S.J. San José Sánchez

No Patient Left Behind: A Decision Framework for Addressing Representation Bias in Open Health Data

A Qualitative Study into the Use of Open Health Data

TU Delft

Source cover image: Vecteezy.com

No Patient Left Behind: A Decision Framework for Addressing Representation Bias in Open Health Data

A Qualitative Study into the Use of Open Health Data

by

S.J. San José Sánchez

in partial fulfilment of the requirements for the degree of

Master of Science

in Complex System Engineering and Management

at the Delft University of Technology,

to be defended publicly on Tuesday, March 12, 2024 at 10:30 AM.

Student number:	4975162	
Project duration:	October 1, 2023 – March 12, 2024	
Thesis committee:	Dr. A.M.G. Zuiderwijk-van Eijk	TU Delft – Chair & 1st Supervisor
	Dr. J.M. Durán	TU Delft – 2nd Supervisor
	Dr. C. Figueroa	TU Delft – Advisor

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Preface

Dear reader,

With the completion of this master's thesis, my time as a student in Delft comes to an end. I am beyond exited to present this research to you, as it has probably been my most challenging yet rewarding journey so far. Throughout my studies at TPM, I had the opportunity to work on my personal and academic growth, evolving into the person I am today. The interesting lectures, discussions and people I have met: I will always cherish them and look back on an unforgettable time here at TU Delft.

First of all, I want to thank my graduation committee: Anneke Zuiderwijk-van Eijk, Caroline Figueroa and Juan Durán. It was your expertise and constructive feedback that has shaped the content of this research and contributed to my academic development. It was a pleasure working with you. Furthermore, I want to thank the interview participants who have dedicated their time and insights to enhance this research with valuable contributions.

In addition, I want to express my gratitude to Ivar who has supported me during the joyful, but also stressful times of not only writing this thesis but throughout my whole study. At times I was probably not fun to be around, but you always continued to get me back on my feet and helped me to believe in myself. I am so grateful for that.

I am beyond lucky to have an amazing and supporting family. Mom, you have always been my rock, I appreciate everything that you have done for me. Jordy, your help and belief in me was a key factor in studying at TU Delft. Michael, our conversations about school, work and the future meant more to me than you know. Lastly, my father, who did not experience this journey with me, but of whom I am sure he is guiding me along the way. Help comes in many forms.

To everyone I have not named but who has been a part of my academic journey in one way or another, I am grateful for your encouragement and support.

Closing off, I want to end with a short, but powerful quote that my mother said to me countless times. It has functioned as a mantra throughout my academic but above all my personal development and will continue to do so in the future:

"Als je wilt, kun je alles"

S.J. San José Sánchez

Delft, March 2024

Executive Summary

With the rise of machine learning applications in the healthcare domain, open health data can serve as a source of training data which is necessary to make the ML algorithm perform. These machine learning models are applied in various healthcare areas: from predictions for diagnosis and treatments to organizational challenges such as hospital occupancy. Open health data is a form of open data that can freely be accessed, used, modified and shared for any purpose, by anyone (Open Knowledge Foundation, n.d.). Open data redefines the perspective on data as an intellectual property of the data owner towards data as a common good (Huston et al., 2019). The addition of *health* to open data refers to the accessibility of information coming from the health domain, such as patient's health histories, clinical trial results, but also governmental health surveillance studies. Since health data is sensitive, it comes with many privacy challenges and is thus not easily accessible, open health data can fill this gap. To ensure that machine learning models trained with open health data provide reliable and accurate outcomes, open health data must be representative of the population it will be used for. Previous research has shown that this is not always the case. Existing open health datasets often contain representation bias due to social, economic and demographic factors, norms and institutions (Simon et al., 2020), which occurs during the data collection phase. Representation bias affects how open health data can be used as training data for machine learning models. A model that is trained on biased data will result in unreliable predictions, potentially leading to unfair and discriminatory practices. This causes harm to the groups that are misrepresented in the data and affects their health equity as there are unfair, avoidable differences among groups of people (World Health Organization, 2021) in how they are treated in healthcare.

Therefore, the objective of this research is to analyse how representation bias is present in open health data and how it can be addressed when open health datasets are created and/or used. This will contribute to filling the gap between the development of representative and diverse open health datasets as training data for machine learning models and their real-world, fair deployment in healthcare. For this, the following research question was formulated:

> *"How can representation bias in open health data be addressed for the training of medical machine learning models?"*

To answer this question, a qualitative research approach was used which consisted of literature reviews, exploratory (seven) and validatory (four) semi-structured interviews, and use cases. The literature review and exploratory interviews contributed to the development of the decision framework as it helped identify challenges, roles and responsibilities when working with open health data. Subsequently, the framework was applied to use cases of open health data collected from intensive care units (ICU). The use cases, as well as the validation interviews served as a base for redefining the final decision framework.

The decision framework, consisting of several guidelines, can assist providers, prosumers and consumers of open health data by incorporating measures for representation and diversity in

the dataset. It consists of six guidelines that can be applied to both existing and to-be-developed open health datasets. It emphasizes the need for diverse design teams and the inclusion of patients to stimulate the incorporation of different perspectives, needs and values. It highlights the importance of evaluating data collection methodologies in an external setting to stimulate representation bias identification. In addition, identifying contextual constraints requires active and reiterated thinking about the generalizability of the dataset. The decision framework stimulates appointing a contact person when publishing the data to improve communication and transparency instead of working with assumptions. Lastly, detailed metadata allows for a better-informed decision-making process for users of the open dataset as it provides insights into the content but also the limitations of the dataset. Applying the decision framework to the use cases showed that there is a varying approach in appointing a contact person in ICU open data and improvement is needed in contextual constraint identification and external validation. Metadata is often included, but the degree and depth of the metadata is variable. The involvement of patients and the composition of the design team are unclear at the time of writing this thesis.

This deliverable of this research is the decision framework that fills the existing literature gap on how representation bias in open health data should be approached to improve its use in real-world machine learning applications in healthcare. As open health data has the potential to be used on a widespread basis when compared to closed data, filling this gap was crucial. Using the decision framework results in informed decision-making for the usability and relevance of the data as it improves awareness of representational limitations and bias of the open dataset. This results in fair treatment of patients, protection of their health equity, and overall contributes to a better understanding of diseases, treatments and diagnoses which could improve health outcomes in our society.

Whereas the decision framework can significantly contribute to improving awareness on how to create and use open health datasets fairly, it also raises questions on how the implementation can be realized without obtaining too many resources in the already scarce environment of healthcare. In addition, the guidelines were designed based on existing open health datasets and exploratory interviews. Due to the early stage in which open health data currently finds itself, the availability and knowledge about open health datasets are relatively limited.

Future research directions were derived from the research findings and limitations and are concerned with exploring the effect of the full open health data chain on the usability of the framework. In addition, applying the decision framework in other healthcare domains such as oncology or paediatrics allows us to further examine the effectiveness of the framework. Lastly, more research is needed into the possibilities and challenges that rise when the decision framework is implemented in real-world situations.

Table of Contents

List of Figures

List of Tables

1

Introduction

1.1. Background

In the past decade, the use of data in the field of healthcare has prevalently increased due to the development of information technologies (Spector-Bagdady, 20222). This can be of great potential as it can reduce healthcare costs and provide the field of medicine with new insights (Ottenheijm, 2015). Data analysis is used in various areas of medicine, including biomedical research, health interventions, and health policymaking (Maastricht UMC+, n.d.). The data can be derived from medical imaging, sensor informatics through smart wearables, laboratory results, hereditary diseases, and can provide information about the characteristics of the patient including their age, gender and socio-economic positions (Martin-Sanchez & Verspoor, 2014).

With the introduction of machine learning (ML), a discipline of artificial intelligence, medical algorithms can be created to identify patterns in large amounts of health data (Alanazi, 2022). The created predictive models can inform caregivers about, for example, a patient's predicted diagnosis or a treatment's success rate. An example of a machine learning application is the prediction of developing diabetic retinopathy, a damaging eye condition, which can evolve in people with diabetes (Alanazi, 2022). As the performance of a machine learning model is dependent on how the model is trained, it is important to obtain a large amount of training data (Wiens & Shenoy, 2018). However, the obtainment of training data in healthcare, often consisting of patient data, is challenging as legal and ethical restrictions limit the use of the data outside of clinical research (Albahri et al., 2023). To overcome this, open health data can be of great value as it can fill in the lack of training data for medical machine learning models.

However, a major impediment is that open health data underrepresents or is biased towards specific groups. An example of this is the UK Biobank, to which participants can share their lifestyle and genetic data. The UK Biobank is not representative of their target group which is the population of the United Kingdom (Fry et al., 2017). This is due to selection bias where healthy individuals are more likely to participate and share their data than unhealthy individuals (Schoeler et al., 2023). If this dataset were used to train a machine learning model, this could lead to inaccurate outcomes where patients could receive a wrong

diagnosis or treatment. In addition, the model would not be beneficial to those who are underrepresented in the open health dataset.

Therefore, it is important to create, but also to use open health datasets that are representative of the population they were designed for. This also has an impact on the outcomes of medical machine learning models, as the trainings data that was used to train the models is representative of the population the machine learning model was created for. The scope of this research is therefore on how representation in open health datasets can be encouraged so that the training of machine learning models with open health datasets does not result in biased and unequitable outcomes for patients. This is critical for the successful deployment of machine learning models (Lee & Viswanath, 2020).

1.2. Main concepts and knowledge gap

This paragraph will introduce core concepts of the use of open data in healthcare. This leads to the identification of an academic knowledge gap for which the main research question of this research will be formulated.

1.2.1. Open health data

Open health data refers to free, accessible information in the healthcare domain. This entails various sources of information, examples are electronic health records (EHRs) with details of a patient's health history, clinical research results on medication studies and health system performances. In addition, governmental health surveillance studies can provide a rich volume of health data (Kostkova, 2016).

Open health data is a form of open data, the latter being defined as publicly available data, which is accessible without any restrictions, and one is free to use, reuse and redistribute the data (Molloy, 2011). The definition as given by the Open Knowledge Foundation (n.d.) defines open data as data that can freely be accessed, used, modified and shared for any purpose, by anyone. This is opposed to governmental data and big data, which are not accessible to the public, as seen in Figure 1. According to Huston et al. (2019), the switch from closed, private data to open data changes the perspective on datasets being the intellectual property of the creator to the dataset as a common good. As a result of opening up data, collaboration between researchers is stimulated and the increased amount of accessible data could result in new, innovative ideas. In addition, open data increases transparency and improvement of decision-making as policymakers can use information from the open datasets (Charalabidis et al., 2018).

Opening up health data is beneficial for our healthcare system as it allows other researchers to work with the data and come up with new and potentially innovative solutions (Hulsen, 2020). According to Kostkova (2016), a better understanding of symptoms and diseases can improve the treatment and the overall health outcomes for patients. Having a broader and detailed view on many levels of healthcare can also contribute to making healthcare equally beneficial, regardless of the economic position of patients or the regions and countries in

which they live. In addition, it is also beneficial for healthcare practitioners as new insights from the data can assist them with daily tasks, such as examining medical images, diagnosing patients and developing treatment plans to enhance the efficiency of their workflow.

The adoption of open health data is not without risks and challenges. Major challenges for open health data are the concerns for data security, quality and privacy. Open health data is freely available, meaning that there is no to little control over who uses the data. This may influence the willingness of data owners to open their data. The quality of open health data varies as the data comes from different sources; meaning that there are different interpretations and data types. Lastly, there are many privacy concerns when it comes to open health data. The privacy of patients is an important pillar as their data is sensitive and should not be shared with outsiders without special considerations. After all, there is doctor-patient confidentiality, where the information you share with your medical practitioner is private. This conflicts with open health data, where such information would be disclosed in a publicly available dataset. Although measures have been taken in the form of de-identifying the data and getting consent from the patient to share their data, it is still a large concern.

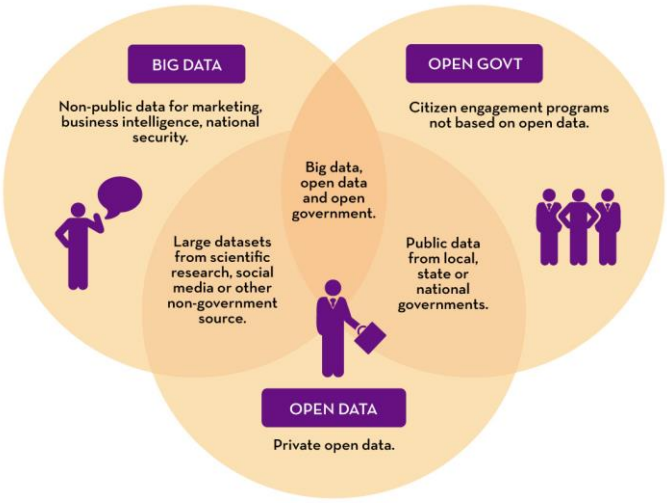

Figure 1 Open data (Gobierno de España, 2017)

1.2.2. Machine learning applications in healthcare

In our daily lives, machine learning has been integrated in many different ways. From Facebook suggesting tagging the friend that is in the photo you have just uploaded to automated detection of money laundering fraud in the banking sector: machine learning is here to stay (Yapo & Weiss, 2018). This also applies to the healthcare domain, as machine learning applications in healthcare can advance the current quality of care (Habehh & Gohel, 2021).

As mentioned before, large amounts of data are collected in healthcare in the form of EHRs, medical wearable devices and clinical research. This data is valuable and could provide many

insights for the healthcare domain. However, it is also voluminous, complex and of different types, making it difficult to process and analyse by hand. A solution to this is machine learning; the computational technique of identifying patterns in large volumes of data. The identified patterns can provide an understanding of phenomena and even predict future outcomes. When translated to healthcare applications, medical machine learning models allow for predictive disease analytics, assessment of medical imaging and patient diagnosing (Javaid et. al., 2022).

Machine learning models require large amounts of data to be trained on before they can be deployed for their intended purpose, such as the concepts listed above. Machine learning aims at finding non-linear relationships between numerous variables and this will work best when it has many data examples to train on (Wiens & Shenoy, 2018). Patient data can thus serve as training data for medical machine learning models. However, the attitude towards sharing patient data varies. This is due to the sensitive nature of patient data, making patients unwilling to give consent for sharing it, as well as being subject to privacy laws. In addition, it is challenging to obtain high-quality data for training machine learning models as legal and ethical restrictions often do not allow clinical research data to be used outside the scope of the research (Albahri et al., 2023). Therefore, it is difficult for researchers, scientists and developers to obtain the required patient data (Khan et al., 2020). He et al. (2019) state that data sharing efforts can contribute to the widespread adoption of machine learning technologies, for which open data is a suitable solution. This highlights the importance of open health data as it can improve the availability and accessibility of patient data.

1.2.3. Bias and its implications in machine learning models

Although machine learning provides the healthcare domain with great opportunities, it also results in ethical concerns. A significant ethical risk of using machine learning models is bias, which is defined according to the Cambridge Dictionary as *'a prejudice for or against one person or group, in a way considered to be unfair'*. This can be viewed from a technical perspective in which the machine learning model contains systemic errors, but from an ethical perspective, bias results in unfair, unreliable and unjust outcomes for patients. (Martinez-Martin & Cho, 2022). It thus affects the reliability of the outcomes by a machine learning model, as these outcomes are skewed towards specific individuals or population groups. (Mehrabi, 2021).

Bias in medical machine learning can lead to unfair and incorrect outcomes and can even cause disparities in the treatment of patients. Bias in healthcare is already a prevalent problem, which can result in the reinforcement of already existing stereotypes in the world (Starke et al., 2021). In addition, McCradden et al. (2022) describe the risk of bias in healthcare when existing social inequalities affect the predictions made by the machine learning model, resulting in further harm to the disadvantaged patients.

Bias can occur during different development stages of machine learning models. In the first step of building a machine learning model, the data must be collected. In practice, model

developers often use existing datasets instead of collecting the data themselves as this is a long process of identifying and sampling the target group and the corresponding features (Suresh & Guttag, 2021). As mentioned previously, the lack of available health data can also be a reason to use existing and publicly available data. It is important to mention that the scope of this research is directed to open health datasets that are collected by a collaborative approach instead of a single person. Figure 2 shows the areas during the data collection phase in which bias can arise and how these build into the development of the machine learning model.

From this figure, it becomes clear that it is important to have representative training data as it will affect the output of the model. This is also emphasized by the European Commission (2020) who state that "biased datasets can create biased algorithms" (para. 1). Bias can occur in both open and closed datasets and the potential bias types are thus similar. However, since open datasets may be reused for many instance as they can provide a substantial source of training data, it is especially important to address bias to prevent the recurrent use of biased datasets. When biased open datasets are used for training the medical machine learning model, this will result in a predictive algorithm with unreliable predictions when deployed in a real-world situation.

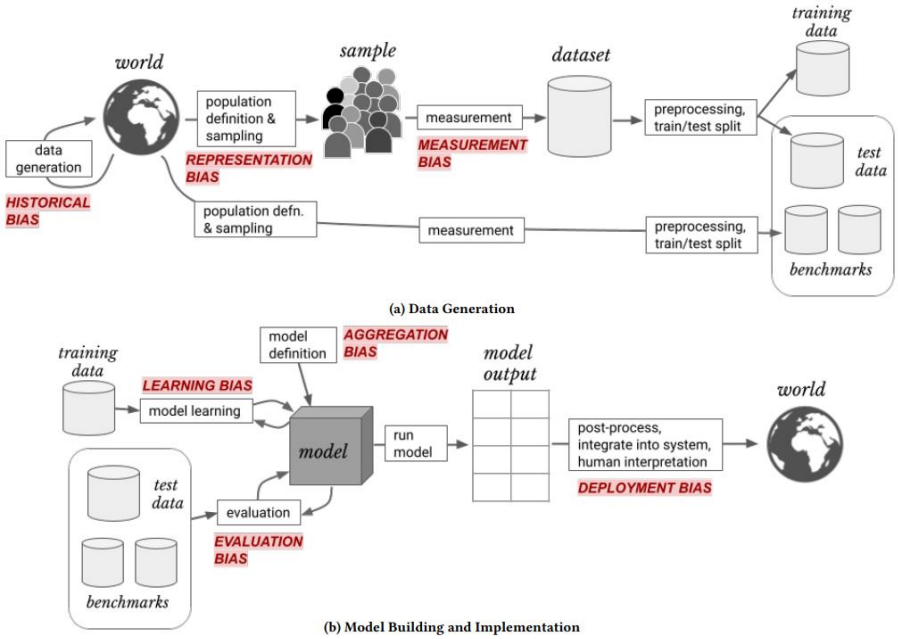

Figure 2 Bias when building a machine learning model (Suresh & Guttag, 2021)

The origin of bias in datasets can have many different causes. For example, a potential form of bias in a dataset that contains EHRs, occurs through the availability of electronic health records (Chen et al, 2021). Anonymized electronic health records (EHRs) are a form of open data that could be used as training data for machine learning models as they contain information about patients' health history. The existence and availability of a patient's EHR illustrates that the patient had access to healthcare in a country or region where EHRs are being used. This is opposed to countries or regions that do not make use of EHRs, which are

often less-developed countries. Eventually, this can cause bias in the developed machine learning model when information in EHRs is used as training data: only people with access to 'modern and digital' healthcare are included in the dataset. In addition, bias in open health data can occur when social media or other methods that require internet access are used for data collection, for example through surveys about patients' health and well-being (Chen et al, 2021). This is supported by Veinot, Mitchell & Ancker (2017) who state that internet-requiring techniques are used more by people with a higher level of education and income, resulting in inequalities between populations that differ in socio-economic backgrounds. Moreover, Chen et al. (2021) and Brewer et al. (2020) argue that bias may occur as specific population data is simply non-existent e.g. due to resource constraints of low-income countries or populations. Even when data of these populations is available, it is often incomplete and noisy, leading to inappropriate training datasets for machine learning models as the predictions are not reliable and can not be trusted (Lee & Viswanath, 2020).

The use of biased training data for training of machine learning models can have serious impediments. The model will be trained on biased data, leading to unreliable prediction outcomes when the model is tested or even deployed (Lee & Viswanath, 2020). These unreliable predictions may harm or are not beneficial to underrepresented groups in the open data. An example is provided by Craig et al. (2022), where a machine learning model was used to schedule hospital appointments and report the appointments to patients through a web-portal. Patients who did not show up to previous appointments according to their EHR, were scheduled in an overbooked timeslot. This resulted in overbooking of patient populations who might not have the resources to access the hospital's web-portal as other groups have, leading to an unintended, discriminatory practice.

The World Health Organization describes equity as "the absence of unfair, avoidable or remediable differences among groups of people, whether those groups are defined socially, economically, demographically or by other dimensions of inequality" and as health is a human right, equity should also be safeguarded in healthcare (World Health Organization, 2021, para. 1). Discriminatory practices in healthcare can therefore influence the health equity of underrepresented or vulnerable groups and should urgently be addressed to prevent this in future practices.

1.2.4. Research scope: representation bias in open health data

Figure 2 shows the types of bias that occur in the data generation stage. The focus of this research is specifically on representation bias, which occurs "when the training data under-represents (and subsequently fails to generalize well) some parts of the target population" (Shahbazi et al., 2023, p. 2). For instance, representation bias can occur when only a small number of people from a low-populated region are included in a dataset. A model that is trained on this dataset will generally perform badly for people living in the low-populated region as the model does not have sufficient instances to learn from and thus cannot provide reliable and accurate predictions. Another example of representation bias is presented in Pozzi's (2023) research, where patients' risk scores for opioid addiction are predicted by an

algorithm. One of the variables that the algorithm made its predictions on, was the travel distance from the patient's home to the opioid-selling location, e.g. the pharmacy. It assumed that those who live far from the pharmacy and thus travelled a longer distance to get their medications, were more likely to be 'drug shopping' and were given a higher risk score for opioid addiction. Pozzi (2023) described that the algorithm did not take into account any other factors for why the patient had to travel a longer distance, for which an explanation could be that they live in a rural area due to forced lower costs of living. People living in rural areas were misrepresented in the dataset. The predictions for people living in rural areas could be inaccurate and are thus not reliable. This shows that representation bias causes the training data to be biased which eventually leads to a biased algorithm (Danks & London, 2017).

Representation bias in open health data can be a result of unconscious and implicit associations and attitudes that all people have, as remarked by Marcelin et al. (2019). This also applies the people who create open health datasets. In addition, they describe that this can lead to exacerbating the existing disparities, which is especially the case when policy and decision-making are based on predictions by a machine learning model. Representation bias can have various causes but as the methodologies for data collection of open datasets are often hidden or unknown; it is difficult to determine its exact origin in a dataset. Since open health datasets can be used on a wide scale, representation bias in the datasets must be addressed as inaccurate outcomes can be a result when the biases datasets are yet being used for training medical ML models (Kumar et al., 2023).

Representation bias in open health datasets may, as previously remarked, occur because of implicit and either conscious or unconscious associations. These associations do not come out of nowhere, they have a base in historical and systemic prejudices and have thus been pre-existing (Simon et al., 2020). These prejudices are often based on pre-existing, social and cultural norms and institutions. This is in line with the bias taxonomy as given by Friedman & Nissenbaum (1996), who state that pre-existing bias comes from before the computer system was developed, where in this case the system refers to open datasets being used in healthcare. These prejudices find their origin in structural discrimination that specific population groups experience in healthcare, which affects their relationship with healthcare to this day. (Webster et al., 2022). Consequently, they feel less comfortable to seek medical care but this does not mean that they are not in need of receiving medical attention. Another prejudice lies in using the male body as the norm for medical research. This affects the effectivity and generalizability of diagnosing and treatment in women, since diseases express differently in males and females (ZonMw, 2022).

In the past few years, the problem of representation bias in datasets has been examined thoroughly as it strongly affects the generalization of machine learning models, making them inaccurate and irresponsible to implement in real-world environments. By choosing and limiting the scope of representation bias, a deeper understanding will be obtained of how this type of bias in open health datasets can be addressed during the data collection stage and by the users of open health data. This will help to fight existing health disparities and reinforcing

systemic and discriminating practices in healthcare. It is important to mention that as of this point, 'representation bias' is interchangeably used with 'bias' in this thesis.

1.3. Knowledge gap

Although machine learning models are increasingly being developed in the healthcare domain, the availability and accessibility of health data to train these models is relatively low due to legal and ethical considerations, as well as patients' attitudes towards data sharing. Although open health data can contribute to solving this challenge, they also come with impediments. Open data has the risk of being biased towards specific groups, such as minority or underrepresented populations.

The importance of having representative datasets is known, it is even addressed by the European Commission and amplified by Spector-Bagdady et al. (2022) who argue that bias should be addressed to prevent biased machine learning models. However, less is known about how to exactly address representation bias to encourage representative collection and use of open health data, which can stimulate responsible and fair deployment of machine learning models in healthcare. This illustrates the need to bridge the gap between addressing representation bias in open health datasets for training medical machine learning models, and the fair and reliable real-world deployment of these models in healthcare. Presumably, no previous research has been found that touches upon addressing representation bias in open health datasets.

1.4. Societal and scientific relevance

The societal relevance of this research is significant: using open health data for machine learning models has the potential to contribute to healthcare that is efficient, precise, and personalized, but when predictive models are trained on open datasets that do not represent the whole population, the harmful consequences of these models are significant as they can lead to the wrong medical interventions and knowledge for the underrepresented groups (Ibrahim et al., 2021). Healthcare should be safe and reliable, hence the integration of machine learning in healthcare must be done securely and safely so it does not discriminate against specific population groups. Addressing representation bias in open health data is therefore a strong requirement to take the understanding of diseases and treatments to a higher level, making it both effective and equitable for all patients involved.

In addition, the scientific relevance of this research lies in enhancing the current understanding of how open health datasets can created and used in a representative and diverse manner. With these insights, representatives open health datasets can be developed and used that contribute to new, innovative research in the healthcare domain. Prior research has stretched the importance of diverse datasets and has called for reconsidering data collection approaches, to which this research will contribute.

1.5. Alignment with CoSEM

This research aligns with the *Complex System Engineering and Management* program as it concerns a multidisciplinary area where technical, institutional and social aspects are at play. It aims to design a solution for the problem of biased open datasets within the healthcare domain: a field with many interests, many concerns and a complicated web of stakeholders. The needs of various stakeholders including patients, healthcare providers and technical developers must be balanced. To address such a problem in a complex socio-technical system, creative employment of methods is required to not only design but also organise and manage the designed solution. The problem of having biased open health datasets affects both public and private values, including high-quality healthcare, well-being and non-discrimination towards individuals and groups.

1.6. Structure of the report

The structure of this thesis will be discussed to guide the reader through the content. In Chapter 1, the research topic was introduced and relevant background information for a coherent understanding of the research area was discussed. Chapter 2 will present the research design consisting of the chosen research approach and methodologies. In addition, the main research question and the sub-questions will be presented as well as a schematic overview of the Research Flow Diagram. Chapter 3 provides the literature review whereas Chapter 4 will present the findings of the explorative interviews, both chapters focus on the bias types in open health datasets and their ethical implications, as well as the social context of open health data. This serves as a knowledge base for the decision framework that is presented in Chapter 5. In Chapter Error! Reference source not found., the decision framework is applied to open health data use cases and Chapter 7 will provide an expert validation of the decision framework, both contributing to a refined version of the decision framework that will be presented at the end of this chapter. Chapter 8 will discuss the findings of the research, as well as identified limitations. Finally, Chapter 9 will present the conclusion of this research, elaborate on the societal and scientific contribution as a result of this research and lastly, future research trajectories are presented.

2

Research design

This chapter elaborates on the research approach and the chosen methods to answer the main and sub-questions.

The knowledge gap highlights that the importance of representative open datasets is known, in particular in the field of healthcare as the presence of bias can impact patients' health equity. Nevertheless, this recognition is not found in real-world applications as open health data contains bias, making them inappropriate to use for training of machine learning models as this would lead to unfair and unreliable predictions when the model would be deployed in a real-world situation. As the use of machine learning models in healthcare offers various oppurtinities, it is important to understand how representation bias in open health data should be addressed. The deliverable of this research will contribute to assessing and stimulating representation of patients in open health data.

2.1. Research approach

There is a current void in the functioning of the system: open health data can serve as training data for medical machine learning models if they are representative and diverse. This calls for a solution design that will address representation bias in open health data. This will be done by designing a decision framework that helps creators of open health data with designing representative datasets. Simultaneously, it assists users of open health data with assessing the usability and relevance of the dataset to ensure that the data is used in an appropriate context to prevent future biased outcomes.

The problem identification has been presented in Chapter 1. The objectives for the deliverable will be identified by obtaining knowledge on the types of representational bias in open health data, their ethical implications and the social context in which open health datasets are designed. This will ensure that the decision framework consists of appropriate guidelines that target both the creators and users of the data. The deliverable of this research is a decision framework consisting of guidelines that address representative collection and use of open health data.

2.2. Methodology

This research will be conducted by using qualitative research methods as seen in Figure 3. Qualitative research is a suitable method as it allows the discovery of urgent issues in a relatively new field of study (Jamshed, 2014), which is suitable for the upcoming use of open health data. Firstly, it is important to build a foundation of existing knowledge by conducting a literature review study. This serves as a base for the additional research methods, allowing comparisons, hypothesis-testing and challenging existing literature. Additionally, interviews are conducted with experts in the field of open (health) data, medical machine learning and ethical challenges in health and technology. Lastly, use cases on open health data will be examined.

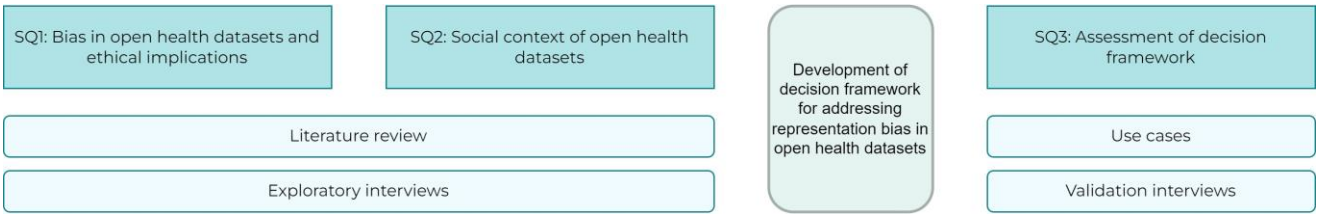

Figure 3 Overview of research methodologies

Literature review

Firstly, a literature review will be conducted into existing literature on bias in open health data. Obtaining a base of objective and scientific knowledge from existing literature will contribute to delineating the theoretical foundation and the context of biased open health data. It also provides an insight into the social context of open health data, the relevant and involved actors, and how the interrelationships occur between the actors. In essence, performing the literature review will contribute to examining the state-of-the-art, as well as contributing to the novelty of the research by identifying shortcomings and opportunities (Knopf, 2006).

Explorative and validation interviews

The interviews will be held with participants in the field of open health data, medical machine learning and health ethics, to gain an understanding of how representation bias is currently addressed in (open) health datasets, and to examine their methodologies and conditions for creating datasets that are diverse and representative. The interviews will be semi-structured as this allows for guidance towards research areas of interest by using the preset questions, but it also leaves space for additional input from the interview participant (Jamshed, 2014). The TU Delft's Human Research Ethics Committee has approved the interview design. It must be noted that interviews take up a lot of time for preparation, execution and evaluation. For this reason, a sharp schedule and approaching interviewees from early on was essential.

The setup of the interviews is two-folded: in the first stage of the research, explorative interviews are held to gain a deeper understanding of the current state-of-the-art, as well as opportunities and challenges when it comes to open health data. After developing the decision framework, validation interviews are conducted to assess the usability and effectivity of the framework from the experts' viewpoints. The contents of the interviews are thematically

analysed by reviewing the transcript summaries and identifying overarching, similar topics between the interview participants such as solution directions, suggestions for improvement and limitations in the solution area. Thematic analysis is a suitable method for analysing interviews as it "allows for flexibility and interpretation when analyzing the data" (Castleberry & Nolen, 2018, p. 808) which is especially important in the evolving field of open health data. This flexibility is seen in the overarching themes that are derived from the interviews instead of being preset and predefined.

Use cases

After performing a literature review and conducting exploratory interviews, the decision framework will be designed. To examine how the decision framework will be deployed and assess its usability, a number of use cases will be explored. The use cases refer to health datasets that have already been opened to the public. The analysis of the use cases will consist of, firstly, applying the decision framework to multiple instances of open health datasets. Subsequently, the performance and potential areas of improvement of the decision framework can be identified. Use cases are often applied in software development to design requirements for the system (Ratcliffe & Budgen, 2005). Overall, use cases aim to explain how a system is used (Hunt, 1999). In the event of open health data, the use cases will demonstrate how the decision framework must be applied to address representation bias in the dataset. The choice for use cases as a research method is made as it can result in requirements, which in this instance will be seen as recommendations, for the validated framework.

2.3. Main research question

The objective of this research is to provide a decision framework to address representation bias in open health data. By doing so, the integration of machine learning in healthcare can be realised more fairly and responsibly. The main research question is therefore:

> *"How can representation bias in open health data be addressed for the training of medical machine learning models?"*

The research approach as discussed previously addresses a suitable design to carry out the research. The formulation of sub-questions in the next paragraph will allow a step-by-step approach to answering the main research question.

2.4. Research sub-questions

To provide a decision framework on how to address representation bias, it is firstly important to know what types of representation bias occur and what their ethical implications are. Therefore, a descriptive sub-question has been formulated:

> *SQ1: What are the different types of representation bias that occur in open health data used for medical machine learning models and what are their ethical implications for those at risk?*

This question will be approached by performing literature reviews on bias types in open health datasets, as well as on the ethical implications of biased open health datasets for individuals and groups at risk. To do so, reliable and valued academic research databases will be consulted. In addition, interviews will be conducted with a variety of experts, such as open (health) data experts, ethicists and medical machine learning developers, to obtain knowledge about representation biases that have occurred in open health data and to gain an insight into how ethics play a role in the development of medical machine learning models. This will result in a theoretical base that identifies bias types and their ethical implications, thus answering the sub-question.

As the application of open health datasets is still in its infancy, it is important to consider the existing social environment that influences the creation and application of open health data. This results in the following sub-question:

> *SQ2: How does the existing social environment influence how open health datasets are created and deployed for real-world application in healthcare?*

The answer to sub-question two relies on literature reviews as well as on interviews with experts in the field of (health) open data. Experts can also be from policy-making fields or researchers from healthcare and machine learning backgrounds. Examining the social context of open health data will result in a broader view on potential interventions, which can be related to for example standardization and interrelations between the actors that are involved with open health data.

After the two sub-questions have been answered, a decision framework will be developed to address representation bias during collection and use of open health data. The framework is based on the examined bias types, their ethical implications and the social context of open health data. The decision framework consists of guidelines and recommendations for representative collection and use of open health datasets. The effectivity and usability of the framework must also be evaluated, for which the third sub-question is formulated:

> *SQ3: How can the developed decision framework contribute to representative data collection and use of open health datasets to protect patients' health equity?*

The decision framework is evaluated by applying it to use cases on open health datasets that were created in intensive care environments. Applying the framework sheds light on the usability and effectivity of the framework and contributes to identifying limitations of the framework. In addition, validation interviews with relevant experts will be conducted to evaluate the decision framework. This will provide a direction on how representation bias can be addressed in open health datasets, which is essential for answering the main research question.

2.5. Research flow diagram

A schematic overview of this research is presented in Figure 4 below in the form of a research flow diagram (RFD).

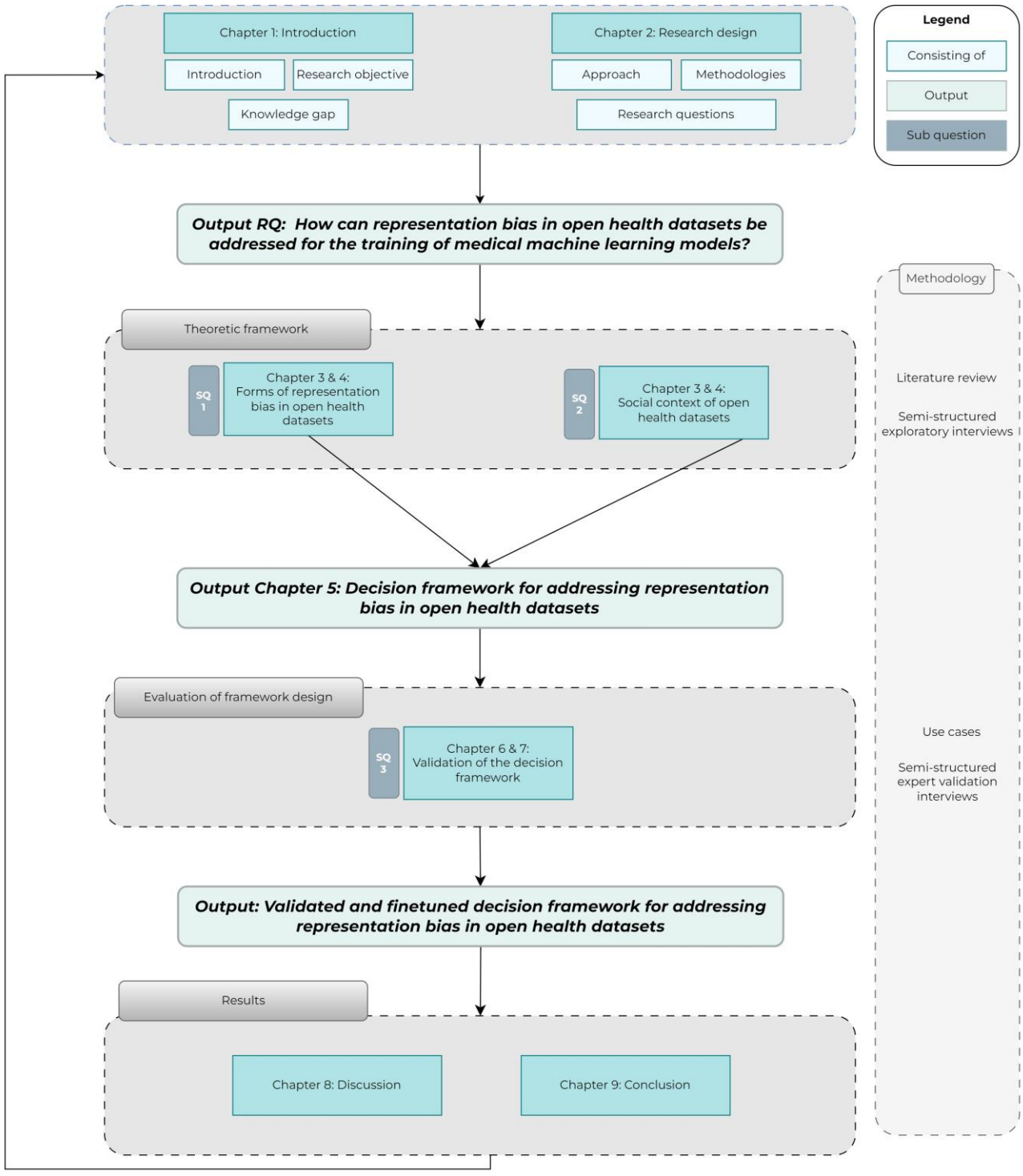

Figure 4 Research Flow Diagram

3

Literature review

This chapter will provide the conducted literature review and explorative interviews for the sub-questions 1 and 2.

3.1. SQ1: Bias in open health data and ethical implications

Firstly, it is important to identify the different types of bias that occur in open health datasets to answer the sub-question: *What are the different types of representation bias that occur in open health data used for medical machine learning models and what are their ethical implications for those at risk?* As there are many different forms of existing bias in datasets, it is essential to examine which ones are relevant for using open data in a healthcare environment. At this stage, it is important to mention that both closed and open datasets can have the same type of bias; but might differ in their occurrence, as the difference between close and open data is the accessibility of the data for the general public. When the types of bias are known, it provides a base for determining the ethical implications the type can have on a particular individual or group. This will also contribute to the use cases and interviews, as this can open up the design space of the interview questions in more specific directions.

3.1.1. Search strategies and selection of literature

Conducting a literature review will provide a structured manner to examine and evaluate relevant literature on existing biases in health (open) datasets.

The literature databases of both PubMed and Scopus were used. The keywords consisted of “data” and “healthcare” and “bias” and “machine learning” and their abbreviations or synonyms, which are shown in Table 1. The goal of the literature search was to find articles and reviews that presented often occurring bias types in datasets used for training machine learning or examined the ethical implications of bias in datasets. This research question consists of two parts: the types of bias in open health datasets, and what the ethical implications of the biases are. Firstly, an initial literature search for common types of bias was done followed by additional techniques, including snowballing and citation searching, to identify ethical implications of such bias types.

Inclusion and exclusion criteria

As open data and specifically open health data are a recent concept, literature was included if it was published in the last five years (2018 – 2023). Solely peer-reviewed articles were included. In addition, the literature must entail the application of (datasets for) machine learning for human healthcare. Some of the collected literature was excluded based on formulated exclusion criteria. This entails literature behind a paywall and papers that are in other languages than English or Dutch.

Table 1 Search terms and synonyms

Search term	Synonyms
Healthcare	Medicine, health sector, health domain, healthcare domain
Open data	Public data, public dataset
Open health data	Medical open data, public health data, published data
Bias	Bias types, bias forms, representation, diversity, representativity,

Figure 5 shows the PRISMA Flow Diagram of this literature search strategy. Subsequently, Table 2 shows an overview of the included literature and its relevancy.

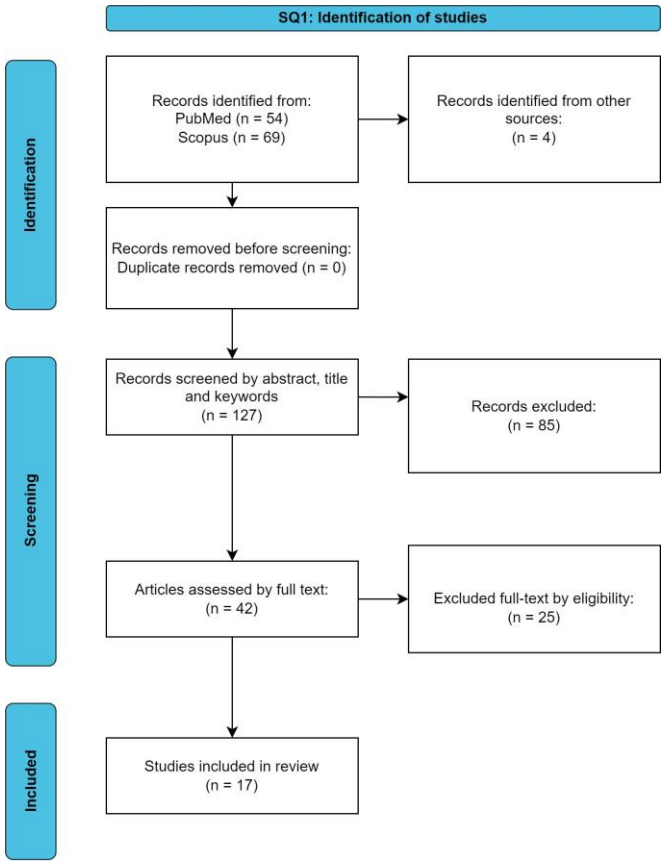

Figure 5 PRISMA Flow Diagram of sub-question 1

Table 2 Overview of included literature for sub-question 1

	Author(s)	Article	Bias	Aim
1	I. Straw & H. Wu	Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction (2022)	Gender	Analysing open dataset for prediction of liver diseases
2	Larrazabal et al. (2020)	Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis (2020)	Gender	Analysing open dataset for prediction of thoracic diseases
3	Norori et al. (2021)	Addressing bias in big data and AI for health care: A call for open science	Gender	Ethical consequences of gender bias
4	Gichoya et al. (2021)	Equity in essence: a call for operationalising fairness in machine learning for healthcare	Gender	Call for reporting gender bias
5	Meng et al. (2022)	Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset	Racial	Evaluating fairness of various prediction models
6	Cerrato et al. (2022)	A proposal for developing a platform that evaluates algorithmic equity and accuracy	Racial, socio-economic, team	Strategies for mitigating bias by reporting
7	Obermeyer et al. (2019)	Dissecting racial bias in an algorithm used to manage the health of populations	Racial	Illustrating the effect of racial bias on health outcomes
8	Polevikov (2023)	Advancing AI in healthcare: A comprehensive review of best practices	Racial, socio-economic	Ethical implications of racial bias and call for improved data collection policies
9	Dehkharghanian et al. (2023)	Biased data, biased AI: deep networks predict the acquisition site of TCGA images	Location	Illustrate the effect of bias in datasets on models

	Author(s)	Article	Bias	Aim
10	Gichoya et al. (2023)	AI pitfalls and what not to do: Mitigating bias in AI.	Location, team	Bias in health AI and mitigating strategies
11	Celi et al. (2022)	Sources of bias in artificial intelligence that perpetuate healthcare disparities: A global review	Location, team	Elaborate on the lack of diverse datasets
12	Burström & Tao (2023)	Social determinants of health and inequalities in COVID-19	Socio-economic	Illustrate the effect of socio-determinants on health outcomes
13	Kaplan & Keil (1993)	Socioeconomic factors and cardiovascular disease: a review of the literature.	Socio-economic	Ethical implications of socio-economic factors
14	Chicco et al. (2022)	A survey on publicly available open datasets derived from electronic health records (EHRs) of patients with neuroblastoma.	Socio-economic	Insight into public open datasets based on EHRs
15	Johnson et al. (2023)	MIMIC-IV, a freely accessible electronic health record dataset	Socio-economic	Insight into public open datasets based on EHRs
16	Knevel & Liao (2022)	From real-world electronic health record data to real-world results using artificial intelligence	Socio-economic	Implications of using EHRs as input data
17	MacIntyre et al. (2023)	Ethical considerations for the use of artificial intelligence in medical decision-making capacity assessments	Socio-economic	Illustrate the effect of bias on health AI

3.1.2. Results from the search

From the included literature, several bias types were identified in open health datasets used for creating medical machine learning models. The types will be discussed in this paragraph, as well as their ethical implications and mitigation techniques as proposed by the authors.

Focus on gender

Previous research has shown that there are instances of gender bias in open datasets. A potential ethical implication of gender bias in open health datasets is that minority genders

will be treated unfavourably. Existing stereotypes of this gender will be maintained and can even be reinforced when open health datasets are used for training medical machine learning models. To illustrate, Norori et al. (2021) argue that women are often misdiagnosed when having a heart attack because of their misinterpreted symptoms. Well-known symptoms of heart attacks are different for men and women, so if an algorithm were to be trained on an imbalanced dataset towards men, the predictions of the deployed model would be inaccurate for women. This will reinforce the existing environment in which women are misdiagnosed, leading to unfair treatment and causing them more harm.

A practical example of gender bias occurred in the Indian Liver Dataset (Straw & Wu, 2022). This dataset consists of Indian patient records of which some have a form of liver disease (for example, cirrhosis and fatty liver disease). The authors concluded that this dataset is imbalanced, as there are significantly more males than females included. The result of this is a possible high number of false negatives for females. False negatives occur when a model incorrectly predicts the negative class: the model predicts the female to not be at risk for liver disease while in reality, she is. Females are thus more likely to have a missed diagnosis when machine learning models are trained on the Indian Liver Patient Dataset when gender bias is not accounted for. This has a large impact on the well-being of females as it affects their physical and psychological state. Another example of gender bias was identified in an open health dataset containing X-ray images to assist with diagnosing thoracic diseases. Larrazabal et al. (2020) trained a machine learning model on this open dataset and concluded that it performs worse if the dataset is imbalanced. In those cases, the algorithm disadvantages the underrepresented group in the data, whereas a balanced dataset showed equal performance for both women and men. This shows that having a diverse open health dataset adds to the performance of the mode. Larrazabal et al.'s (2020) study calls for incorporating diversity measures when designing new open health datasets that go beyond the current existing institutions such as the United States Food and Drug Administration (FDA) which do not determine gender as an attribute of the research population but on the other hand do stress the importance of including gender in the release of new medical devices.

The impact of gender bias and the ethical consequences for minority genders are thoroughly examined. It is thus not surprising that the urgency of addressing gender bias is well known. Although different institutions concerned with digital health technology in the European Union, the US and the UK call for action on addressing and evaluating bias, their proposed actions do not provide sufficiently detailed information about how they can be addressed in a practical manner (Gichoya et al., 2023). This shows an area of improvement for proposing tools that impact how gender bias is approached for open health datasets.

Racial and ethnical prejudices

Another form of bias that occurs in open health datasets is racial bias. Racial bias is a form of systemic bias where the dataset contains (un)conscious prejudices towards specific races or ethnicities, resulting in unwanted consequences.

Open health datasets are also at risk for containing bias towards specific racial groups. To illustrate, this has occurred in the MIMIC-IV open dataset. The dataset consisted of multiple variables, but also protected variables; of which one was the ethnic background of the patient. Meng et al. (2022) note that various predictive models created while using the MIMIC-IV dataset were proven to discriminate between white and black patients. The models used racial variables to generate their predictions but did this unfairly and unequally. For example, black patients would have a shorter duration of ventilation treatment as opposed to white patients. Cerrato et al. (2022) have documented racial bias in predictive models trained by open heath datasets as well. The model was created to assist the US healthcare system with determining which patients are in urgent need of (extra) medical attention and thus require more medical resources. As the resources are limited, the algorithm provided valuable insights into which patients at risk would benefit the most from receiving additional care. The model's outcome was that black and white patients are given the same risk level; although in reality, black patients are sicker when receiving that risk level than white patients are.

In addition to the cases listed above where a clear distinction is made between e.g. white, black and Hispanic patients, there are also instances of indirect racial bias through the use of proxy attributes. Such proxies occur when patient attributes are collected in a dataset of a non-racial nature, however, these attributes indirectly indicate the race of the patient (Cerrato et al., 2022). This is illustrated with an example of a study by Obermeyer et al., (2019) which shows the underlying contribution of proxies to unfair outcomes based on racial biased assumptions. In Obermeyer et al.'s case, the model considered yearly health costs to be a predictor for healthcare attention. Overall, the yearly costs for black patients were lower which would indicate that they do not need extra healthcare attention as they were healthier according to the model. It had turned out that this relation was not true, it was of the utmost consequence that black patients had less access to healthcare and therefore had lower yearly healthcare costs. Since the model did not pick this up, black patients were attributed less extra healthcare which decreased their access to healthcare even more.

The ethical implications of existing racial bias in open health datasets are major. It can amplify current health inequalities, leaving minority groups at a disadvantage. (Polevikov, 2023). The biased outcomes of the model limit the generalizability of the model. This again puts minority groups also at a disadvantage, as the technology is not accessible to them. Obermeyer et al. (2019) state that although having insight into the exact functioning of an algorithm would be beneficial for identifying biases, getting actual access to an algorithm is often difficult or even restricted. A valuable alternative for this would be to have access to the training data of the model as it offers an understanding of how potential biases can arise from the data, and thus finally in the model. This is where open data comes into play as it can offer a public and detailed overview of the training data.

Advantage by location

In addition to gender and racial bias in open health datasets, there is a third type that occurs in open health datasets: geographical bias. This entails that the data is collected in a specific geographic area, resulting in similar characteristics.

When looking at geographical bias in open health datasets, The Cancer Genome Atlas, a public dataset, contains images and clinical reports of cancer subtypes and can be used to assess scans of patients that presumably have a type of cancer. The data was collected by combining data from multiple hospitals within the same area. However, collecting this batch of data can result in geographical bias (Dehkharghanian et al., 2023). This is similar to the UK Biobank that published a dataset about the health of the population of the United Kingdom. In this dataset, only 6% of the entries are non-EU, making algorithms created from this dataset unlikely to be suitable when the target group is outside of the EU (Gichoya et al., 2023). The existence of geographical bias in datasets has been an issue for the use of AI and ML in general, not only for open health datasets. Celi et al. (2022) concluded that AI models are often trained on datasets from either China or the United States. The underlying reason for this is that well-developed countries have the knowledge and resources to realize machine learning initiatives in the healthcare domain.

Therefore, an open health dataset containing geographical bias can have serious ethical implications for the patients involved, as it is desirable to apply machine learning models on a wider scale. Geographical bias can limit this as the model is trained on data coming from a specific country, region or area. The generalizability of the model will be lacking, as an algorithm performing well on U.S. data will not necessarily perform well on Chinese patients and vice versa. This will exacerbate the disadvantage of minority groups, especially in the so-called 'data poor' regions where a low amount of data is collected and used for beneficial purposes, resulting in even larger healthcare inequalities.

Socio-economic determinants

The socio-economic factors that play a role in the daily lives of patients also affect their medical activities. Socio-economic determinants include, but are not limited to, education level, income and job type (Kaplan & Keil, 1993). To illustrate, Burström and Tao (2020) provide insight into the effect of social determinants for COVID-19 infections by showing that people who do not have access to a car and must use public transportation are more likely to be infected. The health literacy of patients has also been shown to influence their use of healthcare services.

Several open health datasets are derived from electronic health records (EHR's). As EHR's are a rather modern development in healthcare, there are many countries that do not have this resources in place and the population of these (often developing) countries will thus not be accounted for when EHR's would be used as input data. However, there are many examples of open datasets derived from EHR's, examples are those of patients with neuroblastoma (Chicco et al., 2022) and of patients admitted to the intensive care unit (Johnson et al., 2023). EHR's provide a rich source of real-world patient data, as they consist of the total medical path

of patients: from visits or calls to their general practitioners, medicine prescriptions and hospitalizations. A key aspect highlighted by Knevel and Liao (2022) is that the density of the information provided in a patient's EHR is dependent on their care-seeking behaviour; which is affected by socio-economic determinants. By using EHR's as input data for a medical machine learning model, the model would likely underrepresent the patients that are averse to seeking medical attention. The density of a patient's EHR can be influenced by a patient's insurance status or ability to finance their medical bills. Patients without or with limited healthcare insurance are less likely to seek care, similarly to patients with low-income. The risk of using EHRs as training data has a risk to be skewed towards the patients who do visit their caregivers, which are often from middle or high socio-economic backgrounds, while patients from lower economic backgrounds will be underrepresented (Cerrato, 2022).

MacIntyre et al. (2023) remarks an additional socio-economic factor: health literacy of patients, which addresses how well the patient is able to take necessary measures to improve their personal health by finding, asking for and understanding relevant information. Patients with low health literacy find many obstacles in seeking medical help, but should not be overlooked when designing open health datasets to ensure diversity of health literacy levels.

As socio-economic determinants play a significant role in people's care-seeking behaviour, those who are of lower backgrounds are less likely to ask for medical help. Misrepresentation of this group in open health datasets can lead to designed algorithms that will provide inaccurate outcomes for this group, causing disadvantages and additional harm to their health equity. Therefore, equal and fair use of machine learning technologies in healthcare should not disadvantage this group. Open health datasets must therefore reflect all patients, not solely those who can afford or seek medical help (Poleviov, 2023).

Although the previously discussed sources of representation bias (gender, racial, geographic and socio-economic) occur through patient aspects, there is an important consideration to make when it comes to the researchers or design group of open health datasets and medical machine learning models and how this may contribute to the biases as identified above. Gichoya et al. (2023) have highlighted the lack of diversity in the research and development team of medical machine learning models and how this can lead to underlying representation biases. Ideally, the development team should originate from different domains and should work simultaneously on whole pipeline of ML model development, from data generation until implementation. Additionally, ethicists and patient representatives should be included during the development of the model. Thus, it is important that not only the open health datasets are diverse; but the teams developing them as well. Representativity in the development team is not only focused on expertise, it also concerns the characteristics of individual developers on the team. Notably, Celi et al. (2022), emphasize that members of the research team are three times more likely to be male than female. In addition, researchers working with medical ML algorithms often come from high-income countries which may leave lower-income countries out of their perspectives.

A non-diverse development team for either open health datasets or ML models may not always fully understand who might be affected by underrepresentation, whereas team members representing minority groups are more likely to incorporate these perspectives (Cerrato, 2022). Having a development team that represents all segments of society is therefore desirable as it could contribute to an open health datasets that is more diverse.

3.1.3. Conclusion

From the literature search, it became clear that different types of representational bias can be identified in open health datasets. Representation bias can occur as a result of gender, racial, geographic and socio-economic disparities in the data. This affects the health equity of patients, since continuous use of the biased dataset for training medical machine learning models will result in unfair outcomes for those who are misrepresented in the dataset. The composition of a development team can also have an impact on the presence of representation bias since the level of diversity within a design team affects the diversity of the dataset or machine learning model.

As open health data contains different types of representation bias, it is urgent that this is addressed to stimulate its adoption and implementation in healthcare environments. This emphasizes the need for defining a framework that assists creators of open health datasets with composing a dataset that is representative and diverse for its intended usage purpose. By designing a framework that touches upon the biased as listed above, it will contribute to a more fair and responsible use of open health data for training medical machine learning.

3.2. SQ2: Influence of the social context on open health data

When looking at the socio-technical aspects of open health data, the social system of open health data refers to the attitudes, values and relationship among the actors that publish and use open health data, whereas the technical system refers to the technology used to transform input to output (Zuiderwijk et al., 2012).

The relevant actors and their roles play an important factor in determining the social context of a socio-technical system (Kroes et al., 2006). These elements serve as a mould for the decision framework that will be designed. This paragraph focuses on the sub-question: *how does the existing social environment influence how open health datasets are created and deployed for real-world application in healthcare?* By examining the social context of open health datasets, areas of improvement can be identified as well as areas that are underexposed.

3.2.1. Search strategies and selection of literature

A literature review into the social context is useful as it will highlight the current stakeholders and their roles in open health datasets, but it also delivers knowledge on what the current state-of-the-art is and where barriers and opportunities arise.

The databases of Scopus, Pubmed and Google Scholar were used. The keywords consisted “open health data”/”open data” and/or “healthcare” and “actors/stakeholders” and/or

"policies/policy." An overview of the used related search terms are shown in Table 3. The goal of the literature search was to find scientific articles as well as grey literature to gain a broader insight in the current social context of open health data.

Table 3 Search terms sub-question 2

Search term	Synonyms
Social context	Benefits, barriers, challenges
Open data	Public data, public dataset
Open health data	Medical open data, public health data, published data
Stakeholder	Stakeholders, roles, actor

Inclusion and exclusion criteria

This sub-question aims at examining the wider context of open health data, therefore literature was included if it published in the last 10 years (2013 – 2023). Articles were solely included if it was peer-reviewed. Grey literature was included from credible and reliable sources, such as governmental organizations. Some of the collected literature was excluded based on formulated exclusion criteria. This entails literature behind a paywall and papers that are in another language than English or Dutch.

Figure 6 shows the PRISMA Flow Diagram of this literature search strategy. Subsequently,

Table 4 shows an overview of the included literature and its relevancy.

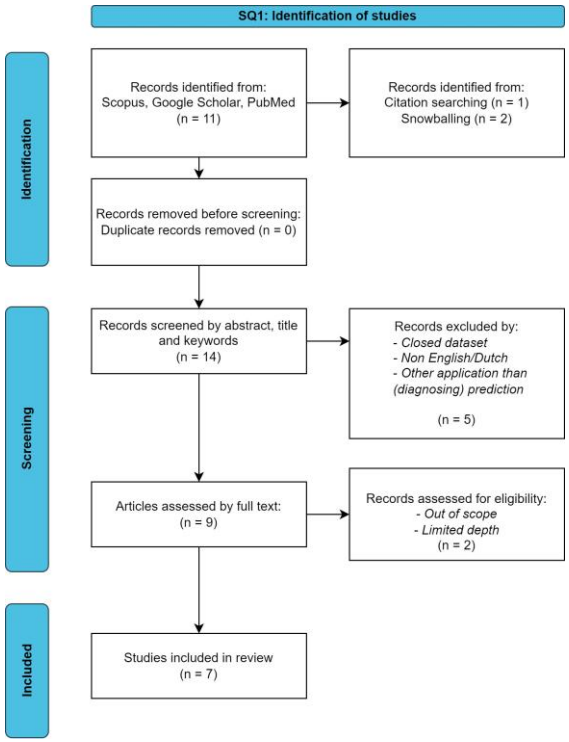

Figure 6 PRISMA Flow Diagram of sub-question 2

Table 4 Overview of included literature for sub-question 2

	Author(s)	Article	Aim
1	Huston et al. (2019)	Open science/open data: Reaping the benefits of open data in public health	Elaborating on the history of open health data and its potential for the public health domain. It also looks into how Canada currently uses open data for public health.
2	Begany & Martin (2017)	An open health data engagement ecosystem model: Are facilitators the key to open data success?	Examining the actors and technologies that affect the publication of open data, as well as the responsibilities and roles they have and obtain.
3	Heijlen & Crompvoets (2021)	Open health data: mapping the ecosystem	Designing an open health data ecosystem for public health data that is managed by governments.
4	Ubaldi (2013)	Open government data: towards empirical analysis of open government data initiatives	Understanding the principles, concepts and criteria that influence the adoption and acceptance of open government data initiatives.
5	Thornton & Shiri (2021)	Challenges with organization, discoverability and access in Canadian open health data repositories	Examining the existing open health repositories in Canada and exploring the potential and limitations.
6	Wu et al. (2019)	Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories	Investigating the user requirements that stimulate the use and findability of data in data repositories
7	Wang et al. (2021)	Quality, reuse and governance of open data	Elaborating on barriers to open data usage and providing a plan of action

3.2.2. Results from the search

The social environment of open health datasets influences how open health datasets are created and deployed as the involved stakeholders have different needs and interests when it comes to working with open health data. Multiple actors in the field of using open health data are identified, this varies from governments, healthcare professionals, and non-profit organizations to researchers (Huston et al., 2019). Ubaldi (2013) describes the different

ecosystems in open governmental data, which relate to data producers and data users. In the context of open health data, these ecosystems can similarly be identified. An actor can be a provider of open health data, making them a 'data producer' that collects the data and publishes the data as well, whereas a consumer of open health data engages with the data that is published by the provider by using it for their (research) purpose. It is important to consider these types of data providers/consumers since their motivation for engaging with the data differs (Begany & Martin, 2017). In addition, it is also possible for the actor to be both an open data provider and consumer, meaning that they can both publish and use open data which is not expected from solely the data providers or consumers.

Within the creation and utilization of open health datasets, we can distinguish three different types of stakeholders (Table 5). Each stakeholder engages with open health data in a different way. These stakeholders should be taken into consideration when designing guidelines on how to address representation bias:

I. Open health data prosumer
II. Open health data consumer
III. Open health data provider

Primary level: open health data prosumer

In the first category, the primary data user is defined as the actor who is both collecting and using the data. This could be a hospital that collects and publishes its data and could also (potentially) use other open health datasets as well. Health data coming from the prosumers can originate from e.g. historical research and clinical trials. The user specifies a problem area and wants to address it by collecting data, potentially for using it in a medical machine learning model to provide solutions or suggestions to the problem. This emphasizes that there is a clear intention for which the data is collected. The data will thus be collected in such a way that it is altered to fit the targeted population of the problem. An example can be a medical team that wants to examine the effect of alcohol on developing liver cancer. How the medical team shapes the research, influences how the data is collected. It determines which variables are selected, what data collection methodologies are used and what ethical considerations should be made when collecting the data.

This actor group has an interesting role since they can produce open health datasets, but at the same time, they can also consume open health datasets. As open health data offers the potential for better health outcomes and treatment, open health data prosumers are also benefitting from representative open health data since they can use other open datasets for other purposes or research. Their motivation and benefit in developing guidelines to address representation bias is therefore in two directions: their dataset can be benefitted but also other datasets that they might use, can improve from these guidelines.

Open data prosumers are essential to involve when addressing representational bias in open health datasets. Since they contribute to the available volume of open health data, practical guidelines should be developed on how open health data prosumers can be stimulated to

address representational bias for improved future use. The guidelines can also assist the prosumers in developing new, representational datasets.

Secondary level: open health data consumer

Subsequently, the actor on the secondary level is identified. This actor is different from the primary level as they do not influence how data is collected but are solely users of the data, meaning that they have no impact or power when it comes to reducing representational bias in the dataset itself. They rely on the open health data that is offered by providers/prosumers to have access to data for their research purposes, which affects how they engage with the data (Ubaldi, 2013). An open health data consumer can be an individual or organization that only uses the data, but does not provide any form of open health data. Examples of actors that may take the role of open health data consumer are academia or university researchers, but also medical institutions that use open data but do not open up their data.

Open health data consumers are relevant in the social context of open health data as they are the users of the data during a later phase: after the open health data has been published by the provider or prosumer. Considering that one of the goals of open data is to stimulate innovation and foster collaboration, this category is filled not only with academia and researchers but also data enthusiasts who might use the data for "recreational" purposes or non-profit organizations. As they are the core of scientific research, it is important to involve open health data consumers in developing guidelines to address and mitigate representational bias.

Tertiary level: open health data provider

Lastly, there is a tertiary actor in open health data; the actor that collects health data but does this with a broader intention than on a primary level (prosumers). There is still an underlying reason for collecting the data, for example, to monitor public health, but as opposed to primary and secondary level actors, tertiary actors do not use open health data themselves. Instead, the core value of this actor is to provide reliable and objective information to open health data consumers and prosumers.

Examples of tertiary users can be governmental organizations concerned with public health. In the Netherlands, there is the Institute for Public Health and Environment, but there are also statistical institutions that collect large amounts of data to provide information about broader social topics. Governmental health data, such as population health programs, are more related to aggregated data and thus provide broader information about regions or larger populations (Heijlen & Crompvoets, 2021). On an EU level, this is the Eurostat data repository with a broad offer of healthcare datasets, e.g. data on vaccinations by people older than 65 against influenza (Eurostat, n.d.).

These organizations collect data continuously, meaning that there is an impact to be made when it comes to reducing representational bias in their dataset. This is dependent on data collection methodologies, which can be addressed by designing suitable guidelines for representative data collection.

Table 5 Overview of open health data actors

Level actor	Examples of actors	Provide open health data	Use open health data
Primary: open health data prosumer	Medical doctors, hospitals, medical institutions	X	X
Secondary: open health data consumer	Academia, researchers		X
Tertiary: open health data provider	Private/public initiatives, governmental organizations, statistical bureaus	X	

As shown in Table 5, the three open health data actors have different roles in the context of creating and/or using open health data and mitigating representation bias. Whereas primary and tertiary actors are both providers of open health data, they are different in the sense that datasets are targeted towards a specific research area and do not necessarily have the intention of being published, whereas tertiary open health data actors collect the data to publish them. Representation bias can be reduced during the data collection phase, but afterwards, it is the responsibility of the data user to correctly address representation bias if present. Open health data consumers are dependent on the prosumers and providers to have access to open health data and thus have limited capabilities to reduce representation bias.

Shortcomings in using open health data

Previous research has been done into how the use of open data repositories can be stimulated and improved. When it comes to using open health data, Wu et al. (2019) advocate for making it easier for researchers (who can be both primary and secondary open health data actors) to judge the relevance and reusability of a dataset. This could be achieved by e.g. showing the data consumer the statistics of a dataset. Such information is important since researchers want to know whether a dataset that they will use for their research is representative. Information on how exactly the data was collected is also desirable. This affects whether they use the open dataset or not. By inspecting the data, it helps researchers decide whether the data is right for their research. Examples of important information that can be included are who collected the data, where the data was collected, when, and under which circumstances. Ideally, this will provide a more detailed insight into the relevancy and usability of the open health dataset.

A large issue with using open health datasets is the incomplete, inconsistent and poor-quality metadata (Thornton & Shiri, 2021). Metadata of open health datasets is essential for searching and re-using data. Wang et al., (2021) show the importance of including metadata in open (health) datasets, however, including reporting of representational bias in this metadata is not included.

This also influences representational bias in the dataset. If the context of the dataset is not known, it is difficult to assess whether a dataset contains representational bias; this will arise

during later stages of model development and validation. To circle back to the identified open health data actors, two actors can contribute to metadata inclusion: the primary and tertiary actors. Primary actors who publish their (research or organizational) data and tertiary actors who collect the data for the goal of publishing have valuable information, such as the composition of the dataset and the used data collection methodologies. This information adds value to the metadata and it improves the quality of it, which is beneficial for consumers of the open health dataset.

3.2.3. Conclusion

The social context of open health data is illustrated by identifying different levels of open health data actors who have different motivations and needs for publishing and/or using open health data. This affects their abilities to intervene and mitigate representation bias in the dataset. In addition, the current shortcomings in using open health datasets are elaborated on since this affects the action space of the open health data actors. The lack of metadata for open health datasets affects the usability of the data for data consumers, but this is the responsibility of data providers and prosumers since they have insight into the information that can be included as metadata.

To elaborate on the action space of the open health data actors, the primary and tertiary level actors can contribute to the availability of metadata. Important decisions are made concerning the collection of the data, such as the choice of data collection methodologies, scoping the research population and taking a significant sample of this population to include in the data. All of this is important and valuable information for data consumers. In addition, they can adhere to improved data collection policies that enhance the diversity of the dataset. This is opposed to secondary-level users, who are left with assessing the suitability of the published dataset for training their machine learning model. Having an increased amount of high-quality, reliable and relevant information on the open health dataset is valuable for the users of the dataset as it elaborates on the usability and relevance of the dataset. In cases where representation bias occurs in the dataset, the data consumer/prosumer will know beforehand. These roles and differences in action spaces must be incorporated into the decision framework.

4

Explorative Interviews

In addition to building a theoretical foundation as provided in Chapter 3, explorative interviews with experts at this stage of the research are a valuable method for gaining a deeper understanding of the problem. The interviews aim to discover pillars for the decision framework for designing diverse open datasets but are also useful as they allow a detailed and applied view of the problem area. It allows the participants to "speak in their own voice and express their own thoughts and feelings" (Alshenqeeti, 2014, p. 1).

4.1. Participant selection

The explorative interviews require participants who are working in the field of open (health) data, medical machine learning or health ethics. The participants were found by identifying relevant organisations, companies and institutions and were subsequently approached by e-mail. There are no prior relationships nor any knowledge between the interviewer and the participants. All interviews were conducted online by using Microsoft Teams as a meeting platform. Table 6 below provides an overview of the participants, their background and their relation to this research. In total, seven experts were interviewed, all working in The Netherlands. The interview protocol that includes policies on privacy, data processing and storage can be found in Appendix A.1.

Table 6 Explorative interview participants

Participant	Job title, background	Background in research area
E-P1	Engineer	Healthcare ethics
E-P2	Machine learning engineer	Medical algorithms
E-P3	Expert in health data	Reuse of health datasets, FAIR data principles
E-P4	Researcher	Health & technology
E-P5	Expert in health data	FAIR data principles
E-P6	Researcher	Ethical and systematic injustice

E-P7	Program manager	Diversity and inclusivity in data research

4.2. Interview scope

During the interview, the participants are asked to share their experience and knowledge of open health data, its potential for machine learning and potential ethical implications. Although the implementation of open health datasets is still in its beginning phase, the participants are asked to share whether they have been working with open health datasets and what the intended purpose of working with the dataset was. They are also asked to shed light on how they perceive the contribution of open health datasets to the development of equitable and fair medical machine learning models.

The ethical aspects of open health datasets are examined by obtaining previous instances of a lack of diversity in data, which resulted in disadvantages for specific groups or affected the health equity of patients. When participants share this information, potential areas of improvement can be identified based on historical events. The participants also share their considerations of how patients' health equity can be protected and ensured when creating open datasets, as well as the ethical challenges they foresee in the use of open health datasets for machine learning. These concepts relate to sub-question 1: *what are the different types of representation bias that occur in open health data used for medical machine learning models and what are their ethical implications for those at risk?*

In addition, participants are requested to share their opinion on the existing legal and institutional boundaries that assist with the development and use of open health datasets or in case of the absence of these boundaries: what the effect is of not having these instruments in place. This touches upon sub-question 2: *how does the existing social environment influence how open health datasets are created and deployed for real-world application in healthcare?* This also entails to what extent the design of open health datasets should be seen as a collaborative challenge on a regional, national or international level and what potential challenges could arise from (the lack of) collaboration. Furthermore, by diving deeper into technical requirements and best practices for the use of open health data, an interesting range of technical improvement areas can be derived. The participants are encouraged to share their ideas but are also asked to share their opinions on prominent recommendations from existing literature.

The approach of the interviews allows for a deeper understanding of the implementation of diverse open health datasets from an ethical, social and technical point of view. This strengthens the theoretical foundation of the design.

4.3. Results

The content of the seven exploratory interviews was analysed to identify directions for answering the sub-questions below. The insights from the interviews will be summarized for the two sub-questions.

4.3.1. SQ1: Bias in open health data and ethical implications

Due to the limited application of open health data, only some of the participants have experience with the use of open health datasets. Therefore there are also mentions of biases in open health datasets that are seen as risks by the participants, that come from their experience with non-open health data. These risks are also relevant to consider as they come from the healthcare/ethics domain and thus touch upon the effect of potential biases in open health datasets.

The participants agree on the potential of open health data and see opportunities for stimulated innovation, broader knowledge of diseases and improved health outcomes. Simultaneously, there is a clear consensus on the negative effects of bias in open health data and these should carefully be addressed to prevent harmful outcomes for patients. One participant even mentioned the effect it can have on data enthusiasts or hobbyists who enjoy analysing datasets, potentially contributing to new pattern identifications.

When it comes to bias types, selection bias was mentioned by the participants. For example, this may occur in data that was derived from trials and tests for new medication types. These tests are often performed on healthy, young people, whereas the medication is targeted for a patient group with a specific disease, e.g. kidney problems, which often arise in older people. To build further on this example, these medications are often tested on participants who do not use other types of medication to ensure the reliability of the trials, but this isolated measurement is not representative of the population that usually needs the medication: people who are older, less healthy and thus often use other medication as well. Another important mention of bias is in the MIMIC-II dataset, according to E-P4, this dataset contains gender bias.

In addition, E-P2 mentions that geographical bias is something that should carefully be considered by those who use (open) health datasets for medical machine learning models. Depending on where data is collected, it is not always possible to reuse the data. An example of hospital divisions is given: in the Netherlands, there are two types of hospitals: academic and general hospitals. Academic hospitals do not solely have a larger patient population, but the cases for which patients visit academic hospitals are often more complex. Data generated from the general, often smaller hospitals may therefore not be representative of academic hospitals, and vice versa. Although open data has the potential of combining multiple open datasets and can thus also contribute to improving representativity, a single open health dataset can show geographic bias on its own.

Furthermore, anomalies can result in representation bias in open health datasets; closely related to sampling bias. When measurement points are strongly deviated from the 'normal' base, they will be misrepresented in the final datasets. These anomalies may occur from different origins. E-P2 and E-P4 both highlighted this form of representation bias. A recent and well-known anomaly is the COVID-19 pandemic; health data that was collected during this hectic and abnormal period is not likely to be representative of the period before or after the

pandemic. E-P2 mentioned that data that was collected during these periods was not used for their machine learning model, since it did not share the characteristics of the 'normal' time. A similar remark was made by E-P4 who worked with data about food consumption. Data that was collected during the end of the year, during Christmas and New Year's, gave very different insights into food consumption than any other arbitrary period of the year. It must be noted that anomalies are not inherently a problem but are dependent on with which goal the open health dataset will be used. Therefore, it is important to consider the context of the problem.

Taking into account the matter of context, bias can sometimes be seen as desirable. An example of this was given by E-P1: "*sometimes bias is relevant, as some diseases might be more occurring in certain racial groups. So in some cases, you actually want the discrimination in the data because it is relevant for the outcome*." As an example, bias could be relevant in the cases of sickle cell disease. Patients with this disease mainly live in Africa, therefore it would be less relevant to look at populations in Europe (Rees et al., 2010). Opposed to bias being desirable and intended, bias can also occur through proxies as highlighted by E-P5. They state that these proxies indirectly indicate an attribute of the patient, for example determining how long their drive to the pharmacy is and linking this to an increased risk for drug shopping while not considering the fact that they might live in a rural area due to their socio-economic status.

The explorative interviews showed that gender, age, geographic and selection bias may occur in open health datasets, either since these biases have been present in open or (for now) closed datasets or because of the perceived risk of the participants. In addition, selection and sampling bias can contribute to misrepresentative open health datasets. Datasets that already exist and will be opened up in the future can already contain the above-identified biases, therefore making it difficult to eliminate these biases. Open health datasets that are yet to be created can address this by advancing data collection methodologies. However, the essence of open health data is that the data should be usable for those who want to work with it as mentioned by E-P1, E-P2, E-P4 and E-P6. This shows that various approached are needed for addressing representation bias in open health datasets.

4.3.2. SQ2: Influence of the social context on open health data

A recurrent topic was the need for context for which the open health data will be used, and to use this context when designing new open health datasets. E-P1 encouraged to discuss with future users what they perceive as important to include in the open dataset. Open health datasets that are mainly valuable for general practitioners are likely to have a different design, with other variables and data points, compared to datasets that will be used for, e.g., cancer predictions; talking to a general practitioner will allow the dataset creator to have a broader view on what to include and exclude in the dataset. Therefore, when new datasets are created to publish them, they should be designed in collaboration with experts from the field it is (likely to be) intended for. This can contribute to mitigating representation bias since the dataset is aligned with the needs and desires of the consumers of the open data. These discussions will provide a contextual understanding of what data is relevant to include and how it should be

included to ensure a strong alignment with the target population, making more room for representativity.

In addition to collaborating with the consumer of open health data, a broader collaboration is highly encouraged by E-P1, E-P2 and E-P4. As said by *P1: "You want everyone that is affected by it [open health datasets] in the room; patients, professionals, policymakers and people from the technical side, so they are also working outside of their own silo."*. Having a diverse design team when creating open health datasets is valuable since it brings together people from different backgrounds. The backgrounds can differ from a professional work field perspective where not only data scientists are included but among others, policymakers, patient representatives and ethicists are also involved. In addition, the personal background of a diverse team is also considered important, for example having different genders and ethnicities in a design team if that is relevant to the context. This will result in an interdisciplinary collaboration between different parties, coming from different backgrounds. E-P2 also emphasizes that the importance of a diverse team is high and is actively considered in their organization. The different perspectives will provide a wider view of representation bias. This is also mentioned by E-P6 and E-P7, who state that the end-user for whom the open dataset is being developed should be reflected in the design team, for example by involving experts by experience, and that this is one of their conditions for providing a grant to medical researchers. In line with this, E-P5 states that it is important to design an open health dataset in collaboration with the patients since they have a different experience and their own ways of talking.

As previously stated, there is consensus on the urgent need for mitigating representation bias in open health datasets and healthcare in general. E-P1 mentions: *"you want an algorithm to contribute to better health, better research and recommendations. That is priority number 1. That means that it should be actually good, without any bias and accessible to those who should be able to access it without the risk that they abuse the data. This could also mean that that there is something that would warn when you want to use data that is not suitable from a bias POV, saying 'sorry but you do not have enough data for what you want to research/there is too much bias in the data' and even naming what types of bias it is."*

New technologies and concepts can contribute to achieving goals such as 'improved diagnosing' or 'enhanced healthcare quality' but this is strongly dependent on the usability of the technology. Biased datasets will result in an algorithm with biased outcomes, and it should therefore be addressed as early in the machine learning model development process as possible. When data is collected, there is a choice between keeping the data closed or open. For collected data that was opened up at a later stage, it is more difficult to reduce the representation bias since it was gathered with a specific goal in mind. This is also emphasized by E-P4: "*Data is collected by a researcher with a specified goal. They are not going to collect extensive amounts of data, which they will not use, simply to make it more representative."* However, this does not mean that the data is unsuitable as an open health dataset. This is dependent on the intention of the open health data consumer.

This strongly aligns with the need for additional metadata in open health datasets. Metadata can be used to provide information about the open health dataset. Multiple participants stretch that there is a serious lack of metadata for open health datasets which significantly affects the usability of the dataset because the data consumer does not know whether the dataset contains representational bias. In the case where the dataset would still be used for training a medical machine learning model, it will result in bias outcomes. In the case where the consumer decides not to use the open health dataset, the utilization of open health datasets will become lower and block its full potential. Therefore, providing data users with metadata can strongly contribute to the usage of open health datasets. The user of the data can determine whether the open health dataset is in line with their intended usage.

This is also advised by E-P3, who is an expert in FAIR data. They amplify that findability is concerned with including metadata as it enhances the reuse of datasets. This can vary from basic documentation on the data: are the data subjects human or not? An anecdote was given to emphasize this: data coming from another researcher was re-used but in the data, many extreme measurement points were present. It was only after contacting the researcher and data collector that it became clear that these data points belonged to horses instead of human subjects. This shows the importance of including metadata when publishing datasets for reuse. Metadata elements that are considered important by the participants are gender, age and geographics, but also elements that are concerned with data collection methodologies. E-P6 mentions that not having metadata for an open health dataset and choosing to use the dataset while making assumptions about the data can result in huge biases. They find that the problem is not necessarily that research was conducted among white males between 40 and 60 years old, but "*the key factor is that this should be included in the metadata so that other researchers know about this.*"

4.4. Conclusion

From the explorative interviews, it became clear that to address representation bias in open health datasets, it is important to include the context of the dataset by implementing metadata policies. Metadata should refer to the characteristics of the data subjects and data collection policies. Furthermore, options to contact the researcher and/or data collector are important to solve conflicting issues in using open health data. These policies will assist and guide data consumers with fair and just usage of open health datasets. This accounts for datasets that are yet to be developed but also for datasets that are already existing and even published. When looking at data collection policies, designers of future open health datasets can implement enhanced collection methodologies to mitigate the risk of gender, age and geographical bias. Since no law or regulation obliges open health data providers with specific criteria for representative data collection, the focus is on emphasizing a two-fold benefit to stimulate representative and diverse design and use of open health datasets. Therefore, it is important that these concepts are considered in the decision framework.

5

Decision Framework

This chapter presents the decision framework for mitigating representational bias in open health datasets. The framework has been derived through shortcomings found in the literature review in Chapter 3, as well as practical, real-world examples of conducted explorative interviews in Chapter 4. A total of seven guidelines are proposed that assist with designing representative and diverse open health datasets and that aim to improve the usability of existing open health datasets. Firstly, the guideline is explained and is followed by a set of recommendations that entails the incorporation of the guideline.

Guideline 1 – Diversity in design

The guideline 'diversity in design' refers to the composition of the design team of open health datasets. The team composition is important as people from different personal and professional backgrounds come together and have the opportunity to share perspectives, perceptions and assumptions among the group. Although this might not directly impact representation bias, it allows for a better understanding of different needs and therefore stimulates thinking outside of your own perspective. The personal background of the members can refer to their education level or gender whereas the professional background refers to the variety of functions such as ethicists and experts on systemic injustice. Important to mention is that '*G1: Diversity in design*' is not focused on filling a mandatory quota.

Recommendations: To incorporate this guideline, it is important to determine the context in which the dataset is created when a new open health dataset is designed and what the research population is. This will help the developer identify which diverse backgrounds are useful to include in the design team. This way, the developer can compose a design team that is representative of the context in which the dataset will be created.

Guideline 2 – Patient involvement

As patients are the subject of open health datasets, their needs and values must be considered during the development of the datasets. This can be achieved through appointing various patient representatives who will take place in open dialogue meetings with the design team where experiences and needs can be shared. For example, since socio-economic factors

have been shown to play a significant role in seeking medical help, it is important that (potential) patients of different levels of health literacy are represented as well as different levels of education and economical status. Their role is thus to provide a more nuanced view of data that can be incorporated into open health datasets. Although this nuanced view is valuable, there is the risk of conflicts between the design team and the patients when there are no agreements to be found. In this case, areas for compromises or alternatives should be explored.

Recommendations: The previous guideline '*Diversity in design*' determines the research population. This population is also applicable to '*G2: Patient involvement*' as this provides information about the patients that should be consulted. The patients, or representatives of patient groups, should be informed about the intended use of their data so they have the opportunity to share their perspectives. Lastly, there must be a retrospective aspect in '*G2: Patient involvement'* to evaluate whether the patients' needs are met accurately.

Guideline 3 – Anomaly prevention

When collecting data, it is important to consider the context in which it will be collected. The circumstances of data collection have the potential to either enrich or discredit the data. Examples of anomalies were given in the form of examining food habits during the Christmas period or examining hospital capacity during a pandemic, but other examples could include external events such as disasters or significant policy changes. *'G3: Anomaly prevention'* stimulates that, depending on the reason for which data is collected, these anomalies should be carefully considered as they limit the generalizability of the data. If the context of the open health dataset requires data collection with anomalies, this should be communicated to the data consumer. If the collection of anomalies is not relevant, they should be avoided to increase generalizability. Designers of open health datasets should identify anomalies that are relevant in their data context. This is different from outliers in the data: where single data points (patients) are in- or excluded from the data. In health data, patient outliers can be considered desirable since healthcare should be accessible to a wide range of patients, including the 'outlier' patients.

Recommendations: While taking into account the purpose of the dataset, identify relevant and potential anomalies that might occur during the data collection. During the data collection, iterative if any additional anomalies arise and if necessary, alter the data collection procedures that are used for the design of the open health dataset.

Guideline 4 – External validation

External validation is a process during the development of algorithms where its performance is determined. After an algorithm has been trained, tested and validated on the used dataset, its performance on external datasets is not guaranteed. The generalizability of the algorithm can be discovered by using the algorithm on external datasets. This concept can be applied to the development of open health datasets as well. Datasets are collected by a provider or prosumer who has used specific data collection methodologies. '*G4: External validation*' prescribes how these methodologies can be externally validated with other data providers or

prosumers who will take up the role of the data validator. The validators can be found in the network of the data provider but also within health organizations or academic networks. The dataset validator must be familiar with the context in which the data provider has collected the data. When looking for validators in other data providers, the willingness to participate as a validator can be improved by emphasizing the mutual benefit for both parties involved. By examining data collection methodologies, one can contribute to improving the quality and usability of open health data and since the validator is in a familiar context as the data provider; the now validated open data may also be more applicable to them. *'G4: External validation'* provides insights into the existence of representation bias in the dataset. This allows the designer of the dataset with courses of action before publicizing.

Recommendations: Identify appropriate validators with regard to the context of the open dataset. Subsequently, evaluate the data collection procedures (identification of research population, sampling methods) and assess the representation of the research population in the dataset to identify potential biases in the data. The findings are reported to the data provider, who can incorporate the suggestions to refine the open dataset.

Guideline 5 – Contacting possibilities

Another pitfall that should be addressed is the lack of possibilities to contact the provider of open health datasets. Although it is in many cases possible to incorporate a way of contacting, this is not always done by the data provider. *'G5: Contacting possibilities'* encourages the data provider to appoint a contact person. This is important in cases where metadata does not provide sufficient information about the data, or when a data consumer requires clarifications or additional information about the dataset or the used data collection methodologies. The data provider can either be a single data owner or an involved representative of the organization that published the data: the main requirement is that the contact is familiar with the dataset. This form of communication will improve the usability of the dataset since the lacking information can be completed by the data provider. This addresses representational bias in datasets since it allows for a complete view of the data but also helps with assessing data collection methodologies, resulting in an improved ability to assess the dataset for utilization for medical machine learning models.

Recommendations: When a health dataset is opened, a contact owner must be appointed as well as a potential substitute in case the contact owner is not able to execute this task anymore e.g. due to job changes. Data prosumers and consumers should be able to contact the owner directly, for example by using a direct e-mail address.

Guideline 6 – Bias reporting

There is a possibility that after all, open health datasets, although they were created with the best intentions, still contain representation bias. *'G6: Bias reporting'* encourages that existing biases must be shared with open health data consumers. This requires providers of open health data to actively think about potential biases in their dataset or to report on identified biases by the data consumers. By communicating which data groups are under- or overrepresented in the dataset, the data consumer has a better understanding of the usability

of the dataset and is thus able to assess the data for suitability to their intended purpose. This will address representational bias as it improves awareness and clarifies the usability of the dataset for data consumers. In addition, the users of the data might discover new biases or limitations of the data. They are encouraged to report this to the publisher of the data.

Recommendations: Determine the intended purpose of the dataset and what lies outside of its scope. The data prosumer/provider should think of potential biases or limitations in the dataset and when these are identified, they must be reported to the data consumer. Vice versa, the user of the data that identifies new limitations or biases should report this to the data provider or prosumer.

Guideline 7 – Metadata Inclusion

When health datasets are opened up and, for example, published in a data repository, the dataset must come with metadata elements. The lack of metadata in open health datasets impacts their usability as it takes away the context of the dataset. When consumers of open health data assess an open health dataset for its relevance, suitability and usability for their intended purpose; it is fundamental to know the details of the data. *'G7: Metadata inclusion'* contributes to the urgent need for presenting metadata with open health datasets. The metadata should at least consist of but is not limited to, where and when the data was collected, by whom the data was collected and the composition of the data subjects (e.g., genders, ethnicities and age). Having a clear understanding of what should at least be incorporated as metadata, from a representational point of view, can help data providers as they do not have to identify the metadata components themselves. This may lower the barrier for data providers to incorporate informative metadata.

Recommendations: The metadata should contain information about where, how and when the data was collected and by whom. It should also report on the population demographics such as size, gender and age. In addition, the metadata should cover the aspects from the six other guidelines meaning that documentation must be included about G1: Diversity in design, G2: Patient involvement, G3: Anomaly prevention, G4: External validation, G5: Contacting Possibilities and G6: Bias reporting.

Decision framework

The application of the seven guidelines is dependent on what level an actor is operating with open health datasets: the primary, secondary or tertiary level. Therefore; a decision framework is created that guides the open data provider, consumer or prosumer in which guidelines to incorporate in their path to designing or using representative open datasets. Figure 7 provides a visualization of the decision framework.

A distinction is made between a preventive and a corrective approach. The preventive approach can be used by data providers and data prosumers when an open dataset has not been developed and published yet. At this stage, the designer of the dataset can take measures to address representation bias during the development of the dataset to 'prevent' representation bias from occurring. On the other hand, the corrective approach is followed

when the dataset has already been developed and published. Since the data has been collected, the preventive measures that influence the data collection are less relevant.

The preventative approach allows for taking measures that address challenges and issues in representation in open health datasets. It incorporates *'G1: Diversity in design'* to encourage representation of the research population in the design team. This brings different perspectives and views together. In addition, *'G2: Patient involvement'* is important as it allows the end user to shed light on how the open health dataset is aligned with their needs and values. Before the data collection starts, *'G3: Anomaly prevention'* prescribes to identify any potential anomalies that might affect the generalizability of the data. The last preventive guideline *'G4: External validation'* stimulates data providers and prosumers to validate their data collection methodologies to identify potential biases in the data and allow them to make alterations to these methodologies and data before publication.

The four guidelines above must be considered before the actual data collection starts. When the open health dataset goes from 'to be created' to 'developed' and data collection has been completed, there is a change from a preventive approach to a corrective approach. The data provider or prosumer should check for alignment with preventive guidelines that stimulate diversity in design, patient involvement, anomaly prevention and external validation.

The corrective approach allows for taking suitable measures that address representation bias in an existing open dataset. '*G6: Bias reporting'* raises awareness of existing representation bias, allowing data consumers to understand the relevance and usability of the dataset for their purpose. In addition, an open health dataset must be published with the contact details of the data provider/prosumer. *'G5: Contacting possibilities'* fosters the inclusion of contact details, e.g. of an individual data owner or the responsible organization. Lastly, *'G7: Metadata inclusion'* stimulates to include data about the data and aims to answer questions including, but not limited to: where was the data collected, how was the data collected, by and with whom was the data collected and validated, who are included in the data, who are excluded in the data? This guideline summarizes the contribution of some of the previously implemented guidelines as it addresses the design team, data subject involvement, data collection periods and validation of the data.

Lastly, the open health data consumer and prosumer use the metadata to assess the usability and relevance of the open health dataset for their (research) purpose. If the dataset is used by them and new biases or limitations are identified, they are encouraged to report this to the publisher of the data who can incorporate this in the metadata. This contributes to the continuous and ongoing cycle of metadata inclusion.

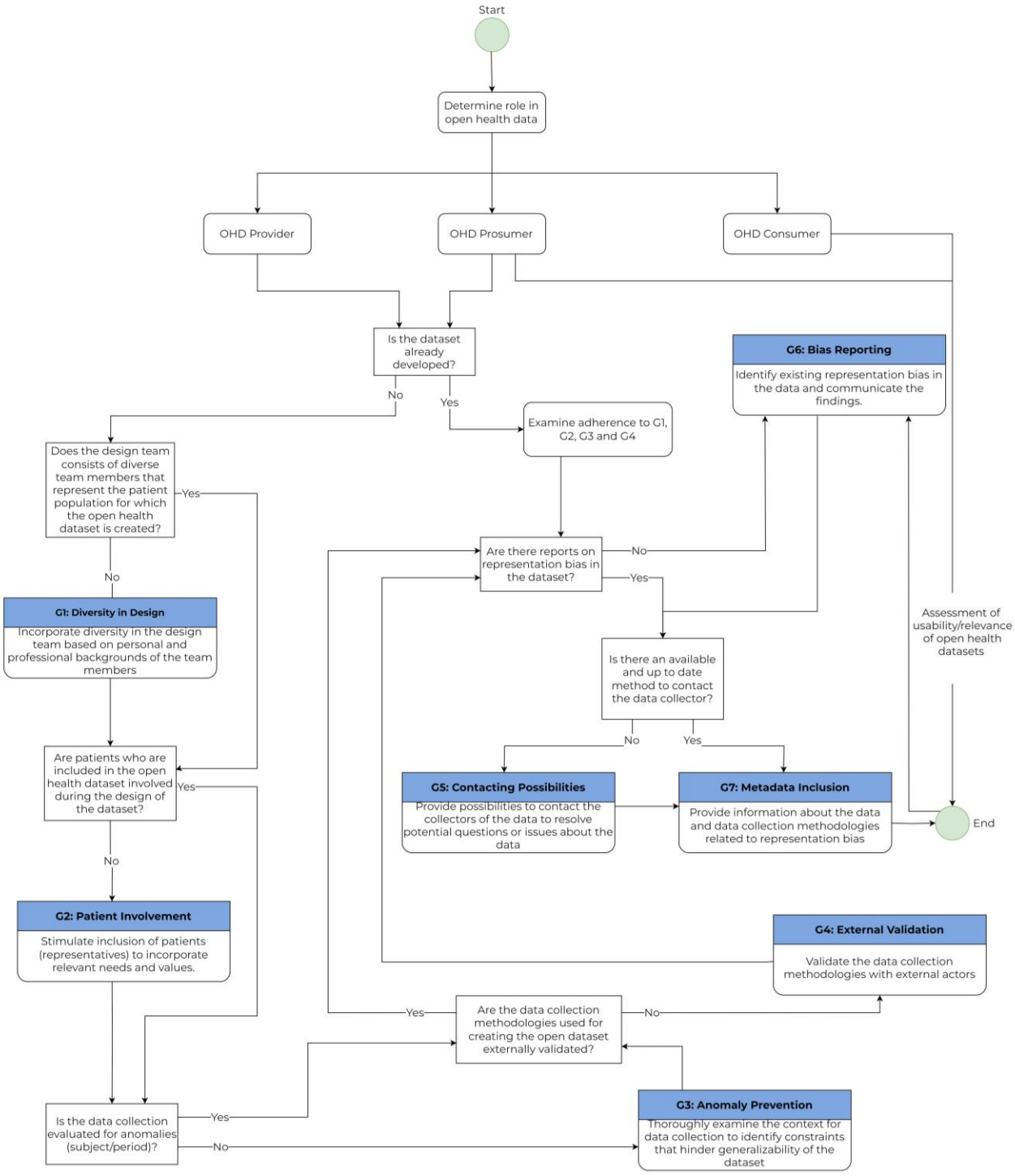

Figure 7 Decision framework for assessing representation bias

The guidelines will assist open health data consumers and prosumers with assessing the usability of the data. To conclude, the decision framework uses both preventive and corrective measures to address representation bias in open health datasets.

6

Use cases

This chapter will apply the decision framework to multiple cases. First, the area in which the use cases are explored will be introduced. Secondly, the identified use cases will be demonstrated and lastly, the decision framework will be applied and its results will be analyzed. The use cases contribute to answering the sub-question: *"How can the developed decision framework contribute to representative data collection and use of open health datasets to protect patient's health equity?"*

6.1. Case introduction

Data-driven decision-making is a rapidly upcoming technique in healthcare, making the application field of open health datasets very broad. A specific healthcare area in which large volumes of data are collected is the intensive care units of hospitals. In intensive care units, high-risk events must be prevented or detected early by monitoring patients strictly (Syed et al., 2021). In addition, the COVID-19 pandemic showed the importance of predicting bed capacity. Therefore, many hospitals have incorporated predictive models that assist with the occupancy of an intensive care unit; since these units provide essential breathing systems for those who are infected (Acosta-Velasquez et al., 2022). The large volume of data that is obtained from patients in ICUs can be used for warning systems in the ICU but also contribute to the prediction of unit occupancy (Bailly et al., 2018).

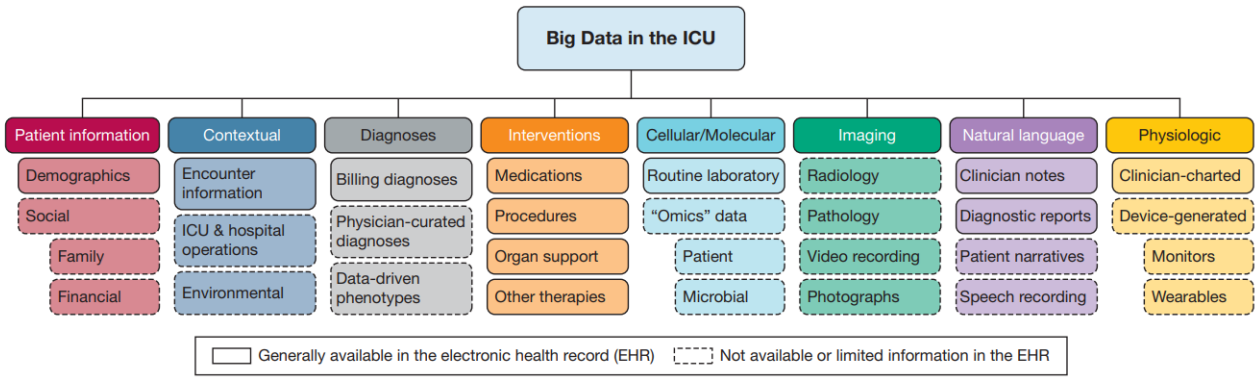

Figure 8 Data sources in ICU (Sanchez-Pinto et al., 2018)

Figure 8 illustrates the various sources of data collected from patients in the ICU environment. A large number of them are often included in their EHR. The large variety of data sources can result in a broader, comprehensive understanding but it is also essential to address potential representational bias. The data collected in ICUs can be published for extended use, making it an open health dataset that will be used for other research purposes. The shows the urgent need to address representation bias when designing and/or publishing an open dataset based on intensive care unit data.

6.2. Data search and collection

Potential use cases were identified by searching the scientific literature databases of Scopus and PubMed for articles that use open health datasets in their research. Search terms used are *("open data" OR "open health data") AND ("intensive care" OR "ICU" OR "critical care")*. This resulted in 30 articles of which the title, keywords and abstract were examined to assess their eligibility based on the use of an open dataset and whether the dataset consisted of data obtained from intensive care units. Duplicates of open health datasets were ignored. In addition, Paragraph 3.1 elaborates on bias types in open health datasets, of which some consisted of data from the intensive care units. Lastly, mentions of ICU open health data were given during the explorative interviews. This resulted in multiple open health datasets, which are visualized in Table 7.

Table 7 Identified open health datasets, in bold: chosen for use cases

Name of open health dataset	Description	Identified from	Authors of dataset usage
MIMIC-IV	Data of patients admitted to intensive care unit of hospital in Boston (US)	Literature Ch. 3	Meng et al. (2022)
RIVM Covid-19 hospital and intensive care admissions	Admissions of patients to hospital and ICU's due to COVID-19 infections in the Netherlands	Interview	RIVM (Rijksoverheid. (n.d))
AmsterdamUMCdb	Intensive care database with admissions and information about lab tests, medication, vitals from hospital in Amsterdam	Database search	Thoral et al. (2021)
MxCov, MxDeath & MxHcare	Dataset on COVID-19 patients, death registries and healthcare capacity in Mexico.	Database search	Núñez & Soto-Mota (2023)
Centro Nacional de Epidemiología del Ministerio de Sanidad de Espana	Hospital admissions of COVID-19 patients in Spain	Database search	Guisado-Clavero et al. (2022)

Name of open health dataset	Description	Identified from	Authors of dataset usage
Paediatric Intensive Care (PIC) database	Admissions of children to intensive care unit of a large Chinese children's hospital.	Database search	Vistisen et al. (2021)
eICU Collaborative Research Database	Data of patients admitted to critical care units in the United States.	Database search	Vistisen et al. (2021)
High time Resolution ICU Dataset (HiRID)	Patient admissions to intensive care unit of Swiss hospital	Database search	Vistisen et al. (2021)

A selection of open health datasets is made which will be used for the use cases. To obtain a diverse set of cases, they each have their origin in different countries. The AmsterdamUMCdb is selected since it is the first ICU open health dataset in the Netherlands. In addition, the HiRID and eICU datasets are chosen, respectively from Switzerland and the United States. Each dataset will be introduced before it will be analysed with the help of the decision framework.

AmsterdamUMCdb

The AmsterdamUMC database is an open dataset that contains intensive care unit admissions with corresponding variables that give insight into patient demographics, medication use and vital signs. It can be accessed here. The dataset is open but requires a training course certificate to ensure appropriate use of the data. It consists of data collected from the intensive care unit of Amsterdam University Medical Center in the period from 2003 to 2016. It provides information about the patients, relating to demographics such as age and gender but also includes medical information. This varies from medication usage to measurements by medical personnel, blood drawing procedures and observations.

eICU Collaborative Research Database

This open health dataset was created from a telehealth program led by Philips, stimulating the use of critical care data to improve patient outcomes. This resulted in a large volume of data, the eICU dataset as found here. Similar to the AmsterdamUMCdb, eICU requires the completion of a training course before the data can be accessed. The dataset was created in collaboration with the Massachusetts Institute of Technology Laboratory for Computational Physiology (Pollard, 2014). It contains a large volume of data about patients admitted to an ICU department in the United States in the years 2014 and 2015. The data is retrieved from blood samples and laboratory results, vital signs measurements and the use of medication prescribed to the patient.

HiRiD

The open health dataset HiRID was created to be able to develop a machine learning model that could predict the risk of circulatory failure when patients are admitted to the ICU. The dataset can be found here. This open dataset also requires the user to first complete a course before the data can be accessed. The data was collected at the ICU department of Bern University Hospital, which has over 6500 patients per year, in the period from January 2008 until June 2016. The dataset was created in collaboration with the Swiss Federal Institute of Technology and consists of more or less 700 variables, collected during over 34 thousand admissions to the ICU department.

6.3. Analysis of cases

The analysis will be done for AmsterdamUMCdb, eICU and HiRID by 'walking' through the decision framework. In Figure 9 below, the green boxes indicate which path must be taken for the cases and which guidelines must be examined that correspond to this path. The exact implementation is discussed in the following paragraphs.

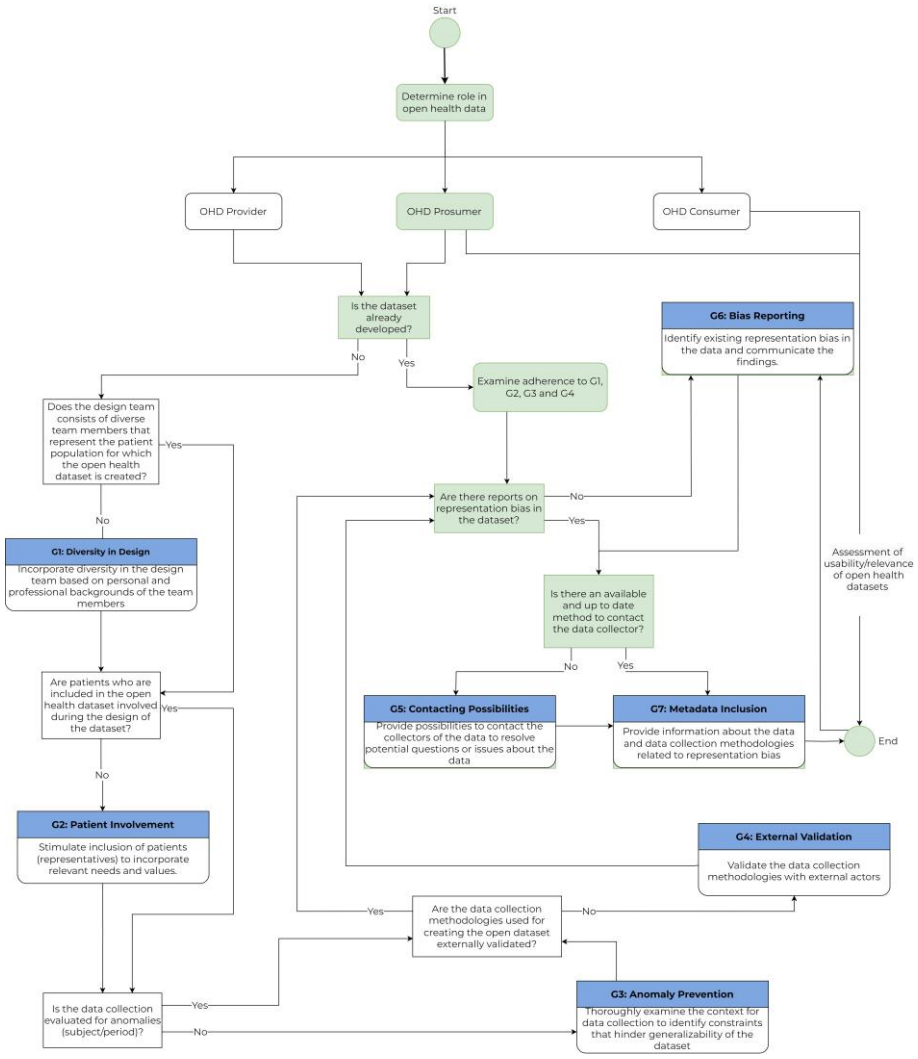

Figure 9 Walkthrough of decision framework for use cases identified as prosumers

6.3.1. AmsterdamUMCdb

AmsterdamUMCdb is the database that is designed by Amsterdam Medical Data Science and led by two medical doctors from the Amsterdam University Medical Center. The data is used by themselves but also opened to the public, making them a data prosumer. Since the dataset is already available, there are only corrective measures that can be taken. The information about the design of AmsterdamUMCdb is provided on their website (Amsterdam Medical Datascience, 2022). This also applies to the corresponding article by Thoral et al. (2021).

G1: Diversity in design

The dataset was created in collaboration with the Research Collaboration Critical Care network of the Dutch Society of Intensive Care Medicine (NVIC). On the website of AmsterdamUMCdb, it is mentioned that clinical ethicists reviewed beforehand the publication of the dataset concerning the usage of de-identified data, to ensure privacy of the patients. Since the data is shared responsibly, no objections were given and thus the data was opened to the public. The AmsterdamUMC also mentions collaborations with other hospitals and the Federation of University Medical Centers on adherence to GDPR. These collaborations are thus focused on legal and ethical aspects of the dataset. To some extent, this provides information about how the design team was composed but the published information is not extensive.

G2: Patient involvement

AmsterdamUMCdb state in their released article that patient involvement was an important pillar in creating their open health dataset. This was done by inviting representatives of organizations concerned with ICU data and patient well-being after ICU admissions and focused on the importance and potential of sharing ICU data. In addition, the perceptions of the public to sharing patient data were examined to gain insights into their attitudes. No mentions are given about the involvement of patients in inclusivity and diversity, for example with patient representatives of minority groups (Thoral et al., 2021). The emphasis of patient involvement merely lays on privacy issues and looks at how the trade-off between sharing the data and the value it delivers should be made.

G3: Anomaly prevention

Since the data was collected during a large period (from 2003 to 2016) and no mentions of anomalies are provided by the publishers, it is assumed that no measures for anomaly prevention were taken and the data that was collected from the different sources were included in the dataset. It is important to mention that since the data covers many years; anomalies could have less impact on the usability of the data if they were to exist. In the case that anomalies were to be identified afterwards and are relevant to the consumer of the open dataset, the 'anomaly data' can be extracted and a large dataset will remain.

G4: External validation

In the public information about the dataset, it is mentioned that independent third parties have audited the project from a legal perspective to ensure adherence to GDPR. Similarly, there are

mentions of external ethics reviews but is unspecified what was examined in these reviews. In the end, approval was given by the Ethics in Intensive Care Medicine Group. Although ethics reviews were performed, it is unknown how the data collection methodologies and strategies are covered. The creators of the dataset emphasize the importance of having a worldwide network of intensive care departments. This connectivity to other departments can help with answering questions and addressing challenges in the ethical domain of open health datasets.

G5: Contacting possibilities

When looking at the possibilities of contacting the designers or creators of the dataset, an important obstacle can be noticed. Although the researchers are published on the website, there are no direct contacting options besides a link to their personal LinkedIn page. This requires a LinkedIn account and a connection to the person, otherwise the message will not appear in their inbox. On the contact page, a general email address is provided. The email address belongs to Amsterdam Medical Data Science, but AmsterdamUMCdb is not the only dataset they have created. Therefore, it is unclear whether substantial questions about the AmsterdamUMCdb, for example concerning the content of the dataset or the used data collection strategies, can be addressed by contacting this email address.

G6: Bias reporting

A data privacy impact assessment was performed to prevent biases in the dataset. AmsterdamUMCdb also decided to work with the 'duty of easy rescue' to safeguard the quality of the data. Obtaining informed consent from ICU patients is difficult since they are likely in a vulnerable and critical state, making them unable to give consent, and also because it concerns many patients that are included in the dataset. Therefore, the authors mention that selection bias is taken into account by not asking for informed consent. It can be assumed that the dataset does not contain selection bias in terms of who is and who is not included in the dataset. In addition, there are mentions of the presence of EHRs in the dataset since this can be seen as a disadvantage for countries that do not have the resources to use EHRs and thus can be seen as a bias in the dataset. Although AmsterdamUMCdb sees the added value of still including them, it is an important indicator for those who will use the dataset in non-EHR areas. There are no additional indications of detected biases or how bias is limited in the dataset given by AmsterdamUMCdb.

G7: Metadata inclusion

AmsterdamUMCdb has a throughout inclusion of metadata. The article provides an overview of the included variables and corresponding data types and also provides visualisations of variable distributions in the dataset. In addition, patient demographics are included but the demographics are limited to gender and age. It is clear where the data is collected and when it was collected. The data was collected in the period from 2003 to 2016. In addition, it is also known how the data was collected. The data is collected from the Amsterdam University Medical Center and comes from different databases as seen in Figure 10. This results in a large number of data types varying from notes by the medical specialist to radiology images of the patient.

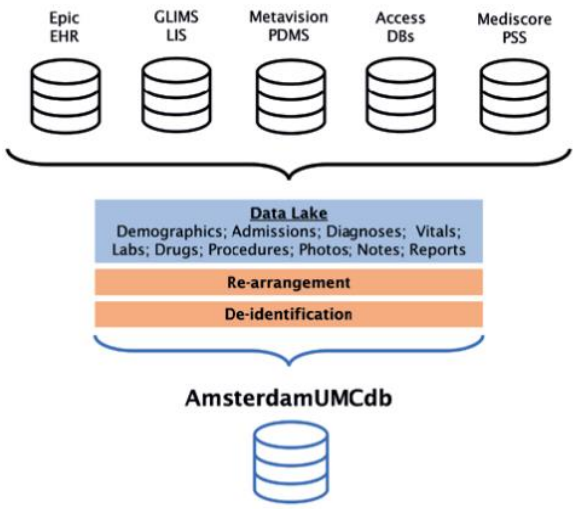

Figure 10 Data sources as input for AmsterdamUMCdb (Thoral et al., 2021)

The metadata is provided in the corresponding article of the dataset, which is easily accessible on the website of Amsterdam Medical Data Science. In addition, the dataset can be found in a GitHub repository which provides extensive descriptions of each variable. This makes it findable for open data consumers and allows them to be better informed about the origin and nature of the dataset.

6.3.2. eICU Collaborative Research Database

The eICU Collaborative Research database was created from the Philips eICU program, where data is collected to provide remote care givers with relevant information about the patient (Philips, 2022). The open dataset is also published with the reason to stimulate research initiatives for improving patient care, thus fitting them into the role of the *data prosumer*. The dataset has already been published. The website of the eICU dataset provides the data and corresponding documentation of the data (Pollard, n.d.).

G1: Diversity in design

From the available information, it is not possible to determine what the design team of the dataset looked like at the time of dataset development. It is known that Massachusetts Institute of Technology collaborated with the eICU Research Institute to create the dataset whereas Philips Healthcare provided the data. Otherwise, there are no mentions of more specific roles or departments within these institutions. This makes it difficult to determine whether the eICU has developed this dataset with a diverse composition of the design team.

G2: Patient involvement

The level of patient involvement during the design of the dataset is unknown. This is likely due to the fact that the data was provided by the Philips eICU Program, which implemented the technology in ICUs where the actual data was gathered. The ICUs decided whether or not to make use of the technology and may have incorporated patient perceptions to make this trade-off, but the decision-making process is not known at the time of writing. This could have influenced what and how the patient data is collected. The developers of the eICU dataset

justify taking a stratified random sample of the dataset provided by Philips to determine which patients are included (Pollard et al., 2018), but since it is unknown to what extent patients were involved during the data collection by the Philips ICU Program and how their needs and perspectives are incorporated, it is also difficult to determine the level of patient involvement for the eICU dataset.

G3: Anomaly prevention

The Philips eICU Program aims to improve patient surveillance at ICUs with a remote monitoring solution and collected the data in the years 2014 and 2015. The article does not mention specific instances that impacted the ICU admissions during these years. Since the eICU dataset is a random sample from the Philips dataset, it is assumed that no trade-offs were made concerning potential anomalies. This can either be because there were no abnormal periods or events, or they were not mentioned in the article. Therefore, it cannot be concluded that the dataset does not contain anomalies and whether the principle of anomaly prevention was taken into consideration for the eICU database.

G4: External validation

When it comes to external validation, it is unclear whether the eICU dataset is externally validated on data collection methodologies. Again, this is more difficult to determine because the data was not collected by the designers of the dataset themselves. The article does mention an acknowledgement for someone who made comments regarding the data, but the content of these comments is not (indirectly) described. The data was validated for integrity when receiving it from Philips, but this does not cover Philip's data collection methodologies. For that reason, the eICU dataset cannot be marked as externally validated.

G5: Contacting possibilities

The creators of the eICU database have provided two methods to get in contact in case of unclarities. First of all, the GitHub repository tab 'Issues' allows users of the data to raise an issue with the creators of the dataset. So far, 207 issues have been raised of which 123 have been resolved. The issues are categorized into seven groups: 'bug, duplicate, enhancement, help wanted, invalid, question and wontfix' (MIT Laboratory for Computational Physiology, 2018). At the time of writing this research, the most recent issue is from December 10th 2023. Important to remark is that not all issues have been answered, meaning that the data consumer may still have questions and potentially make assumptions about the data since their issue remains unsolved.

Besides the issue forum, there is an option to send an email to a general email address, but there are no people linked on the website who can be contacted directly. The website of eICU explicitly states that this email address is solely meant for database errors, such as privacy concerns where patient data is visible and identifiable. More general questions should be addressed as an issue in the GitHub repository. This shows that the eICU creators have made an effort to allow data consumers to contact them.

G6: Bias reporting

For the eICU dataset to report bias, there must be explicit mentions of limitations in the dataset that can result in biased outcomes when using the dataset e.g. for machine learning purposes. There is a remark about removing instances of the dataset when patients were admitted to low acuity units or step-down units. Step-down patients have received high intensive care but are admitted to a decreased, more intermediate level of care (Prin & Wunsch, 2014). This informs future data consumers of this dataset that the instances are limited to high acuity units only, which must be taken into account when their research context does require lower or step-down care units. There are no other mentions of potential causes for bias given by eICU.

Remarkably is the community-based approach that the creators of the eICU database aim to achieve. They encourage the consumers of the database to work as a community, requesting them to document their work and code after using the dataset and sharing this in the GitHub repository. When the data consumer identifies limitations of the dataset, they are motivated to include this in the documentation so it can be used by other researchers. However, this concept is dependent on the effort of data consumers and not on the effort of eICU who published the data.

G7: Metadata inclusion

The website and GitHub repository of the eICU dataset provide detailed information about the content of the dataset. The article from the creators of the dataset provides additional information. It is easily accessible and findable for data consumers who use this dataset. The variables are elaborated on. In addition, the demographics of the dataset are visualized including gender, ethnicity and age. It is known in what period the data was collected (2014 - 2015) and by whom it was collected: Philips Healthcare. The creators have decided to remove the hospital's name and instead opted for a general geographic region in terms of northeast, west, midwest and south (Pollard et al., 2018). This still indicates where the data was collected, but hospital-specific characteristics that could impact the dataset are removed by doing this. It also shows that not all hospitals in the USA are included since regions are missing from the four named above.

6.3.3. HiRID database

The HiRID database is a result of the collaboration between the Swiss Federal Institute of Technology and the ICU at Bern University Hospital. The latter has collected the data for a specific research area and it was not until in a later stadium that it was decided to publish the dataset for external usage. This categorizes the designers of the dataset as data prosumers. The website published the data and provided additional documentation (HiRID, n.d.). Since the dataset was created for a specific research question, the corresponding paper for that research is examined as well.

G1: Diversity in design

Although it is known that the data was collected during routine care at an ICU department, the composition of the design team is unknown. Retrieving the data from a multidisciplinary intensive care department results in many variables for many patients, but specific information about how these variables and patients are determined is not mentioned in the article. Depending on the level of diversity in the design team, variables and patients can either be selected or ignored when the researcher decides to explore a specific research area. This can be influenced by, for example, having an ICU specialist in the design team who knows what should be included, but the HiRID creators do not mention a similar consideration besides mentioning that variable selection is applied (Hyland et al., 2020). Therefore, it appears that there is no substantial claim to be made about the diversity in the design team of HiRID.

G2: Patient involvement

There is no information about how patients were involved during the data collection at the ICU. Hyland et al. (2020) state that there was no indication for obtaining informed patient consent since the data was used for a retrospective study. This could indicate that patients were not taken into consideration when the data was collected. In addition, since the data was collected from the ICU's Patient Data Management System (PDMS), in which all patients are listed that were admitted to the ICU, it is unlikely that patients had any influence on how the data was collected and whether this was reflected their values and concerns.

G3: Anomaly prevention

The dataset documentation does not elaborate on specific characteristics, events or periods that have affected the dataset. This can either suggest that there are no exceptional influences during the period of data collection (2008-2016) or that the design team did not explicitly document these instances. As a result, it is difficult to determine whether the data is representative of 'normal' patterns or whether it is not due to anomalies that are currently not listed.

G4: External validation

The authors of HiRID do not mention that data collection methodologies have been externally validated with similar organizations, for example, another ICU department in a high-density city like Bern in Switzerland and similar demographics. Nevertheless, it is stated that the creators have followed procedures from other open health datasets, such as the AmsterdamUMCdb (Faltys, 2021). These procedures, however, solely relate to anonymization practices to maintain patient privacy and not to its data collection methodologies. It is therefore assumed that the data collection methodologies, which are retrieving the data from the ICU's PDMS, were not externally validated.

G5: Contacting possibilities

The creators of the dataset propose a general email address to get in contact with them. It is unknown if contacting this email address reaches one of the creators and to what extent questions can be answered. Since HiRID also has a GitHub repository, it is possible to flag an

issue to the repository owners. This is similar to the eICU dataset. However, at this point of writing, there are a handful of issues raised but these are not addressed by the authors. The creators of HiRID do not mention that this is a method of contacting them which can be an explanation to why the issues remain unanswered. This shows that there are possibilities to get into contact with the creators, but is unclear who is behind the email address and what their exact role is or was during the design of the dataset.

G6: Bias reporting

There is no information to be found on whether the HiRID dataset could result in biases due to the limitations of the dataset. The consumer of this dataset is thus left to determine whether the characteristics of the dataset could imply potential biases when using the dataset for training a machine learning model.

G7: Metadata inclusion

The documentation that comes with the dataset provides a throughout set of metadata which elaborates on the variables that are incorporated, together with the used data types. As a data consumer, you are thus able to determine whether the variables incorporated in the dataset are relevant to your research question. Something that is missing in the documentation is distributions of important patient variables, such as gender, age and ethnicity. The metadata only explains what variables are included in the dataset but does not provide a visualization or table that elaborates on the distribution of a specific variable. This could hinder data consumers when the data is assessed for, among other things, usability and relevancy. The metadata that covers data collection policies and strategies is included and provides the data consumer with knowledge and empowers them to assess its usability. This concerns metadata about where the data is collected, how the data was collected and by whom the data was collected. This is mentioned comprehensively: the data was collected in a hospital that is called by name, and more specifically the department where it was collected is mentioned. In addition, the period of data collection is mentioned as well as the organizations that have collected the data.

6.4. Conclusion

The application of the decision framework allows us to take up the role of data consumer to evaluate the three open health datasets and how representation bias is addressed by the data prosumers. The scorecard in Figure 11 provides an indicator of how the open health datasets incorporated the seven guidelines from the framework.

From the findings of the use cases, it can be concluded that little is reported on the preventive measures ‘diversity in design, patient involvement, anomaly prevention and external validation.’ Since this could provide us, the data consumer, with valuable information on how representation bias was approached during the design of the dataset, it is unfortunate that little information on the five topics is provided. On the other hand, the documentation of two of the datasets (AmsterdamUMCdb and eICU) does provide information about bias and limitations in the dataset. It is remarkable that for metadata inclusion, all three open health

datasets offer a thorough overview of how, when and under what conditions the data was collected.

Guideline	AmsterdamUMCdb	eICU	HiRID
Diversity in design	Collaboration with medical institutions, no insight in the composition of the team	Not reported	Not reported
Patient involvement	Patient perceptions on data sharing, representatives of ICU organisations	Not reported	Not reported
Anomaly prevention	Not reported	Not reported	Not reported
External validation	Ethics review is performed, its content is unclear	Not reported	Not reported
Contact availability	General email address	General email address and online, interactive forum	E-mail and online, interactive forum
Bias reporting	Reporting on bias/limitations of the data	Reporting on selection bias and community based bias reporting	Not reported
Metadata inclusion	Includes demographics, where/how/when data was collected	Includes demographics, where/how/when data was collected	Includes where/how/when data was collected and selected variables

Figure 11 Overview of decision framework application (red: not reported, yellow: limited inclusion of metadata, green: substantial and clear inclusion of metadata)

7

Expert Validation

In this chapter, the decision framework as presented in Chapter 5 is validated by consulting experts. Incorporating the perspectives of relevant experts is important to gain insight into the efficacy, relevance and usability of the proposed framework. To begin, the interviewed experts are listed as well as the interview protocol that was applied. Furthermore, the results of the validation interviews are analyzed and the validity of the guidelines is presented. The summaries of the interviews are attached in Appendix A.5. This chapter builds further on answering the third sub-question: *"how can the developed decision framework contribute to representative data collection and use of open health datasets to protect patient's health equity?"*

7.1. Protocol for participants

Suitable participants for validation interviews are those who are working in the field of open data, diversity, medical machine learning and health ethics. Table 8 below presents an overview of the interview participants together with their background and how this relates to this research area. In total, four experts were interviewed using a semi-structured interview approach to allow for a deeper and comprehensive understanding when participants have additional information to share.

Table 8 Validation interview participants

Participant	Job title, background	Background in research area
V-P1	Researcher	Open science, transparency in research
V-P2	Assistant professor	Healthcare ethics
V-P3	Assistant professor	Machine learning, healthcare
V-P4	Assistant professor	Medical AI, ethics

To start with, the participant is asked to share their experience with open health data, (medical) machine learning and (representation) bias. This information provides useful insights into how well the participants may be familiar with the content in the decision framework. It also shows to what extent their view and opinions are credible and relevant for determining the validity of the decision framework. Then, the framework is presented. It is presented in the form of the textual guidelines as well as the decision framework for providers, prosumers and consumers of open health data. The materials used for presenting the framework are attached in Appendix A.6. This allows a throughout validation of each guideline by the participants. Subsequently, the participant is asked how well their understanding of the framework is and whether there are components that are unclear to them. In this case, these uncertainties would be addressed by offering a comprehensive explanation of the concepts to ensure that these uncertainties would not play a role in the rest of the protocol.

To continue, the relevancy of the decision framework is examined by the participant. This refers to the efficacy of the framework and how well it addresses and aims to stimulate diversity in open health datasets. The participants are requested to share their opinion on the strengths and weaknesses of the guidelines and whether something critical is missing in the framework, which is important do address before implementing the guidelines in real-world situations. It also refers to what positive or negative outcomes will arise when the decision framework is applied in practice. In addition, the practicality of the framework is validated by asking the participant how well the framework would perform in real-world applications. This refers to how well the guidelines will be useful for the involved actors such as ML developers, researchers and medical specialists and whether the guideline has its limitations when putting it into practice. It also examined whether the framework may raise challenges or issues for actors and how these can be addressed. Participants are also asked if they would make changes to the framework based on missing or redundant aspects and if so, what changes they would make.

To summarize, the validation interviews cover the general understanding of the decision framework, the relevance and usability of the designed guidelines and its practicality. This will result in new insights that can be used for adapting and improving the use of diverse open health datasets.

7.2. Validation interview analysis

After the decision framework presentation, the participants used their personal experiences, backgrounds and knowledge to assess the validity of the guidelines. This paragraph delineates each guideline and reflects on its validity.

Guideline 1 – Diversity in design

The validators see the value of having a team from various backgrounds and professional areas to improve the number of different perspectives in the team. They confirm that having people from different backgrounds in a team allows for more diverse thinking, although they find that the context also determines whether and what need for diversity is necessary in the

group. This is elaborated on by V-P2: *"The diversity in the team must be made up of the same people that the population is made up, so that they are attending to the right kinds of biases that might exist within that population. It should not be a criterium as 'two people from here, two from there."* In cases where this has happened, it had turned out to be difficult to justify why people were part of the team. This is illustrated by V-P1, who gave an example from their own experience: "*A few years ago, it was a criterium that we had to have a woman in the team to be more diverse so I asked a female colleague, to be able to make the application, but she could not make a unique contribution. Since we did not get the funding, we had to apply again the year after. During that time, we found a new [female] collaborating partner who could make a more genuine contribution, so we had to kick the first woman off the grant: how am I going to explain to her that she was in the team because we needed to have a woman on the team?"*. Furthermore, V-P3 emphasizes this by mentioning that in these cases, the names of the person who fulfils these diversity quota criteria are then nothing more than names on paper and they may not be part of the actual design since people do not want to spend time on something if it does not directly benefit them. This shows that simply following a quota can work counterproductive, but substantial motivations for how the composition of a design team should look can contribute to a more diverse open health dataset. Therefore, adhering to the guideline of diversity in design can help designers of open health datasets by creating a dataset that incorporates wider perspectives and safeguards the values of those who are affected by the dataset.

Therefore, it is important to consider the context in which the open health dataset will be developed. To illustrate, V-P2 gives an example of the videogame development industry: *"There is evidence about the importance of having different designers in the room for pilot testing and prototype testing. In the video game, the cameras would not pick up black people's movement as quickly as white people. This is almost directly due to the profile of the average software programmer in the gaming industry; there were no black people in the office to test it against."* This problem was likely to have been identified earlier if the design team had consisted of a more diverse composition. To avoid things from happening from the beginning, as mentioned by V-P3, it is essential to identify the environment and conditions in which the open health dataset will be created. This will result in relevant and accurate measures for the composition of the design team in which the contribution of the team members will be substantial and allow for inclusive, diverse and broader perspectives on how the data should be collected.

Guideline 2 – Patient involvement

When it comes to involving patients during the design of open health datasets, the validators find this to be an important pillar. Patients know what is pertinent for and to them when collecting patient data and are therefore important to consider during the design of an open health dataset. This is amplified by V-P1, who states that *"patients are good at identifying patient-relevant outcomes that should be collected in a dataset."* Although the exact applications of the data are not always set at the beginning, the data that is collected is patient-specific, e.g. it is collected from ICU patients or all patients from a hospital. This

information on its own demarcates the specific set of patients and thus also sketches the context. This is highlighted by V-P2, who states that "*patient involvement is specially important when you create a database about a specific set of patients. Then, it is even more important to invite them to understand how the data is interpreted and used. Not simply for 'having them on the team' but to include their different perspectives that could identify biases that would not be visible otherwise."* Subsequently, this validator makes a comparison to data collection in an urban context where patients are asked to provide input on how they view different parts of a city. The next step could simply be to collect this data and analyse it for the purpose it was collected. However, involving the people who provided the input data can result in a more detailed and comprehensive understanding of why and how they gave their data as input. This way, important variables can be identified that are missing or nuances about existing variables can be made. Without doing so, the data is open to interpretation for those who use the data. Connecting this to open health data, patients should be involved to understand how the data is used and interpreted by the designers of the dataset. This does not only inform the patients about how their data is used and interpreted, it also stimulates to generation of new perspectives and interpretations of the patients. This can be essential for improving the diversity of the dataset.

However, it is mentioned that although patient involvement is a good practice for stimulating diverse thinking in the design team, it may also be difficult to achieve. Patients could not easily be reached since they might not be willing to participate in healthcare research. Although this can have various underlying reasons, V-P4 illustrates cultural consciousness as an obstacle to reaching patients. They state that the African American population has a lot of prejudices against the American health system because they have been victims of health trails where "*they have been used"* which is now part of their cultural consciousness. This is going to make it very difficult for them to participate. It is thus important not only to look at patient involvement but also at how the patients are going to be reached and motivated to participate in the design team.

Guideline 3 – Anomaly prevention

The participants find that preliminary identification of anomalies should be considered during the design of an open health dataset. It is not something that is done at one point during the design, however, there should be continuous active thinking about what anomalies could happen. As V-P3 states: *"it should be an iterative procedure, of course, you have some ideas about what anomaly can happen in the beginning. But along the way you identify things, then you just reiterate and try to make it as clean as possible."* The information about anomalies is valuable for the consumers of the data since they know what exceptional periods, events or incidents have occurred during the collection of the data.

Although V-P2 agrees with the identification of these instances, they are hesitant to call them anomalies. Rather, they see them as *"contextual constraints on the usability of the data, since they will prevent the generalization to other contexts but will also create bias in the dataset towards that context."* They find the constraints highly important to identify and suggest that

this process should be iterative to ensure that no constraints are missed. The two participants, V-P2 and V-P3, highlight the need for including these anomalies in the metadata since it provides information about the context in which the data was collected and therefore informs the data consumer about additional limitations and constraints of the dataset.

Guideline 4 – External validation

A first remark that must be made, is the term 'external validation' in the context of open data with intended use for machine learning models. Since external validation for ML models refers to testing its performance on a new, for the model unseen dataset, it can be confusing when using the term for externally validating the dataset. However, with additional clarification, it did not cause further confusion among the participants.

Externally validating the dataset, which consists of examining data collection strategies and methodologies, was seen as valuable by the participants. According to V-P1, *"it ties in with a broader discussion on how to check the quality of data. In my opinion, we do not really have a lot of good mechanisms to do this today. In an ideal world, we would have some kind of independent mechanisms and standards or a third-party organization that could check and verify data that people put out there, but they are not there. This is a direction that I think is worthwhile."* This suggests that validating your data before publication can improve the quality since generalized or biased assumptions made during the data collection have a higher chance of being identified when this is validated with an external party. In line with this, V-P2 states that the criteria for the external validator should be delineated and the validator checking the data collection process should be related to the context of the to-be-checked dataset: "*you do not want the dataset to be valid in all contexts, otherwise it is useless. What do they need to have in common and what can it be tested against? […] For example, the population upon which it was developed and the population upon which it is being tested must be similar representationally speaking. It is a test for external validity, but not just a generic usefulness test."* This amplifies that generalization is not the primary goal of externally validating the dataset, but it should be tested if it can be generalized to other, similar contexts. By doing so, the biases that have occurred during the data collection will show and thus improve transparency on how the data was collected.

Guideline 5 – Contacting possibilities

The participants find that having a designated contact for an open health dataset is a useful feature. It allows the consumer of the dataset to clear up any confusion or to have their questions about the data answered. As V-P2 mentions, it is *"absolutely important, this would be continuity of the information and being able to go back; more like open science"* since after the publication of a health dataset, it is highly likely that consumers of the data have additional questions or things they would like to see clarified. This is amplified by V-P1, who states that it would be a great addition if it could be implemented.

Although the additional value of having a contact available for an open health dataset is confirmed, it also raises questions for the participants. The questions come from an organizational perspective and entail whether a full-time person should be hired to be available

as a source of information, but also how long this person should be available since it could be unreasonable to expect someone to be available after, for example, ten years of publication although according to V-P1 *"there could be exceptions to this"*. This will require careful consideration of what is seen as an acceptable period and how the responsibilities of the contact are carried out. If this is not done carefully, it could work counterproductive. As V-P1 states: *"it could also put too high of a burden on people and discourage them from data sharing. But still, it is better if they share the data without having someone than don't open the data at all."* This shows that the need for publishing the data is higher than the need for contacting possibilities, but overall, the participants find it to be a valuable and useful contribution when the publisher of the dataset allows for direct communication with the consumers of the dataset.

Guideline 6 – Bias reporting

The importance of reporting existing biases when publishing open health datasets is confirmed by the participants. Having the information about biases in the dataset allows the data consumer to make better-informed decisions. As V-P2 states: *"You are already in a better position than you were before even if the bias still exists, you know what to do and how to work around it. Having this kind of guidelines and having the information available and accessible is sometimes the best and only thing you can do because you cannot eliminate the bias in the dataset."* This statement is strengthened by V-P3, who finds that having more awareness of biases in datasets is important.

It also raises questions about what is defined as bias or simply a feature of the dataset. This is remarked by V-P1, who states that it is difficult to look at a dataset and tell which features are biased and should therefore be reported as potential biases. An example is given by V-P2: *"If I start collecting data at a hospital nearest to us, then we will get a population that is broadly Dutch. You could ask whether that is a bias or more a feature of the data that is unavoidable."* This makes it difficult to determine which features are biased data and in this case, should be reported as bias or as metadata that refers to where the data was collected. Therefore, the same validator states that "*you are not necessarily exposing the biases in the data, you are also as much identifying the limitations of its applicability. You want to identify where those biases might occur so that we are aware of whether or not it is applicable. It is not so much about identifying problems in the data, it's limitations as a dataset and every dataset has those."* Therefore, instead of reporting biases, it is suggested by them to see this guideline as a limitation section of the data. V-P4 illustrates the concept of man-made decisions as a bias by itself, making it difficult to report on: *"data obtainment is an activity where we describe the world, the world is full of information and we pick certain elements of that information. The problem with that is that most of the time the information we pick is in on itself man-made."*

Guideline 7 – Metadata inclusion

The participants see the inclusion of metadata when publishing an open dataset as an essential feature. Currently, it is an ongoing problem where data is used without having the ability to examine its metadata. This is highlighted by V-P4: "*one of the most crucial elements*

of the framework is the metadata. […] Datasets are downloaded to develop machine learning algorithms et cetera, but they have no idea where this stuff [data] comes from: 'who did obtain the data, how was it obtained and what is the data actually about?'" V-P3 amplify this by stating that metadata is very often missing, resulting in the fact that you are not sure what you are working with. This potentially leads to harmful assumptions that are made about the data. As mentioned by V-P4 above, it is important to include metadata concerning who collected the data, how it was collected and also what the data consists of. Nonetheless, the decision-making by the person or team collecting the data determines how the open health dataset is going to look and could therefore indicate biases on its own. Including this in the metadata and illustrating how data was collected "*underscores how much decision-making is present in data obtainment*." The importance of having metadata about what the data consists of is also highlighted by V-P1, who states that *"the metadata features that are particularly important for diversity, such as population characteristics"* should be included in the metadata.

Something that must be considered according to V-P2 is the ongoing lifecycle of including metadata. It can be seen as continuously increasing. In the case where something was missed and is not included in the dataset, the metadata should be adapted. Metadata inclusion should therefore not be seen as a static task but rather as an ongoing dynamic process.

7.3. Validation results

After examining the validity of the separate guidelines, the overall validity of the decision framework was examined with the participants. This paragraph touches upon the understanding, relevance, usability and practicality of the framework. In addition, the challenges of the framework are discussed, as well as the recommendations given by the participants.

When it comes to understanding the terminology used in the framework, the guideline for dataset external validation caused some confusion due to the similarities with a term used in the same domain. Since this required additional explanation to the participants, it affected the clarity of the framework. It is thus necessary to make adjustments in the term for this guideline to prevent future hesitation in understanding the guideline. Those who use the guidelines in the future will likely experience similar confusion if nothing is to be adjusted. Overall, the understanding of the decision framework was consistent among the participants. Other than the issue mentioned above, there were no issues regarding the clarity of the framework.

Looking at the relevancy of the decision framework, it is examined to what extent the participants find the framework suitable for approaching the problem of addressing representation bias in open health datasets. V-P3 found that the difference between the three stakeholder types, data prosumer, consumer and provider, fills a missing spot since there is often a single direction that does not apply to everything. The distinction that is made allows for the identification of relevant guidelines for different stakeholders but still incorporates the

social context in which these stakeholders can be found. Besides solely addressing open health datasets that are to be developed in the future, participant V-P3 find the guidelines also reasonable to label existing datasets and how good they are in terms of representation bias. Often, you do not know what you are working with and what biases are there, although it is known that the biases are likely to be present in the dataset.

Additionally, V-P2 sees open health datasets as an effective solution for researchers as it allows them to effectively reuse datasets instead of continuously accumulating new datasets or even using the same datasets in different ways. Therefore the open health data must be usable and suitable to the researchers. As mentioned by this participant: *"I am all about awareness of complexity rather than trying to reduce complexity, because I don't think you can reduce it in this case."* Since the guidelines touch upon both creating new datasets and assessing existing datasets, they aim to increase the awareness of representation bias in the dataset which provides researchers with adequate information about how the open dataset is suitable for their research. This is also highlighted by V-P3, who mentions that it might not reduce bias on its own, but as the guidelines create awareness, it makes the stakeholder think about the problems and what could be missing in their approach: *"if everybody follows the same thing [the guidelines] and looks at the same points, then we probably have a chance of having less bias."* Lastly, V-P4 highlights that approaching this problem in the form of guidelines is an appropriate and interesting way.

The usability of the decision framework refers to how well the framework is usable in real-world scenarios, whereas the practicality of the framework entails the feasibility of implementing it in the healthcare domain and how this may require additional resources. It is mentioned that the guidelines can be used as a checklist to make sure that when designing an open health dataset, or assessing an existing one to use it, nothing important was missed. On the other hand, they can also be used in a more normative way as proposed by V-P1: "*The guidelines lay a base to check research that has been published and put out there, and to see how well they have followed the guidelines and how they could be improved based on these different metrics. It helps to be concrete and specific. I imagine myself looking at a piece of research, are the guidelines sufficiently detailed, then I can tell if the research meets the criteria or not."* This is also validated by V-P2, who find the decision framework to be usable to determine what dataset has put adequate effort into identifying and possibly removing potential biases in the dataset. In addition, V-P2 mentions that the information generated when applying the decision framework helps to make an informed decision on the usefulness of the data. Adhering to the framework does not necessarily result in bias-free datasets, instead, it provides the consumer of the data with informed usefulness. This is what will increase the usability of the open health datasets. As stated before, the framework consists of guidelines that require resources from the data providers and prosumers, such as incorporating a contact person for the dataset. In addition, involving a larger number of people in the design team also requires financial and organizational resources. This is mentioned by V-P3: *"often, it is not a team of people sitting and thinking 'what sort of database do I want to have?' It is the same people with the same backgrounds, so the resources will be an issue when*

you want to follow diversity in design. I would love for it to happen, not only in healthcare, but in practice it is very challenging because it is extra resources that every organization needs to allocate." In addition, external validation of the dataset will also call for valuable resources since external parties must be identified and tested for generalizability before validation can take place. Although the contribution of these guidelines is validated, they will affect the practicality of the framework.

Challenges and recommendations

The healthcare domain is seen as a challenging domain when it comes to implementing changes in the workflow. This is because people in the medical field are more opinionated than in other fields and the fact that they hold on to the specific ways in which they are doing things, according to V-P3. Therefore, the actual implementation of the decision framework could experience resistance. In addition, V-P4 mentions that, although the guidelines have the best intention, they also complicate medical research: "*I know that doctors already have a lot of burdens, they would be hesitant to embrace and accept the guidelines since they are amongst now other lots of guidelines that regulate for example, artificial intelligence."* Another challenge that is highlighted by V-P3 is the fact that in healthcare, the datasets are collected the way they come in: "*they have the measurements from the patients and it is basically what patients they have and what patients have given consent for the data to be collected from them*." This shows that the decision framework might not always be applicable.

For V-P4, it is clear that although the guidelines can be part of a governance structure, additional steering is needed for this problem, e.g. ensuring that open health data actors adhere to the guidelines. This also builds upon the challenges identified by the validators.

In addition, it was recommended that it would be helpful for the framework to include a checklist or criteria that make it easier to determine whether a guideline is adhered to, both for the designer of the dataset and the end-user of the dataset. Examples were given for criteria on external validation to determine when an external party is similar in context and thus applicable for validation or criteria for how long a contact person should be appointed. In addition, including methods to reach patients for the 'Patient involvement' guideline is seen as desirable for the validators since they have encountered this issue before. Lastly, V-P2 suggests that re-evaluation of the guidelines for bias reporting and anomaly prevention could provide new insights into whether these guidelines can be merged to some extent since both guidelines touch upon the limitations of the dataset.

7.4. Implementation of validation suggestions in decision framework

The decision framework as presented in Chapter 5 will be altered to show the progress that was made during the evaluation of the framework. In Figure 12, the adapted decision framework is shown and includes visualization coding to indicate the changes that were made to the previous framework. It incorporates the suggestions made by the validators and the findings from the use case analysis as underlined text in the figure. A desirable and now implemented function of the framework was to incorporate a checklist-based approach that

guides the user of the framework through the content of the guideline, as seen in purple text in Figure 12. For each guideline, the additions and/or changes are discussed. Figure 13 in Appendix A.6 displays the validated decision framework without the visualization coding.

For '*G1: Diversity in design*', no changes were made. Of course, it remains important to define the purpose of the dataset and to determine the research population. The design team should consist of at least representatives of this research population.

For **'***G2: Patient involvement'*, an important addition to the framework is how patients can be reached for inclusion during dataset design. Patient interest groups and patient federations can mediate by examining their patient database that consists of patients that are willing to participate in medical research.

The guideline 'Anomaly prevention is adapted to '*G3: Contextual constraint identification*' to prevent ambiguity and to provide transparency about its content. These constraints limit the generalizability of the dataset, although this is not always an issue, these constraints must be proactively considered and monitored during the data collection.

For **'***G4: External validation'* it is prescribed that criteria should be formulated to ensure accurate validation of the data collection methodologies. Although this is strongly context dependent, there are criteria that are important to consider in any case, for example sharing similarities with the validator in population size, ethnicities, age and socioeconomic backgrounds. By checking whether these criteria are fulfilled, appropriate external validators can be consulted to identify biases that may have occurred during the data collection.

'G5: Contacting Possibilities' addresses the need for specifying a period for when a contact (and a potential substitute) must be available, which is dependent on the resources of the data provider/prosumer. Having a direct contacting possibility such as e-mail will improve the communication between the data provider and consumer. In addition, lessons can be learned from the use cases, which offered a community-based contact approach where data providers addressed questions and had the answers visible to everyone.

Lastly, the guideline *'G6: Bias reporting'* is merged into '*G7: Metadata inclusion'*, with *'G6: Metadata inclusion'* as the result. This guideline stimulates to report on the adherence to the previous guidelines, on the population characteristics of the open health data and to reflect on data collection methodologies. This guideline has been merged with bias reporting as both refer to reporting on the limitations on the generalizability of the dataset. It is also important to incorporate new limitations as identified by data consumers in the metadata to ensure an iterative process of relevant and accurate metadata.

The above guidelines assist data consumers with making a more informed decision about whether the open health dataset is usable and relevant for their research purpose, in terms of representation and diversity.

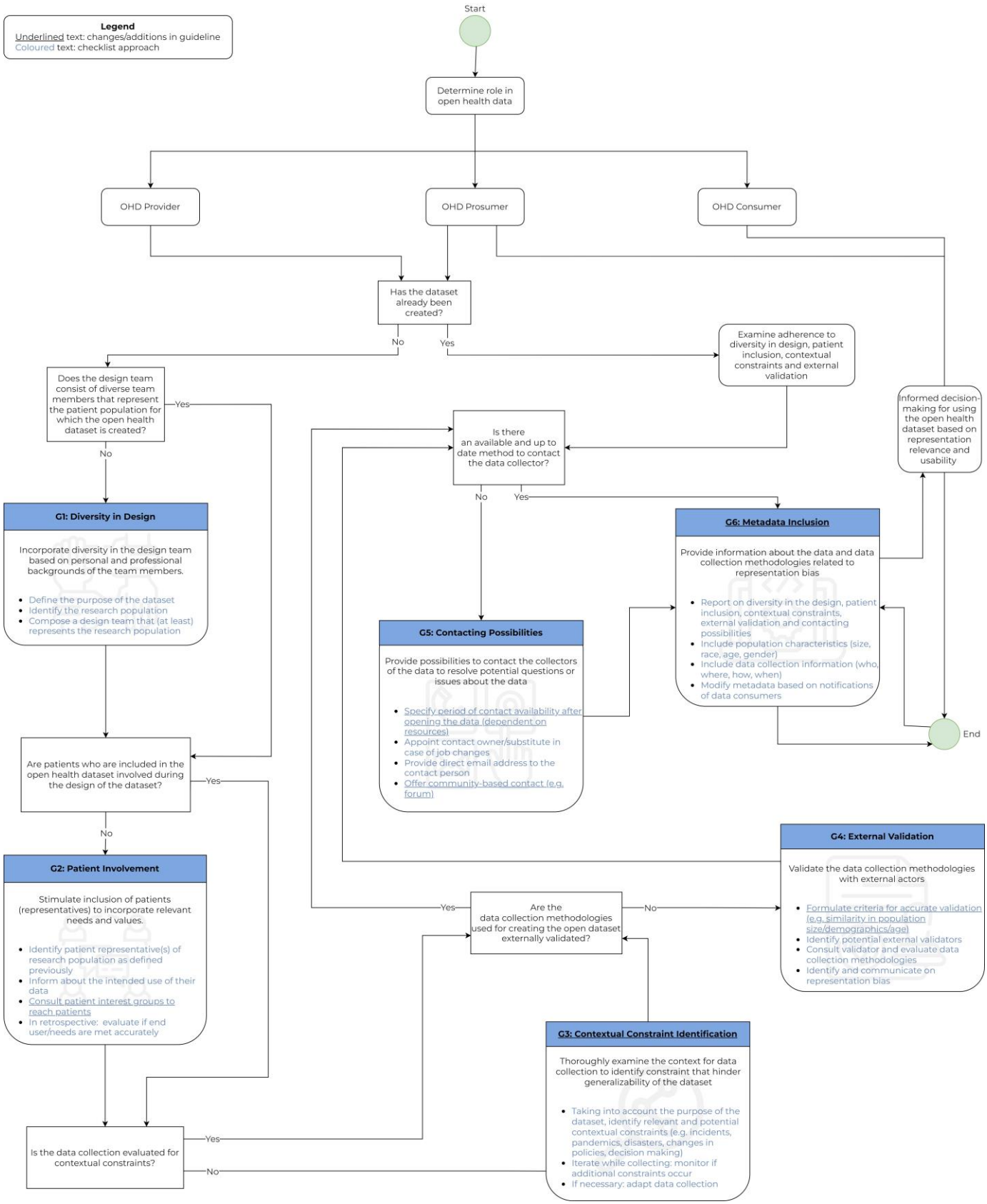

Figure 12 Adapted decision framework as a result of use cases and validation interviews

8

Discussion

This chapter presents a discussion of the findings of this research. Firstly, the results of this research will be discussed and are followed by a reflection on the methodologies used in this thesis. In addition, we determine the validity and reliability of the conducted research. Finally, the limitations of the study are given.

8.1. Results

The literature review was aimed at identifying the types of representation bias that occur in open health data and how this had ethical implications for those that are misrepresented in the data. Four types of representation bias were identified: gender, racial/ethnic, geographic and socio-economic bias influence the fair treatment of patients from misrepresented groups in the data and should therefore be addressed. In addition, by examining the social context, we can determine roles and responsibilities that shape the environment of how open health data is collected, published and used. This draws from research by Ubaldi (2013), who elaborates on the different ecosystems of data providers and users in the context of open government data. This can be extended to the field of open health data, with an additional role of the data prosumer: the actor who firstly uses and sees the data as intellectual property, and later decides to open the data to transform it into a common good. In the context of open health data, the separation between data providers, consumers and prosumers is especially important. It incorporates how health institutions that may already collect and use their data, decide to give the public access to the data, although this is subject to certain conditions such as anonymizing and ethically reviewing the data publication. Although the literature study allows for a useful insight into the concepts of representation bias types and open data roles, there was a gap between the two concepts in existing literature. There are no mentions of how data providers, consumers or prosumers use their power, knowledge or tools to address representation bias in open health data.

Conducting the exploratory interviews allowed for a deeper and more nuanced understanding of how representation bias is manifested in open health data, where challenges arise and how these can be tackled. Something remarkable to mention is that the interviews showed a shift in perspective on bias and diversity in the context of open health data. Whereas it is important to include minority groups and ensure fair treatment and protection of patients' health equity,

the framework should not aim to eliminate bias in general since bias can also be seen as desirable in a medical context. It also highlighted the difficulties in data collection for research. Since data collectors are unlikely to collect additional data that would not be used by them, it is undesirable and unfeasible to incorporate measures that would require them to do so. Although Polevikov (2023) calls for modifying data collection policies to mitigate bias, the exploratory interviews revealed modification constraints that incorporate the context and purpose for which the open health data is intended.

Given the former, the decision framework was designed which incorporated the different open health data roles as well as the representation bias types in the form of guidelines. The decision framework was applied to the use cases, in which three open health datasets were assessed on their adherence to the guidelines. Remarkable is that although the lack of metadata inclusion was seen as a large challenge in existing literature, among others by Thornton and Shiri (2021), and the exploratory interviews, it was not evident in one of the use cases. This showed that all three open datasets adhered well to the inclusion of metadata. This unexpected result could be attributed to the fact that the HiRID dataset based some procedures (such as anonymization) on the AmsterdamUMCdb dataset, whereas the AmsterdamUMCdb dataset took inspiration from the eICU dataset. This could likely have led to an unintended selection bias in which the lack of metadata is not visible, even though Wang et al. (2021) state it is one of the essential pillars for using open data and is often insufficient.

The validation interviews played an essential role in the refinement of the decision framework. The refinements are based on real-world scenarios and instances as illustrated by the interview participants and shed light on the dynamic, complex challenge of addressing representation bias in open health data.

'G1: Diversity in design' is aligned with the need to address the composition of the development team, as highlighted by Gichoya et al (2021). By adherence to this guideline during the development of the dataset, it can target gender, racial and socio-economic bias in the final open dataset. It builds further upon Cerrato's (2022) statement that a team representing minority groups is more likely to incorporate the perspectives of these groups. A nuance to this must be made as this guideline stimulates to sketch the context in which the dataset may be applied and to incorporate those who could be affected by it, it does not determine a specific set of criteria that is seen as diverse.

'G2: Patient involvement' addresses gender, racial, demographic and socio-economic bias as well, and takes into account the perceptions of the patients. This fits under the umbrella of Obermeyer et al. (2019), who identified representation bias in a dataset due to misunderstandings on how people of colour were using the healthcare system and MacIntyre et al. (2022) who concluded that health literacy is affecting how patients express their care-seeking behaviour. Although this guideline contributes to addressing representation bias, it also raises questions on how patients can be reached for involvement in medical research. It is suggested that data providers/prosumers can reach them through patient federation groups as these groups have an overview of people willing to participate in medical research.

However, it must be remarked using such groups can result in selection bias since those who are not willing to collaborate in medical research are unlikely to be part of such groups.

'*G3: Contextual constraint identification'* and *'G4: External Validation'* do not find their origin in the examined literature. The decision to design these guidelines was made based on the findings derived from exploratory interviews with experts in machine learning and research data re-usage. Their credibility and practical knowledge highlight the relevance of including these guidelines in the decision framework since they aim to address selection and sampling bias in open health datasets. As the foundation for these guidelines is not found in existing literature, it is interesting to examine how this can be explained. This could potentially be attributed to the focus on representation bias in existing open health datasets. Since the volume of open health datasets is relatively small, it is not unlikely that this type of representation bias has not manifested itself yet in the studied literature. However, this does not imply that data providers/prosumers should not be concerned with incorporating measures that address selection and sampling bias in their open dataset. Therefore, these guidelines provide a new direction for future research.

The guideline *'G6: Metadata inclusion'* aligns with findings from Wu et al. (2019) who state that the usability and relevance of a dataset should be assessable which is an important factor for data consumers. It stimulates data providers/prosumers to include metadata that refer to representativity and diversity to improve awareness of potential representation bias that is manifested in the data. By doing so, this guideline addresses all the six types of representation bias that have been identified in this research. It is the overarching factor that incorporates the content of the previous guidelines. This guideline fills the gap in Wang et al.'s (2021) research, who touch upon including metadata but does not explicitly refer to representation bias as a part of the metadata. This is similar to the guideline for *'G5: Contact Possibilities*' which serves as an important component of the metadata. It is seen as a separate guideline since it requires the data provider/prosumer to appoint and organize a method of contacting, however, the end-product which is the contact person and their contact details, is to be included in the metadata.

8.2. Limitations

The first limitation of this research is that the field of open health data is in a relatively early stage. Although a significant number of datasets have been opened to the public, there are still milestones to be reached. This has notably impacted the research in the sense that interview participants were difficult to find. This required a creative approach in combining the insights from people with open data backgrounds, medical ML backgrounds and backgrounds in ethics and injustice. In most of the cases, the participants were familiar with the concept of open data but not with its application in healthcare. It is expected that as open health data finds its way into a more general adoption, it is self-evident that more people will become familiar with it.

In addition, the identification is representation bias in the literature review is based on preexisting open health datasets. This resulted in a handful of types of representation bias, but this does not indicate that representation bias is limited to solely these types. As more open health datasets will be released, it is important to consider that this could entail new origins and types of representation bias. Although this research provides a solid base for assessing representation bias, it is encouraged to continuously evaluate for additional representation bias types in open health datasets that will be released in the future. This will allow for a more complete view of representation bias in open health datasets.

Moreover, the aim of this research is focused on the creation and usage of open health datasets. Due to feasibility and time constraints, it was not within the scope of this research to examine the other components that come with open data, such as maintaining, cleaning and storing the data. Since these components could significantly influence the way that data providers/prosumers engage with their dataset, it is crucial to recognize that the decision framework is limited on these components and might need caution, considerations and alterations before implementing it in the full chain of open data to obtain a more complete view. In addition, there is an important remark that must be made when it comes to the scope of open health data in the machine learning pipeline. The decision framework was designed to address representation bias in open health data to train medical machine learning models. It was however not designed to address other bias forms than representation bias, such as the other bias types that occur in the machine learning pipeline as seen in Figure 2 in paragraph 1.2.2. Although the decision framework is a valuable tool for addressing representation bias during the development and use of open health datasets, it does not contribute to removing or addressing other types of bias such as aggregation bias and evaluation bias. It is therefore important that the user of the decision framework is aware of this limitation. The user of the dataset must still address, take responsibility and be aware of bias that may slip into the machine learning model.

Furthermore, the research methodologies open up a discussion for limitations. As briefly mentioned before, the use cases that were examined were influenced by each other. The AmsterdamUMCdb based its approach on how the eICU was created, whereas the HiRID dataset took inspiration from the AmsterdamUMCdb. Potential strengths or flaws in the approach of one of them could have been incorporated into the other, resulting in a misrepresented view of the cases. This research could be improved by assessing open health datasets with the decision framework that are not necessarily designed with inspiration from each other. To illustrate, from the literature study and exploratory interviews it was concluded that the lack of metadata is a large issue in open health data but this was not evident from the use cases. It is difficult to determine whether this is generalizable to open health datasets. This puts a limit on how the guideline for metadata inclusion is seen as valuable and relevant in the decision framework. Something that must be remarked is that the design of the exploratory interviews may have nudged participants towards a specific solution area due to interview questions with a suggestive undertone. For example, mentioning the concept of metadata proposes a direction for a potential solution. Although these directions were

relevant and based on existing literature, it may have limited the creative and problem-solving approach of the participants. This research could benefit from expanding the interview questions and reducing the suggestiveness to open up the design space even more.

To conclude, open health data can serve as a valuable source of training data for medical machine learning when rightly incorporating measures that focus on addressing representation bias. Although the healthcare domain is at an early stage of large-scale adoption of open data, the decision framework provides a valuable tool in representative data collection and usage and will be enhanced when overcoming limitations related to generalizability, completeness and detailedness of the framework.

9

Conclusion

SQ1: What are the different types of representation bias that occur in open health data used for medical machine learning models and what are their ethical implications for those at risk?

To analyse the ethical implications of representation bias in open health datasets, it is firstly important to determine what different types of bias occur. This research question is scoped towards existing open health datasets to identify shortcomings and implications for future dataset design. A comprehensive literature review in combination with exploratory interviews allowed us to analyse the use of open health datasets for machine learning purposes. From this, the following representation bias types were identified:

- Gender bias
- Racial/ethnic bias
- Geographical bias
- Socio-economic bias
- Selection bias
- Sampling bias

Ethical implications of gender bias in open health datasets reinforce the already existing environment in which people with specific genders are misdiagnosed. This can lead to incorrect and thus unfair treatment, potentially causing harm to the patient that is dependent on the machine learning model trained by a gender-biased open dataset. Racial bias leads to maintaining the current system in which people from racial backgrounds are disadvantaged (Polevikov, 2023), keeping up the systemic and historical discrimination. Geographical bias affects those areas that do not have the resources to use open health data. Using, but also collecting, maintaining and processing open health data requires infrastructural, organizational and financial assets. These resources might not be sufficiently present in poorer countries or regions, whereas well-developed countries do not have these obstacles (Celi et al., 2022). This implies that applications of open health data will be focused on populations of prosperous countries, whereas resource-poorer countries may not benefit from these applications and the advantages of open health data in general. Socio-economic determinants can cause bias as these determinants affect the care-seeking behaviour of patients, for example through health literacy of patients. Patients with low health literacy experience multiple obstacles in seeking medical care and are thus less likely to be included in datasets (MacIntyre et al., 2023). This affects the application area of the open health datasets

and eventually exacerbates existing social inequalities. Lastly, selection and sampling bias have been identified. Selection bias can occur when the criteria for data collection are not representative of reality. This could lead to inaccurate representation of patient population groups and disadvantage the misrepresented groups. Sampling bias causes misrepresentation in a dataset, for example by misrepresenting the research population or the overall targeted research period. ML models that are trained on datasets with this bias have unreliable outcomes, as they are not representative of the real-world situation. This may have ethical implications when the sampling bias is not identified and outcomes are seen as trustworthy and reliable, depending on the context in which these outcomes are applied, the consequences can be severe.

SQ2: How does the existing social environment influence how open health datasets are created and deployed for real-world application in healthcare?

After conducting a literature review and several exploratory interviews into the social environment and context in which open health datasets are created resulted in a threefold division of stakeholders with each their own roles, responsibilities and values when it comes to the design and use of open health datasets (Ubaldi, 2013) which influences their motivation for interacting with open datasets. This suggests that a tailor-made approach could touch upon these motivations. Three levels have been identified:

- Primary level: open health data prosumer
- Secondary level: open health data consumer
- Tertiary level: open health data provider

The primary level consists of those who collect data for their own purposes and therefore use the open data, but also decide to open the data. The decision to open the data does not necessarily have to be made from the beginning. Examples of primary actors who find themselves at this level are hospitals and medical institutions. Moving to the secondary level, we find the open health data consumer. This actor depends on the data that is provided by the first level, as well as the third level. Secondary actors mainly include researchers and academia who use the open data for their own research. Lastly, the tertiary level is concerned with open health data providers. The providers are not concerned about using the data themselves, instead, their focus is on gathering the health data for publication. These actors can include governmental organizations or statistical bureaus. The three roles affect how open health datasets are published since providers and prosumers of open health data can influence how their dataset is created, whereas data consumers can alter the existing dataset to their own research context but do not have any influence on how the dataset was created and deployed for public use.

This results in a relation between the data provider/prosumer and consumer, since the provider of the data has published the data with the intention of it being used for future research areas or purposes. The publisher allocated their resources to publish the data, but the assessment of the usability and relevance of the dataset is done by the users of the data; based on what data it concerns, how the data was collected, where it was collected and by

whom it was collected. This information can be included in the metadata, but according to Thornton and Shiri (2021), metadata is often incomplete, insufficient or of poor quality in the context of open health data. Therefore, it is not only the data consumers that benefit from high-quality and reliable metadata, it is also beneficial to the data providers and prosumers since it increases the usability of their dataset and thus justifies their allocated resources.

It is important to ensure that these different levels collaborate as this allows for incorporating the needs and desires of the data consumer, such as high-quality metadata, but also improves transparency on resources and limitations from the side of the data provider. In addition, data providers and prosumers require internal collaboration. This refers to including people from different backgrounds when creating an open health dataset to ensure that the needs and values of those who are affected by the dataset are safeguarded and correctly represented.

SQ3: How can the developed decision framework contribute to representative data collection and use of open health datasets to protect patients' health equity?

The decision framework aims to address the viewpoint of the data providers/prosumers, as well as from the data consumers. From the current issues and shortcomings as identified in the literature interviews, six guidelines have been designed:

- Guideline 1 – Diversity in design
- Guideline 2 – Patient involvement
- Guideline 3 – Context constraint identification
- Guideline 4 – External validation
- Guideline 5 – Contacting possibilities
- Guideline 6 – Metadata inclusion

The framework contributes to improving representative data collection but also allows existing datasets to be evaluated based on their used data collection strategies. This increases the usability of the framework as it is relevant to all three stakeholder levels in this problem domain. Preventative measures cannot be taken for existing datasets unless the data provider decides to re-evaluate and re-design the dataset. In those cases, the preventative measures can contribute to improving the representativeness of the dataset. Corrective measures come into play when the data has been collected and (is ready to be) published. It must be noted that the framework must always be placed in the context in which the open dataset will be designed as this determines the exact interpretation of the guidelines. To illustrate this with an example: when an open health dataset is created for women with cervical cancer, the framework does not aim at having a representational dataset in which men are also included. Instead, it is aimed at representation among women with cervical cancer. This should be ensured in the dataset to allow fair outcomes and protect the health equity of the patients with cervical cancer.

The preventative guidelines (#1, 2, 3 and 4) contribute to representative data collection as incorporating diversity-in-design allows for broader perspectives and decreases the tunnel-vision of the designers. Having people from different professional and personal backgrounds

has shown to be important for diversity in the end product (Gichoya et al., 2023). Involving patients during the design, for example by having patient representatives, contributes to representative data as it results in perceptions, interpretations and nuances of the patients that otherwise would not have been identified. External validation of the dataset with third parties in similar contexts allows for extensive evaluation of data collection strategies and provides transparency on potential biases that have occurred. Reiterating contextual constraints during data collection assists dataset designers with identifying sources of bias and lets them consider these biases.

The corrective guidelines (#5 and #6) can be applied to developing datasets, but also to existing datasets. The inclusion of metadata allows for a better understanding of the open health dataset. This is especially important when the dataset is assessed for its usability and relevance by the user of the data. Metadata that is related to representation and diversity such as data collection strategies (Where? About who? How? When? By whom?), allows the user to make a more informed decision about using the data and eliminates the assumptions that were often made. It also incorporates existing representation biases in the datasets which should be reported on. This stimulates the data provider to think about the limitations and biases of the dataset due to the context in which it was collected. Transparency on these limitations is highly valued. In addition, data consumers are encouraged to share with data providers if limitations have been identified. It must be noted that these guidelines do not target representative data collection on its own, instead they target the use of the data. When the user is informed about what the characteristics, possibilities and limitations of the dataset are, the dataset is likely to be used more responsibly. This will contribute to reliable, fair outcomes that protect the health equity of patients. Lastly, the contacting possibilities allow the user to clarify any remaining issues in cases where the metadata does not provide them with sufficient information. In this case, the user has a direct contact link to the provider of the data to ensure that these issues are solved.

Main research question: How can representation bias in open health data be addressed for the training of medical machine learning models?

Combining the knowledge derived from the sub-research questions, the main research question can now be answered. Representation bias in open health datasets can be addressed by incorporating a decision framework as a tool for designers, but also users of open health data. It stimulates data providers, prosumers and consumers to take preventive and corrective action that affects representation bias. The preventive guidelines help data prosumers and providers target various bias types. First of all, the guidelines '*G1: Diversity in design*' and '*G2: Patient involvement*' aim to target gender, racial and socio-economic through an extensive and critical evaluation of the targeted population. Secondly, the guidelines *'G3: Context constraint identification'* and *'G4: External validation'* contribute to the former bias types as well, it also affects the selection and sampling bias that may occur in open health datasets. The corrective guideline *'G5: Contact possibilities'* encourages data providers/prosumers to be transparent towards data consumers to increase the likelihood of the datasets being used for research purposes and decrease the need for making (biased)

assumptions about the open health dataset. The corrective guideline *'G6: Metadata inclusion'* is relevant for the data providers and consumers as it contributes to a more informed decision for using the open health dataset. All six guidelines have been shown to address representation bias in open health datasets, which empowers the user of the dataset (the developer of the medical machine learning model) to train medical machine learning models on open health datasets that are representative of the context in which they will be used.

9.1. Scientific and social contribution of the research

This research fills the current gap in how representation bias in open health data should be approached to improve use in real-world machine learning applications in healthcare by taking on a pivotal role. Prior to this research, no other research was done into the connection between representation bias and open health data which emphasizes the novelty of this research. Since the guidelines in the decision framework might be seen as common in the field of closed data, they were yet to be connected to open health data. Although there are lessons that can be learned from representation bias in closed health data, there is a difference in how it should be addressed in open health data as it has the potential to be reused by many people with various research purposes. It requires comprehensive thoughts and deliberation considering the large impact it may have on e.g. treatment and diagnosis of patients when representation bias is manifested in open health datasets The decision framework can be used by those involved with opening up health data, such as hospitals, research institutions and statistical bureaus. It allows users of the data to make a stronger informed choice on the usability and relevance of the data and contributes to awareness of representational limitations and bias of the open dataset. It does all this while taking into account the context of the open dataset, meaning that the decisions are made without mandatory quotas or rules but instead offer a method for guidance and support into representation and diversity in open health datasets. Furthermore, this thesis lays a foundation for future research on representation bias in open health data by identifying how the different stakeholders have a role and influence on the representativity of the data. Lastly, the indirect but highly valuable contribution of this research lies in the encouraged use of open health data in a fair manner which allows for innovative and promising research in the medical field.

In addition, this research contributes to an informed, responsible and fair use of open health data for medical machine learning models. This allows the ML models to be trained on representative data and enhances their usability in real-world scenarios. By doing so, the models can provide more efficient and precise healthcare while leaving discriminatory practices behind, which is one of the current dangers as described by Ibrahim et al. (2021). This allows for a more fair treatment of patients in a way that does not jeopardize their health equity. Furthermore, open health data can enhance the understanding of diseases, treatments and diagnoses through ML models that use the open data as input which is desirable for our society as a whole.

9.2. Future research

With this research coming to an end, it opens up a direction for future research into open health data for machine learning models. Based on the limitations and the results of this research, it is highly encouraged to explore these trajectories.

Whereas this research was scoped towards creating and using open health datasets, there are additional components of open data that were not considered within this scope. These components can include, but are not limited to, data maintenance, storage and interpretation. Future research could be focused on how these components influence underlying relations between the data providers, consumers and prosumers and how this affects the feasibility and usability of the guidelines. A more specific research direction could therefore be how maintenance of open health data affects the availability of qualitative metadata and the possibility of having a contact owner such as the data collector.

Furthermore, the use cases allowed us to apply the decision framework and take on the role of an open health data consumer by examining representation bias in the dataset. This showed that although the datasets differed on some of the guidelines, there was a remarkable similarity between the cases. As this could be attributed to the fact that the chosen open health datasets were created with the inspiration of each other, it is encouraged to explore the contribution of the decision framework to data consumers on a variety of open health datasets. As the use case analysis focused on open health datasets derived from intensive care settings, it is stimulated to apply the framework in other healthcare domains such as mental health, pediatrics or oncology care. By doing so, the validity and usability of the guidelines can be evaluated in a wider context and the effectiveness of the framework in addressing representation bias can be refined where necessary.

Additionally, the decision framework serves as a solid base for the development and use of open health datasets that safeguard the health equity of patients. The guidelines act as a foundation for data providers, consumers and prosumers, who are encouraged to implement the decision framework in their current workflow. The effectiveness of the decision framework relies heavily on how the framework is implemented and adhered to by these actors. Therefore, future research is needed to identify bottlenecks, but also possibilities and challenges that arise during the implementation of the decision framework. By identifying these elements, a more nuanced view of the framework will be obtained that creates opportunities for refinement and optimization of the framework. This can be done by e.g. conducting a case study into the development of an open health dataset by hospitals using the decision framework. In addition, a qualitative research consisting of interviews and surveys with stakeholders such as general ethicists, data scientists and medical specialists can contribute to validating the effectiveness and usability of the decision framework.

To conclude, there is a large research area still to be explored in the field of open health data that can contribute to fair, equal and responsible applications of machine learning models in healthcare while keeping one thing central: no patient gets left behind.

Bibliography

Acosta-Velasquez, R., Fajardo-Moreno, W., & Espinosa-Leal, L. (2022). Predicting Intensive Care Unit Admission of COVID-19 Patients with Open Data: Analysis of the First Wave in Colombia. *Preprints*. https://doi.org/10.20944/preprints202212.0330.v1

Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, *30*, 100924. https://doi.org/10.1016/J.IMU.2022.100924

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., ... & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*.

Alshenqeeti, H. (2014). Interviewing as a data collection method: A critical review. *English linguistics research*, *3*(1), 39-45.

Amsterdam Medical Datascience. (2022, December 16). *AmsterdamUMCDB - Amsterdam Medical Data Science*. Amsterdam Medical Data Science. https://amsterdammedicaldatascience.nl/amsterdamumcdb/

Bailly, S., Meyfroidt, G., & Timsit, J. F. (2018). What's new in ICU in 2050: big data and machine learning. Intensive care medicine, 44, 1524-1527.

Begany, G. M., & Martin, E. G. (2017). An open health data engagement ecosystem model: Are facilitators the key to open data success?. *Proceedings of the Association for Information Science and Technology*, *54*(1), 621-623.

Brewer, L. C., Fortuna, K. L., Jones, C., Walker, R., Hayes, S. N., Patten, C. A., & Cooper, L. A. (2020). Back to the future: achieving health equity through health informatics and digital health. *JMIR mHealth and uHealth, 8*(1), e14512.

Burström, B., & Tao, W. (2020). Social determinants of health and inequalities in COVID-19. *European journal of public health*, *30*(4), 617-618.

Castleberry, A., & Nolen, A. (2018). Thematic analysis of qualitative research data: Is it as easy as it sounds?. *Currents in pharmacy teaching and learning*, 10(6), 807-815.

Celi, L. A., Cellini, J., Charpignon, M. L., Dee, E. C., Dernoncourt, F., Eber, R., ... & Yao, S. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health*, *1*(3), e0000022.

Cerrato, P., Halamka, J., & Pencina, M. (2022). A proposal for developing a platform that evaluates algorithmic equity and accuracy. *BMJ Health & Care Informatics*, *29*(1).

Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Höchtl, J., & Ferro, E. (2018). The world of open data. *Public Administration and Information Technology. Cham: Springer International Publishing. doi*, 978-3.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual review of biomedical data science*, *4*, 123-144.

Chicco, D., Cerono, G., & Cangelosi, D. (2022). A survey on publicly available open datasets derived from electronic health records (EHRs) of patients with neuroblastoma. *Data Science Journal*, *21*, 17-17.

Craig, S., McPeak, K. E., Madu, C., & Dalembert, G. (2022). Health information technology and equity: Applying history's lessons to tomorrow's innovations. *Current Problems in Pediatric and Adolescent Health Care, 52*(1), 101110.

Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. In *Ijcai* (Vol. 17, No. 2017, pp. 4691-4697).

Dehkharghanian, T., Bidgoli, A. A., Riasatian, A., Mazaheri, P., Campbell, C. J., Pantanowitz, L., ... & Rahnamayan, S. (2023). Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagnostic pathology*, *18*(1), 1-12.

European Commission. (2020, June 9). *Open data and data bias.* Data Europe. Retrieved May 19, 2023, from https://data.europa.eu/en/news-events/news/open-data-and-data-bias

Eurostat. (n.d.). *Database Health*. https://ec.europa.eu/eurostat/web/health/database?node_code=hlth_care

Faltys, M., Zimmermann, M., Lyu, X., Hüser, M., Hyland, S., Rätsch, G., & Merz, T. (2021). HiRID, a high time-resolution ICU dataset (version 1.1.1). *PhysioNet.* https://doi.org/10.13026/nkwc-js72.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on information systems (TOIS)*, *14*(3), 330-347.

Gichoya, J. W., McCoy, L. G., Celi, L. A., & Ghassemi, M. (2021). Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ health & care informatics*, *28*(1).

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., ... & Purkayastha, S. (2023). AI pitfalls and what not to do: mitigating bias in AI. *The British Journal of Radiology*, *96*(1150), 20230023.

Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current Genomics*, *22*(4), 291.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature medicine, 25*(1), 30-36.

Heijlen, R., & Crompvoets, J. (2021). Open health data: Mapping the ecosystem. *Digital Health*, *7*, 20552076211050167

HiRID. (n.d.). *HiRID - high time resolution ICU data set.* https://hirid.intensivecare.ai/

Hulsen, T. (2020). Sharing is caring—data sharing initiatives in healthcare. *International journal of environmental research and public health, 17*(9), 3046.

Hunt, J. (1999). Use case analysis. *Java for Practitioners: An Introduction and Reference to Java and Object Orientation*, 541-549.

Huston, P., Edge, V. L., & Bernier, E. (2019). Open science/open data: Reaping the benefits of open data in public health. *Canada Communicable Disease Report*, *45*(11), 252.

Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., ... & Merz, T. M. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, *26*(3), 364-373.

Ibrahim, H., Liu, X., Zariffa, N., Morris, A. D., & Denniston, A. K. (2021). Health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health*, *3*(4), e260–e265. https://doi.org/10.1016/S2589-7500(20)30317-4

Jamshed, S. (2014). Qualitative research method-interviewing and observation. *Journal of basic and clinical pharmacy*, *5*(4), 87.

Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, *3*, 58-73.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., ... & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, *10*(1), 1.

Kaplan, G. A., & Keil, J. E. (1993). Socioeconomic factors and cardiovascular disease: a review of the literature. *Circulation*, *88*(4), 1973-1998.

Khan, S. M., Liu, X., Nath, S., Korot, E., Faes, L., Wagner, S. K., ... & Denniston, A. K. (2021). A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health, 3(*1), e51-e66.

Knevel, R., & Liao, K. P. (2023). From real-world electronic health record data to real-world results using artificial intelligence. *Annals of the Rheumatic Diseases*, *82*(3), 306-311.

Knopf, J. W. (2006). Doing a literature review. *PS: Political Science & Politics,* 39(1), 127-132.

Kostkova, P., Brewer, H., De Lusignan, S., Fottrell, E., Goldacre, B., Hart, G., ... & Tooke, J. (2016). Who owns the data? Open data for healthcare. *Frontiers in public health*, *4*, 7.

Kroes, P., Franssen, M., Poel, I. V. D., & Ottens, M. (2006). Treating socio-technical systems as engineering systems: some conceptual problems. *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research*, *23*(6), 803-814.

Kumar, A., Aelgani, V., Vohra, R., Gupta, S. K., Bhagawati, M., Paul, S., ... & Suri, J. S. (2023). Artificial intelligence bias in medical system designs: A systematic review. *Multimedia Tools and Applications*, 1-53.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, *117*(23), 12592-12594.

Lee, E. W., & Viswanath, K. (2020). Big data in context: addressing the twin perils of data absenteeism and chauvinism in the context of health disparities research. *Journal of Medical Internet Research, 22*(1), e16377.

Maastricht UMC+. (n.d.). *Big data in de gezondheidszorg | Gezond Idee*. Retrieved May 5, 2023, from https://gezondidee.mumc.nl/big-data-de-gezondheidszorg

MacIntyre, M. R., Cockerill, R. G., Mirza, O. F., & Appel, J. M. (2023). Ethical considerations for the use of artificial intelligence in medical decision-making capacity assessments. *Psychiatry research*, *328*, 115466.

Marcelin, J. R., Siraj, D. S., Victor, R., Kotadia, S., & Maldonado, Y. A. (2019). The impact of unconscious bias in healthcare: how to recognize and mitigate it. *The Journal of infectious diseases*, *220*(Supplement_2), S62-S73.

Martinez-Martin, N., & Cho, M. K. (2022). Bridging the AI Chasm: Can EBM Address Representation and Fairness in Clinical Machine Learning?. *The American Journal of Bioethics, 22*(5), 30-32.

Martin-Sanchez, F., & Verspoor, K. (2014). Big data in medicine is driving big changes. *Yearbook of medical informatics, 23*(01), 14-20.

McCradden, M. D., Anderson, J. A., A. Stephenson, E., Drysdale, E., Erdman, L., Goldenberg, A., & Zlotnik Shaul, R. (2022). A research ethics framework for the clinical translation of healthcare machine learning. *The American Journal of Bioethics, 22*(5), 8-22.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1-35.

Meng, C., Trinh, L., Xu, N., Enouen, J., & Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports*, *12*(1), 7166.

MIT Laboratory for Computational Physiology. (2018, May 18). *Labels · MIT-LCP/eicu-code*. GitHub. https://github.com/MIT-LCP/eicu-code/labels

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10).

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

Open Knowledge Foundation. (n.d.). *The Open Definition - Open definition - defining open in open data, open content and open knowledge*. https://opendefinition.org/

Ottenheijm, S. (2015). *Big data in de gezondheidszorg* . https://nictiz.nl/publicaties/big-data-in-de-gezondheidszorg/

Philips. (2022). Philips eICU Program: Doing more with less in critical care. In Philips. https://www.documents.philips.com/assets/20170523/91a3d1ad4ff84f98930aa77c016b24d4.pdf

Polevikov, S. (2023). Advancing AI in healthcare: a comprehensive review of best practices. *Clinica Chimica Acta*, 117519.

Pollard, T. (2014). EICU Collaborative Research Database. eICU-CRD. https://eicu-crd.mit.edu/

Pollard, T. (n.d.). *EICU Collaborative Research Database*. https://eicu-crd.mit.edu/

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*, *5*(1), 1-13.

Pozzi, G. (2023). Automated opioid risk scores: a case for machine learning-induced epistemic injustice in healthcare. *Ethics and Information Technology*, *25*(1), 3.

Prin, M., & Wunsch, H. (2014). The role of stepdown beds in hospital care. American journal of respiratory and critical care medicine, 190(11), 1210-1216.

Ratcliffe, M., & Budgen, D. (2005). The application of use cases in systems analysis and design specification. *Information and Software Technology*, 47(9), 623-641.

Rees, D. C., Williams, T. N., & Gladwin, M. T. (2010). Sickle-cell disease. *The Lancet*, *376*(9757), 2018-2031.

Rijksoverheid. (n.d.). Ziekenhuizen in beeld. https://coronadashboard.rijksoverheid.nl/landelijk/ziekenhuizen-in-beeld

Sanchez-Pinto, L. N., Luo, Y., & Churpek, M. M. (2018). Big data and data science in critical care. *Chest*, *154*(5), 1239-1248.

Schoeler, T., Speed, D., Porcu, E., Pirastu, N., Pingault, J.-B., & Kutalik, Z. (2023). Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour 2023*, 1–12. https://doi.org/10.1038/s41562-023-01579-9

Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Computing Surveys*, 55(13s), 1-39.

Simon, J., Wong, P. H., & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, *9*(4), 1-16.

Southall, J. (n.d.). Oxford LibGuides: Data and Statistics for Social Sciences: Data analysis tools & training. Libguides.bodleian.ox.ac.uk. Retrieved June 4, 2023, from https://libguides.bodleian.ox.ac.uk/c.php?g=422947&p=2888387

Spector-Bagdady, K., Rahimzadeh, V., Jaffe, K., & Moreno, J. (2022). Promoting Ethical Deployment of Artificial Intelligence and Machine Learning in Healthcare. *The American Journal of Bioethics, 22*(5), 4-7.

Starke, G., De Clercq, E., & Elger, B. S. (2021). Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Medicine, Health Care and Philosophy, 24*, 341-349.

Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ health & care informatics*, *29*(1).

Syed, M., Syed, S., Sexton, K., Syeda, H. B., Garza, M., Zozus, M., ... & Prior, F. (2021, March). Application of machine learning in intensive care unit (ICU) settings using MIMIC dataset: systematic review. In Informatics (Vol. 8, No. 1, p. 16). MDPI.

Thoral, P. J., Peppink, J. M., Driessen, R. H., Sijbrands, E. J., Kompanje, E. J., Kaplan, L., ... & Elbers, P. W. (2021). Sharing ICU patient data responsibly under the society of critical care medicine/European

society of intensive care medicine joint data science collaboration: the Amsterdam university medical centers database (AmsterdamUMCdb) example. *Critical care medicine*, *49*(6), e563.

Thornton, G. M., & Shiri, A. (2021). Challenges with organization, discoverability and access in Canadian open health data repositories. *The Journal of the Canadian Health Libraries Association*, *42*(1), 45.

Ubaldi, B. (2013). Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris, https://doi.org/10.1787/5k46bj4f03s7-en.

Veinot, T. C., Mitchell, H., & Ancker, J. S. (2018). Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association*, *25*(8), 1080-1088.

Wang, F., Zhu, H., & Wu, Y. (2021). Quality, reuse and governance of open data. Proceedings of the Association for Information Science and Technology, 58(1), 659–662. https://doi.org/10.1002/pra2.522

Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, *66*(1), 149-153.

World Health Organization: WHO. (2021, July 7). *Health Equity*. World Health Organization. Retrieved May 19, 2023, from https://www.who.int/health-topics/health-equity#tab=tab_1

Wu, M., Psomopoulos, F., Khalsa, S. J., & de Waard, A. (2019). Data discovery paradigms: User requirements and recommendations for data repositories. *Data Science Journal*, *18*(1).

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning.

ZonMw. (2022, March 28). The impact of sex and gender on health and healthcare. https://www.zonmw.nl/en/gender

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-technical Impediments of Open Data. *Electronic Journal of e-Government*, 10(2), pp156-172.

Appendix

Appendix A.1 Explorative interview protocol and questions

Thank you for participating in the research for my master's thesis project, which will be focused on designing a decision framework for collecting diverse open health datasets. You are taking part in a semi-structured interview that aims at gathering valuable insights into the field of open data, medical machine learning and health equity. I highly appreciate your time and contributions to this research.

The interview will be automatically transcribed by Microsoft Team's feature and will also be recorded for verification purposes. Data processing policies are detailed in the Informed Consent Form attached to this email. After the interview, which will take approximately 1 hour, a summary will be emailed to you.

---

1. Can you explain your background in open data, medical machine learning, and/or health equity?
    - Open data: Data that is freely available, accessible and can be used and redistributed by anyone without restrictions.
    - Medical machine learning: algorithms developed to analyse medical data, identify patterns and make decisions or predictions on diagnosis.
    - Health equity: Every individual has the opportunity to attain their highest level of health; fight systematic differences in health status/access to healthcare and achieve fair and justified health outcomes for all.
2. Have you ever created or worked with open health datasets?
    a. If yes, can you elaborate on the type and purpose of your work?
3. In your opinion, how can open health datasets play a role in the development of equitable medical machine-learning models?

4. Have you experienced or heard of machine learning applications in healthcare where health equity was affected due to a lack of diversity in data?
    a. If yes, please provide examples.
5. What important considerations do you think should be taken into account when creating an open health dataset to ensure the health equity of patients?
6. What ethical challenges do you foresee in the use of open health datasets for medical machine learning?

7. What (non-)existing regulatory frameworks or guidelines could govern the design and use of diverse open health datasets?
8. Do you believe that the design of diverse open health datasets is an objective that should be addressed at a national or international level?
    a. What potential challenges do you see in achieving this?
9. How do you think diversity in open health datasets can be improved?
10. In your opinion, should any changes be made to the development team of open health datasets/machine learning applications to improve their useability in terms of representability?
11. What technical best practices or guidelines do you recommend for diverse data collection for open health datasets?
12. How does the inclusion of metadata contribute to addressing underrepresentation and promoting diversity in open health datasets?
    a. What specific metadata elements do you consider to be essential for those working with open health datasets?
13. What other suggestions, advice or recommendations do you have for developing a decision framework for diverse open health datasets?
14. Do you have any other input that you want to share?

Thank you for your participation in this research.

Appendix A.2 Informed consent form – explorative interviews

You are being invited to participate in a research study titled 'Bias in Open Health Datasets'. This study is being done by S.J. San José Sánchez, a second-year master's student in Complex System Engineering and Management at TU Delft.

The purpose of this research study is to design guidelines that will promote diverse open health datasets for responsible implementation of machine learning models in healthcare. The interview will take you approximately 60 minutes to complete. The data will be used for the development of fair guidelines on data collection for open health datasets. We will be asking you to discuss experiences, opportunities and challenges concerning medical machine learning, bias and/or open health datasets.

As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by storing the interview recording and transcript, as well as your contact details and informed consent form, on a secured TU Delft institutional storage that only the research team listed below has access to. The transcription of the interview will be sent to you after conducting the interview. The information gathered from the transcript will be aggregated with the other transcripts and presented in an anonymous way in the thesis report.

Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any questions. You are free to request any data to be removed and within 14 days your request will be honoured. All personal data (contact details, job titles), transcripts and recordings will be destroyed one month after completion of the research.

Contact details of the research supervisors:

Chair: A.M.G. Zuiderwijk-van Eijk

Supervisor: J.M. Durán

Advisor: C. Figueroa

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information as listed above or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
3. I understand that taking part in the study involves an audio- or video-recorded interview. The recording will be destroyed 1 month after completion of the research.	☐	☐

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
4. I understand that the study will end in mid February 2024.	☐	☐
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
5. I understand that taking part in the study also involves collecting my email address, job title and organization, and recordings and transcripts of the interview, with the potential risk of my identity being revealed when this data would be accessible by unauthorized entities.	☐	☐
6. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach: using a secured, institutional storage by TU Delft with authorized access for solely the research team. The recording of the interview will only be used for transcription purposes. The included and published transcription of the interview will be anonymised.	☐	☐
7. I understand that personal information collected about me that can identify me, including name and contact details, will not be shared beyond the study team and will be destroyed 1 month after completion of the research.	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
8. I understand that after the research study the aggregated information of the interviews will be part of the results section of my thesis. The thesis will be published for research purposes in TU Delft's repository.	☐	☐
9. I agree that my responses, views or other input can be quoted anonymously in research outputs	☐	☐

Signatures
_____ _____ _____ Name of participant [printed] Signature Date
I, as a researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability ensured that the participants understood what they are freely consenting.
Study contact details for further information: S.J. San José Sánchez

Appendix A.3 Summaries of explorative interviews

The interviews conducted during the exploratory stage of this research are summarized and presented in this appendix. It consists of seven interview summaries.

Interview E-P1

The participant has a background in ethics in healthcare digitalisation. They are not into the data itself, but working on the people side of ethics. They have set up a platform for applying ethics in healthcare. They do not have personal experience with open data, however, in the discussions in their work field the topic of open data has come across. These discussions are about how more can be obtained from existing data and making more data available.

The good thing about open health data is that many people can use it instead of only the people who have the data; which are often only a handful of people/organizations which makes you very dependent. They find it important that people work together: the patient, doctor, researcher and minister.

The concept of bias is often discussed as it has ethical consequences. Of course, you want to have minimal bias. An example that they provide refers to a hospital that did research into the medication use of people with kidney issues. In essence, it is perfectly research what the medication does but it is often tested on people who do not use different medications simultaneously. In addition, it is often tested on healthy people whereas kidney issues are often expressed in people who are older, less healthy and often take more medications. This is also an interesting bias, it is not only based on age and gender but it can also refer to an isolated measurement whereas in reality, this would rarely be the case.

Open health data must not only be available but also easily accessible, easy to create and also give out warnings. These warnings could be given when you want to do a specific thing with the data but the composition of the dataset does not allow this, then a warning could be given such as 'you do not have enough data/too much bias in the data'. Another approach is that the bias could be reported. If you want to create something for young women with a light skin colour and use a dataset of old men with a dark skin colour, and you know about this beforehand you can interpret the data better. If you still use it, you know that there will be outcomes that might not be reliable. It all depends on the user of the data and what they want to use it for. A doctor would want to know what people are included in the dataset, so there could be information included that says something about whether the dataset can be used e.g., only for men or women. Sometimes you also want the bias in your dataset, for example with diseases that only occur in the African continent. The context is thus guiding for whether bias is going to be a problem or not.

They also mention the importance of standards. The AI Act has four categories from 'AI can do everything' to 'AI is dangerous/cannot do anything' and two in between. This act has to be translated and then built into the national law and regulations, it is a top-bottom approach. But since there are too many applications you end up with very general terms about what AI

should adhere to. They think that it is better do to it from the bottom up and identify who is affected by the technology, who is involved etc.

In addition, they find that it is important to have all the different silos of patients, professionals, policymakers and technology people together. In essence, you want everyone in the room who is somewhat affected by open health data. Then you apply a sort of reverse engineering, where you discuss where the data will end up and what will be done with it. Based on this, questions will arise and can be approached by all the parties involved. It is a new form of data and it will also result in questions about how for example doctors should exactly use it and whether it is reliable.

The thing about guidelines is that have to be adhered to by the people using them. In the end, you hope that people will use the guidelines, but this is not always the case. It results in extra rules that have to be adhered to, and sometimes the people working with the technology do not need the guidelines because they already know what to do. By moving this more towards a bottom-up approach, you start with the people who work with it and then work your way up.

Interview E-P2

The participant has a background in IT consultancy and now works on developing healthcare algorithms for prediction purposes. They work at an organization that sees the importance of not only developing data products but also ensuring that the product is relevant and useful for daily work operations.

They do not have experience with open data on itself, but they are familiar with data sharing among hospitals. Academic hospitals in the Netherlands have a partnership and within this partnership, data can be shared among the hospitals for research purposes. This is mostly related to data from electronic patient records. But every data-sharing agreement has to be made specific: what data can we use and what data not? For what purpose are we allowed to use it and how do we have to manage the data?

They think that open health data contributes to more representational healthcare as you have more data and thus also data from other target areas. More data does not only refer to having more rows, so having more of the same data but of more patients. It also refers to more columns, so having more data variables. In the current situation, a lot of data is collected in different areas: the general practitioner, smartwatch, hospital etc. There is a lot of value in getting all of these data sources together as these are now all in separate systems, so the question is how can we get all of this together? The infrastructure is also still fragmented in the Netherlands, there are so different many organizations and so many different systems in which data is stored. And that is where open data can also play a role. It would be interesting to be able to follow the patient throughout the whole care process. If all this information is publicly available, you can learn from each other and make the health system smarter and more efficient. Of course, there are difficulties in connecting patient 100 in system A to patient 20 in system B, not only technical but also from the ethical side where patient privacy should be protected. The technical difficulties lie in different standards that are used, e.g. different

electronic patient record systems and different data type standards such as litres/gallons. There are some standards for this, for example, the International Patient Summary (IPS). Another benefit of open health data that is mentioned by the participant is that data enthusiasts can use open data which may result in new insights or solutions.

The participant mentions that it is really important to think about whether the data you want to publish or use is representative. They provide an example of the COVID-19 pandemic, if data that was collected during this period would be used for training an algorithm, would the algorithm be representative of 'non-pandemic' times? If they have doubts about this, they do not use the data from these periods, therefore it is important to consider it when collecting the data.

Diversity is not only about the patient, it is also about the systems where the data is derived from: not only the hospital but also from other systems. Also, it is about not focussing on solely the Randstad, but on the Netherlands as a whole so they also incorporate zip codes to address the geographical diversity. With using the federated learning technique, data from different hospitals can be combined without having all the data in one big dataset as this also brings risk for privacy and security reasons. By using data from different areas, you also contribute to ensuring diversity and representation in the algorithms. By having a diverse team we also incorporate different views from not only caregivers but also developers and quality managers to ensure that the algorithm complies with regulations such as the EU's Medical Device Regulation.

For each project, the definition of 'what is the bias' is different. It is encouraged to be actively involved with identifying biases, but it is also important to document and communicate these biases so that other people know what can be done with this. Although this is for developing medical algorithms, this can also be applied to open health datasets.

Interview E-P3

The participant is a system biologist and has a background in mechanistic models. They aim to understand the complex interactions within the human body, but as this has a lot of complexity, they quickly delved into data integration, sharing, reuse and modelling. Although the participant is a strong advocate for data reuse, it is very difficult. Openness is often a problem due to ethical and legal obstacles, for that reason the term 'open access' is not in the FAIR principle but accessibility is. Privacy is a big issue in the Netherlands when comparing it to for example China or the US. But using data in the right way also allows us to make people healthier. There are more and more technologies on the rise to make data reusable and the participant thinks that wherever it is possible, open data is key. The more hands-on it is, the better. But also, the larger the dataset and the more parameters it has, the higher the risk of patient identification is.

It is noted that even synthetic data is traceable. The participant illustrates this by giving an example of data about homeowners in Amsterdam, there is one major exploiter: a Dutch prince. When you turn this into a synthetic dataset, do you keep the exploiters in or out? If you

keep them out, you are not able to do anything with it, especially in healthcare the outliers/extremes are important because you also want to be able to treat/diagnose them. If you do keep them in, the model will have better outcomes for the exploiter but the synthetic data is also a part of the Dutch prince. You still get to know more about him, because the data will resemble him. That is also important to consider in open data, you have to ensure that the privacy of the individual patients is preserved. In addition, synthetic data is not always untraceable. There are a few open health datasets on intervention studies (from the pre-GDPR era) which contain variables that are not traceable, so DNA is not included in the open data. Another example of open health data is the COVID dataset from RIVM where they have anonymized test and vaccination data.

The participant has used open datasets before in the form of the personal health environment (PHE, in Dutch: PGOs) as a direct source for access to data and reuse of the data. If someone downloads their health data into their PHE, the patient controls the data which means that you can also ask them for consent to reuse it. This was the basis for one of our projects. Patient consent is thus an important pillar, once you have that you can also work with pseudonymous data. As a proof of concept, we took the information from the RIVM website and put it into fake PHEs while keeping it anonymous, for example by changing the age range back to a specific age to turn it into synthetic data.

Having biased open datasets is highly unethical. The global healthcare system is focused on Western, American/European people and also on people in their late teens/early 20's. Cardiovascular treatments are based on the Framingham score, which was determined on a population that was mostly white Americans. The participant has worked with colleagues from South Korea and illustrates that checking this score on the Korean population (South Korea has data on the entire population). was not at all relevant for Koreans. In the Netherlands, ethnical specificity is gradually increasing. We know that people with a Hindustani background have a higher risk of developing diabetes when compared to Caucasian people. My colleagues in ICT do a lot to prevent bias in an AI model, you do not want ethnicity, gender etc. But from a health perspective, it is not unethical to take this into account, for example, the fact that men and women are biologically different. You need to be aware that your biased dataset may not be suitable for the entire population.

If a model is trained on more data, it is overall a better model. But we also looked at if you train the model on all the data: is it then better for younger or older people? We increased and decreased the number of young and old people in the dataset. It turned out that the model performed better for older people when it was trained on the full dataset, whereas for younger people it was better if the older people were not part of the training data. In advance, you never know when the model will perform best. To be able to determine this, you want to have all the variables/factors in the dataset.

It is an important point that many people do not realise, which is having good metadata. This is also important for the FAIR principles. If the researcher is aware of bias and takes this into account but it is not reflected in the article or metadata, it is likely that the bias will be

extrapolated inaccurately. Therefore, nuance is always important when you look at what to do with the data. This is important in every step of the pipeline. There are specific parameters that are always important as metadata, such as species/gender/how the data was collected but there could be additional metadata that is important to include.

The participant is involved with lifestyle research. In this, there are numerous interventions but the most important one is Christmas. If a study is planned during Christmas, you will never see its effect as it obscures everything. To illustrate, kale is rarely eaten in the summer and strawberries are not eaten in the winter. So when does your research take place? Often it is unclear what is important to the other that is using the data.

In their organization, research is done on mice but the researchers do not specify that the research is on the 'mus musculus' so someone who works with the data could think that they are humans and come up with very strange outcomes. Another example of this is the dataset received from another country on micronutrients in blood to provide patients with advice on vitamins and minerals. There were some strange individuals in the dataset with strange values, after calling the data collector, these values corresponded to horses and not to humans. The importance of contacting the data collector is emphasized. Staying close to the data collector, they are the one who knows what happened. The easier you make it for the researcher to contact the data collector, the better it is.

The participant shows a research database system created by their organization which makes it mandatory before research is published in a database, there are mandatory fields that must be filled in (such as gender, species, and date). It is important to provide the metadata, but you should also make it as easy as possible for the researcher to do so. In addition, they remark that it is not only important to focus on the data collector; the processors of the data are often involved too late in the process. They know what they want to do with it and what is important to include.

Interview E-P4

The participant works as a university lecturer and comes from a medical background, but also focuses on medical AI. They look at questions such as: 'Are we going to operate or not?' and they make predictive models for these issues. In addition, they work on projects concerned with the responsible use of AI, varying from designing to using it (such as ethical principles, representative use or how a doctor follows up on the results of an algorithm).

In the medical world, the MIMIC dataset is very important. Other than that, there are not many open datasets. This is because patient data can not be traceable to the patient in any way. For example, when you open up a dataset from the Erasmus Medical Center in Rotterdam, you can anonymize the data but you still know that the data mostly contains patients from Rotterdam. This makes it very difficult to have open data in the medical domain.

The participant has worked with the MIMIC dataset which is based on intensive care data. The question that was central for publishing the data was: "Can we make our data available so that

other people can do research with it?" This is beneficial for healthcare. What they did for the MIMIC dataset was remove all the patient identifiers but also remove groups that do not often occur. For example, the group of '90 years and older' patients is a relatively small group, not many people are 91, 92 etc. You can request the data when you are working on a medical subject and you also have to explain what you want to use the data for. In that sense, it is open but it still has some degree of access regulation. The result of opening up the data was that a lot of research has been done while using the data with many good results. People who use the data know where it came from, there is a variety of data (structured/unstructured/medical imaging). Although there are other open datasets available, often it is difficult to know where the data comes from and how it was collected. This proposes a barrier to using the data. The potential for open health data lies in documenting how the data was created.

There is still a major challenge in how algorithms should and are developed. Often they are based on historical data but this data is also likely to be biased. Unfortunately, it is difficult to discover where the bias is located. The participant also illustrates two examples where a lack of diversity in the data affected the outcomes of the algorithm:

1) In the Netherlands, everyone has access to healthcare. This is different in the United States where financial reasons are a barrier to access to medical care. People of specific ethnic groups have a lower socio-economic status and therefore do not seek medical care often. The algorithm interpreted these groups are less ill and thus needed less care, but in reality, they did not have the resources to access healthcare.

2) The correction factor for race in calculating the renal function (Gfr; glomerular filter ratio) of a patient was implemented as it was thought that certain races had different kidney functions. It turned out to be that this assumption was also based on discrimination.

Furthermore, the participant highlights the trade-off between guaranteeing privacy for the patients but also knowing where patients come from in the data. If you want to do something good with the data, it is valuable to know where the data comes from. When you make a model, you want to know whether the data was collected in the city of Boston (which is a rich city with an overall healthy and rich population) or in a suburb in the middle of nowhere. Removing attributes such as gender and race in the data is not desirable according to the participant. As an example, when 10 women in a dataset have a bad outcome then the model will keep presenting the women as having a bad outcome. When you remove the gender attribute, you cannot link this to them being a woman. It would be more useful to incorporate questions such as 'Does the dataset contain bias for women?' so you can test it. Therefore it is not desirable to take out the sensitive attributes but leave them in the dataset as they are, although it is likely that other people would argue to take them out due to privacy reasons.

It is also important to take patients' socioeconomic status into account. In the US, education level and insurance level are taken into account but in Europe, they are not, although it does

say something about the composition of the dataset. If something is collected or created in a hospital in Rotterdam, there could be specific environmental factors that you cannot filter out of a dataset. The example is given of living near Tata Steel, where there are environmental factors that affect your health.

If you have developed an algorithm based on a dataset with all patients with a broken hip in the US, you want to look at how the algorithm performs somewhere else. Although a broken hip is a broken hip, the US population in general has more diseases and lives shorter so knowing that the algorithm was based on US data is important. In essence, you want to be able to explain why something works or does not for some patients. If you collect a dataset in Tiel, create an algorithm and then apply it in Rotterdam, you want to know why it works or does not work. For that, the patient population in Tiel is important because you have to compare it with each other. It is really difficult to implement certain fixed rules for this, maybe a direction is to incorporate broader groups similar to the MIMIC dataset where you take a bin from everyone between 20 and 30 years old. Simply said, such rules do not exist just as rules about whether gender and age should be included or excluded from the data. It could be interesting to externally validate the dataset for example by evaluating two cohorts and examining where they differ in content.

It is important to have diverse groups during the design but the participant doubts whether this should apply to data collection as it is more a thing for model development. Data collection is often done in the same way by everyone as you are not reporting men well on purpose but leaving out the women.

Interview E-P5

The participant has a background in systemic injustices and philosophy. They do not have experience in working with open health data.

They mention that with open data, you would reduce the problem in terms of representation but new issues will arise such as privacy. If there is more data available that represents different parts of the population, even better if this would represent minorities that are underrecognized by the system, then open health data has the potential, in general, to contribute to a more responsible and equal machine learning model and have more fair outcomes. But it will always require a trade-off between that and other ethical principles that we value and uphold in our systems.

An example is given of how proxies can result in bias. For example, the risk of drug shopping/addiction behaviour of a patient is determined based on how far they are driving from their home to the pharmacy, all of this is incorporated in the dataset. The algorithm built upon this dataset and identified a relation between the length of the drive and the risk of drug shopping. In reality, this patient might have to drive further as they live in a rural area and do not have a broad supply of pharmacies in their neighbourhood. It is also paradoxical, as someone who lives in the city has many pharmacies nearby and could therefore be undetected by the algorithm. By default, the people living in rural areas are disadvantaged by

the system simply because they have to travel a longer distance to the pharmacy. This entails that a biased view of the problem can lead to a biased outcome. There are a lot of discriminatory outcomes similar to this. This can be through educational backgrounds, gender backgrounds

Some minority groups come from countries where the resources are scarce. This makes it also a question of how this is or should be regulated. If there was a possibility to have this data, would we also have the possibility to access it or should there be a more structural change in terms of how this system functions? In addition, as the data is very sensitive it also makes sense to protect this data as much as possible. It could require a more structural change in how healthcare is provided, so the overarching political thing looks at what we could do with a particular machine learning system.

Of course, the context must be taken into consideration. If the purpose of the end product is only applicable to a subgroup of the population, let's say breast cancer, then it would make sense that male participants are not included in the data. This also relates to the structural injustices in the medical system. For a long time, the standard in medical research was the male body. Now we know that if you get a heart attack and your left arm is numb, that is based on the male body but these are not necessarily the symptoms for women. Bias is rooted in structural inequality, although the male body might not be the standard there are still several pathologies that are only applicable to females and are under-researched.

It is also important to do this together with the patients, either through informed consent or by informing them about what is done with their data. But this is something that might be difficult from a feasibility perspective as there are all sorts of systemic constraints. There should be some kind of mechanism in place to ensure that this is done properly. The same goes for doing it together with different people from different backgrounds. It is sort of the same problem, but also very important as it connects people with different expertise and ways of talking.

Interview E-P6

The participant specialises in FAIR data (Findable, Accessible, Interoperable and Reusable). Fair data can also be seen as ‘Fully AI Ready’ as the data is only valuable when it is not only human readable but also machine-readable. Although the participant does not work with open data themself, their organization provides grants for health research and sometimes uses the condition that new research can only reuse existing data. From that point of view, the use of open health data is stimulated.

They try to approach this by fully focusing on metadata, which concerns describing the data but equally important, describing what is missing in the data. They encourage to use metadata that involves how people and what people were included. Important to mention that this should be done in such a way that the machine also knows what is going on, as the machine is very dumb by itself. It is mentioned that often the computer makes a preselection before the researcher looks at it.

They describe metadata as something that describes the context of the sources, conditions for using the data, information about the content and under what conditions the data was collected. The discussion is often about the data itself and how its sensibility hinders its use in an open context, but the metadata should be enough for a researcher to judge whether the dataset is or is not relevant to them. Their organization is training and encouraging researchers to focus on attaching metadata that is as complete as possible since the metadata can be open to all. When the researcher/data collector is then approached by someone who says 'By looking at your metadata I think that it is very relevant and usable to me, so can we do a collaboration or can you share the data?', the concerns that come with opening the data such as privacy/anonymity will be dealt with later. The scope of the data is therefore on research data that was collected for research that was subsidized by the organization of the participant.

When looking from a diversity point of view, metadata also plays a role as it shows with what intention the data was created, how diverse it is and for what reason the dataset has (a lack of) diversity. If your research was conducted under white males between 40 and 60 years old, that is not a problem at all but the key factor is that it should be included in the metadata so other researchers can use this knowledge. If you solely have the dataset, it results in a huge bias. It also makes other people aware of where the data can be used, if the researcher using the data then still decides to use the data on a different population to compare populations for example; then that is also totally fine.

The discussion about when data is FAIR (Findable, Accessible, Interoperable and Reusable) is not done yet, therefore it is difficult to have mandatory rules. To this point, there are no standards on what should exactly be included as metadata at all times. Their organization is currently researching what different healthcare fields (such as dementia or cancer research) see as valuable metadata and what attributes of metadata are often used by them. By doing so, they can match the inclusion of relevant metadata into the workflow from the healthcare fields.

In addition, they mention that 'standardization' is often seen as a difficult thing. People have their own way of working that benefits them, so why would they do it differently? Instead, the participant aims to see standardization as a method of being able to plot different data against each other. An example is given: if in database A the variable gender has the values 'male', 'female', 'male to female', 'female to male', 'other; and database B has the values 'male'. 'female', other'; then in essence they provide the same information but in a different form. This still makes both datasets usable because you can plot them against each other without necessarily incorporating standards that entail only using male/female/other. But if you want to do research where the values 'male to female' and 'female to male' are important, you know from the metadata that database A is not relevant. Therefore standardization and diversity could sound counterintuitive and is sometimes argued by many, but this is far from the truth. The standardization is focused on comparability but not on losing the freedom of using your own variables.

Currently, there is an ongoing discussion on how the data should be maintained. Research data can only be stored for ten years in most cases, and if your metadata is not sufficient, the researcher is needed to explain what is in the data; but who will stay in the same function for many years? That is unlikely. If the contact person is missing, then the metadata is even more essential as a tool for assessing the usability of the dataset.

The criteria for receiving a research grant enforce the implementation of diversity and inclusion measures (what are you researching and why?) but also involve the end-users. To illustrate, if you do research into MS (multiple sclerosis) you need to incorporate the perspective of people with MS in your research to ensure that your research is relevant to them. This forces the researcher to be aware and to think critically about the diversity in their research and in addition they also have to report on it when applying for a grant. The decisions made by data collectors are thus very important.

Interview E-P7

The participant is a project leader on diversity and inclusivity for medical research. In this, the main question is always: am I considering the needs of different individuals in my research?

The key factor is that every individual is different, the reason for this is inherent and many factors contribute to this: biological, social, cultural et cetera. This should be taken into consideration when conducting new research. The participant mentions that there are still cases in which research is targeted towards a group that is already often targeted and therefore seen as the 'norm.' Other groups are disadvantaged by this as the outcomes of the research are often not applicable to the groups that are already underrepresented.

Research proposals are evaluated on both quality and relevance criteria. Whereas relevance criteria look at 'are we doing the good thing?' meaning that there is sufficient support and the research touches upon societal issues, quality criteria look at 'are we doing the thing right?' These criteria are equal in every research, but the exact fulfilment of these criteria differs based on the nature of the research. This incorporates looking at how the research group is composed and whether they have the right experience and expertise. If you want your research to be broadly applicable, then it is a must to have a diverse research group as the different perspectives of the team members can be of positive value to the end product.

In addition, including the end-user is also important. For example, when research is done into socioeconomic status neighbourhoods, the end-user (which is the resident) must be a part of that research group. When research is about improving the quality of life for people with a disability, it can be seen as a requirement to involve an 'expert by experience' in the research group. By doing so, the chances are higher that the end product corresponds with the needs of the target group. It is also important to check afterwards whether their needs were indeed met. When looking at open health data, it is important that the data is documented in such a way that it allows the user of the data to assess the usability of the data, and whether using the data does not lead to unfair outcomes or disadvantages for specific patient groups. In the end, you want the research population to be a reflection of the to-be-served population. If this

population happens to be middle-aged white males, that is not an issue but the difference is that there was active thinking by the researcher about ‘did I incorporate the right research population for what I want to achieve with this research?’ and that is what makes the difference in the end.

By looking at the inclusion of patients it is also important to have the right reflection of the patient population. This is not always easy to achieve because how can you reach the right people? If there are a hundred people needed for a research, and after a difficult time there are finally a hundred people that want to participate, it is tempting for the researcher to go to the next step of the research. In reality, the patient population might not be a good reflection of the end population but due to time and financial constraints, it is easier to move on. In addition, it is often difficult to find the right people, such as experts by experience. To address this issue, patient associations or patient interest groups can be approached as they have a member databases with patients who want to participate in medical research.

Appendix A.4 Explorative interview protocol and questions

Open health datasets provide great opportunities, and having access to more health data can enhance innovation. Medical machine learning models provide better outcomes, improving patient care. But, many of these datasets are not diverse and the risk of representation bias is severe. The goal is to make open health data more diverse, so it is relevant and applicable to many different patients. I am specifically looking into representation bias since it is a form of bias that occurs already within the data collection stage and is a result of many aspects. I have encountered different forms of bias: gender, geographic, racial, socio-economic, sampling and selection bias.

Therefore, I have created a decision framework which applies to actors involved in open health data. Actors that either already have data that can be opened up or data that is collected in the future that will be opened up. On the other hand, some actors use the data for training machine learning models. Concerning the framework, I would want to discuss the following questions.

- Introduction
  1. Ask for an introduction from the participant.
  2. What is your experience or familiarity with open health data, medical machine learning and/or bias?

- Understanding
  3. How well do you understand the decision framework?
  4. Is there something about the framework that is unclear to you, causes confusion or would need further clarification?

- Relevancy/coverage
  5. In your opinion, is this framework relevant for addressing representation bias in open health datasets; is it effective? Does this address and aim to mitigate representation bias in open health datasets?
  6. Do you think that aspects of representation bias are missing in the framework or need further development?
  7. What do you think are the strengths of the framework and what are the weaknesses?
  8. What positive or negative outcomes do you see when the framework is applied during the design/opening of a health dataset?

- Practicality
  9. How practical do you think this framework is in a real-world scenario? Would it be useful for the actors involved?
  10. What challenges do you think will arise for actors or other stakeholders when applying this framework?

- Other
  11. What is missing or redundant in the framework?

12. What would you personally add (or remove) from the framework?
13. Do you have anything else you want to add or share?

Thank you for your time and valuable insights!

Appendix A.5 Summaries of validation interviews

The interviews conducted during the validation stage of this research are summarized and presented in this appendix. It consists of four interview summaries.

Interview V-P1

The participant has a background in neuroscience. They have been working with open data in various capacities: teaching, application, and developing resources. They have experience with clinical prediction models.

The participants mentioned that it is hard to think of something about their personal qualities that would help to increase diversity in data. The participants illustrated a real-world example, where they had to make an application for a research grant. The criteria said that you needed to have a diverse team, therefore they asked a female colleague to join the team since they did not have a woman on the team yet. They have already worked out the research proposal but they simply needed the woman to apply for the grant. The woman said yes and the application was made. They did not get the grant (due to other reasons) therefore the next year, they applied again when the proposal was further developed. In the meantime, they have found a new collaboration partner who could make a genuine contribution to the research. The first woman was kicked off the team and they put in the second woman, which the participant found to be awkward as it was an uncomfortable situation to explain to the first woman why they had to kick her off the team. The woman did not have a unique contribution to the team other than being a female. This felt more like filling a quota, which is the issue with mandatory diversity according to the participant. If the argument is to make data more diverse because the objective of the data is to improve the chances for advancement of marginalized groups or underrepresented groups, then it is understandable. If the argument is that the data will become more diverse because of the leading team being diverse, then more elaboration is needed. Another personal example is that people who study the health outcomes of LGBT people often LGTB people themselves, the same thing happens if you go to a conference on the neuroscience of music, then most of the researchers will be musicians.

The participants find that patients are good at identifying relevant outcomes that should be collected in a dataset. The anomaly prevention guideline is seen as important, but also a highly general measure.

The guideline ‘external validation’ scores highly as it ties in with a broad discussion on how to check and review the quality of data that are being shared. In their opinion, there are not a lot of good mechanisms to do this to this day, either because they do not exist or because they are not good enough procedures. Therefore, they would like to see more of this. It is mentioned that in an ideal world, there would be independent mechanisms and standard bodies that govern this, or a third-party organization that could check and verify data that people publish but they are not currently there. The direction that this guideline is pointing to is therefore very worthwhile.

It is also discussed whether 'bias reporting' is not just a part of data features. As an example, the participant explains that collecting data from a hospital near us will give us a population that is broadly Dutch. The question is then, is that a bias or a feature of the data that is unavoidable? So how do you know, as a publisher of the data, when features are biased data and should be reported by you as a potential bias? It could be useful to have some kind of criteria for what is bias. In addition, the participant likes to distinguish between bias in the data and bias in the model.

The 'contact possibilities' is labelled as a great guideline, but they see some difficulties in it being implemented. Many open datasets are unlikely to have someone at present who is still maintaining them and can answer questions. The downside to this is that it could discourage data sharing as it can put too high of a burden on the people publishing the data. They mention that it would still be better to have the data without someone available than not having the data at all. On an organizational level, e.g. for a hospital director, questions are raised such as 'How long should we hire someone, and do we do this full time?' because if that is the case, people are less likely to do this.

An important and great point according to the participant is the 'metadata inclusion' guideline. What would improve the framework is to give further statements on what metadata features are particularly important for diversity, for example, population characteristics but also other things such as where/when/how the data was collected.

The participant mentioned that there are lots of guidelines for lots of purposes and they have tried to use guidelines themselves as a way to make sure they have not forgotten something during their work. In that case, the guidelines are used as a kind of checklist. It is mentioned that they can also be used in a more normative way for assessing and reviewing research. In that case, the guidelines would lay a basis for checking research that has been published to see how well they follow the guidelines and how they could be improved on these metrics. Overall, the important thing is to be concrete and specific and to have detailed guidelines so you can tell if the research meets the criteria or not. With diversity it is a bit more difficult, therefore criteria or examples could help.

Interview V-P2

Participant 2 is working in the field of ethics of (healthcare) technology. They do not work directly with datasets but have encountered the problem of needing datasets for research and not having them. They work in research ethics in the context of introducing medical evidence and how it is evaluated, but also in using new tools or expanding the scope of current tools in healthcare and evaluating how useful these tools are to the clinician in a practical setting.

After presenting the framework, they comment that what is missing from the framework is that in the end: you can have all this information (as described by the framework) and the dataset can still be quite biased but also relevant and useful. The framework contributes to informed usefulness instead of assuming it is generally relevant and useful. As for medical research, you do not want one big generic database as it will make the results less and less

useful for individual patients. This is not clear from the framework. They do find that the framework identifies well at what different levels of bias might occur.

For the external validation, it is mentioned that it is effective to test it in another context to see if any bias reveals itself if they are biased towards the initial context or examine its validity in another context. However, it is important to establish what the other context is to show that the partner makes a valid testing ground, as having the same context can also reinforce and repeat the bias. Therefore, criteria need to be established or further research goes into what these criteria might be.

It is discussed that the aim of the decision framework is not to eliminate bias or expose the bias to eliminate it, it is merely aimed at identifying the limitations of the applicability of the dataset. You do not end with an unbiased dataset, you identify where the bias is to make sure we are aware of whether or not the dataset is applicable. Every dataset has limitations, but the step-by-step approach gives an identification of the representational constraints; which are biases, but do not necessarily have to be eliminated.

With this, you stimulate diversity but that is not the solution to bias although it is one way to it which is good. This also applies to the 'diversity in design' guideline, if you are a researcher and you want generic results then diversity is important. For example for testing equipment on medical datasets that are real medical datasets but not necessarily perfect, but the datasets are still useful. There is nice evidence about the videogame development industry where the importance of diversity in the designers in the room is important, especially for prototyping and pilot testing. It has happened where the cameras would not pick up black people's movement as quickly as white people, which is a result of the profile of the average software programmer who is not black and thus there were no black people to test it against. For this research, diversity in design is not necessarily going to reduce bias. If you want to reduce bias, then you need the team to be made up of the same people that the population is made up of so that they are attending to the right kind of biases that might exist within that population, but that is not the goal here to say 'we need to people from here, one from there.'

Patient involvement is especially important when you are trying to create a database about a specific set of patients. Input data always has its interpretations to it and involving patients will help you understand how the data is being interpreted and used. Not in the sense of having a more diverse set of people involved, but more to include a perspective that might identify biases that would not have been visible otherwise.

For external validation, the nature of the dataset becomes important as generalizability is not desirable in medical datasets. In this particular case, generalizability refers to 'to what other context can this dataset be generalized?' and to what must it be generalizable: similar populations, different contexts? Either identify parameters or motivate that parameters should be established. The participant sees it as valuable that this guideline improves transparency and sometimes increasing the transparency is all you can do. External validation and articulating this might be a step towards improving transparency.

Metadata is useful for any adjustments you make. It gives you an indication of what adjustments you should make before you can use the data in your context. The contact availability is essential for being able to go back, it refers to Open Science and these two points go together really well.

Bias reporting and anomalies are part of your metadata, you find them after you start using the data. It is an ongoing cycle of continuously increasing the metadata: being able to go back to the contact person if something is missing about the metadata. Therefore it would be useful to be able to add new biases to the metadata once they have been identified later on. For the anomalies, which are a type of bias, they can also be part of the metadata instead of a separate thing. In addition, the word choice is important as they are not necessarily anomalies but more special or contextual constraints on the usability of the data. They prevent generalization to other contexts but also create biases in the context it is used. Lastly, specify that it is an ongoing process and mention who should be working on continuously updating the metadata.

The real-world application of these guidelines can be used for determining whether an open health dataset is 'good' or 'bad' when researchers make use of open data for medical research. As obtaining medical data for medical research is difficult, open datasets can be a solution. It contributes to a standardization of what counts as a database that has made adequate effort to remove or identify biases so researchers can use it. To the participant, the transparency that the decision framework can bring would already be a big achievement as it sometimes is the best thing you can do since you cannot always eliminate bias.

Interview V-P3

The participant has a background in artificial intelligence with a focus on medicine and healthcare in the form of treatment interventions. They do use a lot of datasets but not specifically open datasets. Normally, their research is in collaboration with hospitals so the data is then provided by the hospitals where you sign an NDA and only use the data yourself.

The participant has a good understanding of the framework and points out that there are points that are very interesting to be considered when using the guideline. They point out from their own experience that adhering to the guidelines is often very difficult. In most cases, in healthcare databases are not created because they want to have databases, they are just collected the way they come in. They provide an example of needing patient measurements in the data and show that they are dependent on what patients they have and what patients have given consent for collecting the data.

They express their concern that most databases will be rejected when following these guidelines as most of the data cannot be developed by going through these steps. If there are no possibilities to amend the dataset, what happens with the dataset? After discussing that the decision framework can be used for both creating new datasets but also assessing existing datasets, the participant finds the framework to be reasonable as it allows for comparison with the points that you have here and also on how to approach developing a new

dataset. In addition, they mention that these guidelines could also apply to any other type of datasets that are not open. They do confirm that the difference in open data is that you did not collect the data yourself and therefore assumptions could be made by the user which affects how it is used on a wider scale.

The strong point of the framework is the consideration of different stakeholders. Often, there is a focus on one single direction making it not applicable to everything. However, having the three different types of stakeholders and each of them using the guidelines in a way that is relevant to them is seen as useful. The participant also mentioned that this approach is sometimes missing in the current situation.

The participant has experienced a lack of metadata when working with medical datasets. It affected them by not knowing what they were working with and being unaware of biases that may be in the data. As biases almost always exist in data, you know that you do not have full information and therefore it was difficult to avoid the bias.

When it comes to diversity in design, the participant mentioned that it was never one of their concerns and also had never been an issue to this point. However, they do agree that it is a good addition to the framework as it helps with identifying things from the beginning and therefore avoid 'bad things' from happening in a later stage. It is important to see the identification of anomalies as an iterative procedure. Although at the beginning there might be ideas about what anomalies can happen at the beginning, there can be other things along the way that come up. Therefore it is important to reiterate so in the end you can make it as clean as possible.

The participant remarks that it is unlikely that in the end, you can guarantee that you will end up with a dataset the way you think it should be or the way you want it to be. Many datasets differ from each other. But what the framework does correctly is create awareness as you have to think about the problems and about what is missing in your approach. That is what the participant sees as valuable about it. They stress the importance of awareness, as researchers you want to think that you are not biased but in the end, since your research is also based on your own opinion you can still have biases. If everybody follows the same thing (the decision framework) and looks at the same points, then we have a chance of having less bias.

They also see obstacles in putting the decision framework into practice. Often it is not a team of people that thinks about how a new database should look. It is merely the same people doing the same thing. This will put a burden on resources especially when you want to establish diversity in design guidelines. This would be an issue in every domain, it is good to have different opinions and different perspectives but in practice, it is challenging to achieve as it requires extra resources that an organization needs to allocate. There are steering committees that can be filled to obtain a research grant, but in reality, the names on the paper are solely used for getting the grant. In addition, the healthcare domain is an opinionated field,

more than others. There are certain ways in which patient data is collected and how they think about things should be done.

Also, there is no control over who comes in as a patient and that makes the data biased by default. It is also about the consent of the patients. Different cohorts of people have different opinions about participating in medical studies, they might not accept it. This also puts a psychological aspect to it which makes it especially difficult in the healthcare domain.

Interview V-P4

The participant has a background in FAIR data sharing, open data and scientific integrity. After presenting the framework, they highlight that it is very interesting to approach this problem in the form of a framework with guidelines.

Rapidly the topic of metadata is confirmed by the participant. From their own teaching experience, when machine learning tools/algorithms are developed, people often have little knowledge about where the data they use comes from. This entails questions such as 'Who obtained the data, how it was obtained and what it is about?' This problem is also seen when data is reused and published with a new machine learning tool because uploading the code is again done without consideration of including metadata. There are no explanations given on the context in which the model was created and what it is exactly about. But again, it is highlighted that this finds its origin in not knowing what the data consists of that was used for building the model. It is seen as a crucial aspect to be more transparent about the origin and sources of data, how it has been obtained and who did this. This allows for a better evaluation of the data and how reliable it is.

The diversity in design guidelines is an interesting aspect as a norm for data collection. They describe that it is not inherently in the team and the people that obtain the data, so therefore it can have an impact. One of the problems with bias is that you do not precisely know or that we are not aware of having bias which makes it difficult to report on bias. We sometimes might believe that data obtainment is an activity where we see the world as full of information and we pick certain elements of that information. The problem with that view is that most of the time the information we pick is in itself man-made, and thus biased.

This is illustrated with an example by the participant about life expectancy: there are no biological markers that tell how old I exactly am. There might be cells that can give a rough indication, but the only thing that we humans use to determine how old someone is is birth and death certificates. So people who obtain data on life expectancy will choose to do this based on these certificates and assume that they are a representation of reality, which informs them that there is an X number of people who were born in January 1953. But this is not necessarily a description of reality: it is the definition that was given to life expectancy with man-made practices. Sometimes, doctors might not be completely honest when giving out these certificates. An example is Georgia and other Eastern European countries where it was discovered that men had a longer life expectancy than West Europeans although this is paradoxical due to their lower-quality health systems. Firstly, researchers thought that the life

expectancy was longer due to genetics, but it turned out that doctors were faking the ages on the certificates to prevent the males from serving in the Soviet army. The bias that we have in data does not have to be intentional, in some way they can even be biased in a twisted, double way. It can be biased because the people who obtained the data had a blind spot on what to look for, causing proxies of human practices to create that data. The participant therefore sees a limitation in this guideline but does emphasize that focusing on bias awareness makes it interesting. Something crucial as suggested by the participant is to think about what an algorithm is supposed to do and to reflect on what it is intended to do. This way, you can rethink historical procedures and standards that were seen as normal. Then the tool can become less biased but it also may serve a different function.

What is often forgotten, is that if there were a representative sample of the population, then we would be able to guarantee if whatever random person walks into a hospital and wants to be diagnosed for a certain thing, the chances are higher that that person will be diagnosed correctly. But health is seen as a system: the question of who is going to walk into a hospital is not only determined by how successful the diagnosis is but it is also determined by how much access that person has to healthcare in the first place. Can they reach the hospital, can they afford it if the hospital is there? This shows that the health system is in itself biased.

Another remark that is also important to consider is what should be seen as a valid data point in a dataset, which is a choice that must be made before collecting the data. Sometimes people believe that data speaks for itself but it does not as there is so much decision-making by the person who collects and analyses the data. All these things should be integrated into the metadata. There should be more education and teaching in statistics to underscore the importance of decision-making in data obtainment.

When it comes to the relevancy of the framework, the participant thinks that it would be beneficial to assess bias in an open health dataset. A nuance is made as the guidelines might be one pillar of governance, but governance requires more than solely the guidelines. This could potentially be intertwined with statistics education in medicine or clinical technology. This all affects how health data is collected and used. The participant stresses that although the guidelines are useful, they can also complicate medical research as doctors already experience many burdens. This is all part of the trade-offs that are made when using any AI tools in healthcare but also promising more efficient healthcare. All the extra work and constraints in your research freedom might be an obstacle in implementing the guidelines. As well as the participants that should be found for research as it can also be a cultural thing to not participate in medical research, for example, due to historical and systemic discrimination that specific population groups have experienced in healthcare.

Appendix A.6 Validated decision framework

In Figure 13, the decision framework is visualized after it had been validated by the use cases and validation interviews. It is displayed without visualization coding.

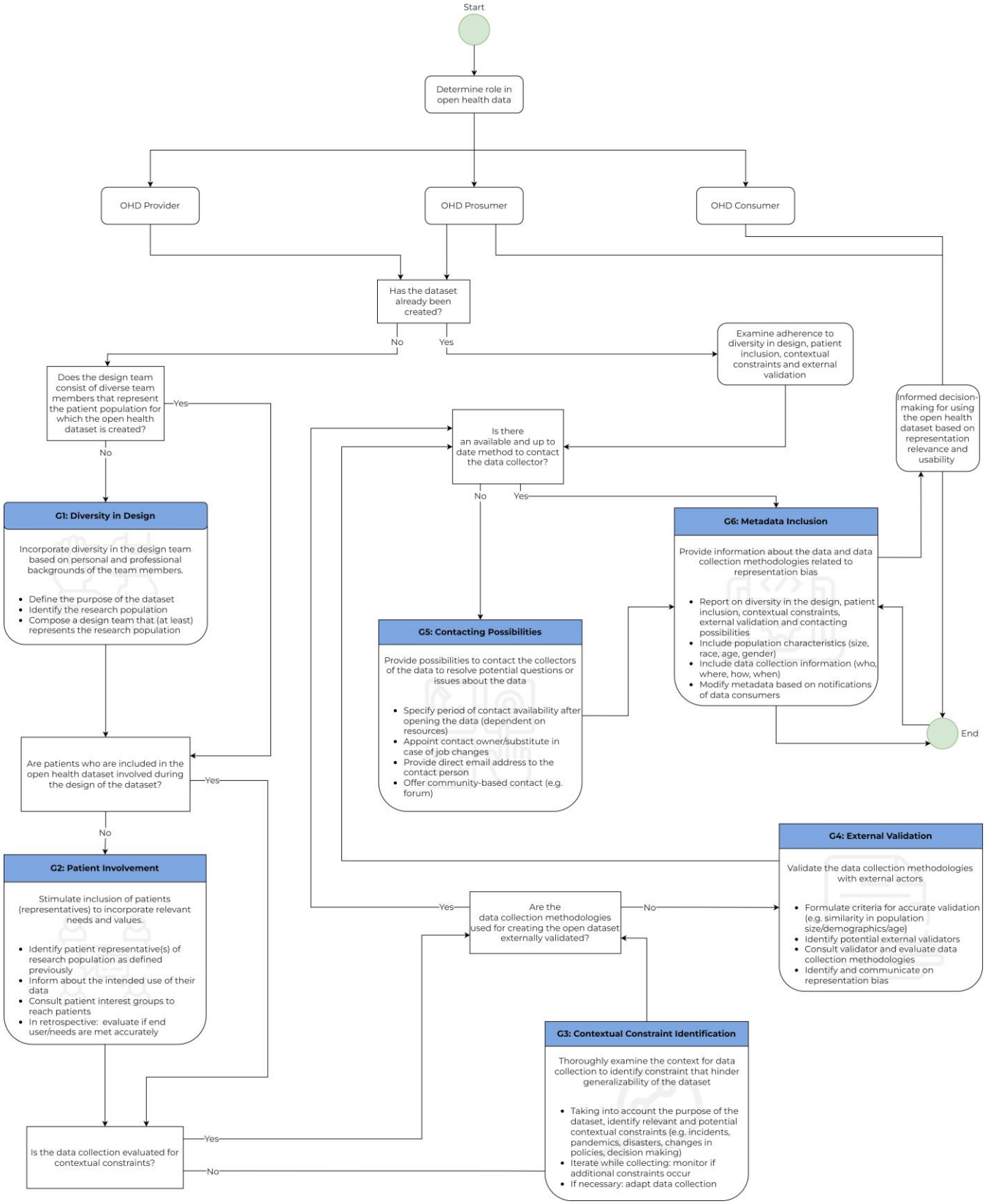

Figure 13 Overview of decision framework