatal epidemiological associations. Although such statistical methods often proof privacy-preserving, their application to temporal and high-dimensional data is restricted in terms of resemblance and utility (Theodorou et al., 2023).

In recent years, synthetic EHRs have evolved from static patient data to longitudinal timeseries (J. Li et al., 2023; Z. Zhang et al., 2022). Longitudinal EHRs are specifically valuable, as these patient journeys that describe health conditions enable new applications to disease progression (Yan et al., 2022). The literature review showed that these EHRs are generated by AI-based models. This section discusses the different generation methods for such synthetic EHRs. From the numerous synthetic data generation methods that have been proposed, GANs have gained substantial part of attention (Yan et al., 2022), and are therefore explained in more detail (§4.2.1). Other, more recent proposals also include variational autoencoders (§4.2.2) and large language models (§4.2.3).

4.2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) is a promising method for generating synthetic health data, as it can model the complex distributions of high-dimensional health data (Hernandez et al., 2022; Kokosi & Harron, 2022; Murtaza et al., 2023). Initially, it was developed to generate image data (Z. Zhang et al., 2021). More recently, GANs have also been successfully applied to generate text, audio and structured data (Z. Zhang et al., 2021). This is reflected in literature, as most of the proposed generative models are based on GAN (Murtaza et al., 2023) (see also Table 9). Generative Adversarial Networks (GANs) are seen as one of the most important breakthroughs in deep learning, or AI research in general, since their establishment in 2014 (Hernandez et al., 2022; Murtaza et al., 2023; Tsao et al., 2023). Figure 9 presents a (simplified) architectural depiction of GANs. GANs use deep learning methods to model multi-dimensional distributions of training data (Murtaza et al., 2023). The unsupervised model memorises the distribution of input data which can be used to generate new data with the same input distribution. GANs consist of two 'competing' neural networks, a generator and a discriminator model (the adversary) that are trained in an adversarial training process to produce more accurate predictions (Hernandez et al., 2022). They are based on a game theory, often in a zero-sum game setting, meaning that the one model's win is the other model's loss (Venugopal et al., 2022). The goal of the generator is to generate outputs that could be confused with real data, whereas the goal of the discriminator is to identify whether an output of the generator is real or synthetic data. The generator produces synthetic data, which are, along with real-world data, presented to the discriminator. Depending on how well the discriminator classifies data as real or fake, and how well the generator produces data that is fooled for real data, the model parameters are altered (Venugopal et al., 2022). This way, the generator iteratively discovers the patterns and distributions of real-world data that can be used to generate synthetic data (Venugopal et al., 2022).

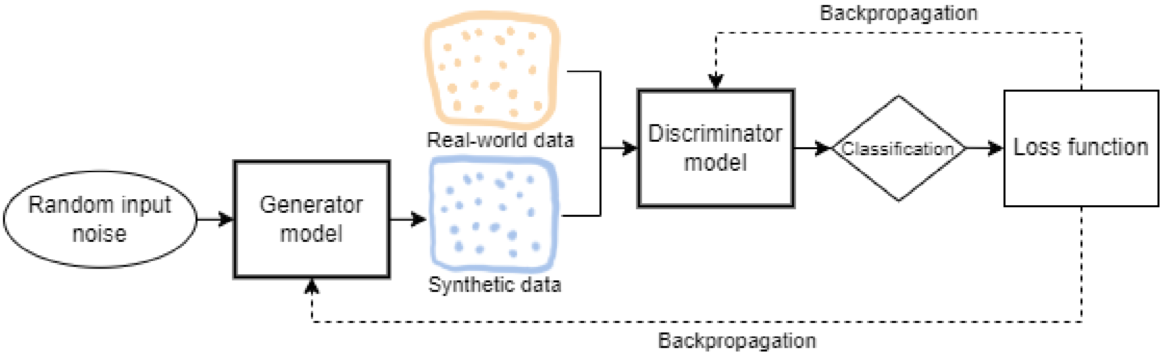


Figure 9. Simplified architecture of GANs (based on Venugopal et al., 2022; Yan et al., 2022)

Since GANs introduction in healthcare in 2017 through medGAN (Choi et al., 2017), GANs have been applied broadly, covering various data types and other specialisations. MedGAN synthesises longitudinal data from EHRs into aggregated snapshots (Choi et al., 2017). Choi et al. (2017) have extensively assessed their model on privacy risks via membership and attribute disclosure (§4.4.2) (Choi et al., 2017; Murtaza et al., 2023). The tests showed minimal privacy risks, labelling the model as privacy-friendly. As a result, many of the researchers that built on medGAN skipped the (necessary) privacy assessments under the assumption there were little privacy risks (Murtaza et al., 2023). Although the aggregation of EHRs imposes minimal privacy risks, a downside to this approach is that the aggregated data blurred correlations between variables. This generation of GANs are criticised for their static character, neglecting temporal and sequential dependencies of EHRs (J. Li et al., 2023; Murtaza et al., 2023; Z. Zhang et al., 2021).

As an answer, several researchers have successfully produced models that specifically address the correlation and temporality issues for EHRs (J. Li et al., 2023; Theodorou et al., 2023; Torfi & Fox, 2020; Z. Zhang et al., 2021). Yoon et al. (2020) propose the ADS-GAN model, also applied by Shi et al. (2022), which allows for an explicit trade-off between utility and privacy of data. Their model better captures multivariate relations in comparison to previous methods, however, still do not support longitudinal data. These models are specifically interesting to study for their built-in privacy guarantees (discussed in §4.4.3).

Torfi and Fox (2020) propose the CorGAN model, which combines GAN with a deviating type of neural network to account for correlation issues (Convolutional Neural Network instead of multilayer perceptron). They noted this method better captures the correlations and learns the temporality of data. As will be discussed in the next section (§4.4), they have extensively addressed privacy threats.

Zhang et al (2021) proposed the SynTEG model to generated timestamped diagnostic events. The model learns in two stages: first, the model learns the timestamp and the expected diagnosis, which is used as a condition to train the generator model in the second stage. The model produces longitudinal EHRs conditioned on the expected diagnosis per visit. Where most of the researchers develop models for specific data types, Li et al. (2023) innovatively propose the EHR-M-GAN model, a GAN-based model to generate mixed-type longitudinal data to account for this characteristic of EHRs, including timestamped biomedical signals and lab test results.

4.2.2 Variational autoencoders

Similar to GANs, variational autoencoders, also introduced in 2014, are a popular generative AI approach (Jadon & Kumar, 2023; Kingma & Welling, 2014; S. Sun et al., 2021). Variational autoencoders (VAEs) are generative models that integrate autoencoders with variational inference (Jadon & Kumar, 2023). Figure 10 presents a (simplified) architectural depiction of variational autoencoders. An autoencoder is a neural network that tries to learn the representation of input data (Murtaza et al., 2023). The autoencoder consists of an encoder network learning the probability distribution of input data into a latent space, and a decoder network trying to re-generate the real-world input data (Jadon & Kumar, 2023). Based on the distribution of features in the latent space, the VAE can generate different synthetic datasets with the same distribution (Hao et al., 2024). By using variational inference, the disparities between the real-world data and generated data are minimised.

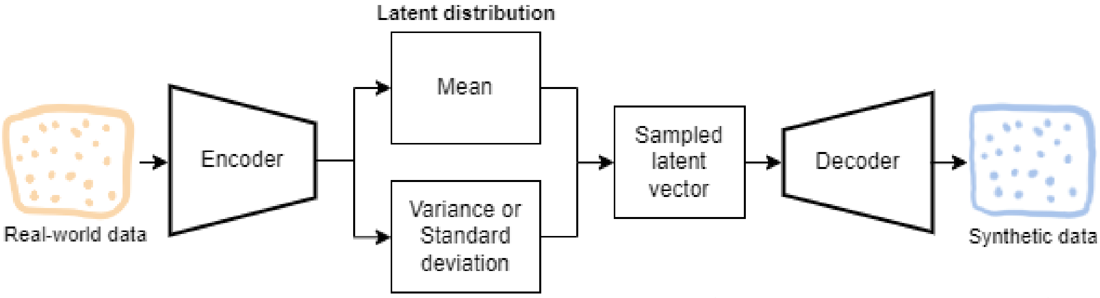


Figure 10. Simplified architecture of a Variational autoencoder (based on Jadon & Kumar, 2023)

Biswal et al. (2021) use (conditional) variational autoencoders to generate longitudinal discrete data of patient events, such as diagnoses, medication and procedures. Conditional variational autoencoders are used to generate data of predetermined events, to account for different use cases (Biswal et al., 2021). Autoencoders are also combined with other approaches (Murtaza et al., 2023). MedGAN, for example, applies autoencoders to address the issue that GANs could not generate discrete values (Choi et al., 2017; Z. Zhang et al., 2021). As mentioned in the previous paragraph, S. Sun et al. (2021) proposed LongGAN, a GAN-based model that applies autoencoders to generate sequences. Similarly, Li et al. (2023) have proposed a EHR-M-GAN, a GAN-based model. To overcome challenges of mixed-type data, they apply variational autoencoders to capture the distributions of heterogeneous features into a latent space with fewer dimensions (J. Li et al., 2023). Yoon et al. (2023) propose a model that simultaneously trains an autoencoder and a GAN, called EHR-Safe, to generate mixed-type EHRs.

4.2.3 Large language models

More recently, methods for natural language processing have found their way in generating synthetic health data (Theodorou et al., 2023). Most of the GAN-based approaches struggle with the high-dimensionality of EHRs due to aggregation of visits and medical codes, and for example, remove rare codes, reducing data utility (Theodorou et al., 2023). To compare this method with GAN-based approaches, LongGAN and EHR-M-GAN offer conditional and sequential capabilities for laboratory values but are constrained to a number of dimensions (J. Li et al., 2023; S. Sun et al., 2021), CorGAN is constraint to a limit number of medical codes (Torfi & Fox, 2020), and SynTEG combines and removes medical codes (Theodorou et al., 2023; Z. Zhang et al., 2021). Theodorou et al. (2023) have proposed a mixed-method EHRs generation model to predict timestamped visits of patients, their records and respective codes. In natural language generation, the model learns the probability distribution of languages by predicting the next word considering the preceding words (Theodorou et al., 2023). To capture the complexity of EHRs, Theodorou et al. (2023) translate this framework to the healthcare context by predicting patient visits based on past visits, resulting in a probability output. Moreover, their approach allows the modelling of sequences of binary variables per record, with high-dimensional synthetic data as a result.

4.3 Critical considerations of synthetic data generation

The previous section gave insight into the reach of synthetic data generation methods and highlighted some technical difficulties of synthetic data generation that some models address better than others. This section critically discusses considerations that reach beyond model performance: data protection (§4.3.1) and ethical considerations (§4.3.2).

4.3.1 Considerations regarding privacy

While privacy-preserving data sharing is one of the primary use cases of synthetic health data (Jadon & Kumar, 2023)), practice can be different (Chauhan et al., 2023). Researchers have addressed the relation between data protection and use of synthetic data. In comparison to other anonymisation techniques, the promise of synthetic data generation lays in its resemblance to the original and preservation of privacy (Yale et al., 2020). Resemblance and privacy, however, are conflicting aims (Murtaza et al., 2023; Wang et al., 2021; Yan et al., 2022). One the one hand, the precision of data is compromised by data synthesis if it should have a small representativeness, however, privacy is preserved, as the synthetic records show little similarity to the real-world dataset (Wang et al., 2021). On the other hand, there is a risk of 'overfitting', which, in this context, means that data generated by the model is too similar to the real-world data, compromising in the model's pursue of precision (Tsao et al., 2023). By overfitting data, models can have high resemblance to the original dataset, but can result in information disclosure because models remember data points and may generate real data points (Yale et al., 2020). Overfitting could result in partially duplicated data records or entire rows (Z. Zhang et al., 2020). To account for the risk of overfitting, model developers can evaluate the similarities between generated data and the original data (Yale et al., 2020). For example, the real-world dataset is too small in comparison to the dimensions of the generation model, personal information may be disclosed (Chauhan et al., 2023).

There may still be a risk to privacy when outliers in synthetic datasets are similar to the real-world data (Kokosi & Harron, 2022; C. Sun et al., 2023; Tsao et al., 2023; Wang et al., 2021). Concerns rise when there is a small number of rare diseases, with similar social and geographical characteristics, or when, purely by chance, the synthetic data shares the same characteristics as patients, resulting in a risk of identification (Wang et al., 2021). The literature showed, however, that with recent generation methods, these concerns are less prevalent

(Hernandez et al., 2022; Tsao et al., 2023). The overview of methods showed that researchers have proposed various models that specifically address patient privacy (J. Li et al., 2023; Shi et al., 2022; C. Sun et al., 2023; Venugopal et al., 2022; Yoon et al., 2020). §4.4 analyses how these privacy risks are evaluated by researchers.

Closely related to privacy, is the principle of transparency about how personal data is processed. The complexity of multi-layered neural networks, such as GANs, VAEs and LLMs, make that it is unclear how patient data is transformed into synthetic data (Wang et al., 2021). In a context where understanding in the algorithm is required, such as the healthcare context, black-box AI models should be avoided to gain trust in synthetic data generation (Wang et al., 2021).

4.3.2 Ethical considerations regarding data

The quality of the generation model depends on its input data, known as the Garbage In, Garbage Out principle (§4.1.2). As with other AI models, synthetic data generation models may perpetuate biases from the datasets they are trained with, causing datasets that are unfair for patient (groups) (Jadon & Kumar, 2023). The impact hereof for synthetic data may be specially large: because of being considered anonymous, they are presumably to be reused more often than personal data. Moreover, health data is dynamic in practice: EHRs are for example, are frequently updated and interpretation of values may change over time as we gain more knowledge on diagnostics and diseases (Chauhan et al., 2023). Synthetic data is (supposedly) irreversible, meaning that individuals cannot be linked to the synthetic data anymore. Therefore, statistically relevant synthetic data can become obsolete as real-world health data continues to evolve, and generation models may not account for this (Chauhan et al., 2023).

Hao et al. (2024) identified the lack of ethical and legal restrictions during the creation process of generation models as a source for these concerns. A lack of restrictions stems from the "pacing problem", which means that technological innovation, such as generative AI in this thesis, outpaces the ability of laws to adapt to new developments (Downes, 2009). Rules change step-by-step, while technology changes exponentially (Downes, 2009). The questions rises how these concerns can be addressed in an application framework (or not).

An initial thought would be that via careful selection and continuous evaluation of appropriate datasets and uses, these ethical risks can be mitigated.

4.4 Evaluation of privacy preservation

As noticed by Yan et al. (2022), there is little consensus on the evaluation metrics to assess synthetic EHRs, albeit important for evaluating synthetic models in practice (Hernandez et al., 2022). Therefore, this section discusses the prevailing metrics to asses privacy of synthetic EHRs.

To stress the importance of addressing privacy explicitly: researchers have demonstrated some synthetic datasets can be reverse-engineered to the real-world data used to train the model – a risk specifically common to outliers (Chauhan et al., 2023; Stadler et al., 2022). The identification of such risks shows the relevance of expanding the focus from researchers on evaluation of utility to evaluation of privacy. Chapter 5 analyses how synthetic data generation relates to the institutional data protection context. This requires understanding of how synthetic data generation contributes to data protection, as users of synthetic have to evaluate what privacy levels are acceptable. But, as stated by Chauhan et al. (2023, p. 134): "there are no robust methods to determine if the synthetic data generated is truly anonymous". The previous sections explained the concept of synthetic data generation and its various uses. With understanding of these technologies, this section further dives into how these researchers have assessed privacy preservation of their models and primarily, what their understanding of a privacy-preserving model is to narrow this literature gap. Based on the metric analyses of Murtaza et al. (2023), Hernandez et al. (2022), and Yale et al. (2022) complemented with additional literature, Figure 11 presents a structure of the most prominent privacy metrics discussed in the literature in scope of this literature review. This section follows this structure, first explaining the types of privacy measures (§4.4.1), the privacy threats they address (§4.4.2), and how researchers assess (or assume) and interpret privacy (§4.4.3-4.4.4).

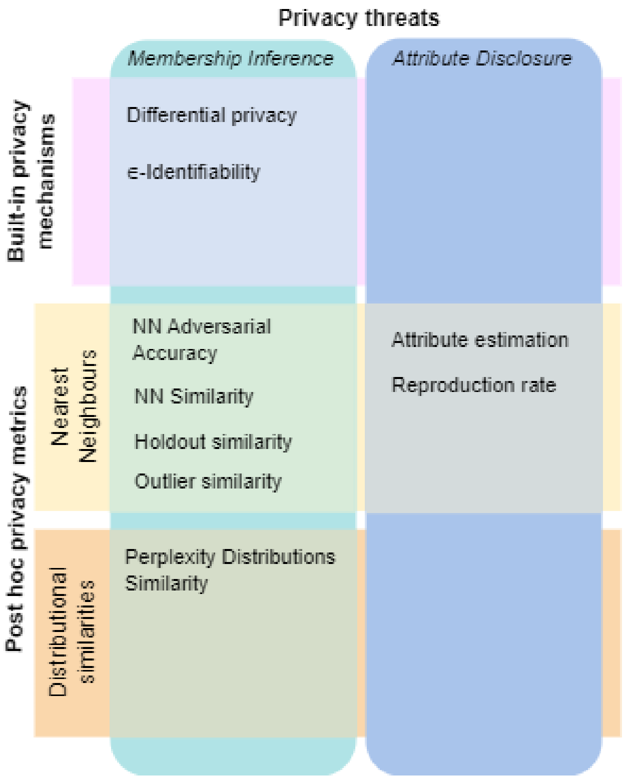


Figure 11. Privacy evaluation of genertation models (based on Hernandez et al., 2022; Murtaza et al., 2023; Yan et al., 2022)

4.4.1 Types of privacy measures

To understand how researchers evaluate privacy, it is important to understand there are various ways to achieve and evaluate privacy of a generation model. Murtaza et al. (2023) have categorised privacy-related evaluations in three type of metrics:

First, there are models that are presuming privacy preservation, as the models they built on or techniques they use have established privacy guarantees. For simpler generation methods, for example statistical models based on the probability density function, privacy concerns are less prevalent, because the synthetic data are newly created data points to fit the model (Thomas et al., 2022). Other than the relationships between variables and descriptive properties, the synthetic data generation model does not learn from the real-world data (Thomas et al., 2022). Thomas et al. (2022), argue such models are privacy preserving for continuous variables, as they are computationally derived. A downside to this approach, however, it that they are unable to capture the high-dimensional and temporal characteristics of health data, decreasing resemblance and utility of real-world data (Theodorou et al., 2023). Regarding more complex, researchers have shown that the presumption does not always hold true: there are still privacy risks in synthetic data generation models (Murtaza et al., 2023). An explicit evaluation of privacy is therefore a requirement for safe application of synthetic data.

Second, some models have integrated privacy measures and metrics, or privacy mechanism, as referred to in this thesis. For example, so-called *wrappers* are privacy measures that perform privacy-preserving transformations during the data synthesis, such as noise injunction with differential privacy. Embedded privacy metrics are built in the objective functions of the models. Researchers trust on the theoretical guarantees of these techniques to preserve privacy (Yale et al., 2020).

Third, models that are evaluated after they are developed with post-hoc privacy metrics. Yale et al. (2019a) have proposed several privacy evaluation metrics, recognised by researchers (Gwon et al., 2024; Venugopal et al., 2022).

4.4.2 Privacy threats

To better understand what privacy metrics are used, it is important to know there are two main privacy threats to synthetic datasets: Membership Inference and attribute disclosure attacks.

Attribute disclosure (or inference) attacks occur when the adversary can derive narrow attribute boundaries of real records, such as diagnoses or medication use, based on a subset of the attributes they know about the individual (Choi et al., 2017; Murtaza et al., 2023; Yan et al., 2022). Information known to the attacker are for example demographic attributes, such as age or gender (Yan et al., 2022). The goal of the attacker is to gather knowledge of a patient by observing similar synthetic data (Choi et al., 2017).

In membership inference attacks, an adversary tries to find out what records were used to train the generation model by feeding data to the model and observe outputs (Yale et al., 2019a). Such attacks are important to address, because if the membership of a patient to a certain dataset can be established, it provides more information about this patient. Yale et al. (2019a) illustrate this as follows: if the datasets concerns diabetic patients and the membership of an individual can be established, the individual is likely diabetic. In other words, membership inference risks refer to the ability of the adversary to infer membership of a targeted record (Choi et al., 2017; Yan et al., 2022; Z. Zhang et al., 2020). Under the assumption that the attacker already has some information on the records of the original dataset, tests with membership inference attacks can indicate the fraction of real data points that can be identified after completing missing data of the original dataset (Yale et al., 2019a).

When evaluating these privacy threats, it is important to consider 1) the number of real patient records known to the attacker, and 2) the volume of synthetic data generation volume (Torfi & Fox, 2020). Usually, the more information known to the attacker, and the higher the volume of synthetic data, privacy risks increase (Torfi & Fox, 2020; Z. Zhang et al., 2022). This section explains how these attacks relate to privacy metrics as summarised in Figure 11.

4.4.3 Built-in privacy mechanisms

The privacy built-in used by authors in this literature review is primary differential privacy. Differential privacy is a mathematical notion of privacy that finds its application in statistical analyses (J. Li et al., 2023). Differential privacy achieves this by introducing controlled random noise to the dataset while ensuring that the inclusion or exclusion of any individual's data does not significantly impact the output of queries or statistical data analysis applied to the dataset (Dwork et al., 2006). Using differential privacy in the training process can safeguard privacy of training data (Gwon et al., 2024). By adding mathematically designed noise to the training dataset, the exact contributions of individual data points for the discriminator are obscured by noise, thereby aiming to mitigate the risk of re-identification (Gwon et al., 2024; J. Li et al., 2023; Murtaza et al., 2023).

Some of the researchers assume privacy of models with differential privacy (Su et al., 2023), others have also performed a post hoc privacy evaluation (Gwon et al., 2024; J. Li et al., 2023). An evaluation of various GAN-models by Yan et al. (2022) showed that models with differential privacy did not perform better on post-hoc privacy metrics. Thus, the application of differential privacy as measure for privacy-preserving synthetic datasets is not irrefutable in comparison to other non-differential privacy models. An argument for the use of differential privacy, however, is that it allows for an explicit trade-off between privacy and utility by controlling the noise ratio (J. Li et al., 2023; Su et al., 2023). Moreover, differential privacy can still serve as a mitigating measure to prevent disclosure of information in the model development process.

Another method to regulate the trade-off between privacy and utility is proposed by Yoon et al. (2020). By embedding the nearest neighbour metric into the loss function (referred to as ‘ϵ-Identifiability’) of their GAN-based generation model, it can be ensured the synthetic data and real-world data are sufficiently separate. Users of the generation model can define a weight vector for various features of the dataset, i.e. they can determine what distance they deem acceptable (Yoon et al., 2020). This method is based on the idea that there must always be a minimum distance between the synthetic data and real-world data to prevent against membership and attribute inference (§4.4.2). Shi et al. (2022) applied this mechanism, setting the identifiability at 0.008%. They benchmarked the identifiability metric of their synthetic data against the identifiability of a randomly generated dataset, which gives a value close to zero.

4.4.4 Post hoc privacy evaluation

Finally, the literature study revealed that researchers perform post hoc privacy evaluations and how they interpret them. Post hoc refers to the assessment of privacy risks after synthetic EHRs have been generated. Interpretation, here, means that researchers explain why the outcomes of their evaluation offers sufficient safeguards for protecting health data, i.e., how they define certain thresholds to be acceptable. This analysis is given in appendix G. For research question 3, it suffices to present the conclusions from this analysis, to grasp how synthetic data generation contributes to protecting health data.

Determining how synthetic data contributes to data protection, requires understanding of how developers of generation models, or primarily users of these models, would define a privacy-preserving model. Many of the researchers label their model as privacy-preserving after their post hoc privacy evaluation, but what is the intended meaning of this? A first finding of the literature review hereto is that none of the articles define the concept of privacy, nor what privacy would mean in the application context of synthetic EHRs. Researchers primarily seem to view privacy as the degree to which data is anonymised, i.e. a model is privacy-preserving when the re-identification risks are ‘low’. Privacy is expressed quantitatively, by calculating numerical digits for various privacy metrics. However, this definition differs from the legal definition of data protection (§3.1.2), where data protection encompasses more than re-identification risks.

We have concluded that a privacy evaluation is important. However, in line with the first finding, the literature review exposed the large amount of articles that did not evaluate privacy explicitly. Specifically in comparison to the extensive focus on utility metrics, privacy evaluations are overshadowed. The selection process (§2.4.2) demonstrates that more than a half of the 71 search results did not perform a post hoc privacy evaluation in their research. This finding is supported by the preliminary research concerning taxonomy of synthetic data metrics in medicine of Kabaachi et al. (2023): only a quarter of the 92 studies performed a post-hoc privacy evaluation, 15% of the generation models had additional privacy guarantees in their model, but no explicit privacy evaluation, and more than 40% of the models did not have privacy guarantees nor privacy evaluation. This is concerning, considering the privacy threats (§4.4.2).

Regarding the researchers that have performed privacy evaluations, however, there seems to be no consensus on what privacy metrics should be used; this resulted in a fragmented technical landscape that is hard to navigate through. This calls for a standardisation of post hoc privacy evaluations. To further standardise the use of privacy metrics, lessons can be drawn from observing correlations between various metrics. For GAN-based models generating synthetic data, Yan et al (2022) calculated the correlations between privacy metrics, with low correlations showing the need to include multiple privacy metrics.

Moreover, the interpretation of these metrics still leaves something to be desired. Appendix G presents an overview of privacy metrics discussed in literature. To give an idea about what privacy metrics are used, metrics such as the ‘reproduction rate’ are easily interpretable: users can easily understand privacy risks associated with duplicated records. Metrics such as the ‘Nearest Neighbour Adversarial Accuracy’, however, are not self-explanatory as they are mathematically more complex. Therefore, some metrics require some more explanations than other metrics. With Yale et al. (2020) and Yan et al (2022) as exception, researchers rarely relate the numerical values to their context. Choi et al. (2017), for example, acknowledge the importance of attribute and membership inference attacks, but fail to explain why their outcomes preserve privacy. Z. Zhang et al. (2022) contributed to measuring membership inference attacks risks, however, other than comparing risks for partially and fully synthetic data and playing with the precision threshold, do not interpret these, i.e. evaluate what this would mean for individuals' privacy. The interpretation of such privacy metrics requires the definition of a certain threshold of what privacy risks are acceptable. Except from El Emam et al. (2020), none of the authors have defined the acceptable privacy risks in consideration of its application context. As indicated by Rajotte et al. (2022), such interpretations are important, as scores that seem good could still present a great privacy risk for a small amount of people. Therefore, such a contextual interpretation should be provided by researchers. But, researchers do not interpret such scores in their context, for example regarding compliance with organisational procedures or regulations, or the characteristics of the real-world dataset.

4.5 Conclusion

This chapter aimed to clarify how synthetic data generation and evaluation contribute to protection of personal data, specifically focusing on synthetic EHRs. Answering this question requires understanding of synthetic data generation and how this relates to privacy.

The literature review showed there are various approaches to generate synthetic data. To illustrate, there are statistical methods that generally produce privacy-preserving synthetic datasets with a lower utility than other generation methods as they cannot capture the complexity and high-dimensionality of health data. More recent developments in AI approaches, including GANs, VAEs, and LLMs, better preserve these characteristics. After gathering understanding of the types and characteristics of generation methods, it can be concluded that synthetic data generation *can* offer a privacy-enhancing way to share data for research: there are various approaches that have little privacy risks. With the emphasis on *can,* as researchers have demonstrated that some models that presumed privacy had actual re-identification risks, such as inferring whether an individual is part of a dataset or disclosing attributes of an individual by comparing the synthetic dataset with the real-world training data. Such false presumptions are accompanied with data protection breaches, as the health data that is shared is presumed to be anonymous, while it actually concerns personal data. In the context of sharing health data for research, such risks should be addressed carefully via an post hoc privacy evaluation.

An evaluation of privacy can inform users about the risks associated with the use of generation models, which is especially important considering the lack of transparency that results from the complexity of the generation methods. Specifically post hoc privacy evaluations are important to address remaining re-identification risks, therefore, contributing to health data protection.

What complicates this, however, is that there is no standard approach yet to evaluate privacy of synthetic EHRs generation models or datasets, complicating the adoption of synthetic data generation. An explanation is that there exists no consensus (yet) on what privacy metrics should best be used to evaluate privacy. This complicates the comparison of various metrics and selection of privacy metrics. Another explanation is the wide variety of generation methods: some metrics, specifically attack frameworks, are developed specifically for certain generation methods, such as GANs.

What further complicates the evaluation of privacy risks, is that researchers hardly explain why their metrics show privacy-preserving results; acceptable privacy thresholds remain largely undiscussed, which complicates the interpretation of complex, probability-based privacy risks. In addition, a privacy score is relative and should be interpreted in relation to the characteristics of the real-world dataset to account for privacy risks. However, most of the researchers only stated their model preserves privacy after the privacy evaluation, without discussing what privacy thresholds they deem appropriate, considering legal requirements or taking into account the real-world dataset. As a result, it is complicated to assess how synthetic data contributes to health data protection.

In conclusion, this chapter has shed light on the role of synthetic data generation and evaluation in safeguarding personal data within data sharing scenarios. It has emphasised the importance of understanding different synthetic data generation methods and their implications for privacy, as well as an interpretation thereof. Chapter 5 will assess how the (lack of) privacy evaluation relates to the institutional data protection environment.

5 Institutional environment of synthetic data sharing

To build a framework that describes the potential impact of synthetic data generation on health data sharing, this chapter studies how synthetic data generation relates to the institutional data protection environment. This chapter aims to answer the following research question:

> *RQ3 Which data protection-related factors influence how synthetic data generation enables health data sharing for research?*

Via the same structure as Chapter 2, following the IAD components, this chapter relates the institutional analysis of health data sharing for research to synthetic data generation. First, the concept of synthetic data is analysed in view of the contextual components of the IAD framework: the rules-in-use (§5.1), the physical and material conditions (§5.2), and the community attributes (§5.3). This serves as a first analysis of the interaction between synthetic data generation and its institutional environment. Based on these contextual components and the outcomes of the analysis on privacy evaluation of synthetic EHRs in Chapter 4, synthetic data generation is studied within the action arenas from Chapter 3. As a result, barriers, drivers, and solution directions influencing the impact of synthetic data generation on health data sharing are elicited for each action arena (§5.4). These findings are converted into a framework that structures these factors during the various phases of synthetic health data sharing (§5.5).

5.1 Rules-in-use

This section discusses how synthetic data relates to the rules-in-use. Generally, there are two ways of looking at synthetic data generation. First, synthetic data generation can itself be seen as a processing activity that should comply with data protection law. Here, the central question is the extent to which synthetic data generation falls within the scope of personal data protection law (§5.1.1). Second, synthetic data generation can be seen as a technical measure to ensure compliance of health data sharing as a personal data processing activity, by contributing to data protection principles (§5.1.2).

5.1.1 Scope of data protection rules

To determine the applicability of the rules-in-use to synthetic data, the material scope of data protection rules should be considered (Gal & Lynskey, 2023), i.e. whether synthetic data generation concerns the processing of personal data (General Data Protection Regulation, 2016, art. 2(1)). Synthetic data generation should be decoupled in three data processing activities: the development of synthetic data generation models, the generation of synthetic data and the use of synthetic data (Beduschi, 2024).

First, the development of synthetic data generation models based on real-world (personal) datasets can be considered as processing of personal data. As seen in §4.1.2, some models are trained with publicly available anonymous data – this would not concern a processing activity subject to the GDPR. Accordingly, the training of models with personal data is a processing activity under the GDPR. Developers of synthetic data generation models or

AI models in general must adhere to the data protection rules “from day 1 of the development” (Brauneck et al., 2023, p. 11). Privacy must be considered in all phases of the design process, from the initial design, the development and deployment of models that process personal data (Brauneck et al., 2023; General Data Protection Regulation, 2016, art. 25). For example, by including additional privacy measures such as differential privacy in the training process (§4.4.3). According to the GDPR roles, patients remain data subjects and the technology developer is considered a data controller, as it determines the purpose and means of processing.

Second, the act of generating synthetic data from real-world datasets is subject to the GDPR, as this adaption to personal data is considered a processing activity (General Data Protection Regulation, 2016, art. 2). The party in control over the synthetic data generation, the healthcare provider, must respect data protection principles, such as purpose limitation, lawfulness, fairness and transparency as well as safeguarding data subjects’ rights when generating the data. §5.4 further specify these rules.

Third, the extent to which data protection rules apply to the generated datasets, depends on the characteristics of the generated dataset. For example, partially synthetic data is subject to the GDPR, as individuals are re-identifiable based on the synthetic dataset (Beduschi, 2024). Logically, datasets that are fully comprised of synthetic data, should be considered anonymous when individuals are no longer identifiable. Synthetic data can be considered as anonymous data if the generation method can be considered an effective anonymisation technique as defined by the Article 29 Working Party (Article 29 Working Party, 2014, p. 29; El Emam, 2020) (§6.3.3H.2). Processing activities with anonymous data, such as sharing them with third party researchers, fall outside the material scope of the GDPR (General Data Protection Regulation, 2016, art. 2) and are therefore not subject to the data protection principles (General Data Protection Regulation, 2016, rec 26). Action arena 2 will further substantiate what constitutes an effective anonymisation technique (§6.3.3H.2).

Unfortunately, this classification is more difficult than suggested here. Although the GDPR explicitly refers to anonymous data and anonymisation techniques have been a topic of discussion by EU DPAs for decades, it is unclear what anonymisation means in practice (Burt et al., 2021). The bar for classifying as anonymous data is high – the GDPR has a remarkably broad understanding of personal data (Beduschi, 2024), as explained in §3.1.2. Moreover, as demonstrated by Stadler et al. (2022), there remains a risk of re-identification of individuals of (presumed) fully synthetic datasets. Thus, the question arises how these privacy risks relate to the legal definition of anonymous data.

I propose Figure 12 to illustrate how synthetic data compares to other health data types.

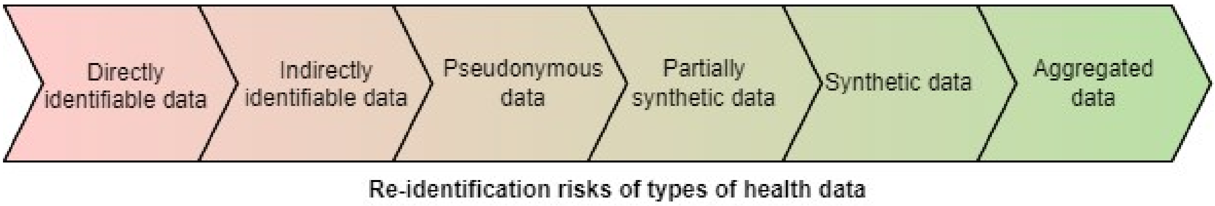

Directly identifiable data
Indirectly identifiable data
Pseudonymous data
Partially synthetic data
Synthetic data
Aggregated data
Re-identification risks of types of health data

Figure 12. Spectrum of identifiability of health data (based on Bellovin et al., 2019)

As shown in Figure 12, re-identification risks are high for directly identifiable data, as the data contains concerns data with direct identifiers, for example, how EHRs are used within hospitals, including patient names and contact details. Data subjects can be identified in indirectly identifiable data by combining attributes or in conjunction with additional information. Pseudonymous data (§3.1.2) can only re-identify data subjects with additional information kept separately. Partially synthetic datasets concern of both pseudonymous data and synthetic data, decreasing the re-identification risks as the distinguishment of real from synthetic data is complicated. Synthetic data has low re-identification risks (§4.2) but can still impose risks for patients on record level. Aggregated data is put as example of data with lowest re-identification risks, as it only identifies larger groups of people, thereby reducing re-identification risks.

5.1.2 Synthetic data as data protection compliance measure

Regardless of the legal classification of synthetic data generation, it should be emphasised that synthetic data can be seen as a privacy-enhancing technology (PET). PETs are generally referred to as technical means that intend to protect privacy of individuals, while preserving data use (Brauneck et al., 2023; Heurix et al., 2015). In terms of the GDPR, PETs can be viewed as technical measure or safeguard to mitigate privacy risks for patients (for example, to comply with General Data Protection Regulation, 2016, art. 24(1) and 89(1)).

The integration of synthetic data in health data sharing practices can promote data protection principles. Synthetic data can be applied as measure to ensure secure processing of personal data, data minimisation and data accuracy (Gal & Lynskey, 2023). Replacing personal data in data sharing processes with synthetic data offers an "additional layer of security to personal data" (Gal & Lynskey, 2023, p. 29), as fewer people and systems can access the personal data. At the same time, the value of using synthetic data for research is high (Wang et al., 2021; Yan et al., 2022), therefore, balances data protection and research interest in health data sharing. As illustration, the Norwegian Confederation of Sport published personal data from 3.2 million citizens online due to an error during tests with a cloud solution – a test that could have been achieved by processing synthetic data, according to the Norwegian DPA (European Data Protection Board, 2021). Moreover, data minimisation can be achieved as the various types of synthetic data decrease the amount of personal data shared for research purposes (Bellovin et al., 2019). Therefore, synthetic data generation may be applied as technical measure to protect personal data and meet the requirements of article 89(1). Via data augmentation, missing or inaccurate values can be replaced by synthetic data, increasing data quality (Gal & Lynskey, 2023).

To conclude, as visualised in Figure 12, the use of synthetic data where possible remains a less-intrusive processing activity in comparison to using real-world data, as re-identification risks are lower.

5.2 Physical and material conditions

As supplement to the description of the physical and material conditions of current health data sharing practice (§3.2), the objective of this section is to show how synthetic data generation changes these conditions (or not) in the use case. The technical resources (§5.2.1) and institutional means (§5.2.2) are serve as input for a more substantive analysis of the impact of synthetic data in the action arenas.

5.2.1 Technical resources

This section discusses synthetic data generation models as technical resources in data infrastructures. Synthetic data generation is usually performed by third party technology developers (interview 6). In practice, healthcare providers either develop their own synthetic generation methods, for example based on open resource models (DS1-H), or they can purchase this service from external technology providers. In view of the models that are currently publicly available, the latter is preferred, as is the case for the use case studied in this thesis. As software can run locally on in the environment of healthcare providers, at least with the technology providers questioned in this thesis (TP1, TP2), the personal data is not exposed to the technology provider. In their tool, technology providers present the outcome of quantitative privacy and utility metrics to the healthcare providers. Hence, synthetic data generation models are tools deployed and initiated by healthcare providers. Regarding the necessary human capacities to work with synthetic data, healthcare providers need humans with analytical and technical skills to construct or select, and interpret generation models (TP1).

5.2.2 Institutional means

This section discusses how synthetic data generation relates to the institutional means described in §3.2.3. Sharing synthetic data between healthcare providers and research institute still requires the conclusion of research agreements, specifically to distribute research outcomes that follow from synthetic data analysis (LC-H). The provisions regarding data protection can be decreased, considering risks for patients are limited (DS1-H). Other institutional routines, such as the need to create DMPs for research projects and conduct DPIAs to identify and address privacy risks, will be the same. Within healthcare providers, DPIAs must be conducted when there are significant privacy risks in the conducted research (PO-H). Whether such an assessment is necessary for synthetic data sharing in research activities, should thus be considered on a case-by-case basis, depending on the privacy risks concerned with the synthetic dataset and the research circumstances.

5.3 Community attributes

The community attributes in §3.3 showed that in general, actors involved in data sharing processes are positive towards improving the health data sharing process. This section further discusses their beliefs towards synthetic data generation and how they judge the beliefs of others. All actors involved in the use case belief synthetic data can ease current health data sharing processes. However, there are some differences in concerns regarding adoption of synthetic data generation in healthcare.

A healthcare provider's data steward understands the potential benefit of health data (DS1-H). Yet critical questions arise regarding the technical boundaries of synthetic data generation in healthcare. For example, the data steward raises concerns regarding the auditability or evaluation of generation models: specifically regarding differential privacy-based models, as these require technical considerations in its deployment. Moreover, assessing whether synthetic data meets the quality standards of specific use cases necessitates individual evaluations, requiring training to elucidate processes, advantages, and limitations (DS1-H). While synthetic data may offer partial solutions, its applicability varies: simpler inquiries, such as inter-organisational pattern comparisons, can be carried out with synthetic data. More complex inquiries, however, that involve causal modelling for health-related research or intricate machine learning architectures may prove unsuitable due to the inherent complexities of generating (DS1-H).

On the contrary, technology providers see little technical challenges, stating their models can suit many use cases (Appendix C). Instead, they see particular challenges in the way synthetic data is applied. For example, one technology provider indicated that they are not (yet) experts in the application of synthetic data (TP1). They are experts in generating synthetic data, but not necessarily in the customer application domain, such as healthcare. The application of synthetic data concerns a grey area: the technology is so new that few people in an application domain know about the technology and how to implement it (TP1). This results in misunderstanding about the potential of synthetic data (TP1), with two consequences. First, synthetic data generation is (perceived as) an abstract concept by user groups. As a result, they find it difficult to imagine its output. They also have questions regarding security; for example, "*how can synthetic data generation be privacy-preserving but at the same time contain the same (statistical) properties as real-world data?*" (TP1). This lack of comprehension presents a second challenge: user groups do not know how synthetic data can still provide valuable information while being privacy-preserving, leading to a narrow view on possible use cases (TP1). This belief is confirmed by a legal councillor, who points out that it is difficult to understand how synthetic data has the same value as personal data for research (LC1-RI). Another technology provider does not immediately recognise these challenges, by stating that only those people that work with synthetic data need to understand it – and they usually do, considering they know what synthetic data is by approaching the technology provider (TP2). They find the primary challenge lies in the identification of suitable applications by users of synthetic data, as healthcare providers tend to start with wanting to implement the most complex use cases that require strong privacy guarantees. To account for this, data owners must implement synthetic data incrementally (TP2).

To conclude, overall, the consensus is that synthetic data benefits health data sharing. However, their beliefs do not align at all points. Users of synthetic data, both from the healthcare provider and research institute, belief there are challenges that the technology providers do not necessarily acknowledge. The differences in identified problems show that clarifications are needed to smooth the adoption of synthetic data. These clarifications are specifically needed to decrease information asymmetries between technology providers and healthcare providers regarding the use and application scope of synthetic data, and to guide users in their applications.

5.4 Synthetic data generation in action arenas

This section makes insightful how synthetic data generation unfolds in the action arenas that Chapter 3 identified for sharing health data. As such, barriers, drivers, and solution directions that influence the impact of synthetic data generation on health data sharing are identified, per action arena. These are indicated as [Bi], [Di], and [Si], respectively, with i=1…N. The identification is the conclusion from an analysis of interviews and literature from academia, governments, and industry. This analysis is presented in Appendix H ; this section only presents a brief recap of each action arena and the factors identified for each arena. The framework in §5.5 will structure these factors.

Figure Ā̊ visualises the changed institutional environment, showing the changed interaction in comparison to the current institutional landscape (Figure 7) and the interactions that cause uncertainty for sharing synthetic health data, highlighted in blue.

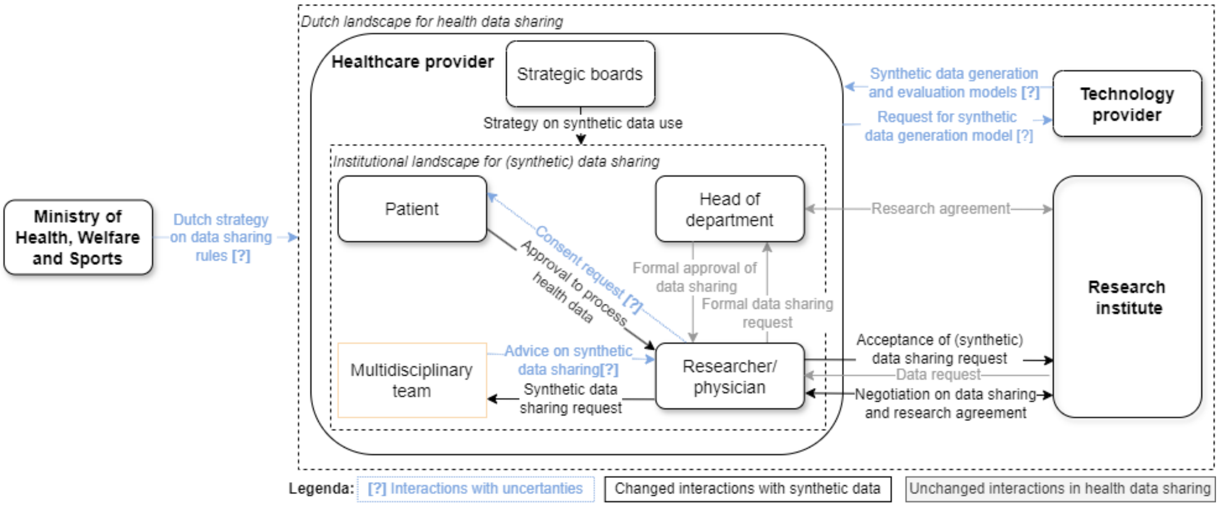

Dutch landscape for health data sharing
Healthcare provider
Strategic boards
Strategy on synthetic data use
Institutional landscape for (synthetic) data sharing
Patient
Head of department
Consent request [?]
Approval to process health data
Formal approval of data sharing
Formal data sharing request
Multidisciplinary team
Advice on synthetic data sharing[?]
Synthetic data sharing request
Researcher/ physician
Ministry of Health, Welfare and Sports
Dutch strategy on data sharing rules [?]
Synthetic data generation and evaluation models [?]
Request for synthetic data generation model [?]
Technology provider
Research agreement
Research institute
Acceptance of (synthetic) data sharing request
Data request
Negotiation on data sharing and research agreement
Legenda:
[?] Interactions with uncertanties
Changed interactions with synthetic data
Unchanged interactions in health data sharing

Figure 13. Proposed interaction patterns with synthetic data generation in on-premise software architecture

First, interactions discussed in this section relate to a changed actors playing field – including the multidisciplinary team and the associated uncertainties – that change the health data sharing process (§5.4.1 and H.1).

Second, there are still uncertainties regarding the legal definition of personal data and anonymisation, causing issues for the legal qualification of synthetic data (§5.4.2 and H.2). This perpetuates in the Dutch strategy on data sharing rules. This, along with the issues related to the interpretation of privacy metrics, creates uncertainties for the privacy evaluation of synthetic data by technology providers and actors within healthcare providers.

Third, there are uncertainties regarding the legal base for synthetic data generation, caused by ambiguities in the Dutch strategy on sharing rules (§5.4.3 and H.3). As a result, healthcare providers may have to rely on consent again.

5.4.1 Action arena 1: Operational decisions in health data sharing

This section concludes on the factors that influence the impact of synthetic data generation on health data sharing behaviour from the analysis of action arena 1. Action arena 1 concerned the operational decisions that are made regarding health data sharing activities. These activities include the initiation of health data sharing, the interactions to obtain access to health data, and the factual sharing of health data. The outcomes of these interactions were lengthy processes to conclude health data sharing agreements, poor privacy due to a lack of due diligence, and the lack of (central) infrastructures that can disclose health data to (external) researchers in a privacy-enhancing manner. The analysis (Appendix H.1) investigated the changing positions of actors in data sharing practice, their interactions, and how synthetic data relates to the current resources to provide health data, data infrastructures and agreements.

Factors and solution directions

A barrier regarding the application of synthetic data is the need for technical expertise [B7]. Following the community attributes, one should not presume the actors involved in health data sharing process are aware of the potential and drawbacks of synthetic data generation [B13]. This challenge occurs in the phase of selecting appropriate generation models as well as the evaluation phase of the synthetic datasets. This organisational challenge is the concern of healthcare providers and research institutes. As data controller, healthcare providers are responsible for compliant sharing of (synthetic) health data. Therefore, they must be able to evaluate the technically complex models and metrics. A proposed solution (direction) for healthcare providers is to organise multidisciplinary support teams that asses the technical, legal and data management facets of synthetic health data sharing [S12]. Either employees with knowledge of AI models need to be involved in the selection and evaluation process, or staff needs to be educated to gain appropriate knowledge about synthetic data generation process, advantages and limitations [S16]. Although healthcare providers are responsible for the evaluation of synthetic data, technology providers should make these models and metrics interpretable for healthcare providers. Moreover, to increase awareness among researchers from research institutes, healthcare providers should inform them about the generation and evaluation process of synthetic data and the properties of the real-world dataset [S20]. To gain trust in its potential to protection personal data, technology providers should allow for independent verification and evaluation of synthetic data [S21].

Synthetic data could ease health data sharing processes by providing an avenue for further standardisation of data sharing agreements, shortening sharing procedures [D1]. This can also be seen as an opportunity to convince external researcher of the value of synthetic data [S22].

Another enabler is that, regardless of the legal classification of synthetic data, synthetic data generation implies lower re-identification risks for patients in comparison to pseudonymous data [D2] and contributes to the principles of data protection [D3]. This could drive healthcare providers to integrate synthetic data into their health data sharing process. As pseudonymisation is currently not governed and often falls to doctors or researchers themselves, intensive collaborations with technical experts that help to transform health data into synthetic data is another safeguard for patients' data protection [S12]. Moreover, lower re-identification risks expand the possible use cases for health data sharing, benefitting data-driven research [D4].

Lastly, it can be concluded that some unresolved issues with synthetic data remain. Naturally, similar to other data types, common data models and vocabularies are still required to make synthetic data interoperable, such as the International Statistical Classification of Diseases and Related Health Problems, or NOMED CT – this is not different for synthetic data (Wiedekopf et al., 2021). Moreover, the data sharing process is still ad hoc: researchers of external organisations still need to know what people could provide them with synthetic data. Thus, the data findability issue does not change with the use of synthetic data [B15]. There remains a need for organisational or central catalogues to access synthetic (meta)data [S25] and adherence to the FAIR principles for synthetic data as well (§E.1) [S24]. Such systems could be either initiated by individual healthcare providers (DS1-H). The development of such catalogues could be seen as a collective action problem, as actors in the healthcare sector would benefit from increased data sharing, however, the development of these systems requires (individual) investments or extensive cooperations and it would be disadvantageous for individual healthcare providers to share their metadata. This could be an explanation for the largely undeveloped catalogues, that have not been developed on a national level yet (Hansen et al., 2021). That said, these solutions should be supported to improve synthetic data findability and utility for researchers (DS1-H-VAL), and are, therefore, considered as factors that enable or inhibit synthetic health data sharing.

5.4.2 Action arena 2: Decisions in definition of personal data

Action arena 2 showed research often involves pseudonymous health data sharing. The bar for datasets to qualify as anonymous data is high. Moreover, EU member states interpret pseudonymisation (and hence anonymisation) differently. This interpretation ranges from considering data after pseudonymisation as anonymous for some parties, to always considering such data as personal data. The latter strict approach is followed in the Netherlands. This, along with the changing definitions of these concepts in case law, results in uncertainties for researchers to determine the applicable rules for health data sharing. Also, an approach that is strict per se, may disproportionality limit health data sharing due to the high administrative burdens described in action arena 1.

Analysis of this arena consisted of exploring the influence of decisions on higher abstraction levels, focusing on how the legal interpretation of anonymisation at the Dutch and EU level relates to the synthetic data generation methods and evaluations. This was done in three parts. First was investigating the implications of a lack of interpretability for the privacy evaluation of synthetic data generation; being able to conduct this evaluation is a prerequisite for determining whether synthetic data qualifies as anonymous data (§H.2.1). Next was exploring the relation between the legal definition of anonymisation relates and the application of synthetic data generation. The legal definition of anonymisation is closely related to re-identification risks (§H.2.2). Therefore, the anonymisation analysis was expanded with important guidance from the EU and related these to the evaluation of EHRs (§H.2.3).

Factors and solution directions

From the analysis of this action arena, we learn that anonymisation is not a binary characteristic of data: it is a process of finding the right balance between re-identification risks and utility of datasets. Acceptable risks depend on the context and nature of the data, such as the sample size in comparison to the population size, the existence of additional data protection safeguards, or the consequences for data subjects when they are re-identified (Agencia Española Protección Datos, 2021; European Medicines Agency, 2016). Therefore, technology providers should not presume privacy of their synthetic data generation models, as this could lead to unaccounted for privacy risks [B8]; technology providers should perform post hoc privacy evaluations in any case [S13], especially in research.[7]

The Article 29 Working Party provides some general guidelines on how to assess anonymisation on a case-by-case basis [S15]. To evaluate whether synthetic data can be considered anonymous, healthcare providers should evaluate synthetic data based on quantitative and qualitative assessments, ensuring re-identification risks are accounted for [S18].

A barrier for applying synthetic data is the lack of interpretability of synthetic data generation and evaluation methods. This is accompanied by a lack of standards to evaluate privacy [B10], to benchmark generation models [B5], or to define the utility/privacy trade-offs in model parameters [B6]. This complicates the selection of appropriate generation models [B4] and evaluation of privacy risks [B11], decreasing confidence of healthcare providers and research institutes in use of synthetic data. To account for this, solution directions for technology providers are to provide insight into how synthetic health data is generated [S10], to provide an explicit privacy evaluation in an understandable manner [S13], and to provide sector-specific guidance or standards to benchmark generation models [S11] and privacy evaluations [S17]. This will provide technology providers clearance over how to quantitatively demonstrate privacy preservation and help data owners to navigate in the complex field of privacy metrics. Moreover, a policy opportunity would be to support opensource initiatives of synthetic data generation by Dutch or EU governments [S9] (DS1-H-VAL).

[7] “Especially in research” refers to the lack of privacy evaluations in technical synthetic data generation literature. The technology providers in the use case already performed post hoc privacy evaluations.

Another barrier is that the legal definitions of personal data and anonymisation are associated with uncertainties. Personal data is based on identifiability, an abstract concept that is hard or even impossible to define in technical quantitative privacy metrics [B12]. This complicates the implementation of harmonised and compliant evaluation methods of synthetic data by technology providers [B13a]. There is a policy opportunity for the EU and Dutch DPA to operationalise the reasonableness standard [S19].

Additionally, regardless of the legal uncertainty, technology providers do not try to analyse how their privacy evaluations meet the legal requirements [B9]. A solution direction to further intertwine computer science practice with the legal understanding of anonymisation, is to interpret proposed privacy metrics and attacks considering risks of singling out, linkability, and inference, to demonstrate the effectiveness of anonymisation [S10]. Synthetic health data literature does not address the evaluation of linkability of different synthetic datasets – a missed research opportunity. Moreover, technology providers should provide qualitative explanations to facilitate interpretation of quantitative privacy metrics [S14]. Interpretability can be achieved by explaining the metrics in an understandable manner; explaining why certain thresholds are used, based on qualitative arguments; and where possible, by linking quantitative metrics to contextual privacy risks. However, there remains a research gap, as technology providers lack user domain knowledge, e.g. which privacy metrics can best be applied, and how these can be presented to users in a specific domain such that they comprehend them. A promising finding is that one technology provider performs their privacy evaluation according to these risks (TP1, TP1-VAL). This eases the evaluation process of whether quantitative privacy metrics are compliant with data protection rules for healthcare providers.

In conclusion, to demonstrate the impact of synthetic data generation on health data sharing, we must collate technical privacy metrics and legal guidance on anonymisation.

5.4.3 Action arena 3: Decisions in legal base of health data sharing for research

Action arena 3 concerned the decisions on policy level regarding the legal base for processing health data for research. The interactions in this arena resulted in a fragmented and contradictory application of the principles of lawfulness, transparency and fairness and purpose limitation, at the EU and Dutch level. In the Netherlands, this has led to a consent-by-default approach, to the dismay of the healthcare sector. Due to the lack of guidance on applicable legal bases and presumption of compatibility with the purpose limitation principle, the health data sharing process compromises on patients' data protection and involves greater compliance burdens for healthcare providers and research institutes.

The analysis for this action arena was structured as follows. The generation of synthetic data is a processing activity subject to the GPDR, which therefore should rely on a legal base (§5.1.1). An important success factor for synthetic data would be that no consent need to be sought under the consent-by-default approach. Therefore, the possible legal bases that the generation of synthetic health data may rely on were explored, starting with the GDPR (§H.3.1), followed by a discussion of synthetic data in the consent-by-default approach (§H.3.2).

Factors and solution directions

The overall lack of governmental guidance regarding the presumption of compatibility and the appropriate legal base is a barrier for synthetic data; it is still unknown to users of synthetic data what legal base they may rely upon [B2]. When sharing synthetic data generation is also subject to the consent-by-default approach, therefore for pseudonymous data sharing, the motivation for adopting this privacy-enhancing technology may be lower [B1]. Accordingly, an important policy opportunity for governmental organisations is to clarify the legal bases for health data sharing for research and for anonymisation [S3]. The first step herein is to acknowledge legal bases other than consent more strongly (such as art. 9(2)(j) GDPR, see §3.1.3) [S4].

This is reinforced by the barrier in action arena 2 regarding the ambiguous and broad scope of personal data, as it is hard for healthcare providers to classify synthetic data. As a result, it is difficult to assess the presumption of compatibility of anonymisation techniques. Also, when the classification as anonymous data cannot be established, research institutes still need an additional legal base for receiving and analysing the synthetic health data.

What underpins this barrier is that the Dutch approach towards consent-by-default and the presumption of compatibility for scientific research is disproportionality strict. Remember that the GDPR follows a risk-based approach (§3.1.1). Hence, as synthetic data generation imposes fewer privacy risks in comparison, I would argue that a looser interpretation of the presumption of compatibility would be suitable in the Netherlands, reaching to cross-organisational contexts. This does not infer with EU regulations. The primary purpose of direct healthcare delivery and the secondary use of health data for (synthetic) data-driven research to improve healthcare are, in my view, close – the application domain is similar, and the patient may benefit from research findings in future visits. Moreover, a broader scope of this presumption would stimulate the implementation of synthetic data generation, as it is easier to determine the applicable rules for synthetic data generation. This stimulation is beneficial for patients, as synthetic data generation sharing comprises lower risks than pseudonymisation.

Practical guidance

These solution directions only consist of policy opportunities. The question that follows is how users can now share synthetic data in the current institutional landscape. What can be learned from the interactions in action arena 3 between EU institutions, EU governments, the healthcare sector and research institutes, is that, in the Netherlands at least, the presumption of compatibility of scientific research is not and should not be interpreted as a concept to evade an additional legal basis; it requires a case-by-case analysis to determine what legal basis may apply and whether the original base may be relied upon (COREON, 2022; European Data Protection Supervisor, 2020).

To guide healthcare providers and research institutes in their assessment, I propose the following. Following the approach proposed by the EPDS and the Netherlands, the healthcare provider should consider the link between the original purpose and the secondary purpose to determine the compatibility of a secondary research purpose [S2]. This requires an analysis of the context in which the personal data have been collected, focusing on the reasonable expectations that patients have regarding the secondary use of their health data. Important considerations are the relationship with the data controller, the sensitive nature of the personal data, and the existence of appropriate safeguards during initial and secondary use (European Data Protection Supervisor, 2020; General Data Protection Regulation, 2016, art. 6(4), rec 50).

The selection of a legal base requires careful risk assessments of the real-world dataset. This requires a trade-off between privacy and utility of synthetic data, depending on its application; this is often complicated [B3]. In their risk assessment, healthcare providers should start with an exploration of the intended synthetic data use, as this affects the legal requirements and privacy risks (Vallevik et al., 2024) [S5]. For example, with whom is the synthetic data shared and for what purposes? Then, analyse the real-world dataset and how it represents the real-world population, e.g. to assess the amount of outliers and sensitivity of information [S6]. This helps to assess the potential impact of data inference from synthetic data, and aids striking the balance between privacy and value for researchers. Lastly, assess whether synthetic data generation is the right tool for the job, e.g. by considering how complex the inquiry or modelling exercise is and suits the analytical needs [S7]. This guides the weights of privacy and utility dimensions. According to the privacy risks and use of the real-world dataset, healthcare providers should implement mitigating measures [S8], such as pseudonymisation of health data before it is disclosed to the generation model. Other measures used by technology developers are, for example, differential privacy (see §4.4.3). Less abstract measures could for example be a secure processing environment in which access to the original personal datasets is restricted to only those people that need to know the real-world data to develop the generation model. Lastly, similar to the EDPS, the Netherlands explicitly value the contribution of research to the public interest (Uitvoeringswet

Algemene Verordening Gegevensbescherming, 2018, art. 24). Therefore, healthcare providers and research institutes would still need to establish the societal value of synthetic data sharing for research (COREON, 2022) [S1].

Depending on how healthcare providers show such criteria, an additional legal basis may not be necessary, according to the GDPR's presumption of compatibility for scientific research.

5.5 Privacy-enhancing application framework

With the knowledge of the interaction between synthetic data and its institutional environment while including the challenges surrounding health data sharing for research, this section structures this information into a framework. This framework supports synthetic data sharing in a privacy-enhancing way, such that it accounts for the uncertainties regarding the classification of anonymisation.

5.5.1 Framework formulation

The synthetic data generation process generally consists of four phases of applying synthetic data in research: the selection of a dataset, training of a generation model and data synthesis, the validation of synthetic data, and its use (Health Community of Interest, 2022). These phases are defined in Figure 14. The difference with the process in Figure 8/Chapter 3, is that in this section, I assume a generation model is trained; the focus of this thesis is the application of synthetic data in cross-organisational health data sharing, and not the development of generation models.

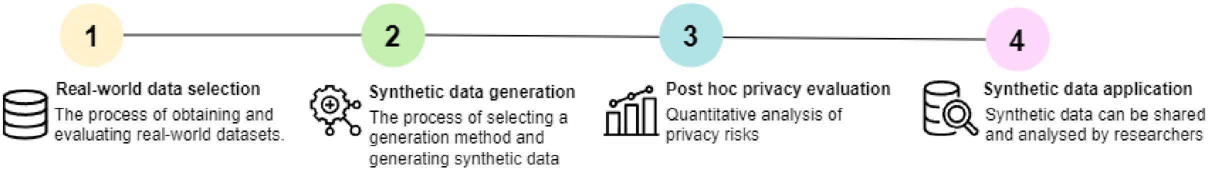

1
2
3
4
Real-world data selection
The process of obtaining and evaluating real-world datasets.
Synthetic data generation
The process of selecting a generation method and generating synthetic data
Post hoc privacy evaluation
Quantitative analysis of privacy risks
Synthetic data application
Synthetic data can be shared and analysed by researchers

Figure 14. Process of synthetic data sharing

It is important to structure the data protection barriers and drivers of synthetic data sharing according to these phases, as these phases are accompanied with different types of privacy risks and institutional challenges. Selection of a real-world dataset for generation is concerned with privacy risks unique to that dataset and the uncertainties regarding the appropriate legal base. Synthetic data generation is concerned with the selection of a model that suits the privacy needs of the use case. Post hoc privacy evaluation is concerned with the legal and technical issues regarding the definition of personal data and the inexplicable quantitative metrics. The application of synthetic data is concerned with needs from research institutes and concluding appropriate agreements for data sharing. The resulting framework is shown in Table 5 and is structured following these four phases.

To further synthetise the findings, the framework assigns the owner of the solution direction and its type as guide through the diverse measures. The owner is the actor responsible for implementing the solution direction. I propose the following types of solution directions:

- **Operational**: practical guidance for health data sharing process;
- **Technical**: alterations to synthetic data generation models, methodologies, and evaluations by technology providers;
- **Political**: opportunities for governmental bodies to clarify existing data sharing policies and stimulate privacy-enhancing health data sharing;
- **Institutionalễ** propose a change in interaction patterns *between* actors;
- **Collective action:** institutional measure that requires shared efforts from actors to meet common objectives that are otherwise not met.

Table 5. Framework with barriers and solution directions for synthetic data sharing oplet

Barrier and drivers	Solution direction	Owner	Type
Phase 1 – Selection of real-world dataset			
[B1] Consent-by-default approach disproportionally restricts health data sharing [B2] Uncertainties regarding presumption of compatibility of anonymisation and scientific research prevent application of synthetic data	[S1] Establish that synthetic data is generated to contribute to scientific research (safeguard for patients)	Healthcare provider; Research institutes	Operational
	[S2] Consider link between the original purpose and the secondary purpose, focusing on the reasonable expectations of patients regarding secondary use (safeguard for patients)	Healthcare provider	Operational
	[S3] Clarify the need for additional legal base for compatible purposes (policy opportunity)	Ministry of HWS; EDPB	Political
	[S4] Harmonise the GDPR risk-based approach regarding a legal base and Dutch consent-by-default approach (policy opportunity)	Ministry of HWS; EPDB	Political
	Same as [S19]	EDPB	Political
[B3] Privacy and utility of synthetic data are difficult to balance	[S5] Following [S1], start with exploration of the intended synthetic data use to whom the data is disclosed (data protection measure)	Healthcare provider	Operational
	[S6] Analyse the real-world dataset and how it represents the population (data protection safeguard)	Healthcare provider	Operational
	[S7] Explore the analytical needs of the intended synthetic data use (utility/research interest)	Healthcare provider	Operational
	[S8] Following [S5-7], adopt de-identification measures to prevent inherent privacy metrics in real-world dataset (data protection)	Healthcare provider	Operational; Technical
Phase 2 – Synthetic data generation			
[B4] Lack of transparency complicates selection of synthetic data generation model	[S9] Support development of opensource alternatives for synthetic data generation (policy opportunity)	Governmental bodies	Political
	[S10] Provide explainable model and privacy evaluations considering legal requirements (interpretability)	Technology provider	Technical
[B5] Difficult to benchmark synthetic data generation models on privacy due to lack of standardisation	[S11] Develop sector-specific guidance or standards to benchmark generation models and datasets.	Technology provider; Research institutes; Ministry of HWS	Collective action; Technical
[B6] No standard for defining the utility/privacy trade-off in model parameters			
[B7] Need for technical expertise to select and evaluate synthetic data generation models	[S12] Form multidisciplinary teams with people with legal, statistical and technical knowledge.	Healthcare providers; technology providers	Collective action; Institutional
Phase 3 – Post hoc privacy evaluation of synthetic data			
[B8] Presumption of privacy by technology providers risks unintentional data disclosure	[S13] Perform a post hoc privacy evaluation	Technology provider	Technical
[B9] Acceptable levels of re-identification risks are not related to context, decreasing the interpretability of quantitative metrics [B10] Lack of standards to evaluate privacy of synthetic data generation models [B11] Lack of interpretability of privacy evaluation limit trust and usability of synthetic data	Same as [S10]	Technology provider	Technical
	[S14] Provide healthcare providers with quantitative privacy metrics with contextual explanation to support the definition of acceptable thresholds	Technology provider	Technical; Operational
	[S15] Evaluate whether the synthetic data meets the legal requirements on a per-case basis	Healthcare provider	Operational
	[S16] Educate users of synthetic data so they can understand processes, advantages, and limitations	Technology provider; healthcare provider; Research institutes	Institutional
	[S17] Develop sector-specific standards for privacy evaluation of synthetic data generation models	Technology provider	Technical; Collective action
[B12] Difficult to meet the legal requirements for anonymisation [B13a] Legal ambiguity complicates standardisation of privacy evaluations	[S18] Combine qualitative and quantitative methods to determine an appropriate privacy risk level	Healthcare provider	Operational
	Same as [S14]	Technology provider	Technical
	[S19] Specify the identifiability-criterium and reasonableness standard to clarify the scope of GDPR	EDPB	Political

Barrier and drivers	Solution direction	Owner	Type
Phase 4 – Synthetic data application			
[B13] Users of synthetic data are unaware of the value of synthetic data	[S20] Inform external researchers about the generation and evaluation process of synthetic data and the properties of the real-world dataset	Healthcare providers	Institutional
	[S21] Enable independent verification and evaluation of synthetic data	Technology providers; Research institutes	Technical; Operational
	[S22] Standardise data sharing agreements	Healthcare providers; Research institutes	Institutional
[D1] Shorten health data sharing processes	Same as [S22]	Healthcare providers; Research institutes	Institutional
[B15] Data findability remains an issue in the ad hoc process of health data sharing	[S24] Follow FAIR principles for data sharing and open science for synthetic data	Healthcare provider	Institutional
	[S25] Develop a central catalogue for metadata of (synthetic) health data	Healthcare provider; Ministry of HWS	Collective action; Technical
[D2] Synthetic data imposes lower risks in comparison to current practice		Healthcare providers; Research institutes	Institutional
[D3] Synthetic data generation supports data protection principles		Patients	Institutional
[D4] Synthetic data generation expands application scope of health data		Healthcare providers; Research institutes	Institutional

5.5.2 Analysis of framework

The validation of the framework presented in Table 5 is provided in Appendix I. In short, the interviewees confirmed the barriers related to interpretability of synthetic data generation models and evaluations, the definition of personal data and the legal base for synthetic data generation. The operational and institutional measures suit the current institutional environment as similar measures are applied for sharing pseudonymous health data (DS1-H-VAL). Moreover, by including evaluation of privacy in the different phases of synthetic data sharing, starting with an analysis of the real-world dataset, a post hoc evaluation of privacy risks, and effective governance, patient privacy can be preserved (DS1-H-VAL, TP1-VAL). Both interviews confirmed there is a need for technical standards. But, emphasised that the process of developing standards is difficult, due to the lack of consensus among technology process; the development of standards is a collective action problem on itself.

The framework shows that there are at least fifteen data protection-related barriers that inhibit the impact of synthetic data on health data sharing. This means that presenting synthetic data generation as silver bullet oversells its utility. These barriers require effort (and thus money and time) from multiple actors, that rely on each other's knowledge and experiences. While the framework only mentions organisations, within them many roles are involved, further complicating implementation of the solutions. Moreover, these organisations must assure long-term resource allocation for their implementation: the barriers apply to four phases, and many reoccur for each new instance wherein data is shared.

5.6 Conclusion

This chapter formulated a framework with barriers that should be addressed to enable synthetic health data sharing in a privacy-enhancing manner (Table 5). Accordingly, the framework provides tangible and structured solution directions to healthcare providers, research institutes, technology developers, scholars, and policymakers alike.

To start with the drivers, synthetic data enables health data sharing by allowing for further standardisation of data sharing agreements, shortening the process of obtaining access to health data. Synthetic data generation implies lower re-identification risks for patients in comparison to pseudonymous data and contributes to the principles of data protection, enabling data sharing in a privacy-enhancing manner. The lower re-identification risks expand the possible use cases for health data sharing, which benefits data-driven research.

Yet, there are barriers in various phases of synthetic data sharing. To summarise, the unclear legal concepts of personal data and anonymisation make it difficult to define synthetic data generation as a anonymisation technique. Also, the ambiguous interpretation of the required legal base for synthetic data generation in the Netherlands, inhibit the impact of synthetic data on health data sharing. In terms of technicalities of synthetic data generation, there is still work to be done regarding privacy evaluations. The lack of interpretable and standardised privacy evaluations complicate the assessment of privacy risk for patients by users of synthetic data. These factors have a negative effect on the possible impact synthetic data generation may have on health data sharing for research – and should be addressed accordingly.

The framework proposes solution directions that followed from the institutional analysis that help to enable synthetic data sharing.

6
Conclusion and discussion

6.1 Research findings

To bridge the gap between the technological reviews of synthetic data generation methods and data protection law and practice, this thesis studied the following research question:

> *How could synthetic health data be shared for research purposes in consideration of the institutional data protection environment?*

This answer to this research question is formed via three research questions:

RQ1 What is the institutional data protection environment of health data sharing for research?

The institutional analysis of secondary use of health data showed the institutional data protection environment is complex and multifaceted, involving various actors at different levels of governance. The studied use case comprised of healthcare provider, research institute and the Ministry of HWS. The institutional landscape, governed by formal regulations such as the GDPR, UAVG, and WGBO, and non-binding rules from collectives, is characterised by uncertainties regarding the legal definition of personal data and anonymisation, contradiction, regarding the use of consent as default legal base in the Netherlands, and time and labour-intensive procedures to share health data. The challenges occurred at various levels, where the interaction patterns regarding use of consent and definition of anonymisation are decided at the national and EU level, data sharing procedures emerge at inter-organisational level. Yet, these challenges are linked, as the data sharing procedures are formulated on the strict approach towards health data sharing for research in the Netherlands. In essence, the institutional data protection environment presents significant barriers to the secondary use of health data for research, limiting its potential for advancing medical knowledge and improving patient outcomes. In addition, one may question whether the data sharing resources provide sufficient safeguards for patients, considering they are focussed on liability instead of data protection. provided by data sharing agreements, of current processes. Addressing these challenges requires a coordinated effort to streamline data sharing processes, clarify regulatory ambiguities, and ensure alignment between formal regulations and practical implementation.

RQ2 How could synthetic data generation and evaluation contribute to health data protection?

Synthetic data generation is seen as a privacy-enhancing technology to increase data availability for research. From the analysis of types and characteristics of generation methods, specifically for health data in the form of EHRs, it can be concluded that synthetic data generation can offer a privacy-enhancing way to share data for research – there are various approaches with little privacy risks that can model the complexities of EHRs. However, it is important to note that some models that presumed privacy still contain privacy risks. For instance, researchers have identified instances where synthetic datasets inadvertently revealed individuals' inclusion or disclosed specific attributes when compared with real-world training data. While newer AI-based techniques (e.g. GANs) show promise in better preserving privacy in comparison to other anonymisation and statistical synthetic generation techniques, there remains a need for formal privacy evaluations due to potential re-identification risks. Furthermore, the lack of consensus on privacy metrics and the scarcity of explicit privacy evaluations poses significant challenges for interpreting privacy risks. This makes it harder for data owners (e.g. care providers) to know when the privacy risk of sharing synthetic health data is (un)acceptable. The same applies for researchers wishing to use synthetic health data for research while warranting the original patients' data protection. Moving forward, collaborative efforts to establish standards for safe and privacy-preserving synthetic data generation are crucial to ensure effective data protection in research contexts.

RQ3 Which data protection-related factors influence how synthetic data generation enables health data sharing for research?

The factors that influence the impact of synthetic data generation on the data sharing behaviour of actors are identified by combining the institutional environment (RQ1) with the knowledge on synthetic EHRs (RQ2). These findings were converted into a framework that structures barriers and drivers during the various phases of synthetic health data sharing. Based on the institutional analysis, solution directions were proposed to address these barriers. These are seen as important factors to enable synthetic health data sharing, and hence, to answer the research question of this thesis.

Synthetic data generation can aid compliance with the institutional context. It can contribute to the data protection principles of data security, data minimisation, and data accuracy, by reducing the need to share genuine personal health data with higher re-identification risks for patients. Also, synthetic data broadens data sharing possibilities, as there are significantly less privacy risks in comparison to pseudonymised data. Regarding the time-intensive process of concluding data sharing agreements, synthetic data enables further standardisation of templates.

Also, sharing synthetic data is accompanied with legal barriers. First, it is difficult to assess whether data protection regulations apply to generating and using synthetic data. This question is difficult to answer, considering the very broad definition of personal data. It can even be questioned whether this definition is technically feasible, considering that there always is a chance of re-identification, depending on the time and computational resources attackers have to re-identify patients. On top of that, the legal framework around anonymisation is characterised by poorly defined concepts, meaning it is unclear which standards need to be met in order to classify synthetic data as anonymous; this makes a case-by-case evaluation of synthetic data compliance inevitable, which slows down data sharing processes for research. There is a policy opportunities for Dutch governmental bodies and EU bodies to clarify concepts on the intersection of data protection law, synthetic data, and anonymisation. I propose a contextual approach to harmonise the technical and legal definition of anonymisation.

Furthermore, this thesis showed that a privacy risk assessment of synthetic data in the process of sharing health data is complicated by the way synthetic data is now evaluated: academics, if they perform an assessment of privacy at all, mainly provide quantitative metrics based on membership inference and attribute disclosure attacks – metrics that are difficult for less technical literate people to translate into concrete privacy risks that are needed to determine whether an anonymisation technique is effective. Adding to this, the literature provides limited support for the interpretation of these quantitative metrics for EHRs by qualitative arguments, for example by placing them in the context of data sharing or relating them to the legal definition of anonymisation. This is important, considering that synthetic data could still pose privacy risks that must be evaluated to implement additional safeguards. As a result, both healthcare providers and research institutes struggle in assessing privacy risks of generating and sharing synthetic health data. This limits the application of synthetic data, and hence, its impact on health data sharing. Such evaluations should be based on legal requirements, to the extent possible, given the above legal barrier. An explicit and explainable evaluation, based on thresholds that fit the context in which data is applied, should better enable healthcare providers and research institutes to evaluate synthetic data with confidence; their confidence is seen as barrier for synthetic data sharing. This helps to better protect patients' privacy, by ensuring there are no privacy risks that are unaccounted for. There is a role for technology providers to actively foster users' understanding of synthetic data and accompanying privacy measures, and with heads of healthcare and research institutes, to equip their support staff and users with sufficient resources and knowledge.

A second group of legal barriers arises regarding the Dutch approach towards consent: the friction between the consent-by-default approach, the presumption of compatibility of scientific in the GDPR, and the presumption of compatibility of anonymisation poses uncertainties regarding the question whether generating synthetic data requires a separate legal base. Following the wording of Dutch law, the generation of synthetic data based on health data for research purposes must rely on its own lawful base, meaning consent must be sought for this act. This slows down the process of sharing synthetic health data for research. I presented how the current approach goes against EU guidance and the GDPR's leeway to balance data protection and public research interest, calling for a policy revision.

Naturally, the identification of the factors identified in RQ3 result in an answer to the main research question. In comparison to the current health data sharing process, synthetic data can enable health data sharing by embodying the principles of data protection law, easing the current health data sharing process, and increasing research opportunities.

Moreover, this thesis showed that each stage of health data sharing differs in types of problems and actors that can be involved to attempt to mitigate the barriers. The resulting framework contained over fifteen data barriers that disprove claims that synthetic data generation can be a silver bullet for sharing health data for research. While solution directions exist, there require time and money investments from multiple actors that are codependent for others' knowledge, experience, and long-term buy-in. Even within organisations, many actors play a role. Especially the long-term commitment is important, as the barriers apply across phases, with many reoccurring for each new synthetic data sharing initiative. By addressing the barriers in the framework, and implement the solution directions, the potential of synthetic data can be better attained. An important limitation herein is that some factors relate not necessarily with the development of synthetic data, but rather play out on a political level. Therefore, the extent to which synthetic data generation can enable health data sharing is dependent on EU and national governmental bodies to take the proposed policy opportunities.

Answering the main research question of this thesis, requires assessing how actors involved in health data sharing can work around these uncertainties caused by legal ambiguities and a lack of interpretability, to share health data in a privacy-enhancing manner. How researchers and health institutes can work around this, is further detailed in the practical contributions (§6.3.2).

6.2 Limitations

To value the research findings, this section discusses the limitations regarding methodology, the dynamics of the regulatory landscape, practical application, and scope of the proposed framework.

6.2.1 Methodology

First, the DSR approach underlying this research calls for frequent iteration between the phases, to repeatedly inform the designed artefact with learnings from the literature and empirically studied context (Hevner et al., 2004). Since the information regarding the institutional environment was limited, the primary focus was to generation knowledge on this context and provide a solid knowledge base to compare this with synthetic data. This prioritisation choice, however, led to less elaborate research activities regarding the design of a framework and evaluation thereof. As a result, because of the limited time wherein this thesis was conducted, evaluation was not feasible beyond the two evaluative interviews. The conceived framework would benefit from further review with practitioners and scholars to tweak it further.

Second, because interviews are time-consuming (in identifying suitable interviewees, arranging an interview, conducting it, and analysing it), the breadth of the research suffers (Knott et al., 2022). There might be other use cases that have experience with synthetic data generation. Their stories could be relevant for improving synthetic data generation for health data reuse, too, but there was no time in this research to broaden the scope to that extent. This extension would also aid the generalisability of the framework: because the research is situated in one use case, problems or guidance identified here might turn out differently in other use cases.

Similarly, the scope of this thesis further limits generalisability: the institutional analysis mainly focused on rules and actors in the Netherlands; an institutional environment that is quite unique for its implementation of the GDPR for health data sharing for research. This complicates generalising the results to other countries. As the identified barriers are primarily founded in unclear boundaries of the core concepts of the EU data protection framework (the presumption of compatibility for research and the definition of personal data). Therefore, the barriers caused by the EU data protection framework are considered uncertainties for all actors involved with synthetic data, also those located outside of the Netherlands. Barriers related to the strict consent-by-default approach, however, apply in the Netherlands only.

Similarly, only organisations that use synthetic data generation were interviewed, so the results might be biased to be positive about SDG, because of the interviewees' investments into SDG. Investigating healthcare providers or researchers that considered synthetic data but refrained from using it, might be interesting participants for future research, to hear which barriers they encountered. Moreover, only a single interview was conducted with an interviewee involved in policy making of secondary use of health data. The interview provided limited insight in why current consent and anonymisation policies are ambiguous. There might be reasons explaining the policy gaps and mismatches, which remain unclarified in this thesis.

6.2.2 Dynamic regulatory landscape

The analysis of this thesis focussed on the current institutional environment. In interpreting these results, it is important to know the regulatory landscape is changing with the EU's plan to create a EHDS to promote the free flow of health data (Proposal for the European Health Data Space, 2022). The regulatory proposal introduces a lawful base for secondary use of health data for various predefined purposes, among which scientific research and innovation activities (Proposal for the European Health Data Space, 2022).. The proposal eliminates the use of consent as lawful base for secondary use of health data, solving issues that are identified in this thesis.

Hence, there are limitations regarding the relevance of examining the current institutional framework and analysing the Dutch consent-by-default approach, knowing these interaction patterns will drastically change in the future. What should be noted here, is that the EHDS regulation is still being written – there is no agreement yet on the final text (interview 10). In addition, it is a very ambitious project, one that significantly changes the data infrastructure and access arrangements. It is reportedly going to take years to finalise and then implement this legislation (interview 4, interview 10). At the same time, the demand for data availability and the changes required for it is high, and demand that can be answered in part by the development of synthetic data. Given that synthetic data is currently making its way into healthcare practice, with more and more hospitals using this technology, it is relevant to encourage a privacy-preserving implementation of synthetic data in the near term – an overview of the current institutional environment helps in this regard (interview 8). Finally, insights into current practice can provide insight into the potential of synthetic data for application in the EHDS (interview 10). Therefore, this is considered an acceptable limitation.

6.2.3 Scope of framework

As articulated on in this thesis, the application of synthetic data requires a trade-off between utility and privacy. Because this thesis focused on analysing privacy metrics in detail, utility metrics remain largely undiscussed. To properly assess how to balance utility and privacy in different applications, an understanding of both metrics is needed. Therefore, in this thesis, no substantive statements could be made on the interpretation of privacy metrics, merely on how they were presented. This limitation is justified by time constraints, and prioritisation of privacy is justified by the literature gap of privacy evaluations of synthetic data in socio-technical research. However, this results in the absence of substantive contributions to important technical questions, such as how privacy metrics can be standardised. Future research could further dive into this trade-off.

6.3 Contributions of research and future research

This section discusses the scientific contributions to regarding methodology, the dynamics of the regulatory landscape, practical application, and scope of the proposed framework.

6.3.1 Scientific contributions

This thesis presents empirical knowledge of how synthetic data generation is used for health data sharing for research. Therefore, the thesis provides the literature with real-world experiences, including a discussion of specificities and privacy concerns that characterise this domain, and demonstrating the misalignment of technical and legal literature and practice. I showed what data-protection related barriers exist, with possible solution directions, for synthetic data sharing. This section discusses how these findings fit certain areas of research.

First, this thesis contributes to the synthetic data literature in the field of computer science, by showing that the quantitative privacy evaluations are non-existent or diverse, showing the need for developing standards. Moreover, this thesis shows that the interpretations of quantitative privacy evaluations are limited; acceptable privacy risk thresholds are not substantiated with qualitative arguments and if an interpretation is given at all, the scores are not interpretated within its application context. Therefore, I propose the following solution directions for researchers: 1) perform a post hoc privacy evaluation; 2) include argumentation for the acceptable thresholds; and 3) qualitatively relate results of the evaluation to the application context of their proposed models. Researchers should consider the legal definition of anonymisation in their argumentation for acceptable thresholds and the application context. The implications for these scholars thus relate to how they perform their privacy evaluations.

Second, this thesis contributes to the legal debate of the scope of EU data protection law. The debates revolves around the definition of personal data in light of current and future technical developments. While some researchers argue the GDPR is up for this challenge (see e.g. Cruz, 2023), I join the researchers that argue that the current definition of personal data makes EU data protection law the “law of everything” (see e.g. Purtova, 2018). At least with the current interpretation of the identifiability criterium and reasonableness standard, I provide an empirical argument for how the broad definition of personal data complicates the synthetic health data sharing; this while data protection law should actually stimulate the interpretation of such privacy-enhancing technologies.

Third, this thesis contributes to DSR research. I took the freedom to apply the three cycle approach of Hevner et al. (2004) to design a framework that identifies barriers for synthetic data sharing based on an institutional analysis. By doing so, I showed how the analysis of the environment can be completed with the IAD framework by Ostrom. This novel approach can guide the design of artefacts that have institutional requirements.

Lastly, this thesis provided an agenda for scholars researching how health research can be facilitated, or how data protection regulations apply to secondary use of health data for research. Multiple knowledge gaps have been identified that should be addressed in further research, to come closer to attaining the goals of furthering research and therefore public health. These knowledge gaps are discussed in §6.3.3.

6.3.2 Practical contributions

§6.1 highlighted the many policy gaps complicating sharing synthetic health data for research in a legally compliant way. But that does not mean that researchers and healthcare providers must sit and wait until governmental organisations address these gaps. This research has numerous practical contributions that aid actors that want to engage in such collaborations in this ambiguous institutional environment.

The proposed framework can be applied to update data policies and manage synthetic data in a privacy-enhancing manner. The operational guidance in Table 5 can be summarised in a reformed application process of synthetic data that preserves privacy of patients in health data sharing for research (Figure 15). Considering the challenge that synthetic data in its current form cannot not eliminate all privacy risks, it is up to the data owners to adopt safeguards to protect patients that ensure that synthetic data does not unintentionally or without consideration, disclose personal data. This thesis argues for a considerate approach towards synthetic data sharing, making explicit the privacy risks at each. The authors emphasise the importance of an evaluation framework rather than relying solely on quantitative analysis of privacy metrics. Privacy should not be considered as an afterthought – instead, privacy should be evaluated in the various stages of sharing synthetic data. Therefore, application of synthetic data requires understanding of ethical and legal implications, in addition to how synthetic data is generated and (quantitatively) validated.

In Figure 15, these considerations (next to the operational guidance in Table 5) lead to a risk-based evaluation process, that requires the definition of the intended use and context of data sharing, an analysis of the characteristics of the real-world dataset that is transformed, a post hoc privacy evaluation, as well as an institutional analysis that combines these analyses. By following these phases, appliers of synthetic data can demonstrate they account for data protection risks that suit the application context. This framework should guide users of synthetic data through the regulatory ambiguity.

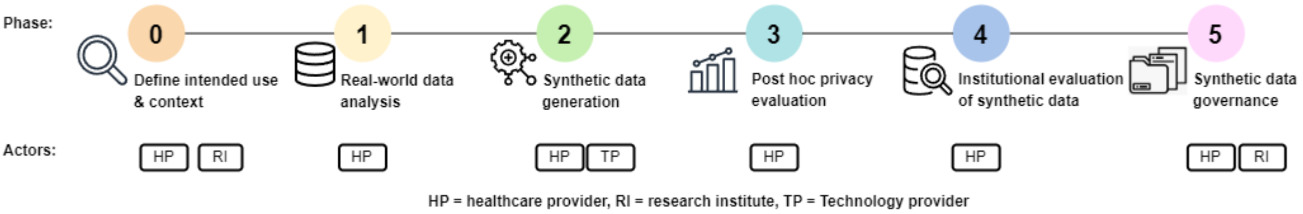

Phase:
0
Define intended use & context
1
Real-world data analysis
2
Synthetic data generation
3
Post hoc privacy evaluation
4
Institutional evaluation of synthetic data
5
Synthetic data governance
Actors:
HP
RI
HP
HP
TP
HP
HP
HP
RI
HP = healthcare provider, RI = research institute, TP = Technology provider

Figure 15. Process of Privacy-Enhancing sharing of synthetic data in healthcare (based on Table 5)

Furthermore, synthetic data generation technology companies should focus on making their technology understandable for users. Users can then better understand the context-specificity, and that synthetic data generation is not a one-size-fits-all solution for all (health) data reuse matters. There is a role for data stewards to be played here, too: organisations should employ support staff capable in understanding the technology, practice of researchers, and legal nature of the data processing at hand. This should also alleviate the burden on researchers to obtain technical knowledge (about synthetic data generation) and legal knowledge (about the (mis)matches with privacy regulations).

Similarly, this thesis highlighted the need for thoughtful governance of synthetic data. Examples include metadata documentation about how the synthetic data was generated and validated, and who it represents (e.g. with a standardised metadata framework); the need for independent verification; and the need for preventing synthetic and authentic data from being mixed up. Synthetic data generation companies, users, and legal experts could sit together to jointly discuss their needs and define standardised practices and frameworks to these ends. The companies could then investigate the feasibility of implementing these in their software, for instance such that generating synthetic data requires that a metadata template is filled out.

Finally, policy makers and supervisory authorities are encouraged to harmonise their different interpretations of data protection and privacy law. Chapter 3 demonstrated that the multiple legal frameworks that govern synthetic data generation in healthcare applications, at both an EU and national level, are at times contradictory. If researchers adopt an interpretation that later appears to conflict with that of a supervisory authority, even though it may be dominant in public discourse or in other EU Member States, this could result in large fines. Consequently, researchers might fear breaching these rules and therefore abstain from using synthetic data at all, meaning the benefits of data reuse for research are not attained.

6.3.3 Future research

The thesis also generated recommendations for future research.

First, the interviews with practitioners showed that there is a lack of (interpretations of) sector-specific standardised privacy evaluations, making it hard to assess privacy risks when sharing or reusing synthetic health data. A first recommendation is therefore to conduct scientific research in how privacy metrics can be standardised, taking into account the legal interpretation of anonymisation. In addition, this thesis calls for (technical) synthetic data researchers to take an example from El Emam et al. (2020) to relate the evaluation results of synthetic data generation to their context. A non-technical recommendation herein, as articulated in Appendix G, is how the technical metrics can be explained to non-technical users, so they feel confident in making decisions regarding the synthetic data use. This could include a qualitative or quantitative study, that lets non-technical people try to explain what metrics are presented to them and measure the confidence in synthetic data sharing, depending on the metrics presented to them. Presentations of metrics can be varied, for example in visual support, different complexities of explanations, to find the most understandable way to present privacy metrics.

Secondly, while privacy is a crucial value to consider in health data reuse contexts, it is not the only public value at stake. In addition to utility, as explained in the limitations, examples of other relevant values that the framework could be extended with are fairness, transparency, and accountability. To initiate further discourse, it is recommended that the guidelines are elaborated on to embody the principles of transparency, accountability, and fairness. Accountability involves the establishment of well-defined procedures to hold accountable those involved in the creation of synthetic data models and generation of synthetic data (Beduschi, 2024). The thesis already proposed guidelines regarding documentation of generation and evaluation to promote transparency. Research can be conducted to reflect this guideline into concrete steps in the synthetic data generation lifecycle. Fairness encompasses assurances that the generation and utilisation of synthetic data do not engender adverse impacts on individuals or society, such as perpetuating existing biases or introducing novel ones. Although some fairness challenges are touched upon in the context of data protection, the societal impact and potential detrimental effects of synthetic data requires more research.

References

Abay, N. C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., & Sweeney, L. (2019). Privacy Preserving Synthetic Data Release Using Deep Learning. In M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, & G. Ifrim (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 11051, pp. 510–526). Springer International Publishing. https://doi.org/10.1007/978-3-030-10925-7_31

Abedi, M., Hempel, L., Sadeghi, S., & Kirsten, T. (2022). GAN-Based Approaches for Generating Structured Data in the Medical Domain. *Applied Sciences*, *12*(14), 7075. https://doi.org/10.3390/app12147075

About Coreon: Purpose and mission. (n.d.). *Coreon*. Retrieved 28 January 2024, from https://www.coreon.org/about-coreon/

Adams, W. C. (2015). Conducting Semi-Structured Interviews. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of Practical Program Evaluation* (1st ed., pp. 492–505). Wiley. https://doi.org/10.1002/9781119171386.ch19

Agencia Española Protección Datos. (2021). *10 Misunderstandings related to anonymisation*. https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf

Alloza, C., Knox, B., Raad, H., Aguilà, M., Coakley, C., Mohrova, Z., Boin, É., Bénard, M., Davies, J., Jacquot, E., Lecomte, C., Fabre, A., & Batech, M. (2023). A Case for Synthetic Data in Regulatory Decision-Making in Europe. *Clinical Pharmacology & Therapeutics*, *114*(4), 795–801. https://doi.org/10.1002/cpt.3001

Appenzeller, A., Leitner, M., Philipp, P., Krempel, E., & Beyerer, J. (2022). Privacy and Utility of Private Synthetic Data for Medical Data Analyses. *Applied Sciences*, *12*(23), Article 23. https://doi.org/10.3390/app122312320

Article 29 Data protection working party. (2011). *Advice paper on special categories of data (“sensitive data”)* (res(2011)444105). https://ec.europa.eu/justice/article-29/documentation/other-document/files/2011/2011_04_20_letter_artwp_mme_le_bail_directive_9546ec_annex1_en.pdf

Article 29 Working Party. (2014). *Opinion 05/2014 on Anonymisation Techniques* (0829/14/EN). https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Azizi, Z., Lindner, S., Shiba, Y., Raparelli, V., Norris, C. M., Kublickiene, K., Herrero, M. T., Kautzky-Willer, A., Klimek, P., Gisinger, T., Pilote, L., & El Emam, K. (2023). A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health. *Scientific Reports*, *13*(1), 11540. https://doi.org/10.1038/s41598-023-38457-3

Becker, R., Chokoshvili, D., Comandé, G., Dove, E., Hall, A., Mitchell, C., Molnar-Gabor, F., Nicolás, P., Tervo, S., & Thorogood, A. (2022). *Secondary use of Personal Health Data: When is it ‘Further Processing’ under the GDPR, and What Are the Implications for Data Controllers?* (SSRN Scholarly Paper 4070716). https://doi.org/10.2139/ssrn.4070716

Beduschi, A. (2024). Synthetic data protection: Towards a paradigm change in data regulation? *Big Data & Society*, *11*(1), 20539517241231277. https://doi.org/10.1177/20539517241231277

Bellovin, S. M., Dutta, P. K., & Reitinger, N. (2019). Privacy and Synthetic Datasets. *Stanford Technology Law Review*, *22*, 1–39.

Benfeldt, O., Persson, J. S., & Madsen, S. (2020). Data Governance as a Collective Action Problem. *Information Systems Frontiers*, *22*(2), 299–313. https://doi.org/10.1007/s10796-019-09923-z

Biswal, S., Ghosh, S., Duke, J., Malin, B., Stewart, W., Xiao, C., & Sun, J. (2021). *EVA: Generating Longitudinal Electronic Health Records Using Conditional Variational Autoencoders*. *149*, 260–282. Scopus.

Boudewijn, A., Filippo Ferraris, A., Panfilo, D., Cocca, V., Zinutti, S., De Schepper, K., & Rossi Chauvenet, C. (2023, November 1). Privacy Measurement in Tabular Synthetic Data: State of the Art and Future Research Directions. *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*. https://doi.org/10.48550/arXiv.2311.17453

Boyd, M., Zimeta, D. M., Tennison, D. J., & Alassow, M. (2021a). *Secondary use of health data in Europe* (p. 38). Open Data Institute.

Boyd, M., Zimeta, D. M., Tennison, D. J., & Alassow, M. (2021b). *Secondary use of health data in Europe*. 38.

Braddon, A. E., Robinson, S., Alati, R., & Betts, K. S. (2023). Exploring the utility of synthetic data to extract more value from sensitive health data assets: A focused example in perinatal epidemiology. *Paediatric and Perinatal Epidemiology*, *37*(4), 292–300. https://doi.org/10.1111/ppe.12942

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brauneck, A., Schmalhorst, L., Majdabadi, M. M. K., Bakhtiari, M., Völker, U., Baumbach, J., Baumbach, L., & Buchholtz, G. (2023). Federated Machine Learning, Privacy-Enhancing Technologies, and Data Protection Laws in Medical Research: Scoping Review. *Journal of Medical Internet Research*, *25*(1), e41588. https://doi.org/10.2196/41588

Breyer, Reports of Cases ___ (Court of Justice of the European Union 2016). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62014CJ0582

Burt, A., Rossi, A., & Stalla-Bourdillon, S. (2021, July 15). *A guide to the EU's unclear anonymization standards* [IAPP]. https://iapp.org/news/a/a-guide-to-the-eus-unclear-anonymization-standards/

Carballa-Smichowski, B., Duch-Brown, N., & Martens, B. (2021). *To pool or to pull back? An economic analysis of health data pooling* (JRC126961). Joint Research Centre. https://joint-research-centre.ec.europa.eu/system/files/2021-12/jrc126961.pdf

Charter of Fundamental Rights of the European Union, Official Journal of the European Union C 364/1 (2000).

Chauhan, P., Bongo, L. A., & Pedersen, E. (2023). Ethical Challenges of Using Synthetic Data. *Proceedings of the AAAI Symposium Series*, *1*(1), 133–134. https://doi.org/10.1609/aaaiss.v1i1.27490

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), Article 6. https://doi.org/10.1038/s41551-021-00751-8

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. *Proceedings of the 2nd Machine Learning for Healthcare Conference*, 286–305. https://proceedings.mlr.press/v68/choi17a.html

Commission. (2020). *A European strategy for data*. https://ec.europa.eu/info/sites/default/files/communication-european-strategy-data-19feb2020_en.pdf

COREON. (2018). *COREON Statement Wetenschappelijk onderzoek*. https://www.coreon.org/wp-content/uploads/2020/04/COREON-statement-wetenschappelijk-onderzoek-v1.6-17-12-2018.pdf

COREON. (2022). *Code of Conduct for Health Research*. https://www.coreon.org/wp-content/uploads/2023/06/Code-of-Conduct-for-Health-Research-2022.pdf

Coutinho-Almeida, J., Rodrigues, P. P., & Cruz-Correia, R. J. (2021). GANs for Tabular Healthcare Data Generation: A Review on Utility and Privacy. In C. Soares & L. Torgo (Eds.), *Discovery Science* (pp. 282–291). Springer International Publishing. https://doi.org/10.1007/978-3-030-88942-5_22

Coyle, D., Diepeveen, S., Wdowin, J., Kay, L., & Tennison. (2020). *The value of data: Policy implications*. Bennett Institute, University of Cambridge; Open Data Institute. https://www.consilium.europa.eu/media/46496/st11481-en20.pdf

Cruz, H. G. T. D. (2023). Exploring the Tenability of the GDPR Becoming the 'Law of Everything'. *Amsterdam Law Forum*, *15*, 58.

Dahmen, J., & Cook, D. (2019). SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors*, *19*(5), 1181. https://doi.org/10.3390/s19051181

Data Protection Commission. (2019). *Guidance Note: Guidance on Anonymisation and Pseudonymisation*. https://www.dataprotection.ie/sites/default/files/uploads/2019-06/190614%20Anonymisation%20and%20Pseudonymisation.pdf

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

de Mul, M., Alons, P., van der Velde, P., Konings, I., Bakker, J., & Hazelzet, J. (2012). Development of a clinical data warehouse from an intensive care clinical information system. *Computer Methods and Programs in Biomedicine*, *105*(1), 22–30. https://doi.org/10.1016/j.cmpb.2010.07.002

Determann, L. (2020). Healthy Data Protection. *Michigan Technology Law Review*, *26*(2), 229–278. https://doi.org/10.36645/mtlr.26.2.healthy

Dikici, E., Bigelow, M., White, R. D., Erdal, B. S., & Prevedello, L. M. (2021). Constrained generative adversarial network ensembles for sharable synthetic medical images. *Journal of Medical Imaging*, *8*(2), 024004. https://doi.org/10.1117/1.JMI.8.2.024004

Diller, G.-P., Vahle, J., Radke, R., Vidal, M. L. B., Fischer, A. J., Bauer, U. M. M., Sarikouch, S., Berger, F., Beerbaum, P., Baumgartner, H., Orwat, S., & for the German Competence Network for Congenital Heart Defects Investigators. (2020). Utility of deep learning networks for the generation of artificial cardiac magnetic resonance images in congenital heart disease. *BMC Medical Imaging*, *20*(1), 113. https://doi.org/10.1186/s12880-020-00511-1

Doutreligne, M., Degremont, A., Jachiet, P.-A., Lamer, A., & Tannier, X. (2023). Good practices for clinical data warehouse implementation: A case study in France. *PLOS Digital Health*, *2*(7), e0000298. https://doi.org/10.1371/journal.pdig.0000298

Dove, E. S., & Phillips, M. (2015). Privacy Law, Data Sharing Policies, and Medical Data: A Comparative Perspective. In A. Gkoulalas-Divanis & G. Loukides (Eds.), *Medical Data Privacy Handbook* (pp. 639–678). Springer International Publishing. https://doi.org/10.1007/978-3-319-23633-9_24

Downes, L. (2009). *The laws of disruption: Harnessing the new forces that govern life and business in the digital age*. Basic Books.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi & T. Rabin (Eds.), *Theory of Cryptography* (pp. 265–284). Springer. https://doi.org/10.1007/11681878_14

El Emam, K. (2020, February 26). Accelerating AI with synthetic data. *IAPP*. https://iapp.org/news/a/accelerating-ai-with-synthetic-data/

El Emam, K., Mosquera, L., & Bass, J. (2020). Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *Journal of Medical Internet Research*, *22*(11), e23139. https://doi.org/10.2196/23139

European Commission. (2020). *A European Health Data Space: Harnessing the power of health data for people, patients and innovation* ((Communication) COM(2022) 196 final).

European Data Protection Board. (2020). *Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak*. https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202003_healthdatascientificresearchcovid19_en.pdf

European Data Protection Board. (2021, June 15). *Norwegian DPA: Norwegian Confederation of Sport fined for inadequate testing | European Data Protection Board*.

https://edpb.europa.eu/news/national-news/2021/norwegian-dpa-norwegian-confederation-sport-fined-inadequate-testing_en

European Data Protection Supervisor. (2020). *A Preliminary Opinion on data protection and scientific research*. https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf

European Medicines Agency. (2016). *External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use* [EMA/90915/2016]. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use-first-version_en.pdf

Filgueiras, F., & Silva, B. (2021). Assessing Data Policy by Institutional Analysis Development Framework. *5th International Conference on Public Policy*, 1–25. https://www.researchgate.net/publication/351945808_Assessing_Data_Policy_by_Institutional_Analysis_Development_Framework

Friese, S., Soratto, J., & Pires, D. (2018). *Carrying out a computer-aided thematic content analysis with ATLAS.ti* (Working Papers WP 18-02; pp. 1–29). Max Planck Institute. https://www.mmg.mpg.de/62130/wp-18-02

Frischmann, B. M., Madison, M. J., & Strandburg, K. J. (Eds.). (2014). *Governing knowledge commons*. Oxford University Press.

Gal, M., & Lynskey, O. (2023). Synthetic Data: Legal Implications of the Data-Generation Revolution. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4414385

Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *Npj Digital Medicine*, *6*(1), Article 1. https://doi.org/10.1038/s41746-023-00927-3

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(1), 108. https://doi.org/10.1186/s12874-020-00977-1

Gonçalves, M. E. (2020). The risk-based approach under the new EU data protection regulation: A critical perspective. *Journal of Risk Research*, *23*(2), 139–152. https://doi.org/10.1080/13669877.2018.1517381

Groos, D., & Van Veen, E. (2020). Anonymised Data and the Rule of Law. *European Data Protection Law Review*, *6*(4), 498–508. https://doi.org/10.21552/edpl/2020/4/6

Guan, J., Li, R., Yu, S., & Zhang, X. (2018). Generation of Synthetic Electronic Medical Record Text. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 374–380. https://doi.org/10.1109/BIBM.2018.8621223

Gwon, H., Ahn, I., Kim, Y., Kang, H. J., Seo, H., Choi, H., Cho, H. N., Kim, M., Han, J., Kee, G., Park, S., Lee, K. H., Jun, T. J., & Kim, Y.-H. (2024). LDP-GAN: Generative adversarial networks with local differential privacy for patient medical records synthesis. *Computers in Biology and Medicine*, *168*, 107738. https://doi.org/10.1016/j.compbiomed.2023.107738

Haendel, M. A., Chute, C. G., Bennett, T. D., Eichmann, D. A., Guinney, J., Kibbe, W. A., Payne, P. R. O., Pfaff, E. R., Robinson, P. N., Saltz, J. H., Spratt, H., Suver, C., Wilbanks, J., Wilcox, A. B., Williams, A. E., Wu, C., Blacketer, C., Bradford, R. L., Cimino, J. J., … the N3C Consortium. (2021). The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association*, *28*(3), 427–443. https://doi.org/10.1093/jamia/ocaa196

Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., & Nakayama, H. (2018). GAN-based synthetic brain MR image generation. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 734–738. https://doi.org/10.1109/ISBI.2018.8363678

Hansen, W. J., Wilson, P., Verhoeven, E., Kroneman, M., Verheij, R., & van Veen, E.-B. (2021). *Assessment of the EU Member States' rules on health data in the light of GDPR*. Publications Office of the European Union. https://doi.org/10.2818/546193

Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z., & Tang, H. (2024). *Synthetic Data in AI: Challenges, Applications, and Ethical Implications* (arXiv:2401.01629). arXiv. http://arxiv.org/abs/2401.01629

Health Community of Interest. (2022). *Policy Considerations for the Use of Synthetic Healthcare Data*. ACT-IAC. https://www.actiac.org/system/files/2022-01/VA%20Synthetic%20Data_0.pdf

Health RI. (2022, November 8). *Building blocks of a health data infrastructure: Together with and for data users and data holders | Health-RI*. https://www.health-ri.nl/en/news/building-blocks-health-data-infrastructure-together-and-data-users-and-data-holders

Hendolin, M. (2022). Towards the European health data space: From diversity to a common framework. *Eurohealth*, *27*(2), 15–17.

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, *493*, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053

Hess, C., & Ostrom, E. (2006). Introduction: An Overview of the Knowledge Commons. In C. Hess & E. Ostrom (Eds.), *Understanding Knowledge as a Commons* (pp. 3–26). The MIT Press. https://doi.org/10.7551/mitpress/6980.003.0003

Heurix, J., Zimmermann, P., Neubauer, T., & Fenz, S. (2015). A taxonomy for privacy enhancing technologies. *Computers & Security*, *53*, 1–17. https://doi.org/10.1016/j.cose.2015.05.002

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105.

Hevner, A. R., & Wickramasinghe, N. (2018). Design Science Research Opportunities in Health Care. In N. Wickramasinghe & J. L. Schaffer (Eds.), *Theories to Inform Superior Health Informatics Research and Practice* (pp. 3–18). Springer International Publishing. https://doi.org/10.1007/978-3-319-72287-0_1

Hildebrandt, M. (2019, June 2). 5. Privacy and Data Protection. *Law for Computer Scientists*. https://lawforcomputerscientists.pubpub.org/pub/doreuiyy/release/7

Hordern, V. (2022, February 4). *Defining and using health data*. TaylorWessing. https://www.taylorwessing.com/en/insights-and-events/insights/2022/01/defining-and-using-health-data

Hutchinson, T. (2016). The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law. *Erasmus Law Review*. https://doi.org/10.5553/ELR.000055

Iacob, N., & Simonelli, F. (2020). Towards a European Health Data Ecosystem. *European Journal of Risk Regulation*, *11*(4), 884–893. https://doi.org/10.1017/err.2020.88

Information Commissioner's Office. (2023). *Privacy-enhancing technologies (PETs)*. https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies-1-0.pdf

Jadon, A., & Kumar, S. (2023). *Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy*. 2023 International Conference on Smart Applications, Communications and Networking, SmartNets 2023. Scopus. https://doi.org/10.1109/SmartNets58706.2023.10215825

Johannesson, P., & Perjons, E. (2014). *An Introduction to Design Science*. Springer International Publishing. https://doi.org/10.1007/978-3-319-10632-8

Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, *110*(9), 2819–2858. https://doi.org/10.1257/aer.20191330

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022, May 6). *Synthetic Data—What, why and how?* arXiv.Org. https://arxiv.org/abs/2205.03257v1

Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Prasser, F., & Raisaro, J. L. (2023). *Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics* [Preprint]. Health Informatics. https://doi.org/10.1101/2023.11.28.23299124

Kable, A. K., Pich, J., & Maslin-Prothero, S. E. (2012). A structured approach to documenting a search strategy for publication: A 12 step guideline for authors. *Nurse Education Today*, *32*(8), 878–886. https://doi.org/10.1016/j.nedt.2012.02.022

Kamel Boulos, M. N., Kwan, M.-P., El Emam, K., Chung, A. L.-L., Gao, S., & Richardson, D. B. (2022). Reconciling public health common good and individual privacy: New methods and issues in geoprivacy. *International Journal of Health Geographics*, *21*(1), 1, s12942-022-00300–00309. https://doi.org/10.1186/s12942-022-00300-9

Kamerbrief Visie En Strategie Secundair Datagebruik, 27529, Ministry of Health, Welfare and Sports, 3561469 (2023). https://open.overheid.nl/documenten/ronl-3f08b9fdcb894267976f5b7da1c90d450c7f5e60/pdf

Kent, S., Burn, E., Dawoud, D., Jonsson, P., Østby, J. T., Hughes, N., Rijnbeek, P., & Bouvy, J. C. (2021). Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. *Pharmacoeconomics*, *39*(3), 275–285. https://doi.org/10.1007/s40273-020-00981-9

Kim, Y., & Stanton, J. M. (2013). Institutional and individual influences on scientists' data sharing behaviors: A multilevel analysis. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–14. https://doi.org/10.1002/meet.14505001093

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *2nd International Conference on Learning Representations (ICLR2014)*. http://arxiv.org/abs/1312.6114

Klijn, E.-H., & Koppenjan, J. F. M. (2006). Institutional design: Changing institutional features of networks. *Public Management Review*, *8*(1), 141–160. https://doi.org/10.1080/14719030500518915

Knott, E., Rao, A. H., Summers, K., & Teeger, C. (2022). Interviews in the social sciences. *Nature Reviews Methods Primers*, *2*(1), 1–15. https://doi.org/10.1038/s43586-022-00150-6

Kokosi, T., & Harron, K. (2022). Synthetic data in medical research. *BMJ Medicine*, *1*(1), e000167. https://doi.org/10.1136/bmjmed-2022-000167

Kroes, Q. R. (2023). Noot bij Gemeenschappelijke Afwikkelingsraad/Europese Toezichthouder voor gegevensbescherming. *Mediaforum*, *2023*(3), 104–111.

Lee, C., Luo, Z., Ngiam, K. Y., Zhang, M., Zheng, K., Chen, G., Ooi, B. C., & Yip, W. L. J. (2017). Big Healthcare Data Analytics: Challenges and Applications. In S. U. Khan, A. Y. Zomaya, & A. Abbas (Eds.), *Handbook of Large-Scale Distributed Computing in Smart Healthcare* (pp. 11–41). Springer International Publishing. https://doi.org/10.1007/978-3-319-58280-1_2

Li, F., Zou, X., Liu, P., & Chen, J. Y. (2011). New threats to health data privacy. *BMC Bioinformatics*, *12*(S12), S7. https://doi.org/10.1186/1471-2105-12-S12-S7

Li, J., Cairns, B. J., Li, J., & Zhu, T. (2023). Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *Npj Digital Medicine*, *6*(1), 98. https://doi.org/10.1038/s41746-023-00834-7

Lima, D. M., Rodrigues-Jr, J. F., Traina, A. J. M., Pires, F. A., & Gutierrez, M. A. (2019). Transforming Two Decades of ePR Data to OMOP CDM for Clinical Research. *Studies in Health Technology and Informatics*, *264*, 233–237. https://doi.org/10.3233/SHTI190218

Lynskey, O. (2016). *The Foundations of EU Data Protection Law*. Oxford University Press. http://ebookcentral.proquest.com/lib/uvtilburg-ebooks/detail.action?docID=4310752

Marks, M. (2019). *Artificial Intelligence Based Suicide Prediction* (SSRN Scholarly Paper 3324874). https://papers.ssrn.com/abstract=3324874

Mauboussin, A., & Mauboussin, M. J. (2018, July 3). If You Say Something Is "Likely," How Likely Do People Think It Is? *Harvard Business Review*. https://hbr.org/2018/07/if-you-say-something-is-likely-how-likely-do-people-think-it-is

Ministry of Internal Affairs. (2018, May 29). *Wat staat er in mijn medisch dossier? - Rechten van patiënt en privacy*. Rijksoverheid; Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/onderwerpen/rechten-van-patient-en-privacy/uw-medisch-dossier/inhoud-medisch-dossier

Molnar, C. (2023). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Molnár-Gábor, F., Sellner, J., Pagil, S., Slokenberga, S., Tzortzatou-Nanopoulou, O., & Nyström, K. (2022). Harmonization after the GDPR? Divergences in the rules for genetic and health data sharing in four member states and ways to overcome them by EU measures: Insights from Germany, Greece, Latvia and Sweden. *Seminars in Cancer Biology*, *84*, 271–283. https://doi.org/10.1016/j.semcancer.2021.12.001

Mosquera, L., El Emam, K., Ding, L., Sharma, V., Zhang, X. H., Kababji, S. E., Carvalho, C., Hamilton, B., Palfrey, D., Kong, L., Jiang, B., & Eurich, D. T. (2023). A method for generating synthetic longitudinal health data. *BMC Medical Research Methodology*, *23*(1), 67. https://doi.org/10.1186/s12874-023-01869-w

Mostert, M., Bredenoord, A. L., Biesaart, M. C. I. H., & Van Delden, J. J. M. (2016). Big Data in medical research and EU data protection law: Challenges to the consent or anonymise approach. *European Journal of Human Genetics*, *24*(7), 956–960. https://doi.org/10.1038/ejhg.2015.239

Mostert, M., Bredenoord, A. L., van der Slootb, B., & van Delden, J. J. M. (2018). From Privacy to Data Protection in the eu: Implications for Big Data Health Research. *European Journal of Health Law*, *25*(1), 43–55. https://doi.org/10.1163/15718093-12460346

Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic data generation: State of the art in health care domain. *Computer Science Review*, *48*, 100546. https://doi.org/10.1016/j.cosrev.2023.100546

myDRE platform. (n.d.). *anDREa-cloud*. Retrieved 19 January 2024, from https://andrea-cloud.com/mydre-platform/

NFU. (2020). *Handreiking Hergebruik van zorggegevens voor wetenschappelijk onderzoek* (20.32161). https://elsi.health-ri.nl/sites/elsi/files/2023-01/update%20NFU%20Handreiking.pdf

Nik, A. H. Z., Riegler, M. A., Halvorsen, P., & Storås, A. M. (2023). Generation of Synthetic Tabular Healthcare Data Using Generative Adversarial Networks. In D.-T. Dang-Nguyen, C. Gurrin, M. Larson, A. F. Smeaton, S. Rudinac, M.-S. Dao, C. Trattner, & P. Chen (Eds.), *MultiMedia Modeling* (pp. 434–446). Springer International Publishing. https://doi.org/10.1007/978-3-031-27077-2_34

Nuffield Council on Bioethics. (2015). *The collection, linking and use of data in biomedical research and health care: Ethical issues*. https://www.nuffieldbioethics.org/wp-content/uploads/Biodata-a-guide-to-the-report-PDF.pdf

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511807763

Ostrom, E. (2005). *Understanding Institutional Diversity*. Princeton University Press. http://ebookcentral.proquest.com/lib/delft/detail.action?docID=483578

Ostrom, E. (2011). Background on the Institutional Analysis and Development Framework. *Policy Studies Journal*, *39*(1), 7–27. https://doi.org/10.1111/j.1541-0072.2010.00394.x

Pavlenko, E., Strech, D., & Langhof, H. (2020). Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC Medical Informatics and Decision Making*, *20*(1), 157. https://doi.org/10.1186/s12911-020-01177-z

Pawar, A., Ahirrao, S., & Churi, P. P. (2018). Anonymization Techniques for Protecting Privacy: A Survey. *2018 IEEE Punecon*, 1–6. https://doi.org/10.1109/PUNECON.2018.8745425

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Polski, M. M., & Ostrom, E. (2017). An Institutional Framework for Policy Analysis and Design. *Elinor Ostrom and the Bloomington School of Political Economy*, *3*. https://ostromworkshop.indiana.edu/pdf/teaching/iad-for-policy-applications.pdf

Powell, W. W. (1991). *Expanding the scope of institutional analysis*. University of Chicago Press.

Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space, Pub. L. No. COM(2022) 197 final (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197

Purtova, N. (2017). Health Data for Common Good: Defining the Boundaries and Social Dilemmas of Data Commons. In S. Adams, N. Purtova, & R. Leenes (Eds.), *Under Observation: The Interplay Between eHealth and Surveillance* (Vol. 35, pp. 177–210). Springer International Publishing. https://doi.org/10.1007/978-3-319-48342-9_10

Purtova, N. (2018). The law of everything. Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology*, *10*(1), 40–81. https://doi.org/10.1080/17579961.2018.1452176

Purtova, N., & Van Maanen, G. (2023). Data as an economic good, data as a commons, and data governance. *Law, Innovation and Technology*, 1–42. https://doi.org/10.1080/17579961.2023.2265270

Qu, S. Q., & Dumay, J. (2011). The qualitative research interview. *Qualitative Research in Accounting & Management*, *8*(3), 238–264. https://doi.org/10.1108/11766091111162070

Radboud UMC. (2019, December 19). *UMC consortium brings Digital Research Environment to next phase*. https://www.radboudumc.nl/en/news/2019/umc-consortium-brings-digital-research-environment-to-next-phase

Rajendran, S., Obeid, J. S., Binol, H., D`Agostino, R., Foley, K., Zhang, W., Austin, P., Brakefield, J., Gurcan, M. N., & Topaloglu, U. (2021). Cloud-Based Federated Learning Implementation Across Medical Centers. *JCO Clinical Cancer Informatics*, *5*, 1–11. https://doi.org/10.1200/CCI.20.00060

Rajotte, J.-F., Bergen, R., Buckeridge, D. L., Emam, K. E., Ng, R., & Strome, E. (2022). Synthetic data as an enabler for machine learning applications in medicine. *iScience*, *25*(11). https://doi.org/10.1016/j.isci.2022.105331

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, Official Journal of the European Union L 119/1 (2016).

Reimers, K., & Luo, Y. (2023). On the Economic Nature of Medical Information: Implications for the Development of Information Infrastructures in the Healthcare Sector. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 2809–2817. : https://hdl.handle.net/10125/102977

Ruhaak, A. (2020, May 28). Data Commons & Data Trust. *Medium*. https://medium.com/@anoukruhaak/data-commons-data-trust-63ac64c1c0c2

Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., & Detmer, D. E. (2007). Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*, *14*(1), 1–9. https://doi.org/10.1197/jamia.M2273

Scharpf, F. (1997). *Games Real Actors Play: Actor-Centered. Institutionalism in Policy Research*. Westview Press.

Scheibner, J., Ienca, M., Kechagia, S., Troncoso-Pastoriza, J. R., Raisaro, J. L., Hubaux, J.-P., Fellay, J., & Vayena, E. (2020). Data protection and ethics requirements for multisite research with health data: A comparative examination of legislative governance frameworks and the role of data protection technologies†. *Journal of Law and the Biosciences*, *7*(1), lsaa010. https://doi.org/10.1093/jlb/lsaa010

Schneiberg, M., & Clemens, E. S. (2006). The Typical Tools for the Job: Research Strategies in Institutional Analysis. *Sociological Theory*, *24*(3), 195–227.

Scholte, R., Kranendonk, E., Paardekooper, M., & Ploem, C. (2019). Hergebruik van patiëntgegevens voor wetenschappelijk onderzoek: Op weg naar eenduidige spelregels. *Tijdschrift voor gezondheidswetenschappen*, *97*(3–4), 55–58. https://doi.org/10.1007/s12508-019-0213-y

Shi, J., Wang, D., Tesei, G., & Norgeot, B. (2022). Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation with multiple treatments. *Frontiers in Artificial Intelligence*, *5*, 918813. https://doi.org/10.3389/frai.2022.918813

Single Resolution Board v European Data Protection Supervisor, Reports of Cases 1 (General Court of the European Union 2023). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62020TJ0557

Slokenberga, S. (2022). Scientific research regime 2.0? Transformations of the research regime and the protection of the data subject that the proposed EHDS regulation promises to bring along. *Technology and Regulation*, *2022*, 135–147. https://doi.org/10.26116/techreg.2022.014

Smit, J.-A. R., Mostert, M., van der Graaf, R., Grobbee, D. E., & van Delden, J. J. M. (2024). Specific measures for data-intensive health research without consent: A systematic review of soft law instruments and academic literature. *European Journal of Human Genetics*, *32*(1), Article 1. https://doi.org/10.1038/s41431-023-01471-0

Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic Data – Anonymisation Groundhog Day. *Proceedings of the 31st USENIX Security Symposium*, 1451–1468. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler

Su, B., Wang, Y., Schiavazzi, D., & Liu, F. (2023). Privacy-Preserving Data Synthesis via Differentially Private Normalizing Flows with Application to Electronic Health Records Data. *Proceedings of the AAAI Symposium Series*, *1*(1), 161–167. https://doi.org/10.1609/aaaiss.v1i1.27495

Sun, C., Van Soest, J., & Dumontier, M. (2023). Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of Biomedical Informatics*, *143*, 104404. https://doi.org/10.1016/j.jbi.2023.104404

Sun, S., Wang, F., Rashidian, S., Kurc, T., Abell-Hart, K., Hajagos, J., Zhu, W., Saltz, M., & Saltz, J. (2021). Generating Longitudinal Synthetic EHR Data with Recurrent Autoencoders and Generative Adversarial Networks. In E. K. Rezig, V. Gadepally, T. Mattson, M. Stonebraker, T. Kraska, F. Wang, G. Luo, J. Kong, & A. Dubovitskaya (Eds.), *Heterogeneous Data Management, Polystores, and Analytics for Healthcare* (Vol. 12921, pp. 153–165). Springer International Publishing. https://doi.org/10.1007/978-3-030-93663-1_12

Synthetic Data. (2023, November 13). IEEE Standards Association. https://standards.ieee.org/industry-connections/synthetic-data/

Theodorou, B., Xiao, C., & Sun, J. (2023). Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature Communications*, *14*(1), 5305. https://doi.org/10.1038/s41467-023-41093-0

Thomas, J. A., Foraker, R. E., Zamstein, N., Morrow, J. D., Payne, P. R. O., Wilcox, A. B., the N3C Consortium, Haendel, M. A., Chute, C. G., Gersing, K. R., Walden, A., Haendel, M. A., Bennett, T. D., Chute, C. G., Eichmann, D. A., Guinney, J., Kibbe, W. A., Liu, H., Payne, P. R. O., … Mendelevitch, O. (2022). Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: Results from analyzing &gt;1.8 million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C). *Journal of the American Medical Informatics Association*, *29*(8), 1350–1365. https://doi.org/10.1093/jamia/ocac045

Torfi, A., & Fox, E. A. (2020). *CorGAN: Correlation-Capturing Convolutional Generative Adversarial Networks for Generating Synthetic Healthcare Records* (arXiv:2001.09346). arXiv. http://arxiv.org/abs/2001.09346

Tsao, S.-F., Sharma, K., Noor, H., Forster, A., & Chen, H. (2023). Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review. *Studies in Health Technology and Informatics*, *302*, 53–57. Scopus. https://doi.org/10.3233/SHTI230063

Vallevik, V. B., Babic, A., Marshall, S. E., Elvatun, S., Brøgger, H. M. B., Alagaratnam, S., Edwin, B., Veeraragavan, N. R., Befring, A. K., & Jan F., N. (2024). *Can I trust my fake data- A comprehensive quality assessment framework for synthetic tabular data in healthcare*. https://doi.org/10.48550/arXiv.2401.13716 van Bon-Martens, M., & van Veen, E.-B. (2019). *Handreiking ontsluiten patiëntgegevens voor onderzoek: Werken volgens de regels uit AVG, UAVG en WGBO*. ZonMw. https://www.rivm.nl/sites/default/files/2019-09/Handreiking%20ontsluiten%20patientgegevens.pdf

van der Sloot, B., & van Schendel, S. (Eds.). (2024). *The Boundaries of Data*. Amsterdam University Press. https://library.oapen.org/bitstream/handle/20.500.12657/87784/9789048557998.pdf

Van Gend, T., & Zuiderwijk, A. (2023). Open research data: A case study into institutional and infrastructural arrangements to stimulate open research data sharing and reuse. *Journal of Librarianship and Information Science*, *55*(3), 782–797. https://doi.org/10.1177/09610006221101200

Veen, E. B. van, & Verheij, R. A. (2023). *Further use of data and tissue for a learning health system: The rules and procedures in The Netherlands compared to Denmark, England, Finland, France, and Germany*. MLCF ; NIVEL.

Venugopal, R., Shafqat, N., Venugopal, I., Tillbury, B. M. J., Stafford, H. D., & Bourazeri, A. (2022). Privacy preserving Generative Adversarial Networks to model Electronic Health Records. *Neural Networks*, *153*, 339–348. https://doi.org/10.1016/j.neunet.2022.06.022

Verhoeven, E., Kroneman, M., Wilson, P., Kirwan, M., Verheij, R., Van Veen, E.-B., & Hansen, J. (2021). *Assessment of the EU Member States' rules on health data in the light of GDPR: Country fiches for all EU MS.* Publications Office. https://data.europa.eu/doi/10.2818/09448

Vukovic, J., Ivankovic, D., Habl, C., & Dimnjakovic, J. (2022). Enablers and barriers to the secondary use of health data in Europe: General data protection regulation perspective. *Archives of Public Health*, *80*(1), 115. https://doi.org/10.1186/s13690-022-00866-7

Wang, Z., Myles, P., & Tucker, A. (2021). Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Computational Intelligence*, *37*(2), 819–851. https://doi.org/10.1111/coin.12427

Wee, B. V., & Banister, D. (2016). How to Write a Literature Review Paper? *Transport Reviews*, *36*(2), 278–288. https://doi.org/10.1080/01441647.2015.1065456

Wet van 16 Mei 2018, Houdende Regels Ter Uitvoering van Verordening (EU) 2016/679 van Het Europees Parlement En de Raad van 27 April 2016 Betreffende de Bescherming van Natuurlijke Personen in Verband Met de Verwerking van Persoonsgegevens En Betreffende Het Vrije Verkeer van Die Gegevens En Tot Intrekking van Richtlijn 95/46/EG (Algemene Verordening Gegevensbescherming) (PbEU 2016, L 119) (Uitvoeringswet Algemene Verordening Gegevensbescherming), Staatsblad 2018, 144 (2018).

Wicks, A. M., & St. Clair, L. (2007). Competing Values in Healthcare: Balancing the (Un) Balanced Scorecard. *Journal of Healthcare Management*, *52*(5), 309.

Wiedekopf, J., Ulrich, H., Essenwanger, A., Kiel, A., Kock-Schoppenhauer, A.-K., & Ingenerf, J. (2021). Desiderata for a Synthetic Clinical Data Generator. In J. Mantas, L. Stoicu-Tivadar, C. Chronaki, A. Hasman, P. Weber, P. Gallos, M. Crişan-Vida, E. Zoulias, & O. S. Chirila (Eds.), *Studies in Health Technology and Informatics*. IOS Press. https://doi.org/10.3233/SHTI210122

Wilkinson, K., Green, C., Nowicki, D., & Von Schindler, C. (2020). Less than five is less than ideal: Replacing the "less than 5 cell size" rule with a risk-based data disclosure protocol in a public health setting. *Canadian Journal of Public Health = Revue Canadienne de Santé Publique*, *111*(5), 761–765. https://doi.org/10.17269/s41997-020-00303-8

Writers collective Nictiz, VWS, VZVZ, ZN. (2023). *Nationale visie en strategie op het gezondheidsinformatiestelsel*. Ministry of Health, Welfare and Sport. https://open.overheid.nl/documenten/ronl-36667024db962a4962d0815e7cf2d3c9596d7255/pdf

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2019a). Assessing privacy and quality of synthetic health data. *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, 1–4. https://doi.org/10.1145/3359115.3359124

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2019b). *Privacy preserving synthetic health data*. 465–470. Scopus.

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, *416*, 244–255. https://doi.org/10.1016/j.neucom.2019.12.136

Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., Mooney, S. D., & Malin, B. A. (2022). A Multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, *13*(1), 7609. https://doi.org/10.1038/s41467-022-35295-1

Yin, R. K. (1984). *Case study research: Design and methods*. Sage Publications.

Yin, R. K. (2011). Chapter 1. A (very) Brief Refresher on the Case Study Method. In *Applications of Case Study Research* (pp. 3–20). Sage Publications. https://study.sagepub.com/sites/default/files/a_very_brief_refresher_on_the_case_study_method.pdf

Yoon, J., Drumright, L. N., & Van Der Schaar, M. (2020). Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, *24*(8), 2378–2388. https://doi.org/10.1109/JBHI.2020.2980262

Yoon, J., Mizrahi, M., Ghalaty, N. F., Jarvinen, T., Ravi, A. S., Brune, P., Kong, F., Anderson, D., Lee, G., Meir, A., Bandukwala, F., Kanal, E., Arık, S. Ö., & Pfister, T. (2023). EHR-Safe: Generating high-fidelity and privacy-preserving synthetic electronic health records. *Npj Digital Medicine*, *6*(1), 141. https://doi.org/10.1038/s41746-023-00888-7

Zhang, P., & Kamel Boulos, M. N. (2022). Privacy-by-Design Environments for Large-Scale Health Research and Federated Learning from Data. *International Journal of Environmental Research and Public Health*, *19*(19), 11876. https://doi.org/10.3390/ijerph191911876

Zhang, Z., Yan, C., Lasko, T. A., Sun, J., & Malin, B. A. (2021). SynTEG: A framework for temporal structured electronic health data simulation. *Journal of the American Medical Informatics Association*, *28*(3), 596–604. https://doi.org/10.1093/jamia/ocaa262

Zhang, Z., Yan, C., & Malin, B. A. (2022). Membership inference attacks against synthetic health data. *Journal of Biomedical Informatics*, *125*, 103977. https://doi.org/10.1016/j.jbi.2021.103977

Zhang, Z., Yan, C., Mesa, D. A., Sun, J., & Malin, B. A. (2020). Ensuring electronic medical record simulation through better training, modeling, and evaluation. *Journal of the American Medical Informatics Association*, *27*(1), 99–108. https://doi.org/10.1093/jamia/ocz161

Zhou, N., Wu, Q., Wu, Z., Marino, S., & Dinov, I. D. (2022). DataSifterText: Partially Synthetic Text Generation for Sensitive Clinical Notes. *Journal of Medical Systems*, *46*(12), 96. https://doi.org/10.1007/s10916-022-01880-6

Zorginstituut Nederland. (2023). *Wegiz: Elektronische gegevensuitwisseling* [Webcontent]. Zorginstituut Nederland. https://www.zorginzicht.nl/ondersteuning/wet-elektronische-gegevensuitwisseling-in-de-zorg-wegiz

Zygmuntowski, J. J., Zoboli, L., & Nemitz, P. F. (2021). Embedding European values in data governance: A case for public data commons. *Internet Policy Review*, *10*(3). https://policyreview.info/articles/analysis/embedding-european-values-data-governance-case-public-data-commons

A
Framework for interview questions

Appendix A discusses the interview questions (§A.1) and analysis (§A.2) for mapping the institutional context of current health data sharing practice and how they examine the concepts of the IAD framework.

A.1 Interview questions

Please note Table 6 merely provides the interview questions. Explanations with regard to context and concepts were provided, depending on the expertise of the interviewee. The questions were altered and prioritised to fit the expertise of the interviewee.

Table 6. Interview questions per IAD-framework component

#	Interview question	IAD concept	Explanation
1	**Introduction**		
1a	What does your job entail?	Action arena	Introduction and clarification of exact role to define the actor
1b	What kind of decisions on health data do you make?	Position rules	First indication of the position of interviewee in health data sharing
1c	What kind of health data are shared within your field of work?	Physical and material characteristics	Definition of the nature of service/good that is subject of discussion.
2	**Roles, positions and processes**		
2a	What objectives do you pursue in your role in health data sharing?	Community attributes	Examines members' beliefs and indication of strategy
2b	Could you briefly describe the process of sharing personal data for research purposes?	Physical and material characteristics	Explains how the good or services is provided.
2bi	What resources are currently available to enable data sharing?	Physical and material characteristics	Explains what resources are required to provide the good or service
2bii	Could you explain the role of other actors (within or without your organisation) involved in health data sharing and their pursued objective?	Position rules	Explains the role and position of other members
2c	What level of control do you have with regard to health data sharing in comparison to other actors?	Authority rules	Specifies the power of participants in the health data sharing process, and therefore their ability to change the process
2d	Have you seen cases where the interests or actions of actors misaligned and hampered the sharing of research data?	Community attributes	Examines members' beliefs about other participants
2di	If yes, could you please explain this?	Community attributes	Examines members' beliefs about other participants
2dii	If not, do you think this could happen?	Community attributes	Examines members' beliefs about other participants
2e	To what extent are you willing to change health data sharing practices to increase the secondary use of health data for research?	Community attributes	Examines members' beliefs and indication of strategy

2ei	Are other actors willing to change data sharing practices to increase the secondary use of health data for research? If yes, how could they be motivated to do so? If not, what barriers should be addressed.	Community attributes	Examines members' beliefs about other participants and preferred strategies
3	**Rules of health data sharing**	Rules-in-use	
3a	On the operational level, how do you determine which data may be shared under what conditions?	Position rules (on operational level)	Specifies what
3b	On organisational level, what rules related to health data sharing have been formulated by your organisation? For example, information policies and procedures.	Boundary rules/Authority rules (on constitutional level)	Determine what roles are involved in health data sharing process and what they may do
3c	On network level, what are the (in)formal rules for making decisions on health data sharing between agencies? For example, how are disputes resolved? Is there financial compensation if organisation A shares its data with organisation B?	Aggregation rules/Pay-off rules (on constitutional level)	Maps how decisions are made by various actors and how the costs and benefits are distributed.
3d	Have you encountered rules from other data sharing entities that were misaligned with the rules you are subject to, such that sharing data was limited?	Rules-in-use/Interaction patterns	Indicates whether there are alarming patterns of interactions between actors based on the rules-in-use
4	**Identification of challenges**		
4a	Could you explain whether you think the way data sharing is currently organised is achieving the desired goals? "Desired goals" can be interpreted from your role point of view.	Action arena/Outcomes	Inventory of what the outcomes of health data sharing are
4b	Based on your experience, what barriers to sharing health data for research remain?	Interaction patterns	Inventory of alarming patterns of interactions between actors
5	**Synthetic data generation**	Interaction patterns	This question relates synthetic data generation to health data sharing practice. These questions provide a first inventory of how synthetic data generation will interact with the institutional environment.
5a	To what extent are you familiar with synthetic data generation in the healthcare context?		
5b	If not, after an explanation of synthetic data generation:		
5bi	How can generating synthetic data change health data sharing practice?		
5bii	What do you need to interpret quantitative metrics to assess privacy risks?		
5c	If yes:		
5ci	From your role, what do you think are the main benefits of synthetic data for research?		
5cii	What drawbacks or challenges do you see?		
5ciii	What do you think other actors involved in health data sharing would think of the use of synthetic data generation?		
5civ	What would you need, to apply synthetic data generation?		

A.2 Interview analysis

Table 7 presents the final coding framework.

Table 7. Coding framework interview analysis

IAD concept	IAD specification	Code	Interviews
Rules-in-use	Position rules	GDPR as primary formal regulation	ALL
		WGBO and UAVG as formal primary regulation	DS2-H LC-H LC2-RI
		Organisations' data policies are important to interpret GDPR principles	DS1-H PO-H LC-H LC1-RI
		Non-binding regulations guide interpretation of GDPR principles	DS2-H LC2-RI
		Consent is the legal base for sharing health data for research in NL	ALL
	Authority rules	Tiered governance structure of (formal) decision making competence: heads of departments are authorised to sign agreements	ALL
	Boundary rules	Broad definition of personal data	DS1-H PO-H LC2-RI
	Aggregation rules		
	Pay-off rules	Quid pro quo	DS1-H DS2-H
		Monetary compensation not prevalent	DS1-H DS2-H PO-H
		Compensation via publications	DS1-H DS2-H PO-H R-RI
		Compensation via research results	DS1-H DS2-H PO-H
Physical world	Economic nature	Pseudonymised data is shared for research	ALL
	Resources	Description of technical infrastructure	DS1-H DS2-H PO-H LC1-RI LC2-RI
		The need for technical interoperability standards	DS1-H
		Sharing health data requires the conclusion of data agreements	ALL
		Sharing health data requires a DMP	DS1-H DS2-H PO-H LC-H R-RI
		Sharing health data requires a DPIA	DS2-H LC-H PO-H LC1-RI LC2-RI
Community	Members' own beliefs	Data sharing should be encouraged to spur health research	ALL
		Specifically addresses the interest of data protection in health data sharing, besides research interest.	DS1-H PO-H LC1-RI

IAD concept	**IAD specification**	**Code**	**Interviews**
			LC2-RI
		Willing to change data sharing process	ALL
		Generally, interests of actors are aligned when it comes to sharing health data	DS1-H DS2-H PO-H LC-H LC1-RI LC2-RI R-RI
	Members' beliefs about other participants	Interviewee beliefs that other actors also belief data sharing should be encouraged for research	ALL
Action Arena	Actors	Relevant actor: data steward	ALL
		Relevant actor: principle investigator	DS1-H DS2-H LC-H PO-H
		Relevant actor: Privacy officer	DS1-H LC-H LC2-RI
		Relevant actor: Legal counsellor	ALL
		Relevant actor: Board of directors (or similar)	DS1-H DS2-H LC-H LC2-RI
		Relevant actor: Head of department	DS1-H DS2-H LC2-RI LC-H
		Relevant actor: Data Protection Officer	LC2-RI
		Relevant actor: Security team	DS2-H LC-H LC2-RI
		Relevant actor: Researcher	ALL
	Action situation	Data sharing is an ad hoc process within professional networks	DS1-H DS2-H R-RI
		Description of concluding health data sharing agreements	DS1-H DS2-H LC2-H PO-H LC1-RI
		Process of obtaining consent to share health data for research	ALL
		Process of anonymisation and pseudonymisation of health data	DS1-H PO-H LC1-RI LC2-RI
Outcomes		Poor data protection in data sharing agreements (e.g. due to lack of auditing)	DS2-H LC1-H LC2-H
		Good data protection via data sharing agreements	LC-H
		Poor data findability due to lack of technical infrastructure	DS1-H DS2-H
		Poor data availability due to a lack of technical standards and infrastructures	DS1-H PO-H
		Lengthy data sharing process	ALL

IAD concept	IAD specification	Code	Interviews
		Anonymisation is difficult to achieve with current legal definitions	DS1-H LC2-H PO-H PM-HWS
		Incorrect classifications of anonymous data	LC1-RI LC2-RI R-RI
		Consent is strictly interpreted in NL	DS1-H DS2-H PO-H LC-H LC2-RI PM-HWS
Evaluative criteria		Data findability	DS1-H DS2-H PM-HWS
		Data availability	DS1-H DS2-H LC-H PO-H RI-H PM-HWS
		Data protection	DS1-H PO-H LC1-RI LC2-H PM-HWS
Synthetic data generation		Synthetic data generation could ease health data sharing process	DS1-H DS2-H PO-H LC-H LC2-RI PM-HWS
		Hesitant towards value of synthetic health data sharing	LC-H
		Unresolved technical challenges of synthetic data generation	DS1-H
		Challenges related to interpretation of privacy metrics	LC-H LC1-RI LC2-RI

B
Definitions of common terms

Table 8 presents an overview of commonly used terms, primarily in referred to in Chapter 3. The terms are extracted from the GDPR and tweaked to the scope of this thesis.

Table 8. Definition of commonly used terms and concepts

Term	Definition
Article 29 Working Party	Predecessor of the EDPB (see EDPB)
Data controller	The organisation that determines the purposes and means of processing personal data. In the context of secondary use of health data for research, this could be a healthcare provider, research institute, or other organisation responsible for managing and overseeing the use of the data.
Data processor	A party that processes personal data on behalf of the data controller. This could be a third-party service provider or entity contracted by the data controller to handle data processing tasks, such as data analysis or storage.
Data recipient	The organisation that receives personal data from the data controller or data processor for specific purposes. In the context of health data research, this could be another research institute, a government agency, or a commercial entity involved in collaborative research projects.
Data subject	An individual who is the subject of the personal data being processed. In the context of health data research, data subjects are typically patients or individuals whose health information is being used for research purposes.
EPDB	An independent EU body that is concerned with monitoring of and providing guidance of EU data protection rules.
Personal data	Any information relating to an identified or identifiable natural person. This includes but is not limited to name, identification number, location data, online identifier, or factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that person.
Personal health data	Specific subset of personal data that relates to the physical or mental health of an individual. This may include medical history, treatment information, genetic data, or information about lifestyle habits that impact health.
Anonymisation	The process of irreversibly removing or modifying personal identifiers from data sets in such a way that the individuals to whom the data refers cannot be re-identified. Anonymised data is considered non-personal and can be used for research and other purposes with limited privacy risks.
Pseudonymisation	The process of replacing direct identifiers with artificial identifiers or pseudonyms, so that the linkage between the data and the individual is still possible but requires additional information that is kept separately. Pseudonymised data allows for some level of privacy protection while still enabling certain types of analysis and research.
Data sharing	The act of making data available to others, either within an organisation or to external parties, for specific purposes. In the context of health data research, data sharing involves sharing data with other organisations to facilitate collaborative research projects or to enable secondary analysis.
Data access	Data stays within one organisation. Data access may be granted to individuals or organizations based on their roles, responsibilities, and permissions within a data management system.
Data protection Impact Assessment	A process designed to systematically analyse and assess the potential risks and impacts of data processing activities on individuals' privacy and data protection rights. DPIAs are typically conducted prior to initiating new data processing activities, especially those involving sensitive or high-risk data, such as health data.
European Health Data Space	An initiative of the European Union aimed at creating a secure and interoperable data infrastructure for sharing and accessing health data across the EU. The EHDS seeks to facilitate research, innovation, and healthcare delivery by enabling seamless and

	secure exchange of health data while ensuring compliance with data protection regulations such as the GDPR.

,

C
Interviews technology providers

This appendix describes the results of the exploratory interviews with two technology providers of synthetic data generation.

C.1 Themes for analysis

Informed by the institutional analysis and literature review on synthetic data, the following themes were formulated for analysing the interviews with the technology providers:

- Relevance of health data sharing from technology provider's perspective (C2.1)
- Identification of relevant use cases (C2.2)
- Relevant generation models (C2.2)
- Benefit of synthetic data in comparison to current health data sharing process (C2.2)
- Challenges of synthetic data generation in healthcare (C2.3)
- Position of technology providers in health data sharing process (C2.4)
- Initiation of synthetic data sharing (C2.4)
- View regarding re-identification risks of synthetic data (C2.3)
- Privacy evaluation methods and need for technical expertise (C2.3 & C2.4)

As explained in Table 1, the technology providers are referred to as TP1 and TP2.

C.2 Interview analysis

C.2.1 The importance of sharing health data

TP1: Research can lead to improvements in healthcare and ultimately improved quality of life for people. For example, a use case concerns the quality register. The quality registry keeps track of all orthopaedic interventions, so prosthetics knee replacements and shoulders. They have more than 1 million registrations of patients with practitioners with prosthetics in hospitals with all kinds of characteristics about those patients. So, based on that data, they can do research on which prosthetics and which treatments are most effective and discover relationships, for example.

TP2: In developing AI models, a substantial amount of data is required for tasks such as model calibration, variable selection, and model assessment. The utilisation of data can enhance the predictive power of AI models by unlocking new information that may currently be unavailable. Additionally, it can expedite processes that currently consume significant time. This was the starting point for this synthetic data technology provider.

Conclusion: naturally, both TP recognise the importance of health data sharing. TP1 focuses more on patients interests, whereas TP2 focuses primarily on the potential for AI development.

C.2.2 Benefit of synthetic data in health data sharing

TP1: Patient data contains a lot of value. Because of privacy risks and privacy laws and regulations, this data may not be used just like that. For example, there are issues around storing data, such as where it may (not) be stored and for how long. If you are only allowed to store personal data for two years, you lose long-term trends, for example. These challenges are addressed by generating synthetic data. The generative model generates a completely new dataset that looks and behaves the same as the real data. So it contains all the statistical properties, trends, patterns and correlations. To generate data, GAN and diffusion models are used, as well as LLMs. They are also working on LLMs. Where we started with the GANS, we are now seeing it diminish and now diffusion models are getting bigger and potentially other models like LLMs. Synthetic data is no longer traceable to a specific individual in the original data, because the one-to-one relationship between the data is broken. In comparison to pseudo anonymisation, where only some data is removed, individuals are often still traceable, or data is broken to such an extent that no longer provides value for analyses. For example, synthetic data can be used to test applications for research and analysis for training machine learning. Analysing and researching data is the biggest use case in the healthcare sector for this provider. For use cases where patients still need to be identifiable, synthetic data cannot be used. Federated learning can also be used to generate synthetic data, for example to build a national model. Sharing data with researchers is definitely a relevant use case. With regard to its effect on data sharing for research, agreements around reimbursement and responsibilities will remain with the use of synthetic data. But, privacy is often one of the stumbling blocks, including issues regarding what data may or may not be shared, its justification and how data will be processed. Another stumbling block concerns what security measures must be implemented to share data with a third party. Synthetic data generation can be seen as a technical measure that can eases these privacy and security issues. Hence, synthetic data generation has the potential to speed up data exchange between organisations, something that is currently a bottleneck.

TP2: One of the predominant challenges in research lies in acquiring access to health data, a task entangled in considerable paperwork, coordination efforts, and diverse opinions. The associated bureaucratic processes incur both temporal and financial costs. The software provider addresses this by offering software designed for data generation. Subsequently clients determine its use. The software encompasses models developed in-house, ensuring compatibility with various datasets, including complex data types such as time series and location data. Their software is installed within the client's environment, ensuring data synthesis occurs at the source without the provider's involvement or access to sensitive health data. Although complete removal of data exchange agreements is unfeasible, offering a streamlined version of the contract is plausible, given the reduced stringency necessitated by synthetic data. This streamlined approach translates to heightened efficiency, cost-effectiveness, and the preservation of research momentum, particularly beneficial in scenarios such as research projects where the traditional 9-month duration for data acquisition can impede progress. The value of synthetic data lies in its resemblance to real-world data, allowing for exploration during research and model development. Comparative to alternative techniques such as federated learning or multiparty computation, synthetic data eliminates the need for additional model development and mitigates risks of potential bias arising from external data source – using central models that generate aggregated results complicates the identification of the source of deviations and biases. This can significantly impact research outcomes. Additionally, developing and running such models is often intensive.

C.2.3 Challenges of synthetic data generation in healthcare

TP 1: Given the early stages of the technology, technology providers are not (yet) experts in the application of synthetic data. They are experts in generating synthetic data, but not necessarily in the customer application domain. The application of synthetic data concerns a grey area: the technology is so new that few people in an application domain already know about the technology and consequently, how to implement it. Hence, there is still a lot to learn in the application of synthetic data generation in certain domains. There are three main challenges regarding application:

First, synthetic data generation is (perceived as) an abstract concept by user groups. As a result, they find it difficult to imagine what it is and what the output looks like. They have also questions regarding security, for example, how can synthetic data generation be privacy-friendly but at the same time contain the same statistical properties as real data. Hence, there is still a lack of understanding of the concept of synthetic data generation.

A second challenge lies in the understanding of usability of synthetic data. User groups questions its use, as they do not understand how fictitious data can still provide valuable information. For example, how can synthetic data generation provide the same results when training ML models. This lack of understanding can be solves by zooming in on application possibilities and provide explanations. Additionally, users can be convinced of utility and privacy by means of a report that shows utility and privacy metrics, for example in comparison to utility of real data.

Third, the standardisation of privacy and demonstrating that an application is privacy-friendly remains a challenge. This results in technical metrics that need to be translated into functional risks in a report. There are no thresholds to determine whether it is anonymous data or not. Suppose in a member inference attack, the probability that someone is traceable or was part of a dataset comes out to 0.8 or 12, what does that say about the data? There is an IEEE working group concerned with formulating standards, with a focus on privacy. It is not in dispute that the use of synthetic data generation preserves privacy better than other methods, such as pseudonymisation. However, whether it concerns personal data is still a point of discussion. Moreover, how such metrics are presented to users is also a challenge. For example, privacy can be translated into colour-coded privacy scores, but this is rather arbitrary. Also, the interpretation of these metrics also depends on its baseline. The user is responsible for interpreting these metrics and making a trade-off.

TP2: From a technical standpoint, synthetic data generation poses no challenges and is highly valuable. The primary challenge lies in the identification of suitable applications by users of synthetic data. Organisations often overlook simple yet impactful use cases in favour of exotic ones. Navigating this challenge involves an incremental commencement, starting with modest initiatives rather than embarking directly on ambitious objectives, thereby necessitating a gradual buildup. The allure of exotic use cases demanding heightened privacy considerations can complicate initial implementation. These use cases are supported by the platform, but organisation should not start with this – it would only let all alarm bells ring and require tick marks everywhere.

A very simple application is many universities, for example, give data analytics courses where students get to work with data. In some subjects, they are given personal data. This can easily be replaced with synthetic data. Another trivial use case is that within organisations, data scientists often have access to all data. Here, using synthetic data can be an important privacy safeguard.

Thus, incremental adoption, starting with simpler applications, is deemed more effective in the adoption of synthetic data generation in practice, than pursuing complex use cases from the start. To educate users about the application of synthetic data generation, they are trained to work with the platform and interpret results. Here, it is important that users themselves learn to use the platform; this does not belong to the responsibilities of the software provider.

C.2.4 Implementation of synthetic data in health data sharing

TP1: Implementing synthetic data generation is a shared responsibility. Often, the technology providers along with the innovators who start implementing – innovators are the stakeholder who are open to new technology with associated risks and understand that it can help them move forward to differentiate themselves in the market. The technology providers alone cannot do it. As for the challenge of setting standards, there is a need for a third party, like privacy experts or the Dutch DPA, to comment on that. Other parties, such as consultancies, play an advisory role on, for instance, data innovation, data privacy, technology policy, strategy, and answer questions such as how they deal with this, what technologies are there in the market? Some of the knowledge necessary to implement synthetic data generation is with those innovators with a pioneering role, some of it lies with consulting firms and some of it lies with the technology providers.

In terms of user groups, there are roles specifically concerned with making data better available and accessible. There was an expectation of the technology provider that certain roles, for example privacy officers, would look more actively for solutions in the market to make data sharing more secure. However, it turns out that most privacy officers are quite process-oriented rather than being strategic about data policy. With regard to their view regarding synthetic data generation, user groups really want to understand what is happening with their data, with some exceptions of course.

Usually, data holders are initiating synthetic data generation, as it also runs on their servers. As they want to share data compliantly, they are also the party who will do substantive checks on the data generated. Subsequently, they make data available to other researchers. Then it could be, for example, that data holders get a fee for the use of synthetic data, but that is between the data holder and user. It should be noted that this is presumably similar to compensation for pseudonymised data sharing.

TP2: In its current phase, the software provider focuses on building scalable software applications and does not offer consulting services. The initiative of the application of the generation software primarily rests with data teams, predominantly constituted by data scientists possessing requisite knowledge. Indeed, there are many people who do not know exactly what synthetic data is and how to apply it, but so there is no need. Although occasional external assistance may be warranted, the provider generally refrains from direct involvement, relying on collaborative efforts with partner consultants when necessary. The people who do not know what synthetic data is basically don not have to work with it either.

Clients who approach the software provider for synthetic data integration, typically healthcare providers, research institutes and other organisations, have already recognised the utility of synthetic data and determined to engage concretely in its application. The provider supplies the software once the organisation has determined how synthetic data fits into their data strategy.

When applying synthetic data, the provider presents users with a quality report displaying utility and privacy metrics. Users are responsible for interpreting these reports, although the provider includes training on privacy metrics during the training sessions. In terms of privacy, we measure how close synthetic data is to the original data. Other risks measured are outliers. These are industry standards from literature. Privacy requirements vary depending on the use case, and organisations must conduct risk assessments accordingly. It is not always necessary to guarantee 100% privacy. Sometimes, as in the universities' example, it is already quite an improvement in privacy compared to the situation before. Often, organisations have different measures for this, depending on the type of use case (e.g. a classification of green, orange and red), requiring risk assessments.

D
Economic nature of data

This appendix analyses the economic nature of data, showing that personal health data cannot be clearly classified as one of the good archetypes; it has characteristics of multiple types.

The economic nature of a service or good can be determined by the level of control regarding access hereto (excludability) and the extent to which one party's consumption limits availability to others (subtractability) (Polski & Ostrom, 2017). Health information is traditionally classified as non-subtractable, as its value does not decrease by its use (Jones & Tonetti, 2020; Reimers & Luo, 2023). Access to health data is argued to be highly excludable, generally limited to the institution in control over data collection (Hansen et al., 2021). Current data protection rules increase excludability, as data controllers are restricted to share personal data. Technology may also decrease the overall availability of an economic good (Purtova & Van Maanen, 2023), as the lack of standardised technical infrastructures for health data sharing reflects (Carballa-Smichowski et al., 2021; Hansen et al., 2021). Therefore, data itself can be classified as a toll (or club) good. The excludability of health data is a dynamic characteristic, affected by changing regulations and developing technologies (Carballa-Smichowski et al., 2021; Purtova & Van Maanen, 2023). For example, the EU wants to facilitate health data flows for research in a European Health Data Space (EHDS). The EHDS proposal is built on the idea that health data is of such importance for society, that it should be the norm, rather than an exception to the general prohibition of the GDPR (Proposal for the European Health Data Space, 2022). As a result, the economic nature of health data is increasingly leaning towards common pool resources in the form of data pools (Carballa-Smichowski et al., 2021).

An alternative way of looking at the economic nature of data, is to view data as part of a complex system: a common pool resource (Hess & Ostrom, 2006). Contrary to toll goods, common pool resources are subtractable, but its use is not easily excludable (Purtova & Van Maanen, 2023), such as phishing pounds. Complexity is another characteristic of common pool resources (Purtova & Van Maanen, 2023), as they comprise ecosystems with interrelated and interdependent elements (Purtova, 2017). Viewing data as part of such an ecosystem rather than in isolation from its context with the economic good approach, allows for an analysis of where data comes from and the societal impact of data processing (Purtova & Van Maanen, 2023). Data itself is not a common pool resource but can be seen as part of a common pool resource that is subtractable and hardly exclusive (Purtova & Van Maanen, 2023), such as scientific knowledge or privacy, in terms of control over appropriate data flows (Ruhaak, 2020). Privacy is subtractable in the sense that if someone shares information about another person, this other person's privacy (e.g. their control over the information) diminishes (Purtova, 2017; Purtova & Van Maanen, 2023; Ruhaak, 2020). Also, when persons share information about themselves, this could reveal information about other people with similar characteristics, affecting their privacy (Ruhaak, 2020). Privacy is difficult to exclude, as persons whose data are shared, cannot stop other entities from sharing information (Purtova & Van Maanen, 2023; Ruhaak, 2020). According to Purtova (2017), data ecosystems comprise three elements: 'people' provide 'data' via collection or analytical tools, called 'platforms'.

Concluding from this complex conceptual background, there are various views regarding the economic nature of data. Defining data as an economic good along the axes of subtractability and excludability suffices where the desired result would be data governance strategies that focus on quality and quantity of data. But this approach is inappropriate where data processing contributes to other (shared) objectives, such as preserving privacy, because it does not account for the complexity of data use, and its societal and technical context (Purtova & Van Maanen, 2023). This distinction is important, as different kinds of goods are accompanied with different kinds of governance challenges. For example, common pool resources are inherently related to collective action problems (Ostrom, 1990). Collective action problems come up when individuals act in a way that maximises their own short-term benefits, even though this disadvantages the joint outcomes. To continue with the example of fishing ponds, albeit fishermen individually benefit from catching as many fish as possible, this has disastrous effects on the sustainability of the fish stock (Ostrom, 1990). Solving collective action dilemmas requires a change in individual decisions through governance arrangements (Benfeldt et al., 2020).

Although data ecosystems are not traditional common pool resources as defined in Ostrom's earlier work (Ostrom, 1990), the application of Ostrom's framework for managing shared resources can offer insight in how to govern data (Coyle et al., 2020). In its essence, Ostrom's work is concerned with reaching agreement about access rules to a resource at the expense of a party's individual gain for the benefit of the community (Coyle et al., 2020). As researchers from healthcare providers (data collectors), may sacrifice some private gains or patient sacrifice privacy by sharing information (DS1-H; Coyle et al., 2020)), sharing information potentially unlocks research opportunities for other researchers that ultimately benefit society. Data ecosystems are increasingly seen as commons that should be governed collectively by various parties to increase the value of data in healthcare while protecting data protection rights (Zygmuntowski et al., 2021). By applying Ostrom's framework, this thesis views health data as an important resource to collectively achieve better research results, while preserving privacy. To apply Purtova's terminology of data ecosystems to health data: the fishing stock can be seen as patients in the digital healthcare environment, from who data can be extracted via all kinds of rods, for example the healthcare provider's system for electronic health records, by researchers from healthcare providers and third parties as fishermen of patient privacy (Purtova, 2017).

The analysis shows that the economic nature of personal data is hard to define. What we can learn, however, is that data should not be studied as a standalone good. Therefore, this study analyses data as part of a system with actors, technical infrastructures and legal means to extract value from it.

E
Legal guidance by COREON

This appendix further specifies the concept of scientific research and the broad consent as proposed by COREON. COREON is a professional network that consists of academic hospitals, universities and other research institutes. Its guidelines are followed by Dutch healthcare providers (DS2-H).

E.1 COREON's definition of scientific research

The GDPR mentions, but does not define what constitutes scientific research. The Dutch implementation act specifies this by stating that processing health data for research should serve the public interest (Uitvoeringswet Algemene Verordening Gegevensbescherming, 2018, art. 24). COREON further specifies what should be considered scientific research in a healthcare context. Following COREON's statement of scientific research in healthcare (COREON, 2018), scientific research should:

- Aim to gather generally applicable and new insights;
- Follow methodological standards that apply within that area of research and is reproducible;
- The research output should not directly lead to decisions regarding the data subjects;
- Follow the principles of scientific integrity;
- Always publish the outcomes;
- Follow the FAIR principles; and
- Justify why this research is able to contribute to better healthcare.

Hence, important for the definition of scientific research are the use of recognised protocols and methods to conduct the research, as well as that it should have a clear link with improving healthcare. This understanding provides additional safeguards for patients that their health data is processed in an honest manner and to their benefit.

E.2 COREON's proposal for broad consent

Another guidance document provided by COREON is the non-binding Code of Conduct, that participants of health data sharing usually adhere to – at least to some extent (DS2-H).

COREON acknowledges explicit consent as the starting point for secondary use of health data for research. To render this framework pragmatic for researchers, they delineate types of consent depending on the impact of research on patients. They propose a general type of consent, that generally informs data subjects about the research, for healthcare providers that frequently share health data for research purposes (COREON, 2022). As additional safeguards for patients, research must relate to the areas of the patient's illness, and the personal data must be pseudonymised. When data subjects do not object or respond to the general consent request, their data may be used for secondary purposes. Research with higher risks for data subjects, then, requires explicit consent as defined in the GDPR. Hence, COREON leans towards an opt-out consent model (COREON, 2022). COREON seems to argue that obtaining consent for secondary use of health data that originally has been collected in primary healthcare delivery for research always poses a disproportionate effort, due to the technical infrastructure necessary to inform participants, complexity of explaining the research and reuse, and a potential consent fatigue for patients.

It should be considered, however, that this code provided by the healthcare sector is not referred to in the Dutch government's vision in secondary use of health data (Veen & Verheij, 2023). Also, the Dutch DPA has expressed their concerns regarding the sector's interpretation of consent (Veen & Verheij, 2023). Presumably, neither the Dutch DPA nor the Ministry of HWS would approve of this argument, as the idea of general consent differs strongly from the GDPR, that the UAVG also specifically refers to. Therefore, the interpretation of COREON differs from formal Dutch law.

F
Overview of literature in synthetic data generation literature review

This appendix present an overview of the articles included in the literature review of Chapter 4. The article selection and analysis process is explained in §2.4.2. Synthetic data generation literature generally consists of two types of literature: articles that propose a generation method (Table 9) and articles that review multiple generation methods (Table 10). Table 10 explains what the articles review, where they fall short in the scope of this thesis, and their contribution to the literature review.

Table 9. Included articles that propose synthetic data generation methods

Author	**Title**	**Model of discussion**	**Use case**
(Biswal et al., 2021)	EVA: Generating Longitudinal Electronic Health Records Using Conditional Variational Autoencoders	Autoencoder	Generating timestamped events, such as diagnoses, medications or procedure (longitudinal electronic health records)
(Braddon et al., 2023)	Exploring the utility of synthetic data to extract more value from sensitive health data assets: A focused example in perinatal epidemiology	Regression trees/ linear modelling	Generating snapshots of electronic health records to predict perinatal quantitative variables, such as birthweight and gestational age.
(Choi et al., 2017)	Generating Multi-label Discrete Patient Records using Generative Adversarial Networks	GAN	Aggregated data generation for disease prediction
(Gwon et al., 2024)	LDP-GAN: Generative adversarial networks with local differential privacy for patient medical records synthesis	GAN	Generating EHRS with differential privacy
(J. Li et al., 2023)	Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications	EHR-M-GAN	Generating mixed-type timeseries data mix-typed (longitudinal EHRs)
(S. Sun et al., 2021)	Generating Longitudinal Synthetic EHR Data with Recurrent Autoencoders and Generative Adversarial Networks	LongGAN/ Recurrent Autoencoders	Generating timestamped data with continuous laboratory and medication values for given diseases (with codes) (longitudinal electronic health records)
(Shi et al., 2022)	Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation with multiple treatments	ADS-GAN/ neural networks	Generating snapshots of electronic health records, including mixed data types, such as lab results and medical history, to predict treatment outcomes
(Su et al., 2023)	Privacy-Preserving Data Synthesis via Differentially Private	Normalizing flows	Generating mixed-type snapshots of EHRs

	Normalizing Flows with Application to Electronic Health Records Data	(neural networks)	
(Theodorou et al., 2023)	Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model	SynTEG/GAN/GPT	Generating timestamped diagnostic events with codes, visits and records (longitudinal EHRs)
(Thomas et al., 2022)	Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: results from analyzing &gt;1.8 million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C)	Probabilistic density function	Generating mixed-type snapshots of EHRs, including geospatial and temporal epidemiologic data
(Torfi & Fox, 2020)	CorGAN: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records.	CorGAN	Generating timeseries data for recognizing epileptic activity and aggregated diagnoses.
(Venugopal et al., 2022)	Privacy preserving Generative Adversarial Networks to model Electronic Health Records	pGAN, tGAN, HealthGAN	Prediction of disease (diabetes) and insurance costs
(Yoon et al., 2020)	Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN)	ADS-GAN	Generating snapshots to predict 3-year mortality
(Yoon et al., 2023)	EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records	GAN/Autoencoder	Generating timestamped numerical and categorical features of EHRs, as well as statis numerical and categorical features
(Z. Zhang et al., 2022)	Membership inference attacks against synthetic health data	SynTEG/GAN	Generate timestamped diagnostic events (longitudinal EHRs)

Table 10. Included articles that review synthetic data generation methods

Author	**Title**	**Description**
(El Emam et al., 2020)	Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation	El Emam et al. propose ‘meaningful identity disclosure risk’ as a privacy metric. This is a risk model based on identity disclosure and the ability of an attacker to learn new information about the patient. Contrary to most articles, they provide argumentation for threshold values, which are deemed important to interpret the scores.
(Hernandez et al., 2022)	Synthetic data generation for tabular health records: A systematic review	From their literature review of GANs, Hernandez et al. concluded that there is a lack of evaluation and benchmark methods. The literature gives an overview of developments and used privacy metrics. However, the authors do not go further than a description of the privacy metrics.
(Jadon & Kumar, 2023)	Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy	Jadon and Kumar give an overview of the applications and challenges of GAN and VAE-based generation models. Their overview is a starting point for identifying current challenges in generation methods. However, the challenges require further analysis. Specifically regarding privacy, the authors acknowledge its importance, but fall short in providing convincing argumentation for their statement.
(Murtaza et al., 2023)	Synthetic data generation: State of the art in health care domain	A literature review that gives an overview of synthetic data applications in healthcare and their representativeness and privacy metrics. Also, the

Author	Title	Description
		authors provide some useful categorisation of generation models, which data is generated and metrics. However, privacy remains a descriptive topic, rather than an analysing it.
(Tsao et al., 2023)	Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review	Tsao et al. studied the status of research about governance of synthetic health data. They identified literature gaps regarding governance and evaluations of synthetic data applications and address the need to do so. Their conclusions correspond with this thesis' multi-disciplinary approach and can help to scope the application framework.
(Wiedekopf et al., 2021)	Desiderata for a Synthetic Clinical Data Generator	Wiedekopf et al. provide considerations for application of synthetic data generation in practice. Their research is primarily relevant for their categorisation of generation models and limitations.
(Yale et al., 2019a)	Assessing Privacy and Quality of Synthetic Health Data	Yale et al. develop quality and privacy metrics and evaluate these of multiple synthetic data generation methods for prediction of mortality
(Yale et al., 2020)	Generation and evaluation of privacy preserving synthetic health data	Yale et al. expand their quality and privacy metrics and evaluate these of multiple synthetic data generation methods for prediction of mortality.
(Yan et al., 2022)	A Multifaceted benchmarking of synthetic electronic health record generation models	Yan et al. provide an evaluation framework for GAN-based generation methods based on utility and privacy metrics, a model scoring mechanisms and accommodates for differences in complexities of models. Although extensively analysing the privacy metrics, its interpretation remains undiscussed.

.

G
Evaluation of privacy metrics

The appendix analyses the privacy metrics of synthetic EHRs and their interpretation as part of the literature review in Chapter 4. They are studied within the privacy threats in §4.4.2.

Following Figure 11, this appendix analyses the following metrics: nearest neighbour (NN) (§G.2), NN adversarial accuracy (§G.3), similarity measures (§G.4), and the meaningful identity disclosure risk (§G.5).

G.1 Definition of privacy metrics

Table 11 provides an overview of the privacy metrics that are proposed in literature to measure the re-identification risks of synthetic EHRs models, as well as the privacy threat they address and by whom they are proposed. This serves as guidance through the analysis of the privacy metrics in the following sections.

Table 11. Definition of privacy metrics in literature review

Metric type	**Privacy threat**	**Metric**	**Definition**	**Metric used by**
Privacy mechanisms	MI	Differential privacy	Framework for introducing random noise to (training) dataset to mask sensitive information while maintaining the statistical characteristics.	(Gwon et al., 2024)
		∈-Identifiability	Identifiability metric embedded into the loss function of the generation model to control the distance of synthetic data records to original data	(Yoon et al., 2020), (Shi et al., 2022)
Post hoc privacy metrics	MI	NN similarity	Group of metrics that explicitly applies NN variants to calculate distances to the closest records in the real-world dataset. These can be used to show a normalised distance to the real-world data. Another example is to use it in frameworks that simulate MI attacks; MI metrics show how well an adversary can infer whether an individual was in the real-world dataset, based on distances of the synthetic dataset to the real-world records known to the adversary.	(Yoon et al., 2020), (Yoon et al., 2023), (Torfi & Fox, 2020), (Z. Zhang et al., 2020), (Z. Zhang et al., 2021), (Z. Zhang et al., 2022), (Yan et al., 2022), (Choi et al., 2017)
		NN Adversarial Accuracy	Metric to measure the degree of which models overfit tot its training data, by comparing aggregated distance from a synthetic record to real-world training dataset and to real-world test dataset.	(Venugopal et al., 2022), (Yale et al., 2019a), (Yale et al., 2020), (Yan et al., 2022), (Theodorou et al., 2023),(Venugopal et al., 2022)
		Holdout sample distance	A type of NN approach that evaluates distances, or distribution in research of real-world datasets to synthetic dataset, by comparing distances from the synthetic dataset with the real-world data used for training and a sample from the real-world dataset that is used for evaluation purposes only. Zhang et al. apply this concept to distributional similarity. When the holdout sample is closer to the synthetic data than the training data, it shows the model memorises data to a lower degree, resulting in less privacy risks.	(Yoon et al., 2023),(Theodorou et al., 2023),(Z. Zhang et al., 2021)

Metric type	Privacy threat	Metric	Definition	Metric used by
		Meaningful identity disclosure	A type of NN metric that is based upon the idea that outliers provide additional information for adversaries, and therefore should be evaluated separately. One prominent example is the meaningful identity disclosure risk, which measures re-identification risk by inter alia analysing the rareness of sensitive attributes in the real-world dataset.	(El Emam et al., 2020), (Yan et al., 2022)
		Perplexity Distributions Similarity	A metric that estimates the likelihood that a given record is generated based on similarities of distributions, thus comparing the perplexity of the distributions.	(Z. Zhang et al., 2021)
	AD	Attribute estimation	A metric that shows the attribute inference risk by calculating the degree to which unknown real-world attributes can be predicted based on a known set of attributes from the real-world dataset, based on the synthetic data. This metric clarifies whether synthetic data discloses attributes of individuals.	(S. Sun et al., 2021), (Z. Zhang et al., 2020), (Z. Zhang et al., 2021), (Z. Zhang et al., 2022), (Yan et al., 2022), (Choi et al., 2017), (Theodorou et al., 2023), (Yoon et al., 2023)
		Reproduction rate	A metric that shows the number of duplicated records in the synthetic dataset in comparison to the size of the real-world dataset. Such metrics clarify whether synthetic data discloses attributes of individuals.	(Braddon et al., 2023),(Z. Zhang et al., 2020), (Yoon et al., 2023)

From Table 11, we can conclude that the landscape of privacy metrics is diverse; model developers seem to have no consensus on what metrics best measure re-identification risks of synthetic EHRs generation models. The interpretation of these metrics are discussed below.

G.2 Nearest neighbour metrics

NN metrics measure similarities between synthetic and original data (Murtaza et al., 2023), by measuring the distance from a datapoint in the original dataset to its NN in the synthetic dataset (Yale et al., 2019a). The interpretation of similarity varies per use case and can differ in distance measures, for example Euclidean distance or hamming distance (Murtaza et al., 2023) or the Wasserstein distance (Shi et al., 2022; Yoon et al., 2020). The distance is calculated with either the real-world training dataset, or researchers separate a holdout sample that is not used to train the model. NN metrics can assess a model's vulnerability to membership inference attacks (Yale et al., 2019a). If there is a record with a distance smaller than a given distance threshold, the adversary concludes that a targeted record is also in the original dataset (Yan et al., 2022).

Yoon et al. (2023) set the ideal value of the membership inference metric, defined as the probability of data of the real-world dataset being used to train the synthetic data generation model, at 0.5 (Yoon et al., 2023). Based on hamming distance, for each test sample, that exists of both training and holdout data, they predict whether the real data sample belongs to the training data, based on the synthetic data (Yoon et al., 2023). 0.5 is seen as a random chance, therefore, a result of 0.5 or close to 0.5 is considered as privacy preserving (Yoon et al., 2023).The EHR-safe model scores 0.496 on one dataset and 0.489 on the other, concluding it is “very close to” the ideal value (Yoon et al., 2023, p. 141). Similar approaches are used by Theodorou et al. (2023) and Li et al. (2023). Theodorou et al. (2023) fixed the number of records known to the adversary and calculated the membership inference risk based on distance. The accuracy of the attack is around 0.5, therefore similar to a random guess and preserves privacy. They state that the model nor the synthetic dataset disclose information about patients. Li et al. (2023) implemented a membership inference risk model to assess classification accuracy with different percentages of noise added to training dataset. By using 90% of original dataset, the model was presumed to be robust against membership inference attacks – the classification accuracy was then near 0.5, which is, again, seen as “flipping a coin” (J. Li et al., 2023). Torfi and Fox (2020) assessed the membership inference attack risks base on a variant of a NN metrics referred to as ‘cosine similarity’. This measure is not based on distance between data points, but on its vectors, measuring the angle between them. This measure is specifically relevant to evaluate correlations (Torfi & Fox, 2020). A normative interpretation of used thresholds is not presented.

In the form of simulating a membership inference attack (“presence disclosure”), Biswal et al. (Biswal et al., 2021, p. 274) assessed what the percentage is of training records that be successfully be discovered by the adversary, under the assumption it already knows some records. This sensitivity rate is complemented with the precision rate, which measures what percentage of the number of patients that the attacker thinks are used are actually used. Regardless of how many records are known to the attacker, it can only discover 20% of the patients in the synthetic dataset with a precision of 70%. A similar evaluation is performed by Li et al. (2023). Zhang et al.(2022) propose a membership inference framework based on a contrastive representation learning approach with a proxy for augmentation. Although insight in these frameworks is not relevant within the scope of this thesis, it important to note there are various approaches to perform membership inference assessments.

Regarding attribute disclosure, Choi et al. (2017), Sun et al. (2021) and Zhang et al. (2021) calculate the NN attribute estimation, measuring the likelihood of the targeted attribute. Similar to other NN metrics, they assume an estimation is accurate when it is below a certain threshold. To illustrate this, Sun et al. (2021) compare the estimation accuracy with an baseline based of the known median value of the population. The average estimation accuracy of the attacker was lower than the baseline (respectively 0.2 and 0.26). Since the accuracy was significantly smaller than a random guess, the model presumes privacy. In addition, to evaluate membership inference attack risks, Zhang et al. (2021) estimate the likelihood that a given record is generated based on similarities of distributions, thus comparing the perplexity of the distributions. Yoon et al. (2023) express attribute inference in terms of accuracy of an adversary to predict the value of sensitive features using the synthetic data by correlating the data known to the adversary to the synthetic data– with as sensitive features gender, religion and marital status. Based on a nearest-neighbour classifier, they have compared prediction accuracy of sensitive features based on other features for the real-world dataset and the synthetic dataset. The model is considered privacy preserving when the prediction accuracy of the real-world data is similar to the prediction (Yoon et al., 2023).

Regarding interpretation of the metrics, Yoon et al. (2020) articulate that the application of NN metrics requires a judgement of what distance between real-world data and synthetic data is deemed “different enough”. Acceptable differences between outliers, for example, should be bigger (Yoon et al., 2020). The authors leave it to the responsibility of users to decide on the acceptable thresholds.

G.3 Nearest neighbour adversarial accuracy

Yale et al. (2019a, 2020) developed the concept of ‘nearest neighbour adversarial accuracy’, a prominent metric for resemblance and privacy in terms of ‘privacy loss’. The variant of the NN metric compares the aggregated distances from one point in the synthetic distribution to the nearest point in the original training dataset, and the aggregated distance between the records in the synthetic dataset and in the original test dataset – thus operating at the record level (Yale et al., 2019a). If a synthetic data point is sufficiently distant from the real datapoint, it is considered a true negative for privacy, and if a real data point is sufficiently distant from a synthetic datapoint, it is considered a true positive (Yale et al., 2020). The metric must be interpreted like a balanced accuracy, as it averages the true positive and true negative rates. Therefore, a value of 0.5 means that synthetic data cannot be distinguished from the original dataset.

Privacy loss can be defined by the difference of the NN adversarial accuracy of the test and training dataset. The privacy loss should ideally be 0, as the desired NN adversarial accuracy for both test and training data is 0.5. When the NN adversarial accuracy of the training set is less than 0.5, the model is exposing data, thus the privacy loss will increase (Yale et al., 2019a). To interpret this metric, it should be taken into account that if the NN adversarial accuracies are both higher than 0.5 this indicates lower resemblance to original data, however, the privacy loss can still be close to zero (Venugopal et al., 2022).

In their test results, Yale et al. (2020) seem to suggest some benchmarks for interpreting the NN adversarial accuracy, for both resemblance and privacy. A value is considered an *optimal* value when 0.50 ± .01, a *good* value when 0.50 ± .03, a poor value when the value is out of that range. For privacy loss, a value is considered optimal when it is 0, excellent when ≤ 0.01, good when it is ≤ 0.03, and poor when the value is out of that range. Venugopal et al. (2022) also applied the NN Adversarial Accuracy, resulting in privacy losses of zero or close zero, concluding the evaluation generation methods are preserving privacy. Yan et al. (2022) and Theodorou (2023) also apply the NN adversarial accuracy, following the thresholds set by Yale et al. (2020). As the values were beneath these thresholds, they labelled their models as privacy-preserving.

G.4 Similarity measures

One metric to measure this information disclosure risk is the reproduction rate, which presents the proportion of duplicated data points in a synthetic dataset (Z. Zhang et al., 2020). Braddon et al. (2023) assess privacy on this metric. They concluded a small number of rows was replicated (0.6%). Although 0.6% seems a small percentages, this translates to 680 patients in their synthetic dataset. Therefore, in my opinion, this percentage is not sufficient to show that the generated data poses no privacy risks. Braddon et al. substantiate their argument by stating that the duplicated data “rarely” concerns “unusual cases” (Braddon et al., 2023, p. 297). It can be argued that when duplicates belong to a larger group, the individual becomes indistinguishable from the whole – with lower privacy risks as a result. However, their conclusions leaves the reader guessing about the extent of the risks for outliers. Their research does confirm, however, that there is indeed a risk of identity disclosure for outliers (Braddon et al., 2023). They propose to remove the duplicated outliers, considering it is only a small proportion of the data. This does not seem to be a sustainable mitigating measure for other researchers, considering its implications for research utility. Moreover, it can be questioned whether the reproduction rate suffices as sole privacy metric: what about data rows that partially match, or are close to the original data?

Another metric for assessing disclosure risks is the outlier similarity rate, which measures how close synthetic data records to real outliers (Murtaza et al., 2023). Yan et al. (2022) have quantified this with numerous utility and privacy metrics, showing a trade-off between utility and privacy.

Yoon et al. (2023) define the risk of re-identification as the probability of records being re-identified by matching the synthetic data to the training data. Specifically, they measure re-identification by splitting the synthetic dataset into sub-datasets with only a subset of different features. For each row of synthetic data, the record in the original dataset that is most likely to represent the synthetic record is determined using the NN metric. If one row is mapped to the same real individual across the sub-datasets, the record is considered re-identifiable. There is always some risk of re-identification, for example, when individuals belong to a larger group with similar characteristics (Braddon et al., 2023; Yoon et al., 2023). An optimal value of 0 is therefore not feasible. The baseline value can be determined by calculating the re-identification risk of a real-world holdout sample, i.e. a portion of the real-world dataset that has been set apart and of which we already know has not been used to train the model (Yoon et al., 2023). A result close to the baseline value is considered as privacy preserving (Yoon et al., 2023).The EHR-safe model scores 0.061 on one dataset (baseline value 0.049) and 0.0.085 on the other (baseline value 0.068), concluding it is “very close to” the benchmark value (Yoon et al., 2023, p. 141).

Thomas et al. (2022) also analysed the reproduction rate. 0.37% of the rows were replicated, which translates to 6800 rows. In their case, most of the duplicated records were missing values, which, to their conclusion, mitigates the risks of meaningful identity disclosure. They consider their application privacy preserving. As the dataset do not represent individuals from the original dataset, reproduction rate is low, considering the small fraction of duplicates in the dataset, and these duplicates were non-informative (Thomas et al., 2022). This statement should be subjectively interpreted by the data owners, as they bear the risks. Interestingly, Thomas et al. are part of health research teams and apply externally developed generation models. This should be considered when interpreting their evaluation. It gives insight in their view regarding acceptable privacy risks. Something to take into account is that records may be duplicated due to chance, therefore, another run of the model may decrease identity disclosure risks (Thomas et al., 2022). Moreover, since the data generated by Thomas et al. is somehow aggregated in comparison to the original data, they considered their approach as privacy preserving as the groups, defined in individuals with the same unique combination of attributes, consisted of at least 10 individuals. Therefore, their approach did not pose identifiability risks for outliers (Thomas et al., 2022).

G.5 Meaningful identity disclosure risk

In addition to the membership and attribute inference risk and the NN adversarial accuracy, Yan et al. (2022), propose to the use of the ‘meaningful identity disclosure risk’, referring to the adversary’s ability to identify synthetic records with meaningful attributes to learn something new about the population (El Emam et al., 2020). Fully synthetic datasets should not have one-to-one mapping with records from the original dataset (El Emam et al., 2020). Therefore, this privacy metrics measures the probability that a sample record in the original dataset can be identified by an attacker by matching it with an individual in the population (the information that is known to the attacker), thus measuring similarities. The meaningfulness is measured by determining to what extent the individual is an outlier in the original dataset, and to what extent the synthetic record has a similar value to the real value (El Emam et al., 2020). This metric is based on the idea that individuals who differ from their sample can reveal more about individuals (El Emam et al., 2020).

Most interestingly, El Emam et al. (2020) provided extensive argumentation to an acceptable risk threshold value of 0.09 based on statements from the European Medicines Agency. One way to translate the probability-based metrics to concrete re-identification risks can be found in the publication of statistical datasets in open data science. Organisations concerned with publishing statistical datasets, such as governments, often anonymise data via aggregation (European Medicines Agency, 2016). To determine whether the aggregated data is sufficiently anonymous, they consider whether the number of data subjects that is represented by one data record, referred to as ‘cell size’, is below a threshold (Wilkinson et al., 2020). For example, the research institution of the case study uses 7 individuals as minimum cell size (interview 5); 5 is also often used a minimum cell size (Wilkinson et al., 2020). El Emam et al. (2020) translate this cell size into a probability of re-identification, calculated as one divided by the minimum cell size. The European Medicines Agency (EMA) uses a threshold of 0.09, for example.

Unfortunately, the model proposed by Yan et al. (2022) has yet to be applied by other authors to proof its value for evaluation GAN-based models. Theodoru et al. (2023), for example, followed some metrics by Yan et al. al. (2022), however, did not apply the full scoring framework.

H
Analysis of synthetic data generation in action arenas

This appendix presents the analysis underlying §5.4, i.e. the identification of barriers, drivers, and solution directors for how synthetic data generation unfolds in the three action arenas that Chapter presented for health data sharing. §5.4 thus contains the conclusions of this appendix. The analysis is based on interviews and literature from academia, governments, and industry.

H.1 Action arena 1: Operational decisions in health data sharing

Interactions discussed in this section relate to a changed actors playing field, including the multidisciplinary team and the associated uncertainties (§H.1.1). After concretisation of the impact of synthetic data on health data sharing (§H.1.2).

H.1.1 Changes to actor playing field

The interviews made clear how synthetic data may change the current institutional landscape. This section introduces how actors' positions in synthetic health data sharing change.

First, as visualised in Figure 13, technology providers deliver the technical resources necessary to generate synthetic health data. As the generation model is ran in the healthcare provider's environment, the health data is not disclosed to the technology provider. Although this may seem trivial, this is an important governance choice regarding the data protection responsibilities under the GDPR. A healthcare provider is a data controller, and the technology provider is not subject to the GDPR, regarding the generation of synthetic health data by the healthcare provider at least. The technology provider only delivers the tools to process data – they do not process health data by themselves (General Data Protection Regulation, 2016, art. 4(7)). This software architecture choice does not (further) invade patients' data protection, considering the data is not disclosed to an additional third party. This would be different if technology providers were to run their generation models using a cloud infrastructure (Brauneck et al., 2023). Furthermore, appropriate configuration, evaluation, and application of PETs such as synthetic data generation, requires expertise (TP1). Carrying out such activities with lacking expertise may result in implementation errors (Information Commissioner's Office, 2023). For instance, healthcare providers are to interpret the quantitative model metrics presented by technology providers. When the interpreters lack knowledge on such metrics, this may result in an inappropriate balance between privacy and utility (Information Commissioner's Office, 2023). Hence, to account for such risks, persons involved in synthetic data sharing must be knowledgeable about how the technology works, how to benchmark certain metrics scores and whether this complies with data protection laws, and above all maybe, having the time to do this properly.

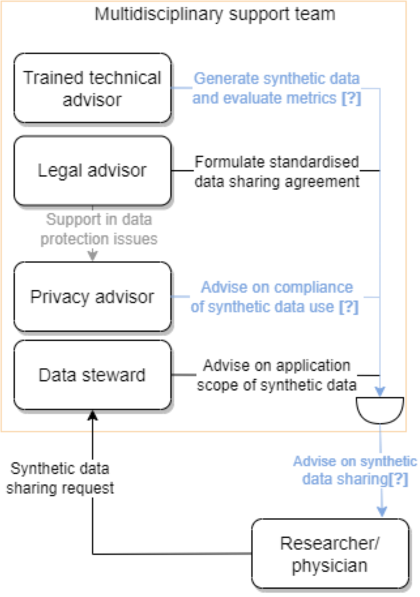

Multidisciplinary support team
Trained technical advisor
Generate synthetic data and evaluate metrics [?]
Legal advisor
Formulate standardised data sharing agreement
Support in data protection issues
Privacy advisor
Advise on compliance of synthetic data use [?]
Data steward
Advise on application scope of synthetic data
Synthetic data sharing request
Advise on synthetic data sharing[?]
Researcher/ physician

Figure 16. Proposed Interaction between multidisciplinary team and

To account for these difficulties, I propose more intensive forms of collaborations between actors within the organisation responsible for health data. The party responsible for interpreting privacy metrics would be the healthcare provider (TP1, TP2). This collaborative effort is referred to as a "multidisciplinary support team". Figure 13 shows how this teams supports researchers in sharing synthetic health data. Figure 16 zooms in on the contributions of each role to advise researchers. This figure specifically refers to roles, as the actors that take on this role may vary. For example, technical evaluations of AI models can be performed by data scientists or data stewards with a technical background (DS1-H-VAL). Another approach would be to train (non-technical) employees via workshops (TP2). To reflect that technical expertise is required, it is explicitly recorded as a separate role in the figure. Furthermore, privacy advisors, a role that is fulfilled by privacy officers, play an important role in guiding researchers to share synthetic data compliantly. Moreover, the legal department that takes on the role of legal advisors, is concerned with (standardising) data sharing agreements for sharing health data, a task that is further clarified in the next section. Data stewards familiar with the technology can carry the role of introducing synthetic data to researchers (DS1-H-VAL). Accordingly, the interpretation of privacy metrics and demonstrating that data are safe enough to share should be a shared responsibility (DS1-H-VAL). Borrowing the symbol for logic AND gates, Figure 13 proposes the roles that contribute to a multi-disciplinary advice on health data sharing to researchers covering the technical, data management and legal aspects of synthetic health data sharing.

When comparing the current institutional environment (Figure 7) with the suggested institutional environment (Figure 13 and Figure 16), it may seem that synthetic data does not significantly change the data sharing process – the same actors remain involved, taking the same decisions. The impact that is not visualised, however, is that the decisions can be made more efficiently, considering the reduced risks of synthetic data in comparison to pseudonymised data and standardisation efforts. The structure remains the same, as the application of synthetic data still requires a risk-based approach, resulting in a case-by-case assessment of the privacy risks. The structure not changing also shows that synthetic data fits the organisational data governance structure: actors can easily be positioned to apply synthetic data. Thereby, no significant changes are required in the organisations' data policies (DS1-H).

H.1.2 Impact of synthetic data on health data sharing

Having understood what actors are involved in synthetic data sharing, the question that remains is how synthetic data enables health data sharing. This contribution is threefold.

First, from a privacy perspective, synthetic data has significantly less privacy risks in comparison to pseudonymised data (DS1-H-VAL). In the current process, researchers of healthcare providers are responsible for pseudonymising data – which often amounts to removing direct identifiers from datasets (DS2-H). When synthetic data generation is used as the default method to share data with other parties, these manual processes are replaced with methods that provide strong technical safeguards for patients (§5.1.2).

A second advantage of synthetic data over pseudonymous data is that the reduction of privacy risks broadens data sharing possibilities, i.e. organisations can share more data with a similar risk appetite. Whereas pseudonymous data may not have been suitable for certain exploratory software testing, sharing with third-party researchers, or educational purposes, synthetic data increases data availability for such uses.

Third, in addition to the broadened application scope, synthetic data generation can simplify the time- and labour-intensive process of sharing health data. In the current process, sharing health data usually takes some time, as the healthcare provider and research institute have to agree on the terms of data use (§3.2.3). The use of synthetic data can speed up the conclusion of data sharing agreements, as agreements can be partially standardised (DS1-H, LC-H). The institutional means showed that data sharing still requires agreement on the acknowledgement of research results resulting from synthetic data analysis, but the legal department can formulate a section on data protection safeguards for most use cases (DS1-H, LC-H). Some sections will remain variable, such as the purposes for which synthetic data may be used.

H.2 Action arena 2: Decisions in definition of personal data

Analysis of this arena consisted of exploring the influence of decisions on higher abstraction levels, focusing on how the legal interpretation of anonymisation at the Dutch and EU level relates to the synthetic data generation methods and evaluations. This was done in three parts. First, §H.2.1 discusses the implications of a lack of interpretability for the privacy evaluation of synthetic data generation. The ability to evaluate synthetic health data is a prerequisite for determining whether synthetic data qualifies as anonymous data. §H.2.2 explains how the legal definition of anonymisation relates to the application of synthetic data generation. The legal definition of anonymisation is closely related to re-identification risks. Therefore, §H.2.3 expands the anonymisation analysis with important guidance from the EU and relates these to the evaluation of EHRs. Lastly, §5.4.2 concludes on the factors that influence the impact of synthetic data on health data sharing.

H.2.1 Challenges related to evaluation of synthetic data

The search for highly usable yet privacy-preserving generation methods for EHRs, has resulted in generation and evaluation methods that comprise extensive ML models to generate mixed-type and high-dimensional data types (§4.2). These models often lack transparency. This issue is inherently present with the black-box models used to generate synthetic data (Gal & Lynskey, 2023; Giuffrè & Shung, 2023). A contributor hereto is that insofar technology providers are third-party commercial organisations, they want to protect their intellectual property (DS1-H-VAL). This lack of transparency complicates the interpretation of generation models (Gal & Lynskey, 2023). Interpretability can be referred to as the capacity to understand how the model works and how it generates outputs (Gal & Lynskey, 2023; Molnar, 2023). Interpretability is especially undermined when synthetic data generation is combined with differential privacy (Stadler et al., 2022), because it is impossible to predict what characteristics and patterns of the real-world datasets are suppressed (Stadler et al., 2022).

These interpretability issues are problematic for the privacy-preserving application of synthetic data. Healthcare providers must assess the privacy risks associated with synthetic data to ensure compliance with data protection rules. To determine what privacy rules are applicable, they need insight into how the synthetic data is generated (DS1-H-VAL). Moreover, the lack of transparency of generation models, especially when combined with differential privacy, can shrink confidence in the use of synthetic data, for example when researchers have to make decisions or draw conclusions from the synthetic data (Gal & Lynskey, 2023; Giuffrè & Shung, 2023; Molnar, 2023). Thus, when relating synthetic data to its healthcare context, a challenge of interpretability of privacy metrics arises.

A strategy to combat interpretability issues is to provide extensive yet practical evaluation frameworks that bridge the gap between AI and its institutional context, so that healthcare providers can confidently generate and share synthetic data (Giuffrè & Shung, 2023). The use of modern generation methods necessitates suitable evaluation methods, as the stakes of data sharing are high for healthcare providers (in terms of liability) and patients (in terms of privacy) (Giuffrè & Shung, 2023). However, there is no standard approach yet to assess the degree to which the synthetic dataset resembles the real-world data and preserves privacy (Chapter 4). The interpretability issue of generation models extends to the privacy evaluation of synthetic data generated by such models. The numerous privacy metrics can be difficult to comprehend due to a lack of standardisation. Additionally, the metrics, based on likelihood and divergence, are hard to translate to concrete privacy risks due to their quantitative nature and a lack of supporting interpretations by researchers (Chen et al., 2021).

To conclude, the interpretability issues complicate the application of synthetic data by clinicians and researchers, e.g. when users want to compare various generation methods or assess the model's privacy risks (Chen et al., 2021; Giuffrè & Shung, 2023). Again, technology providers have great responsibility to present their technology in an understandable manner, presenting the complex methods and metrics to the user in a trustworthy way. This can be supported by standardising privacy metrics via industry initiatives, such as the IEEE work package regarding synthetic data. Although they have not yet published any guidance, they are working on a white paper for industry privacy standards (*Synthetic Data*, 2023).

H.2.2 Challenges related to legal definition of anonymisation

Following the rules-in-use, a question that arises when relating synthetic data to its institutional context is whether synthetic data qualifies as anonymous or personal data under the GDPR (Beduschi, 2024). We have seen that this classification is important to determine what rules apply to synthetic health data sharing and how it may change health data sharing practice. Researchers in the field of synthetic data generation, and computer science in general, express preservation of privacy in terms of re-identification risks (Hildebrandt, 2019) – risks that are closely related to the legal definition of personal data, pseudonymisation, and anonymisation (Mostert et al., 2016). This section discusses the relation between the legal definition of personal data and synthetic data generation.

The definition of personal data is based on identifiability, which requires an assessment of the reasonableness standard. This standard takes into account all means reasonably and likely to be used to identify a person (§3.1.2). As El Emam et al. (2020) articulate, the question is how "reasonably" and "likely" should be defined for synthetic data contexts. Quantitative interpretations of these concepts have proven to be subjective, because these terms leave significant "wiggle room" and thus varying perceptions of when something is "likely" (Mauboussin & Mauboussin, 2018). Case law provided some guidance on the scope of these concepts, showing that the means to identify individuals must be interpreted broadly (§3.1.2). Nonetheless, authors have expressed their concerns, asking for clarification and harmonisation among EU member states and bodies (Groos & Van Veen, 2020; Hansen et al., 2021). The question remains how these concepts can be translated into quantitative values that result from the probability-based privacy metrics (El Emam et al., 2020).

The uncertainties caused by this incoherence perpetuate in the evaluation of synthetic data generation models. Technology providers analyse anonymisation in terms of re-identification risks (§4.4). Still, consensus lacks on how these re-identification risks should be measured and what thresholds for re-identification risks are acceptable (§4.4.4), resulting in a fragmented technological landscape for evaluation of synthetic data. Moreover, apart from (El Emam et al., 2020; Theodorou et al., 2023; Yan et al., 2022), none of the authors provided qualitive argumentation for acceptable re-identification thresholds. Hence, the decisions in action arena 2 that caused a lack of legal clarity also influence the impact of synthetic data on health data sharing – it leaves technology providers guessing about how to evaluate privacy metrics while complying with data protection law.

In addition, the discussion in case law regarding a stricter definition of personal data (§3.1.2) may influence the impact of synthetic data on health data sharing. If pseudonymised data are no longer identifiable to data recipients according to the reasonableness standard, health data can be shared easier as fewer data protection rules apply, contributing to the policy objective of data availability. For this purpose, the value of synthetic data is reduced because these data can already be shared. However, from the policy objective of patient privacy, this development is less desirable because there are fewer data protection safeguards for similar data sharing activities, such as legally binding agreements. To achieve this objective, synthetic data generation can therefore provide an important technical safeguard to protect patient data.

For both issues, a possible solution is to further operationalise the application of the legal concepts (Hansen et al., 2021). Some researchers call for a definition of measurable privacy leakage thresholds for synthetic data (Bellovin et al., 2019). These thresholds can vary to account for sector-specific privacy demands (Bellovin et al., 2019). The benefit of this approach is that it provides legal certainty for developers and users of synthetic data regarding acceptable threshold, which can stir its adoption and development (Bellovin et al., 2019). A drawback of this approach is that it requires a method to define acceptable risk thresholds and metrics that can be applied to different types of synthetic datasets (Beduschi, 2024). Also, it can be questioned whether this is desired from the perspective of patients: a quantitative test may allow users of synthetic data to avoid a risks analysis that carefully balances privacy and utility. The European Medicine Agency highlights that measuring the risk of re-identification primarily concerns a qualitive assessment based on the characteristics of the source data, considering for example the prevalence of the disease, sample size, and number of sites to define an acceptable threshold. At the same time, they encourage the use of quantitative measures (European Medicines Agency, 2016). Moreover, in defining the acceptable threshold, the EMA mentions the existence of mitigating measures as important factor (European Medicines Agency, 2016). Therefore, this thesis calls for a mixed approach to evaluate synthetic data, that comprises quantitative and qualitative assessments.

H.2.3 Challenges related to evaluation of re-identification risks

As if the rules for anonymisation are not yet complicated, the Article 29 Working Party published guidance on anonymisation techniques (Article 29 Working Party, 2014). Like the GDPR, they refer to anonymisation techniques rather than anonymous data to emphasise the remaining re-identification risks that are linked to any anonymisation measure (Article 29 Working Party, 2014). An anonymisation technique is considered effective when it prevents risks of singling out, linkability and inference. Singling out refers to the ability of identifying an individual from a set of records. Linkability refers to the ability to link at records to an individual or group of individuals in one or more datasets, for example based on correlations. Inference refers to the ability to infer attribute values of individuals, for example by predicting values based on other (known) values (Article 29 Working Party, 2014). Overall, the anonymisation process should be "completely" irreversible (European Data Protection Board, 2020; Groos & Van Veen, 2020, p. 3). Moreover, additional measures should be adopted to be considered an effective anonymisation technique, depending on the context and purposes of the data processing activities (Agencia Española Protección Datos, 2021; Article 29 Working Party, 2014). This does not necessarily require technical privacy assessments only: by conducting DPIAs, healthcare providers can evaluate what the impact of certain re-identification risks are, depending on purpose of synthetic data sharing and what type of data is shared (sample size, outliers, uniqueness of information, etcetera).

Authors have expressed their concerns regarding these guidelines, stating that they do not align with the reasonableness standard from the GDPR and *Breyer*. (Groos & Van Veen, 2020; Hansen et al., 2021). Only if "nobody could possibly reidentify, the data can be considered anonymous" according to these guidelines (Groos & Van Veen, 2020, p. 5).[8] However, these guidelines are still referred to by the EDPB (Groos & Van Veen, 2020). To account for this uncertainty, I propose to follow the clarifications provided by the Irish DPA, who views the re-identification risks as guidelines to analyse whether an anonymisation (or synthetic data generation) attempt is unlikely to re-identify individuals, as part of the reasonableness standard (Data Protection Commission, 2019).

To relate these guidelines to the privacy metrics discussed in Appendix G and Chapter 4, quantitative metrics that measure duplication of records can help to demonstrate the singling

[8] This thesis only discusses to the implications of these guidelines, referring to Groos and Van Veen (2020) for a more substantive analysis.

out risk. Other metrics to identify singling out are based on meaningful identity disclosure, typically outlier sensitivity of attributes (Boudewijn et al., 2023). Attribute disclosure metrics help to assess the risk of inference. Membership inference attacks are harder to classify – the information disclosed in such attacks is that an individual is part of a dataset, but not necessarily the value of individual records. As it discloses some information about individuals, it can inform us about singling out or inference risks, depending on the metrics used. Metrics based on nearest neighbour (NN) can be used to quantify inference. Membership inference attacks that predict whether records are real based on a combination of attributes, can be classified as a singling out measure (Boudewijn et al., 2023). Likability risks are harder to relate to specific privacy metrics, as methods proposed in literature are relatively scarce (Boudewijn et al., 2023). Attacks based on NN distances requires information from other sources. Boudewijn et al. (2023) argue that attack frameworks that presume some information from other sources than synthetic data, as is the case with attacks that use NN distances, may be interpreted as a linkability attack.

What can be concluded, is that even though the Article 29 Working Party provides some guidance to the interpretation of privacy metrics, none of the articles in the literature review in Chapter 3 evaluate the privacy risks according to these guidelines. This makes it harder for healthcare providers and research institutes to evaluate whether the synthetic data meets the legal requirements for anonymisation.

H.3 Action arena 3: Decisions in legal base of health data sharing for research

The generation of synthetic data is a processing activity subject to the GPDR, which therefore should rely on a legal base (§5.1.1). An important success factor for synthetic data would be that no consent need to be sought under the consent-by-default approach. This section explores the possible legal bases that the generation of synthetic health data may rely on, starting with the GDPR (§H.3.1), followed by a discussion of synthetic data in the consent-by-default approach (§H.3.2).

H.3.1 Challenges regarding legal base in the GDPR

The presumption of compatibility, which is closely related to the question what legal base may be relied upon for health data sharing was only mentioned in a broader discussion of the legal bases in action arena 3. The purpose limitation principle is more extensively discussed in this section, as guidance by the EPDS regarding anonymisation may affect whether the generation of synthetic data needs an additional legal base.

Personal data shall be collected for specified, explicit and legitimate purposes (‘purpose limitation’) (General Data Protection Regulation, 2016, art. 5(1)(b)) (§3.1.3). When providing healthcare, the purpose for personal data processing is clear: health data must be processed to deliver care to patients. Subsequently, for healthcare providers to lawfully process data for scientific research purposes should be “considered to be compatible lawful processing operations” (General Data Protection Regulation, 2016, rec 50). According to the GDPR, healthcare providers thus not need to define an additional, separate legal basis than the one defined for the original data collection – i.e. delivering care to patients. Scientific research has a broad understanding in the GDPR, practically referring to any research. The EDPS, however, specifies this, arguing that the presumption of compatibility should apply to genuine research only. This means that research must serve the public interest (European Data Protection Supervisor, 2020).

When synthetic data is generated for health data sharing, the purpose of processing health data is not to train generation models or to generate the data. The purpose is to use synthetic data to support research, e.g. to train and test ML models to predict diseases. When the secondary use of health data for research is allowed with the original dataset, generating synthetic data for this research purpose is considered compatible. If not, the healthcare provider needs to formulate new purposes with their own legal basis. The GDPR seems to indicate that scientific research is compatible with the original purpose per se (Slokenberga, 2022).

The presumption of compatibility is expanded by Article 29 Working Party's guidelines on anonymisation techniques (Article 29 Working Party, 2014). The Article 29 Working Party considers anonymisation as an act of further processing that is compatible with the initial purposes, provided that the anonymisation method reliably produces anonymous data conform the guidelines defined in their opinion on anonymisation techniques (Article 29 Working Party, 2014). Based on this, El Emam (2020) argued that the act of anonymisation through synthetic data generation does not require an additional legal base. I belief it is more nuanced than this: the question of whether synthetic data generation may rely on the presumption of compatibility comes down to the qualification of synthetic data generation as an effective anonymisation technique. As shown in action arena 2, this is very difficult to establish, with some arguing it is even impossible from a technical perspective (van der Sloot & van Schendel, 2024). This prevents researchers from simply relying on the presumption of compatibility of anonymisation.

Lastly, there is some controversy about the interaction between the principle of purpose limitation on the one hand, and the principles of fairness, lawfulness and transparency on the other hand (Slokenberga, 2022). The GDPR article that prescribes the legal bases for processing personal data, does not mention a ground for scientific research (General Data Protection Regulation, 2016, art. 6(1)), whereas the article that exempts the prohibition on processing special categories of personal data, among which health data, addresses scientific research explicitly (General Data Protection Regulation, 2016, art. 9(2)(j); Slokenberga, 2022). This leaves room for different interpretations: a cautious interpretation states that secondary use of health data for research requires a specific legal basis in both article 6 and 9 GDPR, regardless of what the purpose limitation states about scientific research (Slokenberga, 2022). Another interpretation is that, based on the purpose limitation principle, the original legal basis may be relied upon for secondary use of health data for research purposes (Slokenberga, 2022). This debate is confirmed by the EDPS, who has stated that recital 50 only has an advisory role (European Data Protection Supervisor, 2020). The EDPS seems to indicate that the definition of a specific purpose and legal basis are two different principles, and should hence be considered separately (European Data Protection Supervisor, 2020). The EDPS has explicitly spoken against the broad interpretation of the presumption of compatibility (European Data Protection Supervisor, 2020). They argue that the presumption is not a free pass for all further processing activities for scientific research purposes.

What is specifically interesting for synthetic data generation, is that the EDPS highlighted the importance of Article 89(1) GDPR to ensure appropriate technical and organisational measures, such as pseudonymisation and access limitations (European Data Protection Supervisor, 2020). Synthetic data generation could be seen as such a measure (§**5.1.2**). Therefore, by implementing such a privacy-enhancing technology, healthcare providers and research institute can more effectively comply with this condition, in comparison to pseudonymisation techniques.

To conclude, similar to the definition of anonymous data, there is some uncertainty regarding the scope of the purpose limitation principle and how this relates to scientific research. This complicates the selection of an appropriate legal base for healthcare providers that want to share synthetic health data. In the context of this thesis, this means that if the cautious interpretation stands, synthetic data generation requires an additional legal base as defined in Article 6 GDPR. However, following the wording of the purpose limitation principle as well as recital 50, synthetic data generation may presumably rely on the original legal basis as long as a contribution to scientific research can be established. Moreover, when synthetic data is seen as a reliable anonymisation technique, the act of anonymising personal data may be considered compatible as well. The EDPB has promised to provide guidance on the scope of the presumption of compatibility of scientific research (European Data Protection Board, 2020, para. 43), which should provide clearance on this matter.

H.3.2 Challenges regarding legal base in the Netherlands

The previous section showed that there is an additional argument provided at EU level for the presumption of compatibility of anonymisation. Unfortunately, it is still uncertain how this relates to the Dutch rules regarding secondary use of health data for research. Action arena 3 showed that the Netherlands maintain a rather strict approach towards processing health data for research, which requires researchers to obtain consent where possible as defined.

First, actors in the Netherlands have interpreted the presumption of compatibility of scientific research strictly. Following the survey study of Hansen et al. (2021), the presumption of compatibility of scientific research only seems to apply within organisations in the Netherlands, presumably because this is in line with reasonable expectations for patients about where their data will flow. Healthcare providers follow a similar approach, stating that the presumption of compatibility is only applicable to data controllers that already process and have access to health data, based on the EDPS' guidance (COREON, 2022). Therefore, research institutes need an additional legal base for receiving the health data. However, following the exact wordings of the EDPS, "personal data collected in the […] healthcare context, […] may be further used for scientific research purposes, by the original *or a new controller*, if appropriate safeguards are in place" (European Data Protection Supervisor, 2020, p. 22). Therefore, applying the presumption of compatibility only within the organisation that already processes data, is not required according to the EDPS, and therefore considered strict. Here too, governmental actors such as the Ministry of HWS or the Dutch DPA have not provided guidance on the interpretation of the purpose limitation principle and compatibility presumption of scientific research.

Second, there is some discussion on how the presumption of compatibility should be interpreted in the Netherlands. Based on the EDPS' guidance, COREON (2022) for example, argues that the process of anonymisation does not need an additional legal base. To apply the presumption of compatibility, the organisation generating the synthetic health data must take appropriate measures to protect patients (General Data Protection Regulation, 2016, art. 5(1)(b) jo. 89(1)). As such, synthetic data generation can be seen as safeguard (§5.1.2). As organisational safeguard, the Code of Conduct by COREON (2022) states that the purpose for which the health data is anonymised must be scientific research. Scientific research should aim to gather new insights, follow recognised research standards and contribute to the improvement of healthcare (Appendix E). With these measures, anonymisation is compatible with the initial collection purposes (COREON, 2022). It should be emphasised that the identified uncertainties apply to the generation process of synthetic data. When the output of this process qualifies as anonymous (or non-personal) data, synthetic data sharing is not subject to the GDPR and healthcare providers and research institutes do not require an additional legal base for sharing and analysing such synthetic health data.

As the process of anonymisation occurs within the organisation that collected the data, the issue that arose with the presumption of compatibility in cross-organisational secondary use of health data does not apply here. Yet, when the synthetic data generation cannot be argued to be an effective anonymisation technique, and qualifies as personal data, researchers fall back on the current rules for health data sharing as discussed in action arena 3. This means that they have to follow the consent-by-default approach and both the healthcare provider and research institute need to rely on an additional legal base to share synthetic health data.

Third, to make things even more complex, there is also a relationship between synthetic data and the WGBO. COREON, for example, believes that anonymous data is not covered by the WGBO (COREON, 2022), whereas the Dutch Federation for Academic Hospitals ('Nederlandse NFU) argues that sharing anonymous data still breaches medical confidentiality, and therefore in principle requires consent from the patient (NFU, 2020). Thus, there are uncertainties regarding the legal bases defined in the GDPR and UAVG, as well as the obligation to ask consent following the WGBO.

The results of the contradictory approaches from EU and national actors create uncertainties for healthcare providers and research institutes regarding the applicable legal base for synthetic data generation. These interactions between synthetic data and its institutional context can best be illustrated with an example from the use case. A healthcare provider's data steward considered that the generation of synthetic data requires an additional legal base (DS1-H). On the contrary, a technology provider assumed synthetic data generation can rely on the original legal base, following the EPDS' guidance on anonymisation techniques (TP2). Another technology provider acknowledged the uncertainties, however, primarily focused on the presumption of compatibility of scientific research for synthetic data generation (TP1). This shows that the uncertainties regarding the applicable legal base are indeed a problem that should be addressed to stimulate a compliant implementation of synthetic data generation.

I
Framework validation

This appendix provides the validation of the proposed framework in chapter 5. The interviewees both play an important role in the adoption of synthetic data, being a data steward of a healthcare provider (DS1-H-VAL) and a technology provider (TP1-VAL). Both interviewees have also been interviewed for the use case. This thesis does not claim to conduct a full case study as validation for the framework. Next steps in the case study evaluation are to develop the framework further, by performing more iterations through the design cycle, and then implement it in practice.

I.1 Confirmation of problems

The interviewees confirm the problems and add the following considerations.

The healthcare provider's data steward from the healthcare provider addressed that the issue surrounding the classification of anonymous data is intricate. Firstly, the GDPR is a risk-based framework. Organisations cannot be simply deemed compliant or non-compliant; they must take all necessary measures to mitigate risks as much as possible. Crucial for data controllers is that anonymous data, as defined in the GDPR, are not considered personal data and therefore do not fall under the GDPR. The ensuing question is indeed: when are data truly anonymous? Aggregated data is often regarded as anonymous, yet even then, there are re-identification risks. The Central Bureau of Statistics can ascertain the identity of an individual with 85% certainty based on a four-digit postcode, gender, and age. Thus, it is not always the case that data is completely anonymous. It is important to consider that this is not a simple yes/no question but requires a case-by-case analysis. There are varying degrees of personal data, each requiring different measures. Synthetic data poses re-identification risks similar to aggregated data, albeit with a relatively higher re-identification risk that demands considerable effort. Regarding the classification of synthetic data as anonymous data, this interviewee estimates that, in practice, synthetic data still requires a risk analysis, such as conducting a Data Protection Impact Assessment (DPIA), similar to pseudonymous data. The advantage of synthetic data over pseudonymous data is that privacy risks are significantly reduced, thus broadening its application possibilities. Therefore, whereas pseudonymous data may not have been suitable for software testing, sharing with third-party researchers, or educational purposes, synthetic data can serve these functions. This issue is also confirmed by the technology provider.

Regarding the Dutch approach to consent, the data stewards states that a this discussion is independent of synthetic data. The perspective currently varies from one organisation to another or even from one privacy officer to another. To answer this question personally, this interviewee first considers the alternative regarding legal basis and whether it can be reasonably assumed that the patient would choose a less risky alternative. Secondly, Dutch legislation is founded on autonomy, the right to self-determination of data subjects, whereas the GDPR also allows for data processing for the public interest and scientific research—this is thus a cultural issue. The Dutch Data Protection Authority (DPA) adopts a highly conservative view, which hinders the Netherlands competitively, for instance, in conducting innovative research. This is something that the Ministry of Health, Welfare, and Sport (HWS) could investigate. This conflict between the GDPR and the Dutch DPA and the healthcare sector is also evident in the fact that the Dutch DPA immediately appeals against EU rulings, whereas most in the Netherlands support the rulings. This issue is also confirmed by the technology provider.

The interpretation problem is confirmed by the data steward (see also the following section). An example of a solution for standardisation is, for instance, the development of benchmark methods, enabling the comparison of the generated dataset with other benchmark datasets or the testing of different models for privacy. The technology provider adds that interpreting quantitative privacy metrics presents a formidable challenge. Their involvement lies within the IEEE Synthetic Data Working Group, where recommending a standard for privacy evaluation stands as the foremost objective. However, within this working group, achieving consensus proves elusive and is expected to prolong. Agreement has yet to be reached on which metrics should be selected, how they should be classified, and how they should be elucidated to individuals outside the information and communication community. The initial white paper maintains a broad focus, delineating synthetic data generation and its general advantages. Regarding privacy, this white paper offers an overview of the remaining risks. Concerning privacy evaluation of the technology provider, concepts such as singling out, inference, and linkability are present, concepts that are now also discussed within the working group. Nevertheless, they continue to grapple with the challenge that these metrics pose in terms of user interpretation.

I.2 .Actors involved in application of synthetic data

It is a challenging task for data users to assess the quality of synthetic data, including privacy aspects. Metrics on re-identification risks, for example, are still relatively new and require thorough explanation. According to the data steward, ultimately, it is up to the data controller to demonstrate that the data are safe enough to share. In addition to technology providers supplying metrics, it may be desirable for hospitals to conduct this themselves—they are ultimately responsible. The same applies to utility metrics. This need arises partly from the non-transparent practices of technology providers: as long as they do not provide insight into how they generate data, it is very difficult to ensure that the model complies with the stringent healthcare regulations (confirming interpretability challenges). Evaluating synthetic datasets is not an easy task; it must be a shared responsibility. Individuals with an AI background, such as those in data science teams, can assist in interpreting quantitative metrics. The application of synthetic data is a shared responsibility: technical data stewards can also assist researchers in implementing synthetic data generation. However, technical knowledge is a prerequisite for application. Additionally, Data Protection Officers and privacy lawyers contribute to the compliance issue, and at the executive level, the Board of Advisors must determine when synthetic data can be used. The technology provider adds to this that government organisations such as the ministry and the Dutch DPA have a significant role in clarifying legal concepts. Parties such as ISO and IEEE have a pulling role in standardising privacy metrics. Technology providers, lawyers and other privacy specialists can then provide interpretation.

I.3 Synthetic data application process and measures

To better deal with interpretability, I propose the process to apply synthetic data, which requires contextual analysis. In response, the technology provider mentions that the proposed process probably contains the components that you will see in practice. However, standardisation of this process is hopefully still possible to simplify these steps. It is added that it is important to look at the opportunities in the application, not just the privacy risks. Looking at the reference point, the way data is shared now, the use of synthetic data can be much more privacy-friendly – despite the fact that 100% privacy cannot be guaranteed.

As for standardisation, everyone hopes for a golden threshold that indicates if you meet it then you are safe. However, there does not seem to be one, nor is there going to be one, as my analysis also shows, because it would then not match the risks of specific use cases. Standardisation can start with agreeing on definitions. In the first phase, in which the use case is determined, you can move towards grading use cases according to privacy risks, for instance depending on whether data stays in an organisation or is shared with external parties. This can also be done for classifying types of data, which can be categorised as having low, medium or high privacy risks, which should be defined in an unambiguous way. Yet, when standardising privacy metrics, you again run into the interpretation issues, because those rates that emerge from the quantitative evaluation must then be scaled. What should a user classify as good metrics? Is a 3% re-identification rate good or bad? You can combine these different classifications, where the acceptable re-identification risk depends on earlier risk classifications.

To convince users that synthetic data is privacy-friendly, the technology provider mentions that quantitative privacy metrics are important on the one hand, and on the other, transparency regarding the inner workings of synthetic data generation models helps. But, there are also plenty of users who want simple information, and more easily assume it is privacy-friendly. In addition, graphs help convince users about the value of synthetic data. Then they can see at a glance how the synthetic data compares to the real-world dataset that has been synthesised. One way to provide users with information to interpret metrics is with white papers on the concepts used in these quantitative evaluations.

The data steward provides no specific comments on the lifecycle, other than that this healthcare provider follows a similar approach for pseudonymised data. The only difference needed in the organisational data policy is then to set the use of synthetic data as default and formulate exceptions for use of pseudonymised or personal data. By including evaluation moments to mitigate privacy risks for patients, synthetic data can be applied in compliance with the GDPR. Such procedures are in place in healthcare providers, so therefore, this framework suits the institutional environment of healthcare providers.

Other than the measures already mentioned, similar to other types of data, synthetic data needs to adhere to common data models and vocabularies (such as the International Statistical Classification of Diseases and Related Health Problems) to increase interoperability between different systems and organisations. Similarly, the findability issue of data is not solved with synthetic data. This requires the development of central (meta)data catalogues, on institutional, national or EU level. Publicly sharing synthetic datasets is not an option, as this will increase chances of privacy attacks. This is not an institutional, but technical question, as stated by the data stewards.