Anonymization of Health Data

Anonymization Approaches, Data Utility and the GDPR

Håvard Wanvik Stenersen

Thesis submitted for the degree of
Master in Informatics: Programming and System Architecture
60 credits

Department of Informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2020
Anonymization of Health Data

Anonymization Approaches, Data Utility and the GDPR

Håvard Wanvik Stenersen
© 2020 Håvard Wanvik Stenersen

Anonymization of Health Data

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo
Abstract

A large amount of health data is gathered through various health institutions all over the world. The District Health Information Software 2 (DHIS2) is a Health Management Information System (HMIS) used in over 100 countries. Each country manages their own installation of this platform and gathers their own data, for their own purposes. There is a desire to be able to use this data for more than was originally intended, however. There are a variety of purposes where this data could prove to be of great benefit, but for research, as well as the further development of DHIS2.

There are, however, barriers which prevent this from happening. To be able to make use of this data, it must first be published, but to do so requires a careful approach. With strict regulations on privacy and data protection being introduced through regulation such as the General Data Protection Regulation (GDPR), failing to comply may lead to the loss of vast amounts of money for anyone handling data, and may even lead to the demise of careless and unprepared organizations. Furthermore, the anonymization process may lead to a significant loss in the utility of data, due to its destructive nature.

This thesis seeks to research the potential for performing anonymization efforts on health data, such that the anonymized data remains useful, while still complying with legislative requirements.
# Contents

1 DHIS2, Anonymization and the GDPR xiii

1 Introduction 1
1.1 Motivation 1
1.2 Aim 2
1.3 Context 3
1.4 Research question 4
1.5 Methodology 4
1.6 Ethical considerations 5
1.7 Discrimination 6
1.8 Risk/utility trade-off 6

2 Background 7
2.1 DHIS2 7
2.2 Privacy and rights 8
2.3 Anonymization 8
2.3.1 Data releases 9
2.3.2 The data in microdata 10
2.3.3 Information disclosure 11
2.3.3.1 Types of disclosure 11
2.3.3.2 Disclosure risks 12
2.3.3.3 Attacks scenarios 12
2.3.4 Methods for anonymization 13
2.3.5 Utility and information loss 14
2.3.6 Utility/disclosure-risk trade-off 14
2.3.7 Privacy models 15
2.3.7.1 k-Anonymity 15
2.3.7.2 l-Diversity 16
2.3.7.3 t-Closeness 16
2.3.7.4 Differential Privacy 17
2.4 General Data Protection Regulation (GDPR) 17
2.4.1 GDPR scope 17
2.4.2 Limitations on processing 19
2.4.3 Takeaway 21

3 Literature Review 23
3.1 Anonymization of health data 23
## CONTENTS

3.2 Improvements of and implementations on privacy models  
3.2.1 Improving upon k-Anonymity  
3.2.2 Implementation of differential privacy  
3.2.3 k-Map, an alternative to k-Anonymization  
3.2.4 \( \delta \) Disclosure privacy, from syntactic to semantic privacy  
3.2.5 \( \beta \) -likeness, limiting attacker sensitive-attribute information benefits  
3.3 GDPR, further examination  
3.3.1 An examination of researcher obligations and legislative GDPR adherence  
3.3.2 A handbook  
3.3.3 A case study on anonymization  

II The Anonymization Process 33

4 Data for testing 35

4.1 Types of data  
4.1.1 Real-world data  
4.1.1.1 Advantages  
4.1.1.2 Disadvantages  
4.1.2 Existing fake data  
4.1.2.1 Advantages  
4.1.2.2 Disadvantages  
4.1.3 Self-produced fake data  
4.1.3.1 Advantages  
4.1.3.2 Disadvantages  
4.1.4 Summary  

4.2 Gathering data  
4.2.1 Choosing a method  
4.2.2 Data quality  
4.2.3 Generating the data  
4.2.3.1 Tools for generating data  
4.2.3.2 Form of the data  
4.2.3.3 Implementation of custom tool for data generation  

5 Anonymization approach 45

5.1 Data release  
5.2 Privacy models  
5.2.1 k-anonymity  
5.2.2 k-map  
5.2.3 l-diversity  
5.2.4 t-closeness  
5.2.5 \( \delta \) disclosure privacy  
5.2.6 \( \beta \) -likeness  
5.2.7 Summary  

5.3 Data attributes  

CONTENTS

5.4 Transformation of attributes .................................. 53
5.5 Measuring data utility ............................................ 55

6 Tools ........................................................................ 57
  6.1 ARX .................................................................. 57
    6.1.1 Anonymization approaches ............................... 58
      6.1.1.1 Privacy models ..................................... 58
      6.1.1.2 Utility metrics ....................................... 58
      6.1.1.3 Transformation methods ............................ 58
    6.1.2 Features ....................................................... 59
      6.1.2.1 Configuration of anonymization process .... 59
      6.1.2.2 Result analysis ...................................... 59
    6.1.3 Usability ....................................................... 59
  6.2 μ-Argus ............................................................... 60
    6.2.1 Anonymization approaches ............................... 60
    6.2.2 Features ....................................................... 60
    6.2.3 Usability ....................................................... 60
  6.3 sdcMicro ........................................................... 61
    6.3.1 Anonymization approaches ............................... 61
    6.3.2 Features ....................................................... 61
    6.3.3 Usability ....................................................... 62
  6.4 Amnesia .............................................................. 62
    6.4.1 Anonymization approaches ............................... 62
    6.4.2 Features ....................................................... 63
    6.4.3 Usability ....................................................... 63
  6.5 Summary ............................................................. 63

7 Execution of anonymization ....................................... 65
  7.1 Configuring anonymization tool .............................. 65
    7.1.1 Input data .................................................... 65
    7.1.2 Attribute hierarchies ...................................... 67
    7.1.3 Attribute utility .......................................... 72
    7.1.4 Privacy models ............................................ 74
  7.2 Executing anonymization ....................................... 75
    7.2.1 ARX efficiency performance ........................... 75
  7.3 Results ............................................................. 76
    7.3.1 Exploring output .......................................... 76
    7.3.2 Analyzing risk ............................................ 78
  7.4 Summary ............................................................. 81

8 Results ..................................................................... 83
  8.1 Transformations and utility scores ........................... 83
  8.2 Equivalence classes ............................................. 84
  8.3 Data quality models ............................................. 85
  8.4 Risk distribution .................................................. 85
  8.5 Attacker models ................................................... 87
  8.6 Data set size ....................................................... 91
    8.6.1 Utility ......................................................... 91
List of Figures

2.1 Differential privacy definition [23] .......................... 17
7.1 ARX after initial import of data ............................... 66
7.2 Configuring attributes ........................................ 66
7.3 Attribute metadata .......................................... 67
7.4 Age hierarchy ............................................... 68
7.5 ZIP code hierarchy ......................................... 69
7.6 Creating a suppression rule ................................... 70
7.7 Health facility location hierarchy ............................ 71
7.8 Hierarchy for time of patient encounter .................... 72
7.9 Attribute weights ............................................ 73
7.10 Utility metric - information loss ............................ 73
7.11 Record suppression limit ..................................... 74
7.12 Executing the anonymization process ....................... 75
7.13 Waiting time for anonymization of large data set ........ 75
7.14 Options for reviewing the anonymization results .......... 76
7.15 Exploring the results ...................................... 77
7.16 Analysis of output data ..................................... 77
7.17 Quality models for the results .............................. 78
7.18 Risk distribution for results ................................. 79
7.19 Risk for quasi-identifiers in the results ................... 80
7.20 Risk in various attack scenarios for results ............... 80
8.1 Risk distribution for result of k-approach on medium data set 86
8.2 Risk distribution for result of δ-approach on medium data set 86
8.3 Risk distribution for result of β-approach on medium data set 87
8.4 Attacker model risk for k-approach on medium data set .... 88
8.5 Attacker model risk for δ-approach on medium data set .... 89
8.6 Attacker model risk for β-approach on medium data set .... 90
A.1 Age hierarchy ............................................... 122
A.2 ZIP code hierarchy .......................................... 123
A.3 Health facility location hierarchy ........................... 124
A.4 Time of encounter hierarchy ............................... 125
List of Tables

5.1 Metadata for data set used for testing ........................................ 51
7.1 Configuration of anonymization approaches ............................... 74
8.1 Hierarchy transformations and utility for the anonymization approach results on the medium data set ........................................ 84
8.2 Equivalence class sizes for the anonymization approach results on the medium data set ........................................ 84
8.3 Quality model metrics for the anonymization approach results on the medium data set ........................................ 85
8.4 Hierarchy transformations and utility loss for anonymization approaches on all data sets ........................................ 91
8.5 Difference in equivalence class sizes in results for small and large data sets compared to medium ........................................ 92
8.6 Difference in quality model metrics sizes in results for small and large data sets compared to medium ........................................ 92
Preface

I want to thank my supervisors Johan Ivar Sæbø and Nils Gruschka for valuable feedback during my work with this project. I also want to thank my family for their support.
Part I

DHIS2, Anonymization and the GDPR
Chapter 1

Introduction

1.1 Motivation

DHIS2 is used on a large scale and gathers a vast amount of data from a large variety of sources. While the data has been gathered for internal use, it might be useful for legitimate purposes beyond that for which it was originally gathered. Thus, to expand its use to potential external application, there is a desire to publish the data. Researchers may want to use it to look for trends in diseases, such as the way they spread and who is susceptible, so that health organizations elsewhere may be able to create better countermeasures for infectious diseases. The social sciences may wish to research how and why public health differs in different countries, and look for trends in socio-economic status in relation to a country’s wealth and prosperity, and welfare-resource available to the general public. Yet again, in particular, with regards to the stakeholders of the DHIS2 platform, training health officials in the use of the platform, as well as testing the platform in development.

None of these purposes can be fulfilled without the data being available, thus, the challenge is how to make it available. Not only from a legal standpoint, but also from an ethical one, it becomes necessary to consider how this data could fall into the wrong hands, and how it could be misused. The data cannot be published in its raw form; Bad actors may abuse private and sensitive health data to harm individuals and organizations alike. Insurance companies may use this data to determine how much they want an individual to pay for their insurance, and even if they want to offer insurance at all; An individual with a history of illness and injuries or pre-existing conditions such as genetic diseases, may for example be a higher risk to the company, and may therefore have to pay more for their insurance. On the other hand, an individual may attempt to blackmail another with health information which could be potentially harmful to their reputation if it were to be made public.

Beyond the moral and ethical, and to the legal: In recent years, individuals’ rights to privacy has gained some attention. Some countries have long
had separate legislation to deal with how personal data may be used and handled. In the USA the Health Insurance Portability and Accountability Act (HIPAA) is a federal law which concerns health information, and the way it must be handled. In Norway, personopplysningsloven concerns the treatment of personally identifiable information. However, in 2016, the European Union (EU) introduced the General Data Protection Regulation (GDPR), which all organization were required to comply with by 2018. This regulation provides uniform rules and guidelines for the handling of personal information regarding citizens and residents of the EU. This, thus, provides a more general framework to refer to when handling personal information, instead of having to comply with the laws of each country when handling their citizens’ data. While this law only regards the information of the residents and citizens of the EU, it is nevertheless a convenient starting point when dealing with personal information. The GDPR introduces hefty fines for violations of its statutes, which makes compliance with it all the more important. It deals with two kinds of fines, depending on the severity of the infraction: The lesser fine can reach up to whichever is amount greater of €10 million or 2% of global annual revenue, while the greater fine doubles that to €20 million or 4% revenue, which for large corporations can obviously be a staggering amount of money.

All of this results in a need for protecting the privacy of any individuals whose data is handled, through complying with the GDPR and ensuring that bad actors will be unable to gain access to sensitive data.

1.2 Aim

Through this project I aimed to research the possibility of anonymizing health data in such a way that it remained sufficiently useful for a wide variety of purposes, while remaining compliant with relevant regulations and legislations, specifically the GDPR.

The specific aim was to create a starting point for DHIS2 researchers and developers to lean on when they want data to be published. With the wide heterogeneity of health data available through the DHIS2 platform, generally and comprehensively researching all possible kinds of data would be an insurmountable task, which I do not intend to accomplish. Neither do I intend to examine all possible approaches to anonymization, as there exists a myriad of options, and comprehensively tackling all of them would beyond the scope of a project such as this. Instead, I wanted to focus on specific and narrow types of data, with a few specific approaches to anonymization, which could then be used as a starting place when doing anonymization, and a pivot point for doing further research on the subject. Thus, instead of having to independently research the breadth and depth of anonymization, system administrators might be able to focus on how their specific use case relates to and differs from the findings of this project, and work from there.

With my findings, I will at the end lay out some guidelines on how to
approach anonymization and what to be aware of in that regard when publishing data; In particular, which risks may be involved from both a moral and legal standpoint. Hopefully, this can help health workers avoid some of the pitfalls of both the anonymization process and the regulations surrounding it, preventing both unnecessary harm to individuals whose data is being handled, and the unfortunate and harsh legal consequences which follow.

1.3 Context

The Health Information Systems Programme (HISP) is a global network hoping to strengthen the Health Information Systems of developing countries. The District Health Information Software 2 (DHIS2) is a platform used by health organizations all over the world, and HISP at the University of Oslo (UiO) develops and maintains its core. It is in use in over 100 countries all over the world [2], and handles a lot of data, including information concerning patients and health organizations. As established earlier, this information can be highly sensitive and will require anonymization efforts to sufficiently protect. This information does not exist in a vacuum however, and what needs to be considered when approaching this anonymization effort is not necessarily obvious.

A simple approach to anonymization is to just remove information which can be used to identify an individual. Remove the name, date of birth, address, phone number, any government-issued identification measures such as social security numbers, etc. So, without all this information, you might think that the job is done, the data is anonymous, and all is well, and sometimes that is the case. However, there exists a vast amount of information on the internet. Companies such as Facebook and Google have trackers on a massive amount of websites and use this data in their advertising businesses. Facebook even reportedly keeps so-called shadow profiles on people not using their services. And Facebook and Google are not the only actors in the advertising business. With the massive amount of data available, seemingly innocuous data might suddenly be used to identify an individual through an attack utilizing background knowledge.

As an example, a healthcare facility might note that a patient who visited on a specific day was diagnosed with a certain disease. There might elsewhere exist information that says only a single patient visited that health facility on that day and who that person is. Thus, simply the fact that the person was diagnosed at that facility at that day becomes enough to identify an individual, and therefore what disease they were diagnosed with. This might be an obvious and simple example, but it illustrates a point: There was no name, phone number, or any other obviously identifiable information present, but re-identification was still possible. Whether the background information was gathered through legitimate or illicit means, the data regarding the patient was ultimately not anonymous. And this
kind of re-identification process could use less obvious paths to reach its goal.

If this logic is stretched far enough, it quickly becomes almost impossible to claim anything but the most severely anonymized. Thus, it is not a question of if information is anonymous or not, but rather, how anonymous it is. If very specific data which is unlikely to exist is required for re-identification, then the information is likely fairly anonymous. If the information needed is publicly available then it’s not very anonymous. For example, if the information needed for re-identification can be found on a person’s Facebook page, on official public documents, then the anonymization process needs to be stricter. However, if most of the information is removed from the original data because it somehow, someway might, potentially, may be used to re-identify someone, then the data suddenly becomes no longer very useful. The solution is obviously somewhere between remove most information and leave most information intact.

1.4 Research question

Considering my motivation of publishing useful health data, my aim of aiding system administrators of DHIS2 in that endeavor, and in the light of the context of strict legislations regarding data protection, where GDPR serves as a good representative, I decided to examine the following:

How can existing approaches to data anonymization be applied to health data to sufficiently comply with privacy and data protection regulations stipulated in the General Data Protection Regulation (GDPR), while preserving utility in the resulting data?

This research question will help me focus on three concerns with regards to data anonymization:

1. Which approaches to anonymization are useful and sufficient when working with health data.
2. Compliance with GDPR when publishing health data.
3. Preserving utility to such a degree that the anonymized data is useful.

1.5 Methodology

To attempt to answer this question I systematically go through the following steps:

Initially, I examine existing approaches to anonymization, how they have been used in previous research, if there are any recent developments in regards to old methods, if they sufficiently anonymize data, and if the loss of information from those uses outweigh their benefit. In addition, I examine how they have been, and could be combined.
The next step involves gathering data that I can utilize to perform the anonymization process. In this project I will not be using real-life health data to test the different approaches to anonymization. Using real-life data would require a great deal of extra work, both in the process of gathering data, as well as storing it securely and safely using the it during the course of project. While some of the results and conclusions drawn from this project would be lent more credence, and likely affect the outcome of the different anonymization methods, I still believe using fake data will produce interesting results and will provide insight into how these approaches to anonymization will perform on real-life data.

Having gathered the necessary data for testing, the anonymization approach must be decided upon. Deliberating the approaches from previously in the thesis, I ultimately select a few methods which I believe may produce promising results for the purposes I am hoping to achieve.

Then, I examine some of the existing free tools are available to use in an anonymization process, in particular, what features they have and how complicated they are to set up and use, before selecting one that I determine fills my needs and allows me to do the research that I want to do.

After selecting my methods, choosing a tool and generating the necessary data, I will then go through the process of anonymizing the data, examining how the different approaches to anonymization may result in different end results. During this process I will gather and sort the results produced, before finally analyzing them and determining how well they hold up against the data protection regulation GDPR while simultaneously examining the utility of the resulting data.

The final section of this thesis will be a consideration of the process as a whole. What went wrong, what went right, both the validity and the shortcomings of the process and results, as well as some recommendations for how to approach the topic of anonymization. Finally, I will present some thoughts on how one might build on this work, as well as utilize it for practical application of an anonymization process.

1.6 Ethical considerations

When working with sensitive data such as health data, ethics quickly come into play. Disclosure of the identity of individuals in data sets pertaining to health information may have serious consequences, as does the disclosure of confidential health information pertaining to individuals. It is therefore vitally important for anyone working with this type of data to have a clear understanding of what consequences their actions may have. In this section I will discuss some of the considerations regarding the ethics pertaining to handling of health information that I will be taking during my research project.

While my work specifically won’t directly utilize such data, it will make use of technologies and techniques which are elsewhere used in the processing
of real data. In addition, I will be providing insights and further work into these technologies and techniques, which may influence how and why other people approach this subject. As such, it is important that I present my findings in a fair and balanced manner, so that none who read this are led into drawing the wrong conclusions and unintentionally applying my findings in a way which might result in harm to individuals and consequently legal action against the one who caused the incident.

It is also important that I not misrepresent what is required in regards to anonymization of data by the GDPR, such that people might believe they are compliant when they are in fact not.

1.7 Discrimination

Performing any automated learning on data has the potential of introducing biases. This can be because the one who implements the learning has inherent biases, or because the data that is learned from has biases. As an example, should a learning algorithm learn of health issues from a data set and extrapolate to the rest of the population, having only people aged between 40-50 will likely make the learning invalid. This same logic applies to many different biases and could lead to discrimination if not careful [23].

Taking this into consideration, it is therefore important to ensure the techniques used in data anonymization does not introduce issues which could lead to this kind of discrimination. Random swapping of for example age in a data set may make it so that issues which are present in specific age groups can no longer be identified, and ignoring economic status of an individual could do the same for particularly vulnerable groups. On the other hand, inserting synthetic data which is based on already biased data could further exacerbate the issue. It is therefore important to be careful when employing different strategies in data anonymization.

1.8 Risk/utility trade-off

The trade-off between risk and utility might seem like an optimization problem depending on the approach taken, but it is important to remember that there are real people behind the data. What this means is that in the case of particularly vulnerable individuals or groups, and particularly sensitive data, even if the risk of disclosure is low, publication may not be justified. Extra utility of data sets at the cost of extra risk may also not be worth it in this case. It is therefore of vital importance that individuals employing these anonymization approaches understand the what the risk metrics mean in a practical sense and that not only consequences in the form of punishments because of regulations like GDPR may occur, but also the consequences the people whose identity is disclosed face, which may be wildly unpredictable.
Chapter 2

Background

2.1 DHIS2

The District Health Information Software 2 (DHIS2) is a Health Management Information System (HMIS) platform used in over 100 countries all over the world [2]. The development of its core is managed by the Health Information Systems Programme (HISP) at the University of Oslo [2]. HISP is a global initiative which aims to empower Health Information Systems in developing countries [35].

The platform is both free and open source, allowing anyone to use and modify with only limited restrictions [2]. Health institutions in various countries manage their own independent installations of the platform, which can integrate with their own related information systems [2]. This means that each country will own and control their own systems and gather and store their own data. The Tracker app for the DHIS2 platform facilitates the processing of individual data over time and the information gathered through its use is of a wide variety, including patient data and information related to education [22]. This information will be gathered for specific purposes of the information system in which the DHIS2 installation is implemented.

With each country implementing their own solutions, this creates a vast amount of data, all kept separately and used exclusively for their originally intended purpose. Various fields of research could greatly benefit from being able utilize this data for a wide array of purposes, however, and the developers of the DHIS2 platform, with the Tracker app in particular, could utilize this data during development and testing to make further improvements on the software. But since the data is kept separately in each installation and the information contained in those installations are of a highly sensitive nature, it is not an option to gather all the raw data from the various systems and utilize it directly.
2.2 Privacy and rights

For a long time, the notion of privacy has been linked to an individual’s rights. The Universal Declaration of Human Rights, ratified in 1948, Article 12, states: “No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.” [59] This concept of privacy has been further established and expanded upon in various pieces of legislation in different parts of the world.

In the United States, for example, the Health Insurance Portability and Accountability Act (HIPAA) of 1996 introduced concerns the specifically privacy regulation of health data [48]. The act establishes various restrictions on the processing of health data, however it provides recourse to avoid such limitations. Specifically, de-identified data is no longer covered by the HIPAA, meaning data which contains no identifiers related to the individual [48]. Other pieces of legislation appear in different parts of the world, but in 2018 the General Data Protection Regulation (GDPR) was ratified by the EU [61]. This legislation established extensive limitations on the processing of personal data, requiring data controllers to ensure the safe context in which such data exists and is used, in addition to establishing several rights for individuals with regard to their own data.

The GDPR, like the HIPAA, includes an exemption for data which can no longer be used to identify individuals, anonymous data. The way in which this property of not being identifiable is established differs, however, concerning not only specific identifiers, but rather whether individuals can reasonably likely be identified [47]. In addition to anonymous data, the GDPR talks about pseudonymous data, for which some restrictions might be loosened, and is data containing information which can be used to identify an individual, but only by making use of additional data which the data controller keeps confidential and sufficiently protected [9]. Thus, there is a need to understand how data may be manipulated to ensure that these requirements may be met. In this thesis, the main concern will be anonymous, rather than pseudonymous data.

2.3 Anonymization

The book “Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections” by Domingo-Ferrer, Sánchez and Soria-Comas provides a nice overview of the field of anonymization and serves as a good resource for an introduction to the topic. This section will be using that book as its primary resource, recapping most of the major concepts on the topic that the book broaches.
2.3.1 Data releases

When doing anonymization for data releases, the desired form of the resulting data will inform one’s approach. In particular, there are three types of data releases that are of interest: Microdata, tabular data and queryable databases [23]. These all have their advantages and disadvantages, and may be useful for different purposes. They also come with their own unique risk-factors when it comes to potential threats.

1. Microdata:

   This is a granular type of data. In this type of data-release, the information usually consists of records where each directly relates to a specific entity [23]. This might for example be individual patient-records of a health-care facility doing diagnoses, or it might be a company record containing information on the salaries and roles of their employees.

2. Tabular data:

   This an aggregated form of data [23]. Typically, microdata will be aggregated into single records to provide statistical information on some topic. In the health sector, this could be a data set containing how many individuals have been infected with specific diseases in specific regions, while in the business sector it could contain information on the average salaries of employees in various industries.

3. Queryable databases:

   Not directly accessible data, but a kind of service. Data released in this way is not directly accessible to anyone in the same way as the previous types of releases. Instead, an interface for interacting with the data is provided, such that you may submit some controlled form of query [23]. This could allow the information in the database itself to remain mostly intact, and only allow queries that will not disclose unwanted information, through for example requiring aggregation of records.

While all of these types of data releases have their uses, the main focus of this thesis (as well as the book used as reference for this section) will be microdata releases. It is a much simpler form of release than a queryable database, requiring no interaction post-release, while still containing as much information from individual records as possible. While a tabular data release could certainly provide interesting information on various topics, its aggregate nature will necessarily be more general, and might be useful for more specific purposes. Microdata, on the other hand, can be used to inform more specific statistics, as well as be used for other purposes, but the necessarily anonymized nature of the data might make each individual aggregation of data less useful or representative.
2.3.2 The data in microdata

To properly anonymize a data set, it is important to recognize the properties of the recorded information. In context of anonymization, these properties are how the data functions beyond the information it contains with a focus on privacy and its potential risks and threats. A piece of information might be a name, an age, a medical diagnosis or a favorite ice cream flavor, and they have their own semantic meaning and have different uses when doing research on a data set. In the anonymization process, however, they have different properties which inform how they must be treated.

A name might identify an individual; an age might not directly identify someone, but could do so in conjunction with other pieces of information; a medical diagnosis is likely not something you can use to identify an individual, but is the kind of information that must be kept confidential, and can be the information being protected by the anonymization process; while a favorite ice cream flavor is likely not considered sensitive, and is probably of minimal use when trying to identify an individual.

These examples demonstrate the four kinds of attributes that a microdata set may contain [23]:

1. Identifiers:

   Data that can be used by an attacker to directly identify an individual [23]. Typical information of this attribute is a name, a social security number or address.

2. Quasi-identifiers:

   Related to identifiers, they can be used to re-identify an individual in a data set, but only in conjunction with other quasi-identifiers [23]. A few examples are age, nationality and gender. Later, when discussing the privacy model k-anonymity, which will also be examined in this thesis, Domingo-Ferrer et al. define a combination of quasi-identifier values as an equivalence class.

3. Confidential attributes:

   Information that must be protected and is typically particularly sensitive data about an individual [23]. Such information could result in negative consequences should it be made public. What is and is not sensitive information may not always be intuitive, as it is likely very much subjective what an individual may not want other people to know, however laws and regulations may specify especially sensitive information that must be protected. An obvious example would be information related to an individual’s health, others are religious belief and sexual orientation [23].

4. Non-confidential attributes:

   An attribute not included in the three previous types, meaning a piece of information that cannot be used to identify an individual,
neither directly nor indirectly, and is neither sensitive nor in need of protection [23]. This could be trivial pieces of information with little importance, but they could also be the focus of specific purposes. While a favorite type of ice cream flavor might be of little importance to most people, an ice cream producer might make use of this information to determine business decisions and research directions.

The two attributes of particular interest when it comes to anonymization are quasi-identifiers and confidential attributes. Identifiers obviously have to be scrubbed, and non-confidential information can mostly be left intact. While identifiers are mostly obvious, quasi-identifiers can be more difficult to decide. Domingo-Ferrer et al. refer to a study which states 87% of the U.S. population can be identified by only a few attributes: a 5-digit ZIP code, their date of birth and gender. And while this information obviously contains information which can hint at an individual’s identity, other seemingly innocuous pieces of information may contain some information, the question then becomes how many such pieces of information are needed. Another study from 2019 notes that 99.98% of Americans can be identified in any data set with 15 demographic attributes [50]. Thus, it becomes a trade-off between the risk of information disclosure and the value of the information which is to be released.

2.3.3 Information disclosure

The goal of the anonymization process is to protect against information disclosure. Information disclosure is when some information that was originally intended to be confidential is extracted from released data.

2.3.3.1 Types of disclosure

There are two types of information disclosure that needs to be protected against: identity disclosure and attribute disclosure [23].

1. Identity disclosure is simply the disclosure of the identity of an individual in the released data set [23]. Here, an attacker has managed to re-identify an individual connected to some record, which would lead to any sensitive information linked to those records no longer being confidential [23]. An example would be records linked to a patient in a data set released from a health institution concerning diagnoses of diseases. Diagnosed diseases could be the confidential information, and an improperly anonymized data set may have only removed the individuals’ names, while leaving quasi-identifiers such as age, residential location, gender and occupation intact. Crucially, identity disclosure does not require all individuals in the data set to be re-identified. Even one re-identified individual constitutes information being disclosed, and a breach of that individual’s anonymity and privacy.

2. Attribute disclosure is the disclosure of confidential information [23]. As broached in the previous section, this is typically sensitive
information, such as health information. The information disclosed does not necessarily need to be exactly accurate, but accurate to such a degree that is effectively disclosed [23]. Domingo et al. mention the example of a salary for a position within a company being upper and lower bounded, thus if the position of an individual included in this data set were to be known, their salary would necessarily fall within the lower and upper bounded range. This type of information disclosure, by extension, can then happen without the identity of an individual actually being disclosed [23].

2.3.3.2 Disclosure risks

With the potential threats for disclosure, a way of reducing risk is necessary. To reduce risk, the risk must first be known, which necessitates some metrics for measuring the risk of disclosure. The following are two metrics for measuring the risk of identity disclosure: Uniqueness and record linkage [23].

1. Uniqueness is a metric which concerns the uniqueness of the data in relation to the original population from which the data originates, that is, the probability of rare values for attributes in the data set also being rare in the original population [23]. If a value is rare in both the data and the original population, the possibility of re-identification is greater than if the opposite were true, given that there is a smaller space of possible identity candidates.

2. Record linkage deals with attempting to re-link anonymized records with the records in the raw data set, using one or more algorithms for re-identifying individuals [23]. The number of correctly linked records provides a measure of how good the anonymization process has been, and may indicate that more strict anonymization techniques need to be employed should the number be high [23].

2.3.3.3 Attacks scenarios

Considering the scenarios during which information disclosure can happen, a 2008 paper focuses on two scenarios related to identity disclosure to be conscious of when anonymizing a data set: the prosecutor scenario and the journalist scenario [26].

1. The prosecutor scenario is the riskiest. In this scenario, the prosecutor, or attacker, already knows that an individual exists in the data set, and The prosecutor knows identifying details about this individual, such as their name, age and address [26]. The data set also contains some confidential attribute which the prosecutor wishes to discover, such as a location at a certain time. What this attack entails is that if only a single individual in the data set match a set of quasi-identifiers, the prosecutor will be able to identify the specific individual they want to disclose confidential information on. It is irrelevant how little these quasi-identifying attributes contribute in and
of themselves. For example, if age is quasi-identifying attribute in the
data set, if only a single record has the attribute with age between 40
and 70 years old, if a prosecutor knows a 50-year old man is in the
data set, he would immediately be identified.

2. The journalist scenario is less risky for each individual, but instead,
it can target the entire data set. In this scenario, the journalist, or
attacker, does not have information that a specific individual is in
the data set, rather, they are interested in discovering the identity of
individuals in the data, by using external information and matching
that information to the quasi-identifying attributes in the data set
[26]. It is here important to consider the rareness in the underlying
population of rare attribute values in the data set, i.e. the uniqueness
metric mentioned in the previous section.

2.3.4 Methods for anonymization

Anonymization is the process of reducing the risk of information disclosure
executed through the utilization of various anonymization methods on the
original data set. The process warps and destroys data to ensure attributes
which could otherwise be used to somehow extract information that is
intended to be protected, can no longer be used to that end.

There are two main categories of anonymization methods: Masking and
synthetic data. [23]

1. Masking is the process of altering or removing data from the original
data set, and may be applied to both identifiers and quasi-identifiers,
as well as confidential attributes [23]. The purpose of masking
identifying information would be to both reduce the possibility of
record linkage and the threat posed by the uniqueness of data, while
the purpose of masking confidential attributes, on the other hand, is
to counteract attribute disclosure [23]. Examples include removing
names from a data set, removing individual records, hiding the last
digits of a zip-code and grouping ages into age-brackets.

2. Synthetic data is false data which simulates real data. Instead of the
resulting anonymized data being an altered version of the original
data set, new data is instead created which attempts to simulate
the original data, preserving some properties and features that the
original set possesses, as well as the statistical information which
may be gleaned from it [23]. The resulting data can take three forms:
Only synthetic data; the original data, with high-risk attributes being
replaced; and a hybrid form of both the original data set with an
added entirely synthetic data set [23]. Information from synthetic
data may be more detailed, because the data does not directly relate to
any real individual, it is only simulated through a model based on the
real data. This might make the data, and any resulting information
gained from it, less accurate, since the anonymized data is likely not
entirely representative of the original.
2.3.5 Utility and information loss

The concepts of utility and information loss have briefly been mentioned in previous sections. Utility of data is essentially the degree to which the data is useful for its intended purpose [23]. Implementing the previously explained anonymization methods is necessarily a trade-off between anonymity and information loss, and by extension utility. The more data that is lost, the less information can be gleaned from it. Generalizing the specific ages of individuals to age ranges reduces the ability of any researcher extracting statistics from the data to gather specific data on ages. Generalizing the location of diagnoses of diseases reduces the ability to track how and where diseases spread.

With the information loss that anonymization efforts bring upon data, it is necessary to be careful and selective if one wishes to preserve as much utility in the resulting data as possible. Part of the problem, however, is that utility is not an objective measure that can be applied to all use-cases: the data likely has many possible purposes, and each purpose has different needs from that data [23]. One study might be more interested in the health of a population based on socio-economic background, while another might be interested in how a disease spreads related to the age of the population. The first might be interested in information related to income and residential area, while the latter will likely be interested in the age information in the data set. If anonymization requires the sacrifice of one, the other, or a part of both, utility cannot be maximized for both purposes [23]. Thus, it becomes necessary to compromise, and anonymize in such a way that both purposes still can be fulfilled to as great an extent as possible. This presents another problem, however: It is nigh impossible to identify and predicting the all possible purposes for the resulting data [23]. There are countless ways in which data is being used in the present, and what the data might be used for in the future will depend on the circumstances and challenges faced then.

With this in mind, Domingo-Ferrer et al. suggests that rather than measuring the utility of the resulting data set, simply measuring the information loss might be a good metric. The technicalities of the methods they present for how to measure this information loss can depend on the type of data, with different methods for numerical and categorical data, but they all in some form or another measure the difference between the original and resulting data set.

2.3.6 Utility/disclosure-risk trade-off

With the risk of disclosure, methods for anonymization and resulting information loss, the next step is determining how to trade off risk and utility. Domingo-Ferrer et al. presents two scores as a metric to measure this trade-off: SDC (Statistical Disclosure Control) scores and R-U (Risk-Utility) maps.

1. SDC scores is a metric which favors optimizing on a single score
which is the combination of the measurement of loss and the risk of
disclosure [23].

2. R-U maps are two-dimensional graphs onto which risk and utility
metrics map, which enables easy comparison of different approaches
to anonymization [23].

2.3.7 Privacy models

There are four main privacy models brought up by Domingo-Ferrer et
al., three of which are closely related, and an additional model used in a
different scenario to the other three.

1. k-Anonymity
2. l-Diversity
3. t-Closeness
4. Differential privacy

2.3.7.1 k-Anonymity

k-Anonymity is a model in which the goal is to guarantee a minimum level
of anonymity by ensuring each combination of quasi-identifier values, an
equivalence class, to be shared by at least k records [23]. In a data set
where this guarantee is fulfilled, it would be impossible to narrow down
an individual to less than a group of k individuals.

This model relies on the assumption that there is a known set of quasi-
identifiers [23]. This may not always be the case, however. If you were to
be aware of all publicly available data which may be used to identify an
individual, you could use that data to construct the set of known quasi-
identifiers. Should there exist any confidential side-data that you are not
aware of, the assumption might no longer hold, as the external data could
be used for re-identification purposes [23].

Domingo-Ferrer et al. cover several different versions of k-Anonymity:
Generalization and suppression based, microaggregation based and prob-
abilistic.

1. Generalization and suppression: Uses the generalization technique
   on data to decrease revealed information. Groups for example ages
   into ranges.

2. Microaggregation based: Uses multivariate microaggregation to
   achieve the same result, to significantly reduce the computational
time as compare to the previous method.

3. Probabilistic: Because k-Anonymity essentially creates a 1/k prob-
   ability of re-identification because of the k records sharing a quasi-
   identifier, this type of k-Anonymity relaxes the requirement of k re-
cords sharing a quasi-identifier to requiring $1/k$ probability of re-
identification.

### 2.3.7.2 l-Diversity

An issue with simple k-Anonymity is that even should there be at least k individuals for each quasi-identifier, it makes no guarantees for the diversity of the confidential attribute values connected to those groups of individuals [23]. The individuals might all come from the same region, have the same medical diagnosis or belong to the same religious denomination. Should this be the case, it is inconsequential that there are at least k individuals, because the confidential attribute is the same, and if a person is known to exist in the data set, the confidentiality of their sensitive information is broken.

To attempt to counteract this drawback of the k-Anonymity model, the l-Diversity model introduces an approach which aims to provide some protection against attribute disclosure by requiring the presence of a minimum number of different values for the confidential attributes within each cluster of k individuals [23].

### 2.3.7.3 t-Closeness

Even with the variance provided by l-Diversity in combination with k-Anonymity, there exists two weaknesses: The confidential attributes tied to the k individuals may be distinct, but very close in value; and the distribution of the distinct values in an equivalence class could have an overrepresentation of one value [23]. An example from a data set containing health data may be weight or a diagnosis, while from a business data set, this could be wage or occupation, with an l-value of 10. For the first weakness, should the weight of 10 individuals with equal quasi-identifiers all fall within a 5kg range, or the salaries of ten individuals fall within a NOK 30,000 range, the difference might be insignificant and be considered a breach of confidentiality. For the second weakness, 90 out of 100 individuals in an equivalence class could have cancer in the health data set or be accountants in the second data set. If the distribution of those values in the original data set as a whole happens to be only 1%, this discrepancy in distribution means the data set is disclosing attribute information about individuals.

A way to combat this situation would be to implement the t-Closeness privacy model. This model seeks to ensure that not only does each group of k individuals have enough distinct values, the values are also at a distribution similar to the data set as a whole, limiting the amount of information an attacker might learn from the data [23].
CHAPTER 2. BACKGROUND

Definition 8.1 $(\epsilon, \delta)$-differential privacy A randomized function $\kappa$ gives $(\epsilon, \delta)$-differential privacy if, for all data sets $X_1$ and $X_2$ that differ in one record, and all $S \subseteq \text{Range}(\kappa)$, we have

$$
Pr(\kappa(X_1) \in S) \leq \exp(\epsilon) \times Pr(\kappa(X_2) \in S) + \delta
$$

Figure 2.1: Differential privacy definition [23]

2.3.7.4 Differential Privacy

Differential privacy deals mostly with interactive settings, but can be expanded to microdata releases [23]. Domingo-Ferrer et al. explain that an anonymization mechanism sits between a database and the querier in the interactive setting. The purpose of the model is to ensure that the information that is gained by the presence of a single individual in the data set is limited. Essentially differential privacy with value $\delta$ is satisfied for a function if all data sets the function is performed on, that only differ by one record, vary by less than $\delta$. Figure 2.1, taken from the book, defines differential privacy [23].

Differential privacy in an interactive setting has a set budget which can be spent before the data can no longer be queried [23]. This is to ensure unintended disclosure does not happen because of too many queries. When used for microdata releases, the purposes for which the data is useful may be limited [23].

2.4 General Data Protection Regulation (GDPR)

The GDPR was introduced as legislation in the European Union (EU) and took effect in May of 2018. The set of regulations introduced through this legislation is the toughest of its kind in the world, instituting strict limitations on the handling of sensitive data, with harsh fines for any breaches of its statutes [61].

2.4.1 GDPR scope

The scope of who and what the GDPR applies to is very clear. Article 3 states that it concerns the treatment of sensitive data, specifically personal data [6]. It applies to processing such data of a data subjects inside the EU. It applies to the processing which is done in the context of a controller or processor in the EU, including when the actual processing is performed outside the EU.

With that said, we need to know what personal data is, who is a data subject is and what processing means, and what is defined as a controller and processor.

First off, the GDPR defines personal data as “any information relating to
an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person” [9]. What this means is that if you are handling any data about a person, it is covered by the GDPR. Conveniently, this definition also provides a description of the term ‘data subject’. Basically, a data subject is any person natural person, meaning an actual person, and not for example a “legal person” [45]. And it covers specifically natural persons located inside the EU, which would include tourists and non-citizens of an EU country. Conversely, it does not apply to anyone outside the EU, including citizens of nations not in the EU and EU citizens living abroad.

What can, and can you not do with this data? The scope specifically mentions ‘processing’ of data. Article 4 provides the following definition: “any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction” [9]. Essentially, anything and everything you might do to the data would be covered under the GDPR. Everything being covered by the GDPR, does not mean that you cannot do anything with the data, rather, it simply means that there are limitations, conditions and extra steps to many of these interactions with the data, which will be broached later.

Who must follow these regulations? Article 3 mentions ‘controllers’ and ‘processors’, which article 4 give definitions of: a controller is “the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law”, while a processor is “a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller”, which boils down to any entity which is responsible for the processing of the data, either through making decisions upon what processing is to be performed, who is to perform the processing, or how such processing is to be done. Additionally, it covers any entity which is performing such processing for someone else. Which means if a company has been contracted to perform data processing, they need to ensure that they are being GDPR-compliant and cannot rely on their client to take responsibility. Article 3 states that processing performed outside the EU is also covered, so long as an involved party is established in the EU.

These preceding paragraphs detail who and what is covered by the GDPR,
but what, specifically is not covered by the GDPR. There are some exceptions which are included in the scope provided above, specifically article 2 lists the following exceptions, the processing of personal data:

1. “in the course of an activity which falls outside the scope of Union law”

2. “by the Member States when carrying out activities which fall within the scope of Chapter 2 of Title V of the TEU”

3. “by a natural person in the course of a purely personal or household activity”

4. “by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security”

The first refers to special activities which the EU doesn’t have jurisdiction over, such as matters of national security [29]. If a nation needs to processing of personal data to protect itself, or other similar circumstances, the GDPR does not place any restrictions on that.

The second has very specific terms, and explicitly stated in the Treaty on the European Union, including activities related to foreign and security policy. [27]

The third means if the processing is done on a personal scale for a purely personal purpose, which is not linked in anyway to any professional or commercial activity, then the processing is exempt. Examples of this type of processing is provided in Recital 18, and includes things like correspondence and social networking. This specifically applies to a person doing this themselves, excluding processors and controllers doing this on behalf of people. [46]

The fourth is straight-forward, activity by the government for the purpose of solving crime.

2.4.2 Limitations on processing

Article 5 of the GDPR [11] concerns the principles which relate to the processing of personal data. It sets many limitations on how data can be collected and the way it must be processed. Data processing must be lawful, fair and transparent; there must be an explicit, legitimate purpose for the collection, and any further processing must fall in line with that initial purpose, except for purposes like scientific research where guidelines for data minimization are being followed; the data must be limited to what is necessary for the purpose they were collected; the data must be accurate and kept up to date or removed; stored for a limited time; the data must be protected with consideration for its confidentiality and integrity.
These are the main points which limit the way in which personal data may be handled. There are further specifics throughout the document, with regard to the responsibility of the processor, as well as the rights of the data subject. The following article, for example, concerns the topic of consent on the part of the data subject for the gathering and processing of such data, which is a part of the lawfulness mentioned in the previous paragraph.

There are a few specific points which are highly relevant to this thesis. Article 9 concerns the processing of personal data of special categories, which includes things like racial and ethnic origin, religious beliefs, genetic and biometric data, data on sex and sexuality and other health-related data. Processing of such data is explicitly prohibited by this article, barring the application of one of the exceptions listed, which includes, but is not limited to: consent from the data subject; processing of data is necessary and obligated to ensure the rights of the controller or the data subject; the data is made public by the data subject; in relation to legal claims; processing is necessary for public interest reasons such as protection against cross-border health threats; processing is necessary for purposes like scientific research, provided guidelines on data-minimization are being followed. [13]

These provisions are important, because they allow the processing of data which would otherwise be prohibited. It is therefore important that there is a solid reasoning for why at least one of these exceptions apply to the processing being performed when such special categories of data are involved, which in the case of this thesis is health data. The specifics of relevant exceptions should be examined before any processing is performed.

Article 32. concerns the security of processing, stating that the controller and processor must take appropriate measures ensure a satisfactory level of security, which can include pseudonymization and encryption of personal data, along with ensuring the continued validity of the data and performing upkeep on their security measures to ensure their continued sufficiency. The measures referred to in the step should be implemented with a context of the current state of the art of technology and the risk posed against the data. [7]

If, in context, the processing of personal data likely presents a high risk to the rights and freedoms of the data subjects, article 35 states that the controller must make assessments of such risk, and must do so while in contact with a data protection officer. [8]

If personal data is to be transferred for processing to another country or international company, the further processing must also be subject to the GDPR. [10]

Another point of particular interest to the topic of this thesis is the way in which some of the limitations on processing may be eased, or even disregarded entirely. Recital 26 of the GDPR[47] refers to anonymized data
in relation to data processing. It specifically states that only data concerning identified or identifiable individuals are covered by the regulations, and that anonymized data explicitly is not covered. This is important, because it will allow any type of information to be used for further processing for any reason, without the need for gathering consent from data subjects or ensuring satisfactory protection of the data. This will allow for the publication of anonymized health data, which is the purpose of this thesis. One issue is that the recital is somewhat vague on how exactly the requirements for anonymization can be assessed. The specific wording of the text is as follows:

To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.” [47]

Article 6 [12] provides some comparatively weaker easing off on data processing limitations based on similar grounds, stating that processing may be done beyond the purpose which the data subject originally consented to if among other factors, “the existence of appropriate safeguards, which may include encryption or pseudonymisation”. Pseudonymisation is further defined in article 4 [9] as “the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person”. While this is only one of 5 factors listed in article 6, it may provide grounds for utilizing pseudonymization instead of anonymization for certain types of processing, though it would still be considered personal data according to recital 26 [47], and therefore likely not be a technique which can be relied upon for the purpose of data release.

This specific point will be further examined in the literature review section of this thesis, along with some more research on the topic of anonymization and data publication in light of the GDPR.

2.4.3 Takeaway

There are a few points to consider when considering doing processing of personal data, and in particular health data, the topic of this thesis.

1. Does the data concern people located in the EU?
2.4. GENERAL DATA PROTECTION REGULATION (GDPR)

2. Is the controller or processor based in the EU?
3. Does it fall under the umbrella of purely personal or household-activity use?
4. Has the data been collected in a way which complies with the GDPR, including purpose limitation, data minimization, data subject consent?
5. Are the premises for the processing lawful?

When performing the actual handling of the data, there are some things to consider with regard to the proper procedure and security.

1. Is there a risk of significant harm to the rights and freedoms of data subjects? Has a risk assessment been made? Has a data protection officer been involved?
2. Are the organizational and technological security measures sufficient to protect the data?
3. Is data being transferred during processing or for the purpose of processing? Are there guarantees the data will be treated according to the GDPR?

Furthermore, when doing further processing beyond the original purpose which data subjects consented, consider the following:

1. Does the new purpose align with the previous purpose?
2. Does the purpose fall under the specific purposes which a limited exception is provided for, such as scientific research?
3. Has there been done extra work, such as pseudonymization and encryption, for the data to align with the original purpose?
4. While pseudonymized data is still considered personal data, sufficiently anonymizing the data will ensure the GDPR no longer applies.

It is important to mention that the regulations presented in the GDPR are extensive and cover much more than what is presented in this chapter. Nevertheless, this serves as an introduction to the topic and includes many of the most relevant concepts and regulations to consider when doing work related to the topic of this thesis. A more comprehensive understanding will be required for anyone doing processing, such that no mistakes occur which may lead to serious negative consequences.
Chapter 3

Literature Review

This section aims to contextualize this project in the field of anonymization with a focus on health data. It builds on the previous introduction to the topic, and presents a variety of works featuring research on both anonymization in the health sector in particular and the field of anonymization in general. Furthermore, some research on the more recent legislation introduced by the GDPR is examined. Through this, an understanding can be gained on current state of the field, as well as interesting research directions.

3.1 Anonymization of health data

There exists a variety of research on the anonymization of health data.

The paper Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies [38] reviews a large corpus of previous work on the topic, mostly articles detailing various anonymization strategies used for the treatment of different types of medical data. Kusida et al. examine the different cases studied in the various articles and review the results with a focus on the strengths and weaknesses of approaches taken, with the purposes of possible application in multicenter research studies. Their conclusions are contextualized by the HIPAA regulations that medical research in the United States is covered by, and find the examined strategies to in be somewhat limited in their ability to adequately deal with all the different types of health information found in relevant medical records. They particularly suggest further work on strategies for handling genetic data.

A French Anonymization Experiment with Health Data [15] is a case study on anonymization of microdata containing health information from a French administrative database on hospital stays. They considered two approaches, making use of two different tools for anonymization, µ-Argus and ARX, and attempted to reach k-Anonymity and l-Diversity, specifically
3.1. ANONYMIZATION OF HEALTH DATA

employing a k-value of 10 and an l-value of 3. They reason that, while a k-value of 5 is a common, a k-value of 10 might provide additional protection from an attacker possessing background data. At the same time, this would make it easier to obtain an l-value of 3, a value chosen to avoid exact value disclosure of sensitive attributes, given the larger groupings which may contain distinct values.

To achieve their privacy model goals, they make use of global recoding, and in their discussion and conclusion find that the loss of utility in the data sets is high. They attribute this to the non-uniformity of the specific type of data they were working with, which included two types of geographic data, and conclude that it is hard to reach a good trade-off between disclosure risk and data utility. They suggest a research direction including perturbation of data, specifically geographic data, to attempt a better risk/utility trade-off.

In A Case Study of Anonymization of Medical Surveys [32], researchers seek to anonymize a data set of medical surveys, which are high-dimensional data sets, and present their methods for evaluating the resulting disclosure risk and information utility. They too, make use of the k-anonymization approach, opting for a k-value of 3, reasoning that the extent to which the data is shared, is limited. They also decide specifically not to use the l-diversity technique which counteracts a risk to attribute disclosure, because of the heavy skew and low modality of their confidential attributes, being a binary presence value with a low occurrence rate.

A very recent paper on the anonymization of structured health data [3] propose a practical methodology including a cryptographic algorithm which preserves privacy through construction. Instead of preserving the semantic value of the raw data, and ensuring privacy through methods such as k-anonymity, this approach rather encrypts the values using a method which preserves properties which can be useful for some types of research. With this encrypted data released, analysis can be performed where the results can be shared with the source of the data, such as a hospital, which can then use the results for their purposes. They compare their approach to an approach utilizing k-anonymity in a previous research paper, and find that their method better preserves the utility of the data.

A less recent paper [44] focuses on a method for anonymization of both structured and unstructured data. Their system attempts to identify and extract sensitive information using a Bayesian classifier from unstructured data, before performing anonymization techniques using k-anonymity and on the extracted information. They find that their system performs effectively when extracting information, and manages to sufficiently protect the privacy of its data subjects while preserving as much utility as possible.
CHAPTER 3. LITERATURE REVIEW

3.2 Improvements of and implementations on privacy models

3.2.1 Improving upon k-Anonymity

A paper from 2017 [39] attempts to improve on the existing k-Anonymization model using generalization by better preserving utility when anonymizing health data. The background for their work is the information loss inherent in the generalization technique, which can severely reduce the utility of anonymized data. They introduce a 3-part method consisting of a utility-preserving model, counterfeit record insertion and a catalog of counterfeit records.

They focus on information loss and data truthfulness: “Data truthfulness implies that each anonymized record corresponds to a single original record.”[39] They argue that truthfulness is important because the usability of results from analysis upon a data set is often dependent on its truthfulness.

Their goal is to create data sets which are useful, privacy-preserving and reliable. To achieve this, their three-part model is used. They introduce a utility-preserving model called h-ceiling, which requires the degree of generalization within a table to be less than or equal to a value h. They also insert counterfeit records to satisfy k-Anonymity in equivalence classes which do not satisfy k-Anonymity. They then keep a record of the number of sensitive information records that have been inserted in each group of equivalent class. By doing this counterfeit insertion in a way such that it is impossible to determine which records are counterfeit by comparing the catalog and the equivalence classes, k-Anonymity is not breached.

As a quality metric they employ Reconstruction Error (RCE), which measures the similarity between original and anonymized records, because traditional metrics such as Loss Metric (LM) which measures direct information loss does not take into account data truthfulness.

They then propose an algorithm which implements this model and analyze the results using the RCE metric, as well as the LM metric for reference. Their results are promising, and they are able to reduce information loss and query error rate compared to standard k-anonymization. However, the score decreases as k increases because of the amount of counterfeit records that need to be inserted.

3.2.2 Implementation of differential privacy

A paper from 2018 [16] introduces a truthful algorithm for data anonymization which they claim has strong privacy guarantees. In the background section for their work they write that differential privacy has been criticized for often being non-truthful.

Their algorithm, which they dub SafePub, is an implementation of an
approach based on earlier works which have shown that random sampling of data followed by k-Anonymity on said data can satisfy differential privacy. Their algorithm comprises a search strategy, an anonymization operator and various quality assessments.

They split the privacy budget of differential privacy for the data between the search and the anonymization. Their anonymization operator is based on generalization and suppression based k-anonymity. They employ quality assessors for the anonymization operator as score functions based on several different established models. Their search strategy then implements the anonymization operator and scoring function to search for good candidates.

Their implementation showed the viability of the strategy of a randomized search strategy using quality assessment together with k-Anonymity using suppression and generalization to create result outputs which are both truthful and have fairly strong privacy guarantees. Their algorithm performed similarly to what they refer to as state-of-the-art related approaches. They note, however, that while their algorithm produces truthful results, it provides a slightly lower privacy guarantee.

3.2.3 k-Map, an alternative to k-Anonymization

In Protecting Privacy Using k-Anonymity [26], an improvement to the k-anonymization privacy model in the form of k-map is proposed. The motivation and basis for this improvement is that k-anonymity is often very aggressive for many of its equivalence-classes. In this privacy model, instead of k-anonymizing a data set, a database of the underlying population is k-anonymized. Subsequently the transformation of individuals is applied to the data set to be released, which effectively enforces k-anonymization because the underlying population which can be used to try re-identify individuals will have at least k alternatives for each quasi-identifier combination. While this type of anonymization cannot protect against all types of attack, specifically a prosecutor scenario, it can be an improvement upon k-anonymity in the case of a journalist scenario. While this scenario does require an underlying population database, the paper notes that while such a database might not be available, there exists methods to estimate the equivalence-classes for it using a sample, which would provide an approximation of k-map.

The paper proposes three alternative approaches to k-anonymization, utilizing the k-map method, and analyzes the results. They find that, while any situation where the prosecution scenario is relevant still requires a k-anonymization method to guarantee sufficiently low disclosure risk, the k-map method can provide significant benefits to both risk of information disclosure and utility in the form of information loss.
3.2.4 δ-Disclosure privacy, from syntactic to semantic privacy

Questioning the benefits and validity of the method for ensuring privacy that k-anonymity and l-diversity relies upon, The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing [17] proposes a new method for protecting sensitive data against attribute disclosure. It presents a metric for semantic privacy, which it calls δ-disclosure, which is the measure of how much information an attacker may learn from an anonymized data set, beyond what it could learn if the data had released all quasi-identifier data and sensitive attribute data separately. This is in opposition to the concept of syntactic privacy that privacy models such as k-anonymization and l-diversity use.

The paper posits that while k-anonymity concerns itself with identity disclosure, and in combination with l-diversity does afford some protection from attribute disclosure, it is a limited form of protection because it only protects against an attacker learn the exact value of an individual’s sensitive attribute value in an equivalence class. Meaning that if an equivalence class has a distribution of a sensitive attribute value that is much higher than the data set as a whole, then an individual who is identified to exist in that equivalence class will have a higher probability of having such a value for that sensitive attribute.

The measurement of privacy is done through several metrics, including how much knowledge an attacker can gain on an individual whose quasi-identifiers are known, as well as how accurately an attacker can predict a sensitive value of an individual using the most common value for the individual’s equivalence class. The paper notes that the latter metric would not be an exactly accurate representation of attribute disclosure, but nevertheless finds that it is useful.

For its measurement of utility, instead of utilizing the common metric of simple information loss, this approach to privacy measures empirical utility for specific workloads. This paper elects a classification workload, emphasizing that the workload with its chosen target attribute must benefit from the presence of sanitized quasi-identifiers, otherwise, the data-set could be trivially sanitized, meaning that all quasi-identifiers could simply be suppressed.

Through experimentation, the paper finds that an attacker has more to gain than a researcher from a data-set anonymized using common privacy models such as k-anonymity, l-diversity and t-closeness over one which has been trivially anonymized. The measurements for the gain for an attacker and the utility gain for the researcher were compared through the metrics on attacker knowledge gain and accuracy mentioned previously. Finally, an artificial data-set is presented, where they find that, depending on the specific workload, gained little to no privacy benefit from k-anonymity and l-diversity over trivial sanitation, while their own metrics allow for perfect privacy and perfect utility.
The paper reaches the conclusion that existing privacy models such as k-anonymity provide a poor trade-off with information utility and attribute disclosure. They also suggest research into the existence of real-world data-sets which match the properties of the artificial data-set they constructed, as well as into the design of better algorithms for both privacy and utility preservation.

### 3.2.5 $\beta$-likeness, limiting attacker sensitive-attribute information benefits

Attempting to create a better model for limiting the information gain an attacker can gain from a sanitized data-set, the paper Publishing Microdata with a Robust Privacy Guarantee [18] presents the notion of $\beta$-likeness. This measure aims to improve upon previous models such as t-closeness, which guarantee attempts to enforce a certain extent of uniform distribution in equivalence classes of quasi-identifiers. The paper notes some properties of t-closeness it sees as shortcomings, such as it not considering the relative distance between a distribution ratio in the whole data set compared to in an equivalence class, rather using an absolute measure of distance. $\beta$-likeness is then presented as a better measure, which implements both relative distance and the distinction between positive and negative information gain on attributes, and uses that to establish a threshold of information gain by an attacker.

Following its establishment of the $\beta$-likeness metric, the paper defines an information loss metric for both numerical and categorical data of each equivalence class, combines them, and establishes the single metric average information loss (AIL) for the entire data set. This is a generalized metric, rather than purpose specific like the metric utilized in the $\delta$-Disclosure model. To achieve the $\beta$-likeness threshold while focusing on the AIL-metric, the paper provides two schemes, one generalization-based scheme and one perturbation-based.

The generalization scheme, named BUREL (BUcketization, REallocation, $\beta$-Likeness) starts by performing a bucketization method which constructs partitions of the original data set into groups, or buckets. These groups are constructed in such a way that one can establish equivalence-classes which satisfy $\beta$-likeness by choosing a number of tuples from each group which is approximately proportional to the size of the bucket in relation to the distribution of a value in the sensitive attribute. Subsequently, it creates these equivalence classes by combining all the buckets into a single root equivalence class. It then splits this root node into two approximately equally large equivalence classes, both for which $\beta$-likeness holds. Each child equivalence class is then split, until it is no longer possible to create equivalence class children for which the $\beta$-likeness metric applies. When doing this splitting, the way in which the tuples are chosen from the various buckets to create equivalence classes ensures that the tuples are as close in their quasi-identifier space as possible. This process creates a number of equivalence classes which are both representative of the
proportions of the sensitive attributes in relation to the entire data set, as well as equivalence classes which contain quasi-identifier values which are close to each other.

The perturbation scheme attempts to alleviate a weakness in the generalization scheme, which might perform poorly in the presence of outlier data. Outlier data might result in the generalization scheme constructing very large equivalence classes, because of the way in which it necessitates a large extent of proportionality of sensitive values in each equivalence class in relation to the entire data-set. To avoid this, instead of creating equivalence classes which adhere to sensitive attribute value distribution of the whole data-set to up to a threshold $\beta$, it perturbs values of the sensitive attributes such that the information gain that an attacker would gain from seeing the anonymized data is no different from if $\beta$-likeness were to hold for a set of equivalence-classes.

This anonymized data does not produce immediately correct results for aggregation queries made against it, thus, proper distribution of sensitive attribute values needs to be reconstructed. This is done by creating a matrix which predicts the perturbation of the different sensitive attribute values in the data-set. The predictions produce an expected number of tuples containing a sensitive attribute value, for each value, based on the number of tuples containing that value in the original data-set. The prediction matrix is released along-side the perturbed data set, so that an approximated reconstruction of the original distribution can be made on the result of a query against the perturbed data. There also exists the possibility of releasing the original distribution of the different sensitive values for the data-set as a whole.

The scheme is then tested against approaches utilizing the $t$-closeness metric. The paper finds that the BUREL scheme outperforms the other approaches with regard to resulting $\beta$ threshold and $t$ threshold, as well as desired AIL, concluding that the $\beta$-likeness metric require a different approach to achieve good results. They also examine the performance of their perturbation approach, however note that it does not have an existing direct competitor, because of the way in which it treats anonymization. Its metric for data utility is measured not by information loss, because no quasi-identifier is altered from the data set. Instead, it focuses on the accuracy of its prediction of sensitive attribute values in the results of queries made against the anonymize data set, and finds that the results are promising.

Finally, its strength in regard to various attacks aiming at identity and attribute disclosure is examined. With the perturbation approach used for the second scheme mentioned above, it is immune to attacks aimed at identity disclosure, because of the nature of resulting sanitized data. While the generalization method BUREL is vulnerable to those types of attack, it is expected to perform very well regarding most types because of the way in and degree to which $\beta$-likeness ensures the protection of identity and anonymity. Against the attacks that the perturbation approach
is vulnerable to, it still appears able to strongly resist.

The paper suggests further research into the use of $\beta$-likeness to extend to numerical data. They make a note of the fact that the $\beta$-likeness model presented in the paper is meant for use on sensitive attributes of a categorical nature, and make some suggestions for things to focus on when doing research on the topic.

### 3.3 GDPR, further examination

#### 3.3.1 An examination of researcher obligations and legislative GDPR adherence

In a journal article [24], the GDPR and what responsibilities rest on health researchers is examined. Beyond outlining the framework of the GDPR, the article discusses the tensions between being able to perform research, especially regarding secondary use of data, beyond its original purpose and protecting the rights of data subject. The different requirements for usage of data is broached, and exemptions to those requirements for specific purposes is mentioned as relevant, such as the exemption for scientific research. With the many uses in research for a secondary use of data, the limiting the data for use in processing to only the original purpose it was gathered would severely restrict much research being performed, but is alleviated by the previously mentioned exemptions for scientific research.

The article goes on to mentioned different requirements for data holders to protect the rights of the data subjects. Specifically, the rights of data subjects for access, correction and erasure of their personal data, including their right to information regarding how and why the data is being used and who is using it. While requirements like consent may loosen because of its exemption status through scientific research, the requirements of transparency mentioned above do not automatically do so.

Subsequently, the article introduces the Irish Health Research Regulations 2018, which is an implementation of data protection in health research aiming to adhere to the GDPR. It provides an interesting showcase for what country-specific data protection legislation can look like as a result of the GDPR requirements. These regulations detail the specifics of how to fulfill the requirements processing health data. This includes safe and reasonable processing of data such that data subjects are not harmed, a detailing of the involved parties and the specifics of the project, deliberate planning of the process and a focus on ensuring correct data protection implementation, data minimization, and finally consent from data subject obtained prior to the start of the research, with a process for derogating the consent requirement through an application. There is an opportunity for researchers to ask for a somewhat generalized form of consent, and not a specific application of processing of health data, however it is noted that this kind of generalized consent cannot be used to circumvent the “essence
of the consent requirements”.

They find that while the GDPR does grant more rights to data subjects and oversight on the processing of data, it places restrictions on what data can be used and how. Despite this, though, it addresses some of these concerns through avenues which allow for exemptions from requirements such as consent in certain circumstances. As a particular drawback, it notes the lack of possible exemption from the right of data subjects to withdraw consent, which presents a problem for some epidemiology researchers.

### 3.3.2 A handbook

In this handbook [52], the author presents the topic of data privacy and provides a comprehensive examination of the GDPR legislation, providing useful introductions to each concept covered, from the scope of the GDPR, to the rights of the data subjects, the responsibilities of the data processors and controllers. He examines applicable organizational and technical security measures for ensuring data protection, how to ensure compliance with the lawfulness requirement of the data processing by correctly gathering data, what to be aware of and possible consequences.

The book serves as an excellent resource for anyone interested in gaining a thorough understanding of each of the topics and concepts included in the GDPR. It also provides a look into practical application and adherence to the GDPR and examines the regulations in the context of social media company Facebook and its massive data collection capabilities. Its final outlook onto the direction of data privacy and digital information provides some interesting food for thought on how our future might look like, and at the ends suggests vigilance in how we treat our digital societies to ensure a better digital future.

### 3.3.3 A case study on anonymization

A recent paper on GDPR considering anonymity [33] talks about some of the legal requirements stipulated in GDPR. It provides an introduction to the data protection requirements stipulated in GDPR in relation to data anonymization models and techniques. They state GDPR regulates personal data, but not anonymous data, i.e. data which does not contain direct or indirect identifiers. However, it also states that the definition of anonymity is non-trivial because of the possibility of combining the data with external data to re-identify individuals, which they refer to as a “background knowledge attack”.

For measurement of anonymity of a dataset, they write that it is considered “anonymous when re-identification is only possible with high effort or unlikely means.” They go on to talk about the consent required from individuals that data is collected from, as well as a few other legal requirements which are not particularly relevant to this project.

The fines for non-compliance with GDPR are, however, very relevant. They
state that fines can reach “up to 10 million Euros or 2% of annual global turnover (whichever amount is higher)” for some types of infringement, and “up to 20 million Euros or 4% of annual global turnover (whichever amount is higher)” for other types.

They mention some of the concepts previously broached in this thesis, in the background chapter, such as attribute types like identifiers, privacy models like k-anonymity and anonymization methods such as suppression and generalization to deal with these regulations. They point out the limitation of the attribute classification, because of the possibility of a wide variety of data possibly being able to be used to identify someone. In addition, there can be a large difference in how much identifying information some values of an attribute reveals, using the example of location with value of a small village and a large city.

The rest of the paper concerns case studies regarding data protection and lessons learned from them. A takeaway from the section on lessons learned is that asking for consent for data processing when gathering data may reduce the amount of data that needs to be anonymized. This method could also be used when gathering health data to increase the utility of the results.
Part II

The Anonymization Process
Chapter 4

Data for testing

4.1 Types of data

Before choosing an approach to use for anonymization, choosing a tool to perform the anonymization and executing the anonymization, the data that is to be tested on must first exist. There are a few approaches to gathering this data available. The first would be to utilize existing real-world data, the second would be existing fake data and the third would be creating test data myself.

4.1.1 Real-world data

Real-world data is data taken from an actual population, which means the data consists of information on real people and would contain sensitive and confidential information. There are several advantages and disadvantages to using this type of data.

4.1.1.1 Advantages

Real-world data would provide an accurate representation of what the kind of data that is the subject of this thesis would like. While there is a large possible variety in the form health data takes, this would provide a true example, and while not representative of the whole, it would still provide extra validity to any results that might be gained from this project.

Another advantage to utilizing existing real-world data is removing the work of generating data myself or examining the validity of fake data. The data would be readily available, and while an analysis of its representativeness in the larger scope of general health data might be a part of drawing conclusions based on the results of testing, the results applicability would be strengthened.
4.1. TYPES OF DATA

4.1.1 Disadvantages

In opposition to the advantages of utilizing real-world data, there are several drawbacks which make utilizing this type of data less practical. The first is the matter of gaining access to this type of data. Real-world data, because of its sensitive nature, is not freely distributed to anyone. It would require an organization which possesses the data to share it for the purpose of this project and to accept any risks associated with that process.

Storing and using sensitive data such as health data is strongly regulated. [51] It is necessary to protect the privacy of any individual which this data concerns and ensure that the data is fit for purpose, as well as ensure the data exists in a safe environment. With regards to the storage of sensitive data, there exists services which can be utilized. Particularly relevant for this project would be the Services for Sensitive Data (“Tjenester for Sensitive Data”, TSD) which is a service available for researchers at the University of Oslo (UiO) and other research institutions to assist with the storage and processing of data which require particular care with regards to sensitivity, privacy and data protection. With the basic package fee covered by UiO [58] this would provide a fairly reliable avenue for protecting the data.

With the other strong regulations on the data required by the GDPR [1], the advantages of being able to use accurate and representative data are quickly outweighed by the limitations placed on what they can be used for and how they must be protected.

Some of these regulations could be avoided by utilizing data not protected by the GDPR [30] and other similar data protection regulations, however this could raise serious ethical concerns. These regulations exist for a good reason, and not following these regulations simply because other groups of individuals are not specifically protected by them would go against the human rights that these regulations are founded upon.

I, personally, have very strong views on the rights to privacy of the individual and believe all people should be protected by regulations such as provided by the GDPR, and therefore do not wish to compromise the privacy of others by skirting regulations that I myself am protected by, simply for the purpose of giving this project a more representative and accurate set of data.

4.1.2 Existing fake data

Utilizing existing fake data is an alternative to real-world data. Existing fake data would be some set of data that contains health information on individuals which is to a certain extent what one could expect to find in a set of real data, but which explicitly does not connect to any real person.
CHAPTER 4. DATA FOR TESTING

4.1.2.1 Advantages

Like real-world data, utilizing existing fake data would negate the work required by producing new data myself. With the data readily available, there would be no need to spend time researching how to create representative data and creating a process for performing the generation.

Another advantage would be if this data was produced with a real data set as a basis, as in it attempts to simulate data found in a real-world data set, this data could have a high degree of representativeness for the specific type of health data that it simulates. While it would still be limited by the same lack of representativeness of health data, as a whole, that real-world data sets are limited by, its specific applicability would still lend credence to any conclusions drawn from analysis performed on it.

4.1.2.2 Disadvantages

As with real-world data sets, there are disadvantages to utilizing existing fake data. As with real-world data, there is still the matter of gathering this data. While fake data technically can be publicly available due to its fake nature rendering it no longer sensitive, data is still a valuable commodity, even fake data, and therefore companies and organization may not wish to make public any fake data sets that they possess. Thus, while there may exist a plethora of fake data sets used for a variety of purposes, such as training of users and testing of systems, many of these data sets are likely kept private internally within the organizations and companies that make use of them. While there exist data sets which are publicly available, the issue would be finding one that is fit for the purposes of this project, or making an agreement with an organization or company which does have an applicable fake data set.

An obvious candidate for a data set to use for this project would be the example data used in the “play” [41] version of DHIS2, basically a version of DHIS2 set up for training and test use, which is publicly available. This data sets includes fake information for Sierra Leone and provides a way for interested individuals to play around with the system and develop applications for it. The data, however, is not an accurately representative version of a raw data set that would be required for the purposes of this project. It is heavily limited with regard to personal data on individuals, and the way in which data is stored in DHIS2 makes it hard to employ as is without significant work going into constructing a more usable data set with the information from it.

While real-world data uses, obviously, actual data, a fake data set would necessarily be different from what an actual data set would look like. Beyond the limitations on general representativeness briefly mentioned in the advantages section, if the data is not directly modelled from real data, and even to some extent if it is, there will be some discrepancies with regards to the accuracy of the data and the representation of its particular
form of health data. This drawback can have a significant impact on the validity and credibility of any conclusions drawn from the results of analysis on the results. Without any real data to compare with, it would be difficult to know just how much the fake data deviates from a real example, further compounding the issue.

4.1.3 Self-produced fake data

Producing fake data, myself, is the final alternative for gathering data test data for this project. The data would be similar, in the physical sense, to existing fake data, but differs in its properties relevant for this project because of the way in which it was both produced and obtained.

4.1.3.1 Advantages

By producing the fake data specifically for this project, it will allow me to avoid needing to rely on other parties for gathering information. This grants freedom with regard to the process in which I gather and utilize the required data, not requiring permission on utilizing the data, neither having any limitations on storage, protection nor purposes.

A definite advantage would be the way in which the data could be created. While existing fake data will have been previously created for some specific reason, self-produced data can be produced for this specific project. This implies that the data can take a form that benefits the purposes laid forth for this project, that best highlights the general goals and challenges that would impact the process of anonymization in health data.

Without needing to gather data from a limited pool of resources, the amount of data produced is only limited by the ability to produce it. With a good and efficient process for producing data, this makes grants the opportunity to generate as much or as little data as required, making it possible to test on as many sets as desired, with varied forms and sizes, if necessary.

4.1.3.2 Disadvantages

To be able to produce this data, it is necessary to first learn the production process of such data, the properties that affect the quality of the produced data and how utilizing fake might affect the results of this project, which would require research into the different topics.

After researching the necessary information, the next challenge will be producing the data. There already exist some tools for simply generating a large amount of data such as generatedata.com [31] and Synthea [55] and services which does some form of data production [40]. Aside from those tools, creating a custom program which produces data is another option, but comes with the drawback of requiring development and testing, as well as proper configuration.
Finally, regarding the validity of the data for the purposes of the project. As with existing fake data, this data will necessarily deviate from real-world data, which comes with the same drawback regarding applicability, usefulness, validity, representativeness and credibility.

4.1.4 Summary

In summary, there are three methods to choose from, real-world data, existing fake data and self-produced fake data, all of which comes with their specific advantages and disadvantages:

Real-world data has the advantage of being accurate and representative of its type, but comes with the drawback of being sensitive because of its real-world nature, requiring special care with regard to gathering storage and use. Gathering the necessary data may prove difficult, in addition to it potentially being on a form that isn’t as applicable as desired to the purpose of this project.

Existing fake data, while settling the issue of special care regarding storage and use, suffers from the same drawbacks regarding difficulty to gather and potential lack of applicability, in addition to deviating from real-world data, causing potential drop in the validity and credibility of any conclusions resulting from the project.

Self-produced data solves the issues of difficulty to gather and applicability to the purpose of the project, because of the fact that it can be created independently and purposefully. Apart from the same drawback as existing fake data considering its effect upon the validity and credibility of any results, it also introduces the issue of researching data production, as well as potentially requiring development of a production utility.

4.2 Gathering data

4.2.1 Choosing a method

The method I ended up choosing is self-producing fake data. The advantages of using real-world data regarding the validity and credibility of results does not make up for the risks involved and the extra work involved with both gathering, transferring, storing and processing the data. The potential consequences of a breach of security because of mishandling of information in need of protection required by regulation could be massively negative. Not only are the fines sky-high, the human cost could be worse. The work being done through this project, and especially the benefits gained through using real-world data in stead of fake data, is, in my opinion, not sufficiently valuable to risk the potential negative consequences.

Utilizing existing fake data would have been the simplest option, if there had been any readily available data sets of sufficiently high quality, which
were also fit for the purposes of this project. However, with the DHIS2 training data set not being ideal, and the work of gathering a data set through another avenue, along with the prospects of the data sets nevertheless not being ideal for the purposes of this project, I instead electing to go through the process of producing the data myself. This allows me to fit the data for the purposes of this project, have as much data as I need and avoids having to rely on and consider extraneous factors.

4.2.2 Data quality

The first step to producing test data is knowing what affects the quality of data used in testing. In general, the basic factors concerning the quality of data is whether the data servers its purpose [36] and its utility measured by its ease of processing and analysis [21]. Thus, our data should both be able to fulfill the purposes this project is founded on and make the process of testing the different approaches to anonymization and the subsequent analysis of result easy.

The purpose of this project is researching approaches to anonymization and assessing their quality as a step in the process of releasing health data sets, following the regulations stipulated in the GDPR. This means that the data must have the same properties which that bars real-world data from being released raw, and the properties which makes the process of ensuring compliance challenging: the properties which constitute a breach of the regulations of the GDPR.

The GDPR concerns only the regulation of personal data, which means that any personal data in a real-world health data set would be what necessitates some form of data processing before release. The test data set must therefore also contain personal information in the sense that health data contains personal information. However, not only must our data contain personal information, it must do so to the extent that the challenges to anonymization faced by real-world data is reflected in the fictitious data set.

The reason anonymization of health data is challenging is because personal data in the health data sets is often what makes the data useful for many of its purposes. The age of a patient can inform a researcher of how disposed certain age groups are to certain diseases. The residential address of a patient can indicate how a disease spreads, and which areas are in need of more adequate medical infrastructure. For the test data to have properties representative of the challenges a real-world data set would face in the anonymization process, the loss of personal information from the test data must constitute a significant loss in utility.

Beyond these points, the data should be both easy to process, and should provide for ease of analysis. With the processing being done by a computer, the data must be on such a form and stored in such a way that processing of applying the different approaches to anonymization can easily done by a
purpose-designed tool, and the resulting data must be easily be comparable with the original data.

### 4.2.3 Generating the data

With an idea of what is required of the test data set, the following step would be to generate the data.

#### 4.2.3.1 Tools for generating data

As mentioned in the section on self-produced fake data, there exists some tools and services which could assist in this process:

“generatedata.com” is a simple website which allows anyone to create a data set with a custom set of columns of different pieces of information, such as name, country, date and number. While the resulting information could be somewhat useful, it is also very simple, and would be very fast and simple to implement to a sufficient degree in a custom program, and is therefore not very useful for my purposes.

Synthea is a much more complex tool, which allows for the generation of realistic health data for entire synthetic populations, based on real-world locations. It can generate a representative population for a specific city of a specified size, and provide a patient history, including medications and medical encounters. This tool is highly useful for working with health data in settings where realistic health data is important. The generated health data from Synthea is on a very specific format, however, and would require extra manipulation and effort to be useful for the purposes of this project. This lack of ease in processing the data is a factor that would reduce the quality of data when measured by the metrics presented in the previous section. While realistic population health data could be useful in this project, it is not the extent to which the data is realistic that matters, but that it is representative in the challenges realistic data would face in the anonymization process.

Finally, there is the Patient Generator service provided by MiHIN. It is a service for generating simulated patients for developing and testing healthcare systems. While data from this service could also be used in this project, it is excessive for the purposes of this project. In the same vein that the realistic data produced by Synthea is not necessary for the test data, the details produced by this service are extraneous. It would, however, be very useful in testing a healthcare application with many different purposes which require a wide range of details about its patients.

Beyond this tools, there is the possibility of creating a utility for the generation of data fit for the purposes of this project. Ultimately, this is the alternative I decided on.
4.2.3.2 Form of the data

The data I decided to create is meant to mimic the patient log of healthcare facilities where diagnoses had been made for the patients. The reason I decided on this type of data is because it is likely to include necessary health data on the patients, several points of personally identifiable information tied to each patient, and relevant healthcare facility information. The generation of this type of information is simple and can be easily implemented in a custom tool. In addition, this type of health data is likely prevalent in the health sector, and of highly desired for public release for several purposes, including research on infectious diseases.

Having decided upon a what type of data to represent, the next step is designing the data set. The data set will require personal information such as name, age and residential location; health information such as diagnosis; and healthcare facility information such as residing health worker, time of visit, and facility name and location.

When creating this data, it is not satisfactory that the data is straightforwardly random. Simple random data, especially in large quantities, is likely to be uniformly distributed. If 100 integers are randomly selected 100.000 times, each integer is likely to be selected approximately 100 times each. A uniform distribution is not likely to occur in the real world regarding health data, and a non-uniform distribution is part of what can make the anonymization process destructive. Therefore, the generator will designate a random weight to the different values of some of the relevant attributes, and then distribute the generation of values according to those random weights. While this will not make the data realistic in a sense that such a distribution is likely to occur in the real world, but rather, the properties which relate to challenges faced when anonymizing real-world data will be represented.

4.2.3.3 Implementation of custom tool for data generation

The tool itself is constructed using python. It starts by reading predefined lists of values for some of the different attributes, including names, locations, hospitals and diseases. It then creates a specified number of patients whose attribute value are distributed along the randomly chosen weights for some values, while others are completely random. A list of healthcare workers is also generated for each healthcare facility with a more limited set of personal details. The patients are then paired with a healthcare facility, along with the information related to the visit, including time of visit and related health worker, and finally the diagnosis the patient received, most of which is distributed according to randomly selected weights for the different values. After generating information on these patient visits, the tool compiles the information and writes it to a file in a machine-readable format. The data may subsequently be imported into a tool used for anonymizing data.

All this data is randomly generated, and it can be used to generate as much
data as needed. This allows for the generation of data sets of varying sizes, which may be useful for examining how well the different approaches to anonymization perform on differently sized data sets. It is important to remember, however, that the data generated through this process is not realistic in the sense that it will be an accurate representation of a real-world log of patient visits and medical diagnoses. It is not intended to be, and while it could lend some credibility to any conclusions drawn, it is not a requirement for testing the different approaches to anonymization.

As a note, the tool draws values from lists of actual names, so while none of the individuals generated are real people, certain instances might coincide with the values of real individuals. Because of the completely random nature of the generation process, this does obviously not constitute any handling of data that is protected by the GDPR. Similarly, the healthcare facilities used in this data generation tool are actual health facilities, but their names and locations are publicly available, and no information tied to them using this tool represents any real-world information, even should a visit coincidentally be similar to an actual event.

The full code for the script used to generate data can be found in Appendix C.
4.2. GATHERING DATA
Chapter 5

Anonymization approach

Having chosen data that can properly display results for the testing which is the purpose of this thesis, the actual process of anonymization needs to be decided. To construct a complete approach to anonymization, there are several components which need to be decided. The first is desired type of data release, which were described early in the background chapter. Subsequently, the privacy model to be used in the anonymization process. There are many possible privacy models which can be used for anonymization, and this chapter will consider the models described in the previous background and literature review chapters. Beyond a privacy model, the specific methods for transformation which can be used on the attributes of the data-set, in conjunction with those privacy models to achieve the goals of those models, need to be decided.

5.1 Data release

There were three types of data releases which were described early in the background chapter: microdata, tabular data and queryable databases. In this project, while other types might be useful, microdata is the type of data release that is most relevant and is the one which will be the goal of the anonymization process.

A microdata release will allow researchers to see data that is as close to the original raw data as possible. Only the information deemed to present a risk for the potential disclosure of confidential information to an attacker will be removed, preserving as much of the data utility as possible. This grants a data set which is flexible in the purposes it may be used for. Part of the reason for a data release is that controllers of the original data do not know all the possible legitimate ways the data may be useful. While they might be able to imagine different statistics, which can be used for various purposes, there likely exists many more, and yet more may be revealed in the future. With an as intact a data-set as possible, this presents a form of future-proofing.
The tabular type of data release mainly includes statistics learned from an original microdata set. This introduces a significant loss in utility for the released data, and presents problems in particular because it means only the specific statistics that were considered before the data release will be available for further usage and limits the usage of the data largely to the purposes considered by the controller releasing the data. Should any new purposes become relevant in the future, they might not be able to gain full utilization of the data by only relying on the statistics revealed in the tabular data release. A legitimate use case for this kind of data release might be a statistical agency deciding to release regular statistical data on various forms of data of public interest, such as Statistics Norway (Norwegian: Statistisk Sentralbyrå, SSB). They produce official statistics related to subjects such as the economy and the population of Norway. An advantage to doing data releases in this manner is that the data backing the statistics will be as correct as the original data is. While a researcher may be able to produce statistics from a microdata release, this data will inevitably vary from the original, meaning the statistics which are produced may be limited to a certain extent, which might mean a tabular data release might be better where statistics need as high an accuracy as possible. This project, however, aims for a more flexible kind of data release than this can provide, allowing for, in particular, researchers all over the world to use the data for any purpose they might think of.

The queryable database release is the closest to a microdata release in terms of utility, and might even provide more utility. In this type of data release, the data itself is not directly released, but rather some kind of interface for interacting with the data. The data backing this interface will remain largely intact, allowing interested parties to send in queries, asking for particular kinds of information. This intercepting interface will perform anonymization techniques on the data resulting from the queries it receives, before sending the data back to the source of the query. In addition, there may be placed restrictions on the kinds of queries that are able to be sent to the query interface, further ensuring no unintended disclosure of sensitive information occurs. This approach to releasing data could provide great utility for people making use of it, allowing for accurate data to back the anonymous data being presented, while at the same time ensuring the protection of confidential data. This, however, requires the upkeep of a service to allow this kind of querying. The data cannot simply be released and then ignored. Ultimately, this may prove costly in both time and money for the data controller. For this project, the desire is to release the data to everyone, and not have to worry about its subsequent usage.

With these types of data releases, the focus of this project will be on microdata releases. This allows for a kind of fire-and-forget type of approach, requiring no more interaction after the release, while also allowing for as flexible a release as possible.
5.2 Privacy models

There are several privacy models described in earlier chapters. There are, in particular, two kinds of privacy models: those which protect against identity disclosure and those which protect against attribute disclosure. These types of identity disclosure are described in the background chapter, and will be considered in light of the privacy models examined in this section. Beyond the type of disclosure they protect against, their approach to protection can be further divided into syntactic and semantic privacy.

The most common privacy model is k-anonymization, which considers a syntactic privacy measure for protection against identity disclosure. k-map, presented in the literature review chapter is a privacy model based on the k-anonymity model, but takes another approach to syntactic privacy measures. k-anonymity can be paired with l-diversity, which is a syntactic privacy measure against attribute disclosure. t-closeness is a further privacy model which, like l-diversity, seeks to prevent attribute disclosure. δ-disclosure is a privacy measure which focuses on protection against semantic attribute disclosure over syntactic. β-likeness is another semantic type of privacy model which protects against attribute disclosure. Differential privacy, a final privacy model which was presented in the background chapter, though an interesting approach to protection against information disclosure, is not particularly useful in the case of this project. The data release aimed for here is microdata, and the differential privacy model deals with mostly interactive settings such as the release of a queryable database, and in the cases where it is used for microdata it may significantly impact the purposes for which the resulting data is useful.

5.2.1 k-anonymity

This approach to anonymity was first introduced by Sweeney in 2002 [54] and has been widely studie, appearing in quite a number of papers and books, some examined previously in this thesis. It guarantees a minimum protection against the disclosure of the identity of any single individual through ensuring that any single combination of quasi-identifiers, an equivalence class, which appears in the data set, appears in at k records. In a prosecutor scenario, described in the background chapter, if a man of age 40 to 50 living who lives in Norway appears in the data set, and these attributes can identify the man, and k is set to 5, there needs to be at least five such men in the data set. Thus, if an attacker were to know of a specific Norwegian man of age 43 in the data set, he would only have a 20% chance of guessing correctly which man that is. Conversely, if an attacker was to take an identity from the data set to match against public information to discover their identity, there would be at least five individuals which would match that identity. This is a syntactic measure of privacy, it looks only at the characteristics of the data set, namely the number of individual records.
which appear similar, rather than the information contained in them. Crucially, this privacy model makes no guarantees for protection against attribute disclosure. It effectively may grant some protection, and likely it will because of the guarantee that several individuals share an identity which will in turn have a possibility of having distinct confidential attributes. This, however, is not a guarantee. All of those k individuals could share the same confidential attributes. In a medical set, during the previously mentioned prosecutor scenario, all five of those 40 to 50 year old Norwegian men may have lung cancer. Should this be the case, the previously mentioned 43 year old Norwegian would have confidential information disclosed. It is, thus, clear that while this privacy model may provide a part of the solution of a method of anonymization, it cannot do so on its own.

5.2.2 k-map

In the literature review section, a model for anonymization called k-map was described. This was effectively somewhat similar to k-anonymization. It relies on the assumption that should there be k individuals in the underlying population which a data set comes from that match a single identity in the released data set, then the privacy guarantees are the same as k-identity. Thus, the model uses the k-anonymity model on a data set of the underlying population, and then applies the identities in that data set to the released data set. While this does grant some protection from identity disclosure, namely a journalist scenario described in the background chapter, it provides no guarantees against a prosecutor scenario where the attacker already knows an individual exists in the data set.

Furthermore, this relies on the assumption that the data controller has access to a database of the underlying population that they can use the initial k-anonymity privacy model on. While there exists estimating a population based on a sample [26], however this raises a question as to their accuracy and if they sufficiently provide the guarantees k-map promises.

5.2.3 l-diversity

Using k-anonymity, a measure to protect against attribute disclosure, it is absolutely necessary to also employ protections to combat attribute disclosure as well. l-diversity is one option, guaranteeing a minimum diversity in the values that confidential attributes within an equivalence class. With an l-value of 3, this would mean that the value of the confidential attribute for the five Norwegian men mentioned in the previous example could not be cancer for all of them. While three of them could have cancer, two of them would have to have for example pneumonia and diabetes, ensuring that there is a variety of at least three values for the confidential attribute.
This provides a syntactic measure of protection. First, the equivalence class could have 300 individuals instead of 5, meaning if 298 individuals share one value for the confidential attribute and 2 individuals have another value, that would still provide poor protection. A different issue is that these values could be close. The three distinct values could be lung cancer, skin cancer and breast cancer in this case, or they could be an annual income of 500,000, 510,000 and 520,000. While these all ensure that there is no chance of exact discovery, there is still significant disclosure of information. Domingo et al. [23] refer to the attacks exploiting these weaknesses attacks as skewness and similarity attacks, respectively.

While $l$-diversity, then, can provide some protection from exact attribute disclosure, depending on the type of data that is being released it may not be an adequate measure. Especially in situations where the confidential attribute is a number such as income or some equivalence classes are especially large.

### 5.2.4 t-closeness

$t$-closeness comes as a response to some of the weaknesses of and potential attacks against the $l$-diversity model mentioned in the previous section. This model examines the distance between the distribution of a confidential attribute in an equivalence class versus its distribution in the data set as a whole. Meaning that if a confidential attribute with a domain of ‘yes’ and ‘no’ has a representation of 10% and 90%, respectively, in the entire data set, but 90% and 10% in the equivalence class, then these have a distinct absolute distance. The exact distance between these distributions can be measured with different metrics, but it is nevertheless the same concept.

Brickell and Shmatikov [17] consider this a syntactic privacy measure, but in the spirit of a semantic one. It is not so much the difference between what an attacker can learn about a value in a confidential attribute that $t$-closeness measures, but what an attacker can learn about the attribute as a whole.

This is a decent measure for attribute disclosure, though. It considers the distribution of an attribute in a data set versus an equivalence class, meaning that any particularly egregious discrepancies in this distribution, such as vastly overrepresenting a rare attribute in an equivalence class, or vice versa underrepresenting a common attribute will be caught by the privacy model.

### 5.2.5 $\delta$-disclosure privacy

Described by Brickell and Shmatikov [17], as presented in the literature review chapter, this is a semantic privacy model which focuses only on attribute disclosure in particular, though also effectively ensuring some form of protection against identity disclosure in the process. This model
5.2. PRIVACY MODELS

focuses on how much an attacker might learn about specific attribute values compared to how much they could learn about such values from a data set completely scrubbed of any quasi-identifiers.

Using a model focusing on each value of confidential attributes, this provides a very robust and representative measure of how much an attacker might learn semantically about from a particular data set, and mostly negates the need for employing another privacy model for ensuring protection against identity disclosure such as k-anonymity. The $\delta$-disclosure already ensures a proper distribution of values within an equivalence class, and as such, an attacker will learn no more about a single person in an equivalence class, regardless of its size.

5.2.6 $\beta$-likeness

A different improvement on the t-closeness privacy model, and similarly to $\delta$-disclosure privacy, $\beta$-likeness, presented in [18] is also a privacy model which focuses on protection against attribute disclosure. This differentiates between negative and positive information gain from an attackers perspective, where negative information gain is when a value is underrepresented in an equivalence class, and vice versa for positive information gain, with a value being overrepresented in an equivalence class. $\beta$-likeness considers negative information gain to enhance privacy, and vice versa for positive information gain, and sets a threshold for this information gain for each value of the confidential attribute.

Currently, the application $\beta$-likeness is limited to categorical data in confidential attributes [18]. This means that the privacy model might be useful for some purposes in the health field such as reporting diseases in patients, but will likely need to be adapted for other areas where attributes with continuous domains are considered confidential.

5.2.7 Summary

Each of these privacy models have their advantages and disadvantages. k-anonymity provides a simple, easy to understand and straightforward approach to protection against identity disclosure. While k-map may prove a less destructive on a data set, it requires knowledge on an underlying population, in addition to the fact that it provides no guarantees for protection against a prosecutor scenario. These, however, offer no guarantees for protection against attribute disclosure, and must be used in conjunction with a further privacy model. l-diversity is another simple privacy model which can be used in this regard, however it offers very little in the way of guaranteed protection against attribute disclosure. t-closeness offers a fairly strong guarantee for syntactic protection against attribute disclosure, which when used in conjunction with k-anonymity and l-diversity likely has the potential to provide robust protection against both attribute and identity disclosure.
Beyond these established models, are privacy models which attempt to provide guarantees for protection against semantic attribute disclosure. $\delta$-disclosure privacy and $\beta$-likeness are likely to give an accurate representation of the actual semantic information disclosure risk of a confidential attribute, more so than t-closeness, and may therefore be more successful in reducing the information loss which results from attempting to achieve t-closeness. They also reduce the complexity of through negating the need for separate privacy models for identity and attribute disclosure protection.

E lecting, then, to test out various approaches to examine which might provide a strong measure of data protection while at the same time preserving as much utility in the data as possible, I will be using (1) $k$-anonymity in conjunction with l-diversity and t-closeness, (2) $\delta$-disclosure on its own and finally (3) $\beta$-likeness on its own.

### 5.3 Data attributes

In the chapter on Data for testing, I produced a data set with several different attributes. The metadata for those attributes are shown in Table 5.1.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
<th>Domain</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient name</td>
<td>text</td>
<td>list of names</td>
<td>random, might include duplicates</td>
</tr>
<tr>
<td>Patient age</td>
<td>number</td>
<td>5-89</td>
<td>random weighted distribution</td>
</tr>
<tr>
<td>Patient biological sex</td>
<td>text</td>
<td>male, female</td>
<td>50/50</td>
</tr>
<tr>
<td>Patient residential location</td>
<td>number</td>
<td>Norwegian zip code</td>
<td>random</td>
</tr>
<tr>
<td>Patient diagnosis</td>
<td>text</td>
<td>18 diseases</td>
<td>random weighted distribution</td>
</tr>
<tr>
<td>Responsible health worker name</td>
<td>text</td>
<td>list of names</td>
<td>random, unlikely to include duplicates</td>
</tr>
<tr>
<td>Responsible health worker occupational role</td>
<td>text</td>
<td>general physician, nurse, specialist</td>
<td>random distribution</td>
</tr>
<tr>
<td>Health facility name</td>
<td>text</td>
<td>Norwegian hospital</td>
<td>random weighted distribution</td>
</tr>
<tr>
<td>Health facility location</td>
<td>text</td>
<td>Norwegian municipality</td>
<td>tied to facility name</td>
</tr>
<tr>
<td>Time of patient encounter</td>
<td>date/month/year</td>
<td>2000-2019</td>
<td>random weighted distribution on month and year</td>
</tr>
</tbody>
</table>

Table 5.1: Metadata for data set used for testing

In the background chapter, there were four kinds of attribute types described: Identifying, quasi-identifying, non-confidential and confidential. Before we can choose how to anonymize this data set, we need to identify what type each of the attributes listed above are.

The goal is to retain as much utility in the information as necessary, without
risking the disclosure of confidential information. Thus, we first define the confidential attributes, which do not increase the risk of information disclosure.

Sensitive data are those deemed by the GDPR to fall under “special categories of personal data”, which includes health data. The patient’s diagnosis is something that is definitively sensitive information and should thus be kept confidential [13]. Beyond that, no other information is particularly sensitive. Beyond that, the rest is personal information, which includes location the patient was treated at and the time of visit [9]. All of which is all protected by the GDPR, but aren’t afforded the same protections as the special categories of data.

The patient name is obviously identifying information. It’s important to remember that the health worker involved is also a protected by the GDPR, and thus, the health worker’s identity must also be protected, which makes their name also identifying. Whether the health worker’s personal data is of such utility that it needs to be included in the final data release, though isn’t as certain, and may end up being removed in favor of data deemed of higher utility.

The patient age, residential location and gender are quite clearly quasi-identifying information, but what about the health facility name and location, and the health worker occupational role? Furthermore, can the time of the patient encounter be used to identify the person? GDPR Recital 26, referring to whether a person is identifiable, that one should take into account all means which are reasonably likely to be used in the process of identifying a person, and further goes on to specify what it means by likely, which includes considering amount of time required, available technology as well as technological developments, meaning where technology will be going and whether that might have an impact [47].

So, what available technology should we consider, and how much time is more than reasonable? An obvious piece of technology that should be taken into account is social media. Users share a considerable amount of information about themselves, and it is not unlikely that a person will share information about going to a health facility, including when and where. The patient might not know prior to the visit that they will end up being diagnosed with a health condition, or at the very least what that health condition could be. Beyond the patient’s technological devices might be recording this information, whether it’s a smart phone, a car or a smart watch. This means that companies which control this information are likely to know where the patient was and at what time.

The question then becomes how much effort, including time and cost, it would take an entity, whether a private person or a company, to gather the data and use it to identify the person. A company which has tracking data can probably quite easily use it to determine if any of their data indicates a presence at that location at that time. For data shared on social media, it is quite well-known that some have used Facebook to gather
data on individuals, and a Business Insider article [34] details some of the information which could be gathered by Facebook apps at one point in time, which includes user location. Beyond such collection, any person connected to the patient on the social media platform is likely to see it.

All this considered, it is likely that time and location fall under “reasonably likely” to be used, and should therefore be considered quasi-identifiers, which would include hospital name. The health worker’s name is potentially public information as well, and should also be considered a quasi-identifier.

Thus, we have:

1. Confidential information
   • Patient diagnosis
2. Identifiers
   • Patient name
   • Responsible health worker name
3. Quasi-identifiers
   • Patient age
   • Patient gender
   • Patient residential location
   • Responsible health worker occupational role
   • Health facility name
   • Health facility location
   • Time of patient encounter

That gives one confidential attribute, two identifiers and seven quasi-identifiers, and no non-confidential attributes. Some attributes in this list are more important than other attributes, however, and will be weighted differently during the actual anonymization step to produce results with high utility. This will be discussed further in the section on configuring the anonymization tool.

### 5.4 Transformation of attributes

To achieve the requirements of the chosen privacy models, specific methods for anonymizing data will have to be applied to the different attributes in the data described in the previous section, specifically the quasi-identifiers. The goal is to maintain as much utility in the resulting data while simultaneously ensuring its protection. This means that the
transformations performed on the data need to be done in such a way as to make sure the conclusions one can draw from it are useful.

In some cases, knowing what groupings of different attribute values are useful in the end result may require domain-specific knowledge, especially some categorical attributes. For example, if occupation was involved in the data set and needed to be generalized, knowing which occupations would be useful to group together might require knowledge specific to the socio-economic domain.

In this case, I make the assumption that grouping together data which is semantically close is more valuable than maintaining smaller equivalence classes and fewer distinct values for their attributes. As an example concerning age in the data set, grouping ages in ranges close in value, such as 10 to 15 and 40 to 50, is likely to allow researchers to draw more sound conclusions about age demographics than grouping specific ages. An age grouping like 13 with 46 and 11 with 49 is less likely to be useful information to a researcher, even if the number of specific ages included in the generalization – 5 and 10 age values, versus 2 and 2 age values – and the number of records included in each equivalence class is lower. The logic of this transformation is extended to all relevant quasi-identifiers.

The identifiers will simply be removed, and the confidential attribute will be left as is.

The patient’s age will be generalized into age ranges, which will be configured as a hierarchy. The exact configuration of this hierarchy can be seen in the chapter on configuring the anonymization tool. The patient’s gender will go from male/female to being removed if it falls below its value threshold.

In the previously described data set, the patient’s residential location is a Norwegian ZIP code. I make the assumption that two ZIP codes which are numerically close are geographically closer than two which are numerically further apart, even though this might not be strictly true in absolutely all cases. With that as a basis, the residential location will be generalized by removing single digits from the zip code in each step of generalization, starting with the least significant digit. As in, 1234, followed by 123*, followed by 12** and so on.

The health worker occupational role will be removed if it falls below its value threshold. This attribute has few distinct values, and performing other generalization on this attribute will not be particularly useful.

The health facility name will be intrinsically tied to its location, so while it is not identifying information (for the patient or the health worker, it does “identify” itself), I consider its utility low, and therefore scrub it from the data set when facility location needs to be generalized. This transformation is essentially generalizing a health facility to its location.

The health facility’s location, initially a Norwegian municipality (Norwegian: kommune), will be generalized to a county (Norwegian: fylke), to a
region of Norway, and finally to Norway.
Time of patient encounter will be generalized from date to month to quarter to year to year ranges.

In addition to the transformation of attribute values in the records, the anonymization process will allow for the suppression, i.e. deletion, of records which are outliers in the data set. Applicable records will be those whose inclusion would produce a reduction of utility in the resulting data set larger than if they were to be suppressed, due to consequential further necessary generalization of attributes.

5.5 Measuring data utility

The utility of data can be measured in many ways, which includes both metrics for what is useful about data and algorithms are used to compute those metrics. As we have seen from some of the previous work on this subject mentioned previously in this thesis, this includes metrics like rate of successful classification of categorical data, information loss and successful prediction of confidential attribute values in perturbed data. Both information loss and the prediction metric are purpose agnostic, while the classification metric bases measures one specific purposeful use of the data and scores the data utility from successfully fulfilling that purpose.

In this project, we will be using a purpose-agnostic method, however the predictive model bases its utility measurement on the assumption that the confidential data of the anonymized data set is not truthful, i.e. it has been perturbed, and checks how well the data’s original information can be predicted. Thus, the predictive metric does not fit the transformations being done to the data in our case. That leaves the information loss metric, which is a fairly objective measure for data utility, while its usefulness will vary depending on purpose, it makes no assumptions about how the data will end up being used.
5.5. MEASURING DATA UTILITY
Chapter 6

Tools

Before going ahead with the actual anonymization of data, we need to choose a tool to assist with this process. The purpose of this chapter is to examine the different options available and finally choose one to use. This choice will have to take into account several different factors. The goal is to choose a tool which will allow me to test the different privacy models and approaches to achieving them to as large an extent and as thoroughly as possible, such that the results of the testing process will be of greatest possible value. The question is, what makes a tool good? The different features that the tool allows includes will be important: what privacy models are supported, what methods for achieving them are supported. How user-friendly the tool is, including how easy it is to install and configure, as well as how it presents important data and if it assists with the evaluation of any results of the anonymization process. An important note is that only free tools will be considered for use in this project.

The tools which will be considered are:

1. ARX
2. $\mu$-Argus
3. sdcMicro
4. Amnesia

6.1 ARX

ARX [14] is an open-source tool for anonymization. It provides a comprehensive suite of features for the anonymization process, including a graphical user-interface for interacting with the tool, various privacy and risk models, transformation methods for the data to be anonymized and methods for analyzing and reviewing the resulting anonymized data set.
6.1.1 Anonymization approaches

ARX includes several preconfigured settings which allow users to employ many different strategies when it comes to anonymizing their data, including privacy models, utility metrics and methods for transforming attributes.

6.1.1.1 Privacy models

It supports all the privacy models that were chosen in the section considering our approach to anonymization, and more [42]:

1. k-anonymity
2. k-Map
3. l-diversity
4. t-closeness
5. $\delta$-Disclosure privacy
6. $\beta$-likeness
7. Differential privacy

In addition to these, it includes several more models which have not been closely examined in this thesis, including $\delta$-presence, average risk, population uniqueness and sample uniqueness. This makes for a fairly comprehensive list of privacy models, not all of which have been examined in this thesis. They do cover the most common risks to information disclosure and attack scenarios against a data set.

6.1.1.2 Utility metrics

ARX supports a host of different models with which to calculate the quality of data resulting from an anonymization process. These focus on different aspects of data, some focusing on the values in each attribute, some on the attributes as a whole, while others focus on the properties of the data set itself. The most important model it includes is information loss, which is the one we are interested in for measuring the utility of our data [20].

6.1.1.3 Transformation methods

Most common methods used in transformation of attributes is also included in ARX. This includes the ability to apply both global and local transformations, generalization, random sampling, suppression, microaggregation and top- and bottom-coding. In our case, we will be utilizing mainly generalization and suppression techniques for achieving the goals of our chosen models [57].
6.1.2 Features

ARX has several features included in its software, which supports a complete flow for working with and anonymizing data sets.

6.1.2.1 Configuration of anonymization process

Through a graphical interface, users may configure the ARX software in several ways which makes it possible to perform the anonymization process in a variety of ways, and because of its included preconfigured settings for privacy models, utility measures and transformation methods, makes the usage fairly simple and straightforward, avoiding in-depth technical knowledge requirements from the user for how the different components work. It allows for the creation of value-hierarchies on the different attributes in the input data, which the transformation methods make use of during processing. This gives the user control of not only what data is important, but also how the data is important [19].

6.1.2.2 Result analysis

After the anonymization process has been executed on the data, the ARX software has several features for reviewing and analyzing the output data. This includes exploring the solution space which the process has created, meaning the variety of possible output data results, showing how they scored on the utility metric along with the disclosure risk metric and how the transformations configured in the previous step were applied. It includes a separate view each for analyzing utility and risk, giving several varieties of visual feedback [28, 49, 60].

6.1.3 Usability

The installation process of ARX is fairly simple. It does run on Java, so that needs to be installed in order to use ARX, beyond that, for Windows, Linux and OSX it is a simple installer, and other options are provided as well, including a runnable JAR file and a java library for use in other systems [25].

Using the tool is simple and straightforward, and while some domain-specific knowledge from the anonymization field is required to operate it, the graphical user-interface is easy to navigate, the tool is responsive and the actual anonymization process is fairly fast, even for large data sets. There is documentation for the various parts of the tool on its website, making usage even more convenient.

The tool is also licensed under the Apache License, Version 2.0, which is a very permissive license for free software, meaning it can be used and reworked very freely.
6.2 μ-Argus

μ-Argus [62] is another freely available anonymization tool. This tool sports a graphical user interface, and several features useful in the anonymization process, in particular when it comes to fine-control over input data. It is lacking much of the automated behavior of ARX, however.

6.2.1 Anonymization approaches

While it does feature a k-anonymity model which an input data set can be measured against, μ-Argus does not feature any preconfigured privacy models which may be applied to an input data set, nor any models for measuring the utility of the data set. Instead, it gives the user several options for exploring their data set and applying different transformations.

The main ways to perform anonymization in μ-Argus, after importing and configuring an input data set, is applying individual transformations to the different attributes in the data set, and then assessing the status of the data set with regard to k-anonymity. Included transformation methods are global recoding according to user-specification, a randomization scheme, a risk-based approach featuring local suppression, top- and bottom-coding, microaggregation and rank swapping [37].

6.2.2 Features

The main feature of μ-Argus is the control it gives users over the anonymization process. It sports an interface which allows the user to import an input data file, manually specify metadata for the data in the file, or include a metadata file in the import process.

This manual configuration of data gives the user fine-tuned control over both the input data file, as well as the properties of the variables. It includes several options for specifying identification information in each attribute, form of the values in the file and other useful properties. This metadata can also be used for any file the user has, meaning the user can use their own way of storing information.

After manually configuring the data of the input file, the user can then explore the data, checking singular and combinations of attributes for properties which indicate risk with regard to data protection, and then applying various transformations to relevant attributes in order to produce a more safe data set. Most of the steps involved in using this tool use k-anonymity as a basis.

6.2.3 Usability

The installation process of this is very simple. This tool, like ARX, runs on java and requires that to be installed. Beyond that, the download
page includes a zip-file which can be downloaded, extracted and then run straight out of the package, no installation necessary.

The tool is simple, but like ARX, requires domain-specific knowledge on anonymization to use. In this case, however, the fine-control it gives the user over the data of the input file also makes its usage a little more complicated. The different features for specifying metadata on the input file make the process of importing the data more technical, and thus somewhat more difficult than with ARX. The lack of any included privacy models for the tool to automatically apply user-specified transformations to work towards makes the usage even more cumbersome. And finally, the limitation of only being able to measure the data against k-anonymity, lacking any attribute disclosure specific measures in particular makes the tool less useful for the purposes of this project.

The tool does feature an extensive user manual, which gives a primer on both the field of anonymization as well as how to use the tool, which is very helpful for a new user.

6.3 sdcMicro

sdcMicro [56] is a tool for anonymization utilizing the R Project framework, a project for statistical computing [43]. Like µ-Argus, it does not support much automated behavior, but sports a fantastic graphical user-interface, functions great as a data visualizer and allows just as much control over the anonymization process.

6.3.1 Anonymization approaches

It mostly supports the same transformation methods as µ-Argus, including recoding, randomization, suppression and top- and bottom-coding. The process of performing the anonymization of the data is different from µ-Argus, but is similar in spirit, choosing a single attribute at a time and applying transformation.

The tool does provide some more metrics regarding the output data set though, including several risk metrics for information disclosure like for l-Diversity and SUDA2, the latter is not examined in this thesis, while also providing the k-anonymity metric. In addition, the information loss metric for utility can be used during the anonymization process to better advise the user on the output data set.

6.3.2 Features

Beyond the granular control the tool provides in the anonymization process, allowing individual transformation steps on the various attributes, the data visualization is impressive in this tool. At every step of the process, various visual aides help the user navigate both the tool and their
data set, facilitating more informed decisions by the user for how best to proceed.

6.3.3 Usability

In contrast to the previously examined tools, sdcMicro is somewhat more complicated to install. The tool is actually a package for the R Project. This means that the R Project needs to be installed, and then sdcMicro needs to be installed as a package through that framework. The difficulty of this process will depend on the technical expertise of this user, particularly if the user has any experience using framework which uses packages to expand its functionality. After installation of the sdcMicro package for the R Project framework, the user needs to start the tool through a command-line interface in the R Project application, which will open a browser-window with the sdcMicro web application, and then the tool is ready for use.

This process allows the tool to utilize the features of the R Project, but it makes the set-up part of the process quite a bit more difficult than ARX and µ-Argus. Once the tool is up and running, though, the user-interface is great. The design is intuitive, streamlining interactions for the user. One issue with the tool is that the visualization parts can be fairly unresponsive, especially when working with large data sets. This is quite a large drawback, seeing as the data visualizations are this tools strong points.

There is available documentation for this work available on github which can assist users experiencing any particular issues with using the tool, providing information on both basic usage and possible error messages. The documentation also provides information on quite a few of its various features, explaining how to utilize them. The software is provided under the GPL-2 license, a free license which requires derivative work to remain under the obligations of the GPL-2 license [56].

6.4 Amnesia

Amnesia [4] is the final anonymization tool we will examine in this chapter. This is the simplest among the tools we have covered, supporting very few features.

6.4.1 Anonymization approaches

This tool supports two methods for anonymization, generalization and randomization. After importing a data set into the tool, the user has the option to create or load hierarchies for the different attributes in the data set. The user then chooses a value for k-anonymity that they wish to achieve, and the tool will apply transformations for the various attributes according to the hierarchies which will produce a solution graph.
including all possible solutions, examining which transformations fulfil the k-anonymity requirement.

### 6.4.2 Features

Beyond those simple functions, there is a visualization tool for some statistics on for the various results in the solution graph. The documentation for the tool [5] claims the tool has some more functionality for showing further statistics and making queries against the result data set, however, I was unable to make those features work.

### 6.4.3 Usability

The tool is easy to install and use. It runs on Java, like some of the other tools, but is otherwise requires just a simple download to install. There is also an online web version of the tool available, which requires not set up on the part of the user.

The features of the tool are straightforward, and likely requires the least amount of domain-specific knowledge to make use of. The design is intuitive and easy to navigate, and the graphical user-interface is responsive to interactions, ensuring a mostly painless experience. I did encounter some issues during the creation of hierarchies for the different attributes, but the tool worked nevertheless, except those features related to statistics and queries mentioned above. The tool appears to work well with large data sets as well.

I was unable to find any licensing information for the tool.

### 6.5 Summary

We have four different tools to choose between ARX, \( \mu \)-Argus, sdcMicro and Amnesia. ARX offers the most built-in features, while \( \mu \)-Argus allows for the most granular control of the data sets. sdcMicro offers a more user-friendly user-interface, and Amnesia offers the simplest approach for anonymizing data.

While Amnesia is very easy to get started with and use, its limited functionality renders it not very useful for the purposes of this project, and while \( \mu \)-Argus and sdcMicro offers granular control of data, the features and ease of use of ARX makes it a clearly better candidate for our purposes. Especially its built-in support for more privacy models and automatic execution of transformations to the various attributes in the data set will make the workflow of the testing phase much smoother. Beyond that, its extensive suite of features for analyzing the results with regard to disclosure risk and information utility will greatly assist in the evaluation of the different approaches to anonymization.
Chapter 7

Execution of anonymization

In the previous chapters we have gathered the data needed for testing, chosen the approach to anonymization to test and chosen an anonymization tool to perform the testing on. Finally, in this chapter, we will perform the actual testing of the various approaches. First, we start by configuring the chosen anonymization tool, ARX, with the necessary settings, followed by executing the tests and taking an initial glance at the results, which we will examine and discuss more thoroughly in the following chapter.

7.1 Configuring anonymization tool

First off, we decide to utilize three data sets of different sizes: 2,000 records, 20,000 records and 200,000 records, all of which have been created using the method described in the chapter on Data for anonymization, and all containing the same type of information on the same form. All of the three data sets will be configured the exact same way in ARX.

7.1.1 Input data

To begin the process we import one of the data sets into ARX, this leaves us with the interface seen in Figure 7.1, and a reminder that this is all synthetic data.

In this user-interface, we can see our fake generated original data set on the left hand side and elements which can be used to configure the anonymization process on the right hand side and menu on top. There are also methods for reducing the data set to a sample, as can be seen on the bottom, and the different tabs called “Configure transformation”, “Explore results”, “Analyze utility” and “Analyze risk” separate the interface into the different stages of the anonymization process, which we will come back to more later.

The first configuration we need to make is to designate the types of the various attributes and which transformation methods should be utilized on the quasi-identifiers, through the interface seen in Figure 7.2.
7.1. CONFIGURING ANONYMIZATION TOOL

Figure 7.1: ARX after initial import of data

Figure 7.2: Configuring attributes
CHAPTER 7. EXECUTION OF ANONYMIZATION

Figure 7.3: Attribute metadata

All the transformations will be left as Generalization in our case.

That leaves us with the configuration see in Figure 7.3.

The colored dots on the left-hand side signify the different types of attribute. Red signifies an identifier, yellow a quasi-identifier and purple a sensitive attribute. In addition to these colors, there is a green color signifying a non-sensitive attribute, however our data set features no non-significant attributes. This interface also provides some more metadata, including the form of the various attributes.

7.1.2 Attribute hierarchies

Further, we designate the different hierarchies for the various quasi-identifiers through the hierarchy creation interface, first off is age, seen in Figure 7.4.

The data this hierarchy will be used on only contains ages from 5 to 90. This is a hierarchy attempting to group relevant ages together, keeping children, adults, the middle-aged and the elderly separate, and grouping them together as necessary. We also make sure to save this hierarchy configuration such that the same hierarchy can be used on the other data sets without necessitating creating the entire hierarchy again.

Following that, we apply a masking hierarchy to the ZIP code attribute, which the tool auto generates given the type of data and the type of masking we want. The hierarchy is seen in Figure 7.5.

As explained in the chapter on our approach to anonymization, this method masks the ZIP code a single digit at a time, starting with the least significant digit. The “#Groups” list on the left-hand side of the interface
### 7.1. CONFIGURING ANONYMIZATION TOOL

#### Figure 7.4: Age hierarchy

![Hierarchy wizard](image)

```
<table>
<thead>
<tr>
<th>Interval</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>[15, 20)</td>
<td>[15, 20)</td>
</tr>
<tr>
<td>[20, 25)</td>
<td>[20, 25)</td>
</tr>
<tr>
<td>[25, 30)</td>
<td>[25, 30)</td>
</tr>
<tr>
<td>[30, 35)</td>
<td>[30, 35)</td>
</tr>
<tr>
<td>[35, 40)</td>
<td>[35, 40)</td>
</tr>
<tr>
<td>[40, 45)</td>
<td>[40, 45)</td>
</tr>
<tr>
<td>[45, 50)</td>
<td>[45, 50)</td>
</tr>
<tr>
<td>[50, 55)</td>
<td>[50, 55)</td>
</tr>
<tr>
<td>[55, 60)</td>
<td>[55, 60)</td>
</tr>
<tr>
<td>[60, 65)</td>
<td>[60, 65)</td>
</tr>
<tr>
<td>[65, 70)</td>
<td>[65, 70)</td>
</tr>
<tr>
<td>[70, 75)</td>
<td>[70, 75)</td>
</tr>
<tr>
<td>[75, 80)</td>
<td>[75, 80)</td>
</tr>
<tr>
<td>[80, 85)</td>
<td>[80, 85)</td>
</tr>
<tr>
<td>[85, 90)</td>
<td>[85, 90)</td>
</tr>
<tr>
<td>[90, 95)</td>
<td>[90, 95)</td>
</tr>
<tr>
<td>[95, 100)</td>
<td>[95, 100)</td>
</tr>
<tr>
<td>[100, 105)</td>
<td>[100, 105)</td>
</tr>
</tbody>
</table>

Aggregate functions: Interval

Function Parameter:  
```
Figure 7.5: ZIP code hierarchy

informs us of how many distinct values there are for this attribute at every step of generalization.

In this data set, though perhaps not very realistic, the health worker attribute only contains three distinct values: General practitioner, nurse and specialist. There is little worth in doing any special kind of masking for this attribute. For the Health worker role and the facility name we simply specify a rule, through the menu, that the attribute should be suppressed entirely when generalized. See Figure 7.6.

This is very straightforward, simply marking the relevant column and clicking the “Create suppression rule” option will accomplish this. A somewhat more tedious hierarchy to create is the hierarchy for the health facility locations, namely from municipality to county to region to country. The interface is not particularly suited to this type of manual grouping of categorical values, which happen to have semantic meaning when it comes to geographical location. The end result ends up looking like Figure 7.7.

The user interface here necessitates the creation of a box signifying a group one at a time, setting each of their properties and values individually. Furthermore, the list on the left can only be interacted with through the buttons on the bottom left, selecting one single value, which can then be moved up or down the list one step at a time.

The list of municipalities was initially sorted in alphabetical order, so to being the organization of the list, each county was processed in alphabetical order, moving each municipality next to their relevant county, also in alphabetical order. The following step was to group the various counties
7.1. CONFIGURING ANONYMIZATION TOOL

Figure 7.6: Creating a suppression rule

into regions of Norway, i.e. the east, the west etc. The user interface allows for no re-ordering of the group categories, however, which necessitated that relevant county categories had to be removed from the interface and recreated in a suitable location, meaning next to a county which exists in the same region. Subsequently, each individual municipality was then moved, one at a time, utilizing the move up and down buttons, to their new location, next to their respective county categories.

As a personal note on this tool, this specific user interface could gain vast improvements through implementing some drag and drop features, and in particular letting the user move several values in the left hand list at the same time, as well as allowing the reordering of the category groups on the right hand side. This could facilitate a much smoother workflow, especially when experimenting with the user-interface. On the other hand, grouping each county by their respective regions initially would likely also have made the process significantly less time-consuming.

Moving on from the geographical location hierarchy, the final hierarchy required for the quasi-identifiers is the time of the patient encounter, which in this case is a date. See Figure 7.8

Here, we specify possible generalization of the date attribute. Its original form was date/month/year, so we select month/year, year, decade and century as further useful generalizations. While more granular generalization steps could be useful, the tool does not appear to support quarter of year together with year, and there doesn’t appear to be any way
CHAPTER 7. EXECUTION OF ANONYMIZATION

Figure 7.7: Health facility location hierarchy
7.1. CONFIGURING ANONYMIZATION TOOL

Figure 7.8: Hierarchy for time of patient encounter

to group years in smaller ranges than decades.

7.1.3 Attribute utility

Having created hierarchies for the relevant quasi-identifiers, we need to decide the weight of each attribute with regard to their impact on the utility of the data set. To weight the different attributes, however, we must make some assumptions about the purposes of the data. This is not necessarily easy to do, but the context in which the release happens and what is being released, we can guess, not necessarily the exact purposes for which the data will be used, but what about the data makes it interesting.

Our data set is specifically a data set about the diagnoses made on patients at various health facilities, for a set of specific diseases. We, therefore, make the assumption that the most interesting information in the data set has to do with the diseases themselves and the patients which were diagnosed with them. The data set includes information contextualizing this diagnosis, however. Including where it took place and what health worker was responsible for the diagnosis. The purposes for information about health workers and the diagnosis they make might be useful for several purposes, including statistical data about how many of the various diagnoses are made by different occupations in the health sector. It might also be interesting to know how the various hospitals differ in their rate of diagnosing the diseases. Another piece of information which might be useful is the rate of infection with the diseases, which would make the time of the diagnosis important, as well as the geographical location of either the hospital or the patient’s residence, and information about the patient themselves, such as age and sex, might also be helpful in that regard.
We are going to make four different levels of weight for the attribute. The patient’s specific diagnosis will not be transformed in any way, to keep the truthfulness of the data. Beyond that, information about the patient will be weighted the heaviest, including their age, sex and place of residence. The next level will be information about the diagnosis itself, including where it was made and when. Third will be information about the health worker, and fourth will be the facility name. The facility name infers its location, and the location of a facility gives some information about what facility it is, which means the information added by the facility name attribute, which has not been provided by its location, is low. We weight the four levels 0.8, 0.6, 0.4 and 0.2 in the weight distribution in ARX, as demonstrated in Figure 7.9.

We also specify the means with which we want to measure the utility of the data set, information loss. We want to measure the information loss of the various attributes utilizing an arithmetic mean, ensuring we properly account for outliers in the attribute distribution. See Figure 7.10.

We can also specify that ARX may perform local suppression to reach the requirements specified in the privacy models. This enables ARX to delete outlier records in the result data set which prevent the data from conforming to, for example, various distribution requirements for the equivalence classes.

We decide to specify a 5% limit, see Figure 7.11 to the amount suppressed records in the output data set. This allows ARX to remove some outlier information which prevents otherwise high-utility solutions from fulfilling
7.1. CONFIGURING ANONYMIZATION TOOL

7.1.4 Privacy models

The final step before applying the anonymization process is configuring the privacy models. We will be using three different combinations of privacy models, one featuring k-anonymity, l-diversity and t-closeness, one feature $\delta$-disclosure, and finally one featuring $\beta$-likeness, more specifically enhanced $\beta$-likeness.

The main objective of these privacy models is to obtain an anonymized data set which sufficiently protects against information disclosure, but is not unnecessarily distorted to reduce the utility of the data set. We therefore choose the values for the different privacy model metrics to aim for this goal.

The configurations will be as seen in Table 7.1.

<table>
<thead>
<tr>
<th>Privacy model</th>
<th>Approach 1</th>
<th>Approach 2</th>
<th>Approach 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k-anonymity</td>
<td>$\delta$-disclosure privacy</td>
<td>Enhanced $\beta$-likeness</td>
</tr>
<tr>
<td></td>
<td>l-diversity</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>t-closeness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configuration</td>
<td>k: 10</td>
<td>$\delta$: 2</td>
<td>$\beta$: 2</td>
</tr>
<tr>
<td></td>
<td>l: 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>t: 0.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.1: Configuration of anonymization approaches

A k-value of 10 should sufficiently protect against identity disclosure, giving an attacker only a 10% chance of correctly guessing an identity, given that the attacker already knows an individual is in the data set. Beyond protecting against identity disclosure, the l-value of 3 will ensure some variety in an equivalence class, making exact attribute disclosure impossible, and a t-value of 0.2 will ensure that even should the equivalence class be large, the amount of information disclosed will
We choose a $\delta$-value of 2 to with a similar strategy to our $t$-value, ensuring the proximity of distributions of sensitive values in the equivalence classes of the anonymized data set to the data set as a whole. It is for this same reason that we choose a $\beta$-value of 2.

7.2 Executing anonymization

We are finally ready to perform the tests of the various approaches to anonymization on the synthetic data. We execute each of the privacy model configurations on each of the data sets of sizes 2,000, 20,000 and 200,000, to see if the size of the data set matters significantly with the approaches we have chosen.

To initiate the process, we use the interface seen in Figure 7.12.

7.2.1 ARX efficiency performance

For the record, the processor of the computer these tests were performed on is an Intel i5-2500k, purchased in 2011 and used heavily since.

When testing on the data set consisting of 2,000 records completed instantly, no waiting time required. The test performed data set containing 20,000 records takes a few seconds, but is still very fast. The test on the data set containing 200,000 records, however, is quite a bit slower, with the test taking about five minutes to complete, as shown in Figure 7.13.
7.3. RESULTS

Figure 7.14: Options for reviewing the anonymization results

Should the data set increase in size on orders of magnitude, it is likely that ARX will use a significant amount of time to complete its task. This is worth considering when designing a workflow around performing anonymization on testing with especially large data sets.

7.3 Results

This execution will perform all possible combinations of transformations given by the established specified transformation methods and their accompanying hierarchies. This produces a solution space which can be explored and reviewed, which will be briefly presented in this section, and further examined and analyzed in the subsequent chapter. The results can be reviewed through three different main categories, pictured in Figure 7.14, where the left-most category is the interface used to configure the anonymization approach.

7.3.1 Exploring output

The ‘Explore results’ category gives several methods for exploring the solution space produced by ARX, including a graph, a list and tiles, displayed in Figure 7.15.

This shows the tile view of the solution space. The color of the data set indicates its score on the utility metric. The color of its border indicates whether the solution fulfills the criteria of the previously specified, green indicating fulfillment, and red indicating failure to fulfill. No invalid candidates are displayed in the above example. The ideal solution, having the highest utility score among successful candidates, is marked with a yellow border. The solution marked by the user is indicated with a darkened color. The lower part of the screen includes a filtering mechanic on the left, a list of all the various transformations in the middle, and some properties on the right.

The output data set can be further explored in the categories ‘Analyze utility’ and ‘Analyze risk’. After executing the anonymization process, ARX will choose the highest utility solution that fulfills the anonymization criteria by default, however each individual output solutions be selected by the user, both those that fulfill the anonymization criteria and those that don’t.

1.3.2 Analyzing utility The ‘Analyze utility’ category has three main views. One shows the original data set and the selected output solution, as well as various statistics concerning their attributes. See Figure 7.16.
CHAPTER 7. EXECUTION OF ANONYMIZATION

Figure 7.15: Exploring the results

Figure 7.16: Analysis of output data
### 7.3. RESULTS

#### Figure 7.17: Quality models for the results

The left side shows the original data set, while the right side shows the anonymized output data set, showcasing the different transformations performed on the various attributes. As can be seen from the picture above, the name in each record has been scrubbed, while the age has been generalized into ranges, with the age range 5 to 50 being displayed in the picture. Using the bottom portion of the interface, the user to inspect can statistics and properties for each attribute, such as amount of suppressed records and size of equivalence classes.

A second view displays the quality of the output data set for the user to review, including overviews for qualities of individual attributes and on a data set level. See Figure 7.17.

The final view isn’t relevant in our case, but allows a user to evaluate the performance of the output data set compare to the original in a classification scenario. This would be useful if a purpose-specific metric was used for data utility measuring classification performance, such as in the paper introducing the $\delta$-disclosure metric [17]. Our utility metric is to a large extent intentionally purpose-agnostic, beyond weighting different attributes, such that the data controller does not need to particularly consider nor be aware of all potential current and future purposes for the data.

#### 7.3.2 Analyzing risk

The final category which assists with the analysis of the output data set is the ‘Analyze risk’ category. This category visualizes different information about the anonymized data set with regard to what risks the data might
be exposed to. It has three main graphical interfaces for visualizing such risks. The first plots a graph which visualizes the distribution of risk over the data set, indicating the amount of records that are exposed to various levels of risk, as shown in Figure 7.18.

Here we can see the original data set on the left, the graph indicating that all records are at high risk of re-identification under a prosecutor scenario. Our anonymized data set on the right, however, indicate that the risk is very low for all records in the data set, with a maximum risk below 5% for all records. The sections on the bottom of the interface display more statistics for both the input and output data sets.

The next view shows information regarding the quasi-identifier attributes of the input and output data sets, shown in Figure 7.19.

For each quasi-identifier and combination of quasi-identifiers the view displays how unique its values are in its domain, unique values being represented by having high distinction, in addition, the degree of separation between the different attributes is displayed, indicating how correlated they are. The displayed quasi-identifiers and combinations there-of can be filtered out using the section in the lower-left corner.

Finally, there is a view for visualizing the general risk of the input and output data sets in relation to different attacker models, or risk scenarios, shown in Figure 7.20.

The attacker models considered are the prosecutor and journalist scenarios examined earlier in this thesis, as well as an additional scenario named marketer attack model which has not been examined. According to the
7.3. RESULTS

Figure 7.19: Risk for quasi-identifiers in the results

Figure 7.20: Risk in various attack scenarios for results
ARX documentation [42], the marketer attack model aims at re-identifying a large number of individuals, meaning that it’s only successful if it accomplishes a high degree of identity disclosure.

7.4 Summary

With this, we have tested three distinct approaches to anonymization, utilizing k-anonymity, l-diversity, t-closeness, δ-disclosure and β-likeness, on data sets of three different sizes, consisting of 2,000, 20,000 and 200,000 records respectively, which provides us with nine sets of results. Some of the results have been briefly showcased here, along with the various features included in ARX which facilitate a process for review and analysis, which will be the focus of the next chapter.
Chapter 8

Results

After completing our testing of different approaches on the data sets of various sizes, 2,000, 20,000 and 200,000 records, there are nine sets of results to analyze. This chapter will focus on analyzing the results from the data set with 20,000 records, examining how the three different approaches to anonymization affected the results regarding data protection and utility. The remaining sets of results will then be used to study the effect the size of the original data set had on the outcome. The subsequent chapter will then discuss these results in relation to the motivation, aims and research question of the thesis.

For brevity, in this chapter, the approach utilizing k-anonymity, l-diversity and t-closeness will be referred to as the k-approach, the approach utilizing δ-disclosure privacy as the δ-approach and the approach utilizing β-likeness as the β-approach. Furthermore, the data set containing 2,000 records will be referred to as the small data set, the data set containing 20,000 records as the medium data set and the data set containing 200,000 records as the large data set.

8.1 Transformations and utility scores

When ARX executed the anonymization process, it applied various transformations on the input data set, in addition to suppressing some outlier records, in order to comply with the restrictions specified by the different privacy models. Each output data set is specified by a set of integers indicating the level in the transformation hierarchy for each variable which was applied, a higher number indicating a more extensive transformation. The hierarchies for the attributes age, patient residential location, health facility location and time of visit have been presented in the previous chapter, and can also be found in Appendix A, figures 1 through 4. The attributes patient sex, health worker role and facility name hierarchies are binary, being either entirely suppressed or unchanged.

The output data sets are scored according to their utility, which in our
testing was measured through information loss, with a lower score being better, indicating less loss of information. In addition, a certain amount of records could have been suppressed in each result, with a maximum of 5% possible suppression. For the three different anonymization approaches, we get the transformations compiled in Table 8.1, accompanied by utility score and remaining records.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Information</th>
<th>Remaining records</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-approach</td>
<td>[6,0,3,1,3,3]</td>
<td>0.346</td>
</tr>
<tr>
<td>δ-approach</td>
<td>[6,0,4,1,2,3]</td>
<td>0.393</td>
</tr>
<tr>
<td>β-approach</td>
<td>[7,0,4,0,1,2,4]</td>
<td>0.398</td>
</tr>
</tbody>
</table>

Table 8.1: Hierarchy transformations and utility for the anonymization approach results on the medium data set

The transformation numbers refer to the following attributes respectively: age, sex, residential location, health worker role, health facility name, health facility location and time of patient encounter.

The results show that the k-approach scores the best, with the lowest information loss of the three approaches, while the δ- and β-approaches score similarly. The transformations performed show that the k-approach favors information regarding the patient’s residential location, while the δ-approach favors health facility location, and the β-approach favors both health facility location, and health worker role, while sacrificing patient age and time of encounter. The β-approach also suppresses more records than the other two.

8.2 Equivalence classes

An interesting part of the results is the equivalence classes that result from the applied transformations. They are the distinct combinations of values for each quasi-identifier attribute in the data set, and are used in the calculations of metrics for the different privacy models. For example, k-anonymity requires all equivalence classes to contain at least k records.

ARX shows the number of equivalence classes and their sizes in the result data sets. It provides a minimum, maximum and average class size, and does this for both the pre- and post-suppressed result, displayed in Table 8.2.

<table>
<thead>
<tr>
<th></th>
<th>Minimum size</th>
<th>Maximum size</th>
<th>Average size</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-approach</td>
<td>40</td>
<td>477</td>
<td>194</td>
<td>100</td>
</tr>
<tr>
<td>δ-approach</td>
<td>39</td>
<td>916</td>
<td>352</td>
<td>55</td>
</tr>
<tr>
<td>β-approach</td>
<td>26</td>
<td>1133</td>
<td>424</td>
<td>45</td>
</tr>
</tbody>
</table>

Table 8.2: Equivalence class sizes for the anonymization approach results on the medium data set
CHAPTER 8. RESULTS

These values are affected by what transformations were applied to the original data set. The k-approach led to the highest number of equivalence classes, sporting the smallest maximum size and average size, while the β-approach led to both the smallest minimum size and the largest maximum size and the δ-approach ends up somewhere in the middle.

8.3 Data quality models

ARX provides several models with which one can attempt to measure the quality of the resulting data sets. In particular, it uses metrics related to the precision of data, the granularity of data, the distribution of values in attributes and the difference of values in cells between the original and anonymized data sets. Statistics are provided for both individual attributes and for the result set in general, and table below displays the results for the entire sets, with higher values indicating a better score. See Table 8.3.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Granularity</th>
<th>Distribution</th>
<th>Cell-value similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-approach</td>
<td>25%</td>
<td>41%</td>
<td>15%</td>
<td>27%</td>
</tr>
<tr>
<td>δ-approach</td>
<td>27%</td>
<td>53%</td>
<td>13%</td>
<td>25%</td>
</tr>
<tr>
<td>β-approach</td>
<td>33%</td>
<td>53%</td>
<td>12%</td>
<td>28%</td>
</tr>
</tbody>
</table>

Table 8.3: Quality model metrics for the anonymization approach results on the medium data set

The different approaches display similar results, with the k-approach scoring a little lower on the granularity of values in the attributes, and the β-approach scoring a little higher with regard to the precision of the data.

8.4 Risk distribution

ARX presents the risks to both the input and output data sets in various ways, and with various metrics. The first is a graph showing the distribution of risk over the data set. Figures 8.1 to 8.3 show the distribution risk over the results from the three different approaches.

The graphs display an incremental cumulative graph of records affected at thresholds of a percentage-based prosecutor re-identification risk, starting at zero. Thus, an initially steep rise in the graph would indicate a lower risk for prosecutor re-identification in comparison with a gentle slope. The prosecutor scenario is strong with regard to re-identification risk and using it as a metric provides a good indicator for data protection.

As we can see, they all display a risk distribution over the result data sets which is beneath 5% for all records. They show minor differences, with the β-approach displaying slightly higher risks at the maximum, and the δ-approach to comes out barely ahead of the k-approach.
8.4. RISK DISTRIBUTION

Figure 8.1: Risk distribution for result of k-approach on medium data set

Figure 8.2: Risk distribution for result of δ-approach on medium data set
8.5 Attacker models

In this thesis we have previously discussed two scenarios in which attackers may attempt to disclose information using a data set, whether identities or confidential attributes: the prosecutor scenario and the journalist scenario. In addition, the marketer attack model was briefly mentioned in the previous chapter.

ARX provides a visualization of the risks an output data set faces under these attack models, providing information about how many records in the data set are at risk of information disclosure, what the highest risk in the data set is, as well as the success-rate of information disclosure attempts. Displayed in Figures 8.4 to 8.6 are these visualizations for the results of each of the anonymization approaches.

This shows a very similar level of risk for each of the listed attack scenarios, for all the different approaches, with no result set exceeding a re-identification risk of 5%, with the β-approach scoring marginally worse than the other two. The success rate displayed alongside the highest risk is successful re-identification effort over the whole data set. This gives an indication as to the general level of identity protection in the data set, as opposed to a worst-case scenario. In a prosecutor scenario, the worst-case scenario is the one that matters the most, because only a single specific individual is being identified, which might be in that
Figure 8.4: Attacker model risk for k-approach on medium data set
Figure 8.5: Attacker model risk for \( \delta \)-approach on medium data set
8.5. ATTACKER MODELS

Figure 8.6: Attacker model risk for \( \beta \)-approach on medium data set
worst-case scenario. For the journalist scenario, and the marketer scenario in particular, the success-rate over the data set is more likely to give an accurate representation of the risks faced by the data set. Even though the \( k \)-approach has the best score and the \( \beta \)-approach has the worst score for the highest risk, the opposite is true for the general success-rate, with the \( k \)-approach scoring the worst and the \( \beta \)-approach scoring the best.

Beyond these risk metrics, ARX provides an interface for examining the separation of quasi-identifiers and the distinctness of attribute values, but the scores are not immediately individually comparable as an indicator of data set risk, and seem to be more useful as a way of acquiring knowledge on a specific attribute risk or gaining an impression of the level of data protection in the result data set.

### 8.6 Data set size

Having examined and compared the results of the different approaches to anonymization on the medium data set, this section examines the significance of the size of the input data set on the result.

#### 8.6.1 Utility

Looking at the resulting transformations for each approach, it is immediately clear that the size of the data set had a significant impact. Transformations which fulfil the requirements of the chosen privacy models are more extensive for the small data set and less extensive for the large data set, resulting in an increase in information loss of close to a third for the small data set and a drop in information loss of a similar amount for the large data set. Table 8.4 shows a comparison of the \( k \)-approach between the small, medium and large data sets.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Info. loss</th>
<th>Suppr. records</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )-approach small</td>
<td>([6,0,4,1,1,3,3])</td>
<td>0.475</td>
</tr>
<tr>
<td>( k )-approach medium</td>
<td>([6,0,3,1,1,3,3])</td>
<td>0.346</td>
</tr>
<tr>
<td>( k )-approach large</td>
<td>([7,0,3,0,1,2,3])</td>
<td>0.226</td>
</tr>
<tr>
<td>( \delta )-approach small</td>
<td>([5,0,4,0,1,3,4])</td>
<td>0.453</td>
</tr>
<tr>
<td>( \delta )-approach medium</td>
<td>([6,0,4,1,1,2,3])</td>
<td>0.393</td>
</tr>
<tr>
<td>( \delta )-approach large</td>
<td>([6,0,3,1,1,2,3])</td>
<td>0.276</td>
</tr>
<tr>
<td>( \beta )-approach small</td>
<td>([6,0,4,0,1,3,4])</td>
<td>0.458</td>
</tr>
<tr>
<td>( \beta )-approach medium</td>
<td>([7,0,4,0,1,2,4])</td>
<td>0.398</td>
</tr>
<tr>
<td>( \beta )-approach large</td>
<td>([4,0,3,0,1,3,3])</td>
<td>0.252</td>
</tr>
</tbody>
</table>

Table 8.4: Hierarchy transformations and utility loss for anonymization approaches on all data sets

Regarding the equivalence classes of the resulting data sets, the results are less obvious. Table 8.5 shows the difference in values in the various
8.6. DATA SET SIZE

size for each approach on the small and large data set compared with the medium.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>k-approach small</td>
<td>+13</td>
<td>-114</td>
<td>-18</td>
<td>-89</td>
</tr>
<tr>
<td>k-approach large</td>
<td>+1</td>
<td>+150</td>
<td>+18</td>
<td>+797</td>
</tr>
<tr>
<td>δ-approach small</td>
<td>-27</td>
<td>-684</td>
<td>-264</td>
<td>-33</td>
</tr>
<tr>
<td>δ-approach large</td>
<td>-17</td>
<td>+315</td>
<td>+34</td>
<td>+438</td>
</tr>
<tr>
<td>β-approach small</td>
<td>+6</td>
<td>-901</td>
<td>-311</td>
<td>-28</td>
</tr>
<tr>
<td>β-approach large</td>
<td>-12</td>
<td>+4</td>
<td>-26</td>
<td>+489</td>
</tr>
</tbody>
</table>

Table 8.5: Difference in equivalence class sizes in results for small and large data sets compared to medium

The only consistent change among the approaches is that the amount of equivalence classes fluctuates in line with the size of the original data set. There is also a change of the max equivalence class size for the different data set sizes, however the increase for the β-approach is not significant for the large size. The difference in minimum and average do not directly correspond to the size of the input data set.

Table 8.6 shows the percentage point difference in values for the various quality models for each approach on the different data set sizes.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Granularity</th>
<th>Dist. similar.</th>
<th>Cell. similar.</th>
</tr>
</thead>
</table>

Table 8.6: Difference in quality model metrics sizes in results for small and large data sets compared to medium

This data displays a mostly consistent increase in quality for an increase in data set size and drop in quality for a reduction in data set size. There are individual scores which contradict this, however. The precision in the result of the δ-approach on the small data set increases and the granularity for the same approach on the large data set experiences a drop. The granularity of the k-approach on the small data set remains unchanged when compared with the medium data set. In addition, there is a clear variance in the significance of the variance in quality in relation to size.

8.6.2 Risk

The risk distribution for the results of the small, medium and large data do not change significantly for the k-approach, experiencing only minor differences. The δ-approach, however, experiences higher risk both for the
small and large data sets, than with the medium data set, with the small data set, in particular, being at a high risk. The distribution of risk for the β-approach is more at risk for the large data set, while the small and medium data sets have similar distributions of risk. The risks for the various attack scenarios reflect this distribution of risk with the δ-approach on the small data set being the data set result most at risk, experiencing a highest risk value of 8.3% and an identification success rate of 1.13%.
8.6. DATA SET SIZE
Part III

Data Utility and GDPR Compliance
Chapter 9

Discussion

Through this thesis, I have attempted to answer my research question:

How can existing approaches to data anonymization be applied to health data to sufficiently comply with privacy and data protection regulations stipulated in the General Data Protection Regulation (GDPR), while preserving utility in the resulting data?

This research question covers three topics:

1. Existing approaches to anonymization
2. The GDPR
3. Utility in health data

In this chapter, I discuss my work and findings in the context of my research question and these three topics in particular. In addition, I discuss some of the limitations and weaknesses of my research and findings, and what effect they may have. Finally, I consider the lessons I’ve learned through working on this project and compile a set of recommendations for how to approach the anonymization of health data.

9.1 Methodology

9.1.1 Information gathering

To answer my research question, I first set out to gain an understanding of the current state of the field of anonymization. I examined several pieces of literature, including books, research papers and web resources, exploring the ways in which anonymization has been conducted over the past few years, including what privacy models have been used, their purpose and their performance. The literature review also covers the recent GDPR legislation, aiming to understand what the GDPR requires for data to be considered sufficiently anonymous.
9.1. METHODOLOGY

Through this research I examined several privacy models:

1. k-anonymity
2. l-diversity
3. t-closeness
4. $\delta$-disclosure privacy
5. $\beta$-likeness

These models formed the basis of my research into anonymizing health data and the subsequent testing phase.

In addition to those privacy models, during my research I examined how to quantify utility in data sets which these approaches were to be used on. Finding a variety of metrics, the strategy for determining utility most fitting for the purposes of this thesis ended up being information loss. The information loss metric is considered purpose-agnostic to a large extent, which is helpful when data is to be released without requiring the data controller to have extensive knowledge on how that data may be used. Particularly, how new purposes could become relevant in the future, adapting to contemporary contexts.

Delving into the GDPR legislation available through web resources, I found a variety of obligations placed upon data controllers. I focused on the obligations most relevant to the topic of this thesis, in particular, its scope, its constraints on processing personal data, and how data being anonymized affects the scope and constraints.

Finding that the GDPR affects all personal data on persons within the borders of the EU, citizens, residents or otherwise, there are strict limitations placed on how that data may be handled. This includes gathering consent from data subjects on how and why the data may be used, securing the protection of the data through technological and organization measures and potentially required oversight by outside regulatory bodies. Beyond this, the GDPR may allow for the loosening of the requirements regarding specifically consent when processing is done for research purposes, however the other restrictions still apply.

Crucially, the GDPR does not consider anonymous data to be personal data, regardless of if it is based on personal data, thus, anonymizing data sets are a way to avoid any and all restrictions introduced by the GDPR. The question is then what the GDPR requires for data to be considered anonymous. The specific requirements for if a data set is sufficiently protected are somewhat abstract, however, requiring individuals to not be identifiable, but stating that how to determine this should consider all means reasonably likely to be used when attempting to identify an individual. How to determine what it means to be reasonably likely further requires consideration of required time and cost investments in the context of currently available technology and technological developments.
That leaves five privacy models used in anonymization and a requirement for determining whether data resulting from an anonymization effort is anonymous.

9.1.2 Preparations for testing

Having established a basis for answering my research question, a method to answer it was required. Testing the various approaches would be necessary, but to that end, various preparations first needed to be made. The most important was gathering data to test on, what to test and choosing a tool with which to perform said testing.

Considering various approaches to gathering data, including what type of data to test on and how to gain access to it, I made the decision to produce fake data myself. This allowed me to create the data set that I wanted and to act independently of external factors. Thus, I created a script which created a data set containing patient encounters resulting in disease diagnoses of a variety of different diseases, with information relevant to the patient and the context in which the patient encounter took place.

This data set was specifically intended not to be representative of health data as a whole, but rather serve as a specific case of health data where the learnings gained could be applied to other cases. The script constructed this data drawing from pre-constructed lists of values for some of the attributes, such as patient names, diseases and hospitals, and attempted to emulate data non-uniformity through randomizing distributions of various attributes, attempting to mimic challenges in real data.

Having constructed the data, the next part was determining what anonymization approaches to test on the data. Having examined various privacy models in the literature review phase, I decided to go with three approaches utilizing privacy models which utilize different strategies for protecting the resulting data. One approach would utilize the established privacy models k-anonymity, l-diversity and t-closeness in concert, the k-approach, which would take measures to protect against identity disclosure and counteract attribute disclosure. In addition, two approaches, one utilizing \( \delta \)-disclosure privacy, the \( \delta \)-approach, and the other utilizing \( \beta \)-likeness, the \( \beta \)-approach, would serve as testing two less established models for privacy, which seek to provide guarantees for data protection focusing on attribute disclosure. Choosing these three different approaches was intended to showcase how choosing different kinds of privacy model could affect the outcome of an anonymization process.

Finally, choosing a tool which could assist in testing the chosen anonymization approaches on the gathered data would be important in order to accomplish my goals. I initially started development on a self-made tool, but later moved away from this approach, opting instead to utilize a free, publicly available tool. This allowed me to take advantage of existing progress in the field, particularly the extensive features available in some of the available tools. I considered four different tools, each with their ad-
vantages and disadvantages, but ended up settling for ARX, a tool with a comprehensive suite of supported privacy models, including all models I wanted to test, and features for analyzing utility and risk in the resulting data set.

9.1.3 Testing the approaches

With all the necessary preparations in place, the final step was to perform the testing of the various approaches on the gathered data using the chosen tool. This involved constructing hierarchies for generalizing quasi-identifier attributes, choosing their weight in a utility sense and configuring the tool with privacy models for each approach. This configuration was done for data sets of three different sizes.

The outcome of the testing process was nine anonymized output data sets, with each data set having an accompanying analysis of utility and risk.

9.2 GDPR

The previous chapter presented the results of the testing, the culmination of the methodology used in this project. This and the following sections will focus on discussing these results in the context of the research question, specifically the three topics its covers, starting with the GDPR.

9.2.1 Disclosure risk

As stated previously, the metric the GDPR presents to decide if a data set is anonymous is somewhat abstract and open to interpretation. None of the risk factors considered by the various privacy models directly address the concerns in the GDPR’s metric. The GDPR metric focuses on the practical and effective risk that data sets are exposed to, taking into account real-world threats making use of various forms of technology and available data to breach the protections applied to a data set. This is hard to measure using any quantitative metrics. Instead, the privacy models focus on the inherent properties of the data itself to evaluate whether an attacker is likely to achieve information disclosure.

While the risk metrics shown in the results of the previous chapter might give an indication as to how anonymous the different results are, especially their degree of anonymity compared to each other, to comply with the GDPR metric, the privacy models need to be informed by a consideration of the practical and effective risks the data may realistically face. The evaluation of compliance with the GDPR, therefore, starts before the results are available. What information is available in the original data set, how an attacker may know this information and use it to identify individuals in the data set, and how much time and money an attacker would have to invest to perform this attack are important factors.
During the design of the three approaches to anonymization which were to be tested, the information in the original data set was examined and several identifiers and quasi-identifiers were identified. This process was based on whether the information in those attributes was publicly available, or if attackers could somehow gain access to such information. In particular, regarding time and location of the patient encounter, we considered the different ways in which companies gather location data through for example GPS trackers in mobile and wearable devices such as smart phones and watches. In addition, whether a patient may publish on social media websites such as Facebook that they were going for a health-related visit was taken into consideration. It might be unclear how likely it is for an attack to happen making use of this information, but it is at the very least plausible, and as such, the information was decided to be quasi-identifying.

Information such as patient name, age, sex and residential location were considered to public information. Name, specifically, was decided to be identifying, while the rest were considered quasi-identifying. Health workers are also entitled to protection under the GDPR, thus, their personal data in the data set was also considered, with their name considered identifying and their occupation quasi-identifying.

Thus, before even seeing the resulting data, assurances have been made that information and technology which is likely to be used by an attacker is taken into account. Had the restrictions of the GDPR not been considered, the way in which to select identifiers and quasi-identifiers might have resulted in a less robust protection against re-identification attacks. The degree of compliance with the GDPR will then depend on the privacy models’ ability to sufficiently protect information from identity and attribute disclosure attacks which make use of the previously considered personal data.

### 9.2.2 Identity and attribute disclosure

The risk faced by the anonymized data set is considered for identity and attribute disclosure. The risk presented in the results of the previous chapter is identity disclosure under different attack scenarios. The attribute disclosure risk is mainly covered by the privacy models which inform the way in which the data set will be anonymized.

As can be seen from the results, all data sets had a highest disclosure risk of below 10%, most ending up below 5% risk. The highest identification success-rate over an entire result data set was 1.13%. This indicates that the protections for identity-disclosure for all data sets are fairly high.

The protections against attribute disclosure provided by the privacy models depend on the values chosen for the different metrics. The l-diversity metric in the k-approach provides fairly soft protection against attribute disclosure. With the largest equivalence class of the result set from the application on the medium data set reaching 477 records, the chosen l-
value of three means very little. The k-approach includes another measure to protect against attribute disclosure, however: the t-closeness privacy model. This privacy model aims to ensure that even if an equivalence class should be large, the distribution of sensitive values within it should be similar to the distribution of those values in the data set as a whole. This final protection is what the δ- and β-approaches focus on, as well. In particular, the δ- and β-approaches provides a guarantee for each of the values of a sensitive attribute conform to a specific threshold of similarity to the general distribution of the data set, while the t-closeness metric only provides a threshold for the similarity of distribution for the domain of sensitive values as a whole within an equivalence class.

9.2.2.1 Prosecutor scenario

Under a prosecutor scenario, the absolute highest chance an attacker would have of identifying a targeted individual for the various data sets would be the highest value of 8.3%. Given the strong measures against attribute disclosure provided by the t-closeness, δ-disclosure privacy and β-likeness privacy models, the risk presented by this scenario is low. This, in combination with the considerations made for practical and effective risks considered in choosing identifying and quasi-identifying attributes prior to the test execution provide strong protections under this attack scenario.

9.2.2.2 Journalist scenario

Under the journalist scenario, the attacker does not have a specific target that they want to identify in the anonymized data set, instead, they wish to identify any individual. This can be done using externally gathered data, either through legitimate or illegitimate means. While the highest risk value for the data set is still relevant, the re-identification success-rate metric is also important. It indicates at which rate of success the attacker can be expected to identify individuals. The highest value for success-rate is 1.13%, with most values below 1%. As with under the prosecutor scenario, the risk values are low for re-identification and together with the protections against attribute disclosure the anonymized data sets possess strong protections under this scenario.

9.2.3 Compliance with GDPR’s requirements

Prior to anonymizing the original data set, considerations were performed following the requirements of the GDPR with regard to anonymized data. After executing the anonymization process, the risk of identity disclosure was low, both for individual records and over the data set, under both the prosecutor and the journalist scenario. In addition, the protections against attribute disclosure provided through privacy models are strong. Considering all of this, the result data sets can be considered to be in compliance with the requirements for data protection regarding
anonymized data in the GDPR.

9.3 Utility

Ensuring satisfactory protection of the data sets by competently applying anonymization techniques, resulting in a reduction of the risk of identity and attribute disclosure is paramount when releasing data. Having ensured the resulting data sets’ compliance with the GDPR regarding anonymous data, however, the next issue will be considering the utility of the resulting data sets. If a data set is unusable because of low utility as a result of the anonymization process, then the publication of data is meaningless. To ensure a high preservation of utility in the data set, it is important to understand what makes the data useful. This section will consider the utility of the data sets presented in the previous chapter.

9.3.1 Health data

The goal of this thesis was to research anonymization and utility in regard to health data. Health data is widely heterogeneous, and instead of focusing on all types of health data, the experiments of this thesis limits itself to a specific kind of health data, with the aim that conclusions drawn may to a certain extent be applicable to other forms of health data. Those conclusions could, thus, be used as a starting place for performing anonymization in other use cases, or as a pivot point for further research.

The specific health data examined in this thesis is a data set containing information on the diagnoses of patients during an encounter at a hospital in Norway. The information contained in the data sets that were tested on was not intended to be realistic with regard to the specific values used for sensitive information such as disease diagnosis, nor the specific information regarding health workers or facilities. Rather, the intent was to use a data set which mimicked challenges that a realistic data set would face when performing anonymization. Specifically, including non-uniform value-distributions of multiple attributes; attributes of realistic type, such as health facility information, patient information and health worker information and combinations of attributes which may present a challenge in such data sets, like timestamps and multiple locations.

By emulating challenges which realistic data would face, the goal is for the conclusions drawn from the results to be of higher credibility and validity.

9.3.2 Information loss

When anonymizing data, to understand the degree to which utility has been preserved in the anonymized data set, a quantitative metric is required. There are several ways in which data utility can be measured
9.3. UTILITY

quantitatively. An important choice to make when deciding on a metric, is whether there are specific purposes in mind that the data is to be used for. If the specific purpose of the data is classification based on the data set, employing a metric which scores the data on how well it performs as a classifier would be appropriate. An issue with this, however, is that not all the potential purposes that the data could be used for are known. Not only could the purposes the data can currently be used for be unknown, there could be more purposes which only become relevant in the future. In addition, various purposes could directly compete for utility in the data set, with one purpose favoring one attribute and a second favoring another. Thus, choosing a metric which maximizes for specific purposes may be counterintuitive. In this project, the metric chosen, information loss, is to a large extent purpose agnostic.

As can be seen from the results presented in the previous chapter, the measured information loss of the various approaches appears to be somewhat similar. The k-approach comes out slightly ahead for the small and medium data sets, while the $\delta$- and $\beta$-approach beats the k-approach for the large data set, with a very slight advantage for the $\delta$-approach. What can definitively be surmised, though, is that the size of the data set has a significant impact upon the utility score of the resulting anonymized data sets, with data sets of larger size resulting in a better score.

9.3.3 Transformations and hierarchies

During the anonymization process, transformations were applied to the various quasi-identifying attributes in the original data set in order to ensure the anonymized data sets meet the requirements of the specified privacy models. These transformations are applied by following a hierarchy of transformations which has been specified for each attribute. In the testing done in this project, all transformations were performed using generalization. For example, ZIP codes were generalized from 1234 to 123* to 12** and so on.

How those hierarchies are specified will obviously affect how transformations are applied. Ensuring that the generalization is done in such a way that the resulting groupings are useful is important, such as grouping semantically close values together. Beyond that, having fewer levels in a hierarchy would mean more generalization per step in the hierarchy, and might lead to a less optimal solutions with regard to information loss.

As can be seen from the results in the previous chapter, the transformations which yield the lowest information loss while simultaneously satisfying the requirements of the privacy models varies. The transformation over all the data sets which yields the best utility score is the $[7,0,3,0,1,2,3]$ transformation on the large data set using the k-approach, which yields a utility score of 0.226, where lower scores are better. Even with this best-case result, the transformation performed are still fairly extensive. The age of the patient has been generalized to the 7th level of its hierarchy, the health
facility name has been removed, the ZIP code has been reduced to a single
digit, the time of patient encounters have been generalized to decades and
the location of the health facility has been generalized to a region of the
country. Only patient gender and health worker occupational role have
been left unaffected.

9.3.4 Attribute weights

To ensure that the transformations which were performed according to
the specified hierarchies maximize utility in a good result, the weights
for the utility of the different quasi-identifying attributes can be adjusted.
This adjustment of weights can be seen as moving away from purpose-
agnostic utility scores, however, it may still prove to be a useful measure.
Specifically, the data set which is being released may have a context which
informs what the interesting pieces of information in the data set are. Our
data set concerns patients and the diagnosis of disease, thus we can
reason that the information which more directly relates to that will be
more useful. As such, we weight the information regarding the patient
higher than information regarding the health facility and health worker,
and the context of the diagnosis, including time and place of diagnosis,
 somewhere between those bounds. Changing the weights of the attributes
will necessarily impact the measurement of utility, so two utility scores
based on different weights may not be directly comparable, which is
important to take into account when experimenting with different weights
to reach a good result.

9.3.5 Quality models

Beyond the information loss metric which talks about the utility of data,
ARX provides several models which measure various qualities of the
anonymized data sets. These models include: a model for the precision of
attributes, related to the degree to which the attribute has been generalized;
a model for the granularity of the generalization intervals; a model for
the non-uniform entropy level of an attribute, meaning the similarity of
distribution of an attribute’s values compared with the original data set;
and a model for the cell-level similarity of the original and anonymized
data sets. While the information loss metric may be a good way of
quantifying the utility of a data set, these extra models may be more
informative with regard to specific properties of the data set which may
be important for some applications.

For our results, the $\beta$-approach seems to score the best, however the
differences are small and may not be significant. When compared for
the different input data set sizes, an increase in data set size seems to
be weakly correlated with an increase in quality score and the opposite
for a reduction in size. Despite the apparent weak correlation, there are
a few scores which do not fall in line with the observation. Some of
the increases and reductions in score are also small and might not be
statistically significant.
9.4. Limitations and weaknesses

The work presented in this thesis is subject to certain limitations and weaknesses. It can be difficult to know the extent to which these limitations and weaknesses have impacted its findings, however, they nevertheless need to be taken into consideration when drawing conclusions.

9.4.1 Test data

As discussed in the chapter on data for testing, the data which is the basis for the testing process of the different approaches to anonymization is not real-life data. The data may attempt to mimic some of the important properties of real-life data, but will inevitably differ. This difference can be found in both the form and the contents of the data, meaning that the various attributes which can be found in the self-produced data set might not correspond to attributes found in real-life data sets, and the values which make up the attributes may differ in both value and distribution thereof.

The process in which the data was created utilized a pseudo-random
seed to create a random distribution for the values of attributes such as patient age and disease diagnosis. This random distribution is not likely to correspond realistically to a non-uniform distribution over those attributes which might be found in a real-life data set. This can have a significant impact on the performance the various approaches to anonymization had with regard to extensiveness of required transformations for the different attributes and resulting utility score for the data set. It is impossible to know the exact impact without testing on a real-life data set, and even then, while a real-life data set might differ from the one produced for this project, that does not automatically mean that the self-produced data is invalid, since one real-life data set does not represent all real-life data sets.

Furthermore, there is a limit on the extent to which this data set represents all health data. As has been stated previously, it is not intended to represent all health data, but this also means that the conclusions drawn from results gathered from this data might not apply to other forms of health data. That is not to say that the conclusions drawn from this project have no validity for other cases, only that they may be limited to a certain extent, and such potential limitations should be considered. In particular, data sets featuring multiple confidential attributes, data sets featuring confidential attributes with small value domains and data sets featuring a large number of quasi-identifiers may find limited use for the findings of this thesis.

9.4.2 Privacy models

Not all existing privacy models have been tested during this project. I have limited myself to a few privacy models which seemed likely to produce good results. They serve as representatives of existing models, utilizing different strategies for ensuring protection from identity and attribute disclosure. In particular, the models examined in this thesis are for microdata releases.

As mentioned in the background chapter, there are two other types of data releases: tabular data and interactive databases. While tabular data releases are not particularly relevant for the topic of this thesis, producing outputs for very different purposes than microdata releases, the interactive databases can produce similar results to microdata releases. That type of release and the privacy models which are relevant have not been extensively examined beyond a brief look at differential privacy, because of the required involvement and oversight post data release. Nevertheless, it may provide strong protections against information disclosure and high utility scores.

9.4.3 Transformation approach

When performing anonymization, the transformations which are applied to the various quasi-identifier attributes in the original data set will determine the final form, and by extension the utility, of the anonymized data set. In our testing, we mainly focused on generalizations utilizing
user-configured hierarchies for the various attributes. This form of
generalization is non-perturbative, meaning the data transformations
remain truthful to the original data. For example, generalizing age 45 into
age range 40 to 50 still tells the truth about the age; however, a perturbative
transformation changing the age from 45 to 46, would make that value
not truthful to the original data. Perturbative transformation methods
employed in a clever fashion may result in different utility in the resulting
data set, however they have not been considered for this project.

Furthermore, the strategy for applying the generalization transformations
is simple. There exist algorithms which employ specific strategies for
attempting to comply with various privacy model requirements. Examples
include the BUREL and perturbative methods briefly covered in the
literature review section related to the $\beta$-likeness privacy model. Those
algorithms, specifically, were designed to ensure compliance with the $\beta$-
likeness privacy model the paper introduced while retaining a high degree
of utility in the resulting data sets. Different algorithms also exist which
focus other privacy models.

I initially considered implementing some such algorithms in a self-created
tool, however ended up moving away from that when I abandoned the
plan of creating my own tool in favor of using a pre-existing tool. As such,
this project does not cover the use of such specific algorithms, which may
or may not influence utility in resulting data sets.

9.4.4 Tool

The only tools considered for use in this project are free, publicly available
tools. I ended up choosing to use ARX, which includes many useful
features and proved highly valuable in my project, however, other non-
free alternatives for data anonymization exist. I have not extensively
researched their capabilities, but they may offer features which could assist
in producing better results than those which I produced using ARX.

ARX is somewhat limited in the way in which it allows users to apply
transformations to the attributes of an input data set. It only supports
generalization, microaggregation and microaggregation with clustering.
In addition, the strategies it allows for applying the generalization
transformation is limited to following a user-specified hierarchy. This
limitation makes it difficult to implement algorithms for optimizing
transformations, such as the ones mentioned in the previous section.

9.5 Recommendations for approaching anonymization of health data

Through working on this project, I have learned a lot about the field
of anonymization, the GDPR, and other related subjects. With my
goal of assisting system administrators of the DHIS2 in anonymization
efforts, I wanted to share the lessons I’ve learned, by compiling a set of recommendations for approaching anonymization of health data.

Appendix B lists the recommendations, and they mainly consist of what to be aware of when doing anonymization: What steps to go through, what considerations to make and what knowledge is required about both the data to be released and relevant legislations, as well as some potential pitfalls. While these recommendations are not, and are not intended to be, a complete set of guidelines for doing anonymization, they can serve as a resource to lean on through the anonymization process.

The recommendations cover the following topics:

1. The purpose for releasing the data.
2. The type of data that is to be released.
3. What makes the data useful.
4. Relevant legislation.
5. What risks might be faced.
6. The identifiability of the data.
7. Privacy models to inform the anonymization process.
8. How to transform the data to achieve the goals of the anonymization process.
9. Consideration of tools to assist with the process.
10. Ethical considerations regarding releasing sensitive data.
11. Potential pitfalls.
9.5. RECOMMENDATIONS FOR APPROACHING ANONYMIZATION OF HEALTH DATA
Chapter 10

Conclusion

The aim of this research project was to establish a starting point which researchers, developers and system administrators of DHIS2 could lean on when wishing to publish data for a variety of purposes. The GDPR serves as a common piece of legislation for the EU which regulates the processing of personal data. That made it a good candidate to contextualize the publishing health data, a type of data covered by its scope. To achieve my goal, I posited the following research question:

How can existing approaches to data anonymization be applied to health data to sufficiently comply with privacy and data protection regulations stipulated in the General Data Protection Regulation (GDPR), while preserving utility in the resulting data?

Through this project, I have worked towards answering this question, building on previous literature, constructing an approach to testing various anonymization methods and doing preparatory work to facilitate such testing, including the gathering of test data, examining privacy models and judging various tools before determining one of which to make use. All of this culminated in the testing of three approaches to anonymization, utilizing several privacy models, which produced results regarding the compliance with the requirements of the GDPR for the resulting anonymized data sets and regarding the utility of the anonymized data sets.

Based on the considerations taken during the examination of the test data set when selecting the identifying and quasi-identifying attributes, as well as the results showing strong protection against identity disclosure under the prosecutor and journalist scenarios and against attribute disclosure provided by the various privacy models, we can conclude that all three approaches can be used to ensure compliance with the GDPR.

Our experiments indicate a similar utility in the anonymized data sets resulting from the three anonymization approaches, but the results are subject to potential weaknesses. They mainly stem from two factors: the
test data consisted of synthetic data and the chosen tool was restricted in the ways it could perform transformations on the original data set to achieve the goals of the various anonymization approaches. Nevertheless, the k-approach preserves utility slightly better for the medium and large data sets, with the δ- and β-approach scoring slightly better for the small data set. Beyond that, a larger original data sets ensures a better utility score.

Finally, to assist system administrators of DHIS2 in anonymization efforts, lessons learned through working on this project have been compiled as a set of recommendations for how to approach anonymization of health data, listed in Appendix B.

10.1 Future research

This thesis has focused on a narrow section of the health data concept, utilized a limited approach to anonymization and experimented using fake test data. Building on my findings, each of those factors could be further explored in future work. Expanding the type of health data to test on could further inform the applicability of the anonymization approaches on in the field of health data. To directly improve upon my findings, however, testing on real data rather than fake data would provide more legitimacy to the results and utilizing purpose-built algorithms for transformations related to the various anonymization approaches could potentially have a significant impact on utility scores, perhaps better showcasing the strengths and weaknesses the different privacy models have compared to one another. Expanding testing to focus on more privacy models could also lead to interesting results.

Finally, this thesis covered compliance with the GDPR, however, it did not establish a minimum requirement for such compliance. Studying the GDPR to determine a more concrete measure of when data is anonymous, rather than an individual not being reasonably likely to be identified, could facilitate less strict transformations on a data set, likely increasing the utility of the resulting anonymized data set.
Bibliography


BIBLIOGRAPHY


[20] Data quality models | ARX - Data Anonymization Tool. URL: https://arx.deidentifier.org/overview/metrics-for-information-loss/ (visited on 07/06/2020).


[30] GDPR – The Data Subject , Citizen or Resident? | CYBER COUNSEL. Library Catalog: cybercounsel.co.uk. URL: https://cybercounsel.co.uk/data-subjects/ (visited on 02/06/2020).


[38] Clete A. Kushida et al. ‘Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies:’ in: Medical Care 50 (July 2012), S82–S101. ISSN: 0025-7079. DOI: 10.1097/MLR.0b013e3182585355. URL: http://journals.lww.com/00005650-201207001-00017 (visited on 03/06/2020).


[40] Patient Generator – MiHIN. Library Catalog: mihin.org. URL: https://mihin.org/services/patient-generator/ (visited on 02/06/2020).

[41] PLAY. URL: https://play.dhis2.org/ (visited on 02/06/2020).


[53] SSBs virksomhet. ssb.no. Library Catalog: www.ssb.no. URL: https://www.ssb.no/omssb/om-oss (visited on 05/06/2020).


[58] *TSD Prices - University of Oslo.* Library Catalog: www.uio.no. URL: https://www.uio.no/english/services/it/research/sensitive-data/access/prices/index.html (visited on 02/06/2020).


[61] *What is GDPR, the EU’s new data protection law?* GDPR.eu. 7th Nov. 2018. URL: https://gdpr.eu/what-is-gdpr/ (visited on 30/05/2020).

Appendix A

Transformation hierarchies for quasi-identifiers
Figure A.1: Age hierarchy
APPENDIX A. TRANSFORMATION HIERARCHIES FOR QUASI-IDENTIFIERS

Figure A.2: ZIP code hierarchy
Figure A.3: Health facility location hierarchy
Figure A.4: Time of encounter hierarchy
Appendix B

Recommendations for approaching anonymization of health data

The following are recommendations for approaching anonymization for data controllers intending to release data:

1. Purposes:

   Before doing anything else, you should consider why you want to release data. Are there any specific purposes you want the data to be used for, or do you think the data is useful and therefore want to publish it so that anyone who wants to might utilize it? Is it important the released data be truthful to the original data, or does the data have some properties which make it useful regardless of its truthfulness?

2. Release type:

   Having considered how your data could be useful, decide upon a type of release that may facilitate those purposes. Does the data need to be in the form of records concerning individuals, or is statistical information about your data set sufficient to fulfill the purposes from the previous point? Do you have the intention and resources to follow up on your data post-release, in which case, do you want to publish an interactive database instead?

3. Data:

   The original data set contains some information or properties which you want to be made available to more parties. What in this data set is important? Examine the different attributes, and how their presence in the data set affects what the data may be used for. Consider if any attributes in the data set could be excluded from the result without hampering its usefulness. If your data is about a patient’s health, perhaps their hair color isn’t important? The more data
which is included in the data to be anonymized, the harsher the transformation of the data is likely to be. Limiting such data may be useful.

4. Legislation:

Make absolutely sure you understand the requirements of relevant legislation, such as the GDPR, and what consequences might follow should such requirements not be met. Different pieces of legislation have different requirements for sensitive data. Consider whether the type of release you are hoping to make is in line with such legislation's allowances. Are you allowed to process the data, even for anonymization purposes? What does the legislation consider to be anonymous data? Is it enough to remove identifiers, or do you need to further consider how individuals in the data set might be identified? The GDPR has strict requirements, but which can be open to some interpretation. The Data Privacy and GDPR Handbook by Sanjay Sharma provides very an extensive, in-depth look at the GDPR, and might prove useful.

5. Risks:

Understand the context in which you are releasing your data. What risks may the data you want to release be exposed to? What bad actors exist in the world, and how could they misuse the data? How advanced are the attackers? Are they acquaintances of the people of the data set, or are there state agencies with extensive capabilities? Or are they perhaps companies which could take advantage of such data at the cost of the individuals?

6. Identifiability:

Understand what information in your data set could be used to identify individuals. What are the direct identifiers, such as name, residential address and government identifiers such as social security numbers? How could other information be used in combination, such as age and gender? Is there any less obvious information which could be used for that purpose? How likely is it for such information to be available to an attacker? How much time, money and effort would they need to expend to get access to such data? Is it public, or available to a limited set of people, such as friends? Or maybe in proprietary databases, which are hard for an intruder to break into? Has there been any data breaches which contain such information? Use all this information to note which information needs to be completely removed, which information needs to be transformed to ensure acceptable levels of risk, which information needs to be kept confidential for individuals, and which information can be left alone.

7. Privacy models:

Choose one or more privacy models which fit the level of protection
which is required for your data set. What measures are included in
the privacy model which protect against the revelation of individuals’
identity, and against the disclosure of information related to such
identities? Are there only measures taken, or do they provide
guarantees? Is a measure sufficient, or is a guarantee necessary?
Understand the difference, and ensure that the privacy model you
choose is in line with the requirements of relevant legislation.

8. Transformation of data:

Having chosen one or more privacy models for your data, decide how
you want to achieve its goals for data protection. There are a variety
of different strategies for transforming data, which each provide
different advantages and disadvantages. Does the data need to
remain truthful, or can values be changed such that they are no longer
valid for specific individuals? Is the semantic closeness of attribute
values important, or is the extensiveness of the transformations of
higher concern? Does the data need to be a version of the original
data, or could you create a new data set which is merely based on
the original, containing synthetic data? Finally, are there any existing
algorithms for optimizing the transformation of data for the privacy
model you chose, and do could they produce results which are useful
for the purpose you are releasing the data?

9. Tools:

Consider using a tool to assist you with the implementation of your
anonymization process. Are there any available which fit your need
for privacy models and the way in which you want to transform data?
Do you need a free tool, or can you make use of a paid tool? Have
they been vetted for security and usefulness? Does the context of
your process require you to make an assessment of these tools to
ensure they fulfill requirements for security?

10. Ethics:

Having considered the legislative aspect of releasing data previously,
seriously consider ethical ramifications of such a release. Does the
data contain information which could have a significant negative
impact on individuals if they are identified? Do the measures you
have taken for data protection adequately protect these individuals,
regardless of any legislative requirements? Can the information
gained from the released data, in and of itself, have a significant
negative impact on society or groups of people, especially vulnerable
demographics? Is it morally right to release such data despite
potential risks?

11. Pitfalls:

During an anonymization process, there are several pitfalls that
might be stepped in.
(a) After the anonymization process, is the data still useful? A significant amount of information may have been destroyed during the anonymization process. If the data is no longer useful for the considered purposes post-anonymization, there may be little reason to proceed with publishing it. Consider changing your anonymization approach to attempt to produce better results.

(b) Have you sufficiently considered the context of your release to identify all possible ways in which information in the data set might be used to re-identify individuals? There exists a vast amount of data on the internet. Some individuals reveal data about themselves, through for example social media, which might also exist in your data, and thus be used for re-identification. Such data might not consider for all individuals, but it needs to be considered.

(c) Have you properly and completely understood the legislation which concerns your data release? Some pieces of legislation are massive and might cover more than is initially obvious. For example, the GDPR concerns information about individuals in the EU. Not just citizens, or even residents, but also travelers and visitors. Furthermore, even organizations outside the EU must treat such data in accordance with the GDPR. Additionally, there is a distinction between anonymization and pseudonymization, and they are treated differently in the GDPR, meaning their utilization must be considered with that in mind.

(d) Never release anonymized data from a single source of data more than once. Two anonymized data sets originating from a single source can be used in concert to re-identify individuals in unexpected ways. Any guarantees for data protection provided by privacy models very likely fall apart if there are more than one anonymized data set from a single source. There are, however, anonymization efforts where a single source of data is released as multiple separate data sets which have provide guarantees for privacy, such as releasing identities and confidential data separately. But after having released one such result, it must not be done with a separate distinct resulting data set.
Appendix C

Source file for generating data

```python
import random
import datetime
import time

# Data used to construct output file
malenames = []
femalenames = []
lastnames = []
gender = []
diseases = []
countries = []
names = []
zipcodes = []
roles = []
facilities = []
facilitiesLoc = []

# Values informing the make-up of the output data
minAge = 5
maxAge = 89
minYear = 2000
maxYear = 2019
numMales = 1000  # Number of males and females constitutes
numFemales = 1000  # the number of records in the output data
numDoctors = 20

# Generated doctor identities
doctorNames = []
doctorRole = []

# Probability distribution for various attribute values
ageDist = []
diagDist = []
monthDist = []
yearDist = []
facilityDist = []
doctorDist = []
```
```python
# Pre-constructed files containing values to fill certain attributes
malefile = "malednames.txt"
femalefile = "femalenames.txt"
lastfile = "lastnames.txt"
diseasefile = "diseases.txt"
countryfile = "countries.txt"
zipfile = "zip.txt"
occupationfile = "occupationrole.txt"
hospitalfile = "hospitals.txt"

# Column names in output file
namefield = "name"
agefield = "age"
sexfield = "sex"
zipcodefield = "zipcode"
diagnosisfield = "diagnosis"
workernamefield = "health worker name"
workeroerolefield = "health worker role"
facilityfield = "facility"
facilitylocfield = "facility location"
encountertimefield = "time of encounter"

# Columns in output data
fields = [namefield, agefield, sexfield, zipcodefield, diagnosisfield,
workernamefield, workeroerolefield, facilityfield, facilitylocfield,
encountertimefield]

# Fill values for relevant data from pre-constructed files, and
# construct probability distributions where relevant
with open(malefile, 'r') as f:
    for name in f:
        malenames.append(name.strip())

with open(femalefile, 'r') as f:
    for name in f:
        femalenames.append(name.strip())

with open(lastfile, 'r') as f:
    for name in f:
        lastnames.append(name.strip())

with open(diseasefile, 'r') as f:
    x = 0
    for disease in f:
        diseases.append(disease.strip())
        diagDist.append(random.randint(1,100))
        if x > 0:
            diagDist[x] += diagDist[x-1]
x+=1

with open(countryfile, 'r') as f:
    for country in f:
        countries.append(country.strip())

with open(zipfile, 'r') as f:
    for zipcode in f:
```

APPENDIX C. SOURCE FILE FOR GENERATING DATA

```python
zipcodes.append(zipcode.strip());
with open(occupationfile,'r') as f:
    for role in f:
        roles.append(role.strip())
with open(hospitalfile,'r') as f:
    x = 0
    for sykehus in f:
        facilities.append(sykehus.split("\t") [0])
        facilitiesloc.append(sykehus.split("\t") [1])
        facilityDist.append(random.randint(1,100))
        if x > 0:
            facilityDist[x] += facilityDist[x-1]
        x+=1

# Construct patients and doctors
for x in range(0,numMales):
    names.append(malenames[random.randint(0,len(malenames)-1)])
    names+=lastnames[random.randint(0,len(lastnames)-1)]
    gender.append("male")
for x in range(0,numFemales):
    names.append(femalenames[random.randint(0,len(femalenames)-1)])
    names+=lastnames[random.randint(0,len(lastnames)-1)]
    gender.append("female")
for x in range(0,numDoctors):
    if r
        firstName = malenames[random.randint(0,len(malenames)-1)]
        else femalenames[random.randint(0,len(femalenames)-1)]
    doctorNames.append(firstName)
    +""+lastnames[random.randint(0,len(lastnames)-1)]
    doctorRole.append(roles[random.randint(0,len(roles)-1)])
    doctorDist.append(random.randint(1,100));
    if x > 0:
        doctorDist[x] += doctorDist[x-1]

# Probability distributions for age and time of encounter
for x in range(0,maxAge-minAge+1):
    ageDist.append(random.randint(1,100))
    if x > 0:
        ageDist[x] += ageDist[x-1]
for x in range(0,12):
    monthDist.append(random.randint(1,100))
    if x>0:
        monthDist[x] += monthDist[x-1]
for x in range(0,maxYear-minYear+1):
    yearDist.append(random.randint(1,100))
    if x>0:
        yearDist[x] += yearDist[x-1]

# Construct records and write to file
with open(outfile,'w') as outf:
    j=0
    for field in fields:
```

j += 1
outf.write(field)
if j<len(fields):
    outf.write(“,"
else:
    outf.write("\n")

i = 0
# One record for each patient
for name in names:
    age = minAge
    agerand = random.randint(0,ageDist[len(ageDist)-1])
    for x in range(0,len(ageDist)):
        if agerand <= ageDist[x]:
            age += x
            break

    sex = ("male" if i < numMales else "female")
    zipcode = str(zipcodes[random.randint(0, len(zipcodes)-1)])

    # Diagnosis according to probability distribution
    diagnosis = ""
    diagrand = random.randint(0,diagDist[len(diagDist)-1])
    for x in range(0,len(diagDist)):
        if diagrand <= diagDist[x]:
            diagnosis = diseases[x]
            break

    # Health worker according to probability distribution
    workername = ""
    workerrole = ""
    workerrand = random.randint(0,doctorDist[len(doctorDist)-1])
    for x in range(0,len(doctorDist)):
        if workerrand <= doctorDist[x]:
            workername = doctorNames[x]
            workerrole = doctorRole[x]
            break

    # Health facility according to probability distribution
    facilityname = ""
    facilityloc = ""
    facilityrand = random.randint(0,facilityDist[len(facilityDist)-1])
    for x in range(0,len(facilityDist)):
        if facilityrand <= facilityDist[x]:
            facilityname = facilities[x]
            facilityloc = facilitiesloc[x]
            break

    # Time of encounter, according to probability distribution
    month = 1
    monthrand = random.randint(0,monthDist[len(monthDist)-1])
    for x in range(0,len(monthDist)):
        if monthrand <= monthDist[x]:
            month += x
            break
    year = minYear
    yearrand = random.randint(0,yearDist[len(yearDist)-1])
for x in range(0, len(yearDist)):
    if yearrand <= yearDist[x]:
        year += x
        break

date1 = datetime.datetime(year, month, 1, 0, 0)
date2 = datetime.datetime.now()
if month%12 == 0:
    date2 = datetime.datetime(year+1, 1, 0, 0)
else:
    date2 = datetime.datetime(year, month+1, 1, 0, 0)

# Random date inside a month
encounterdate = datetime.datetime.fromtimestamp(
    random.randint(int(time.mktime(date1.timetuple())),
    int(time.mktime(date2.timetuple())-1)))

# Write record to output file
outf.write(name + " , " + str(age) + " , " + sex + " , " + str(zipcode) + " , " + diagnosis + " , " + workername + " , " + workerrole + " , " + facilityname + " , " + facilityloc + " , " + encounterdate.strftime("%d/%m/%Y") + "\n")

i+=1