IT UNIVERSITY OF COPENHAGEN

MASTER THESIS

Differential Privacy for Subgroup Analysis: Opportunities and Challenges for Publishing Rheumatoid Arthritis Data

Author:
Ahmet AKKOÇ

ahak@itu.dk

Supervisor:
Raúl JIMÉNEZ,
Andrzej WĄSOWSKI

A thesis submitted in fulfillment of the requirements for the degree of Master of Data Science

in the

Department of Computer Science

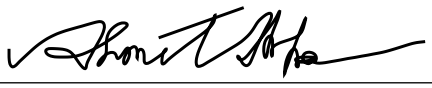
STADS Code: KISPECI1SE

June 1, 2023

Declaration of Authorship

I, Ahmet AKKOÇ, declare that this thesis titled, "Differential Privacy for Subgroup Analysis: Opportunities and Challenges for Publishing Rheumatoid Arthritis Data" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 01/06/2023

IT UNIVERSITY OF COPENHAGEN

Abstract

Department of Computer Science

Master of Data Science

Differential Privacy for Subgroup Analysis: Opportunities and Challenges for Publishing Rheumatoid Arthritis Data

by Ahmet AKKOÇ

In recent years data privacy regulations have restricted our ability to share scientific data. Differential privacy has come into prominence as a technique to ensure individual-level privacy in results, presenting a workaround. Even so, this comes with concerns in regards to precision and the preservation of subgroup-trends in results. In this thesis, I apply differential privacy to synthesize data over a cohort of Rheumatoid Arthritis (RA) patients. I examine the considerations on how we can generate unique data with similar outcomes to the original data. Of the marginal-DP synthesis algorithms considered, I empirically find AIM (Adaptive and Iterative Mechanism for Differentially Private Synthetic Data) to generate less distant data and more correlated data to that of the of original. Furthermore, I find that preserving the trends of subgroups demands a higher privacy budget allocation than preserving the trends of the overall population.

Acknowledgements

I would like to express my deepest appreciation to Raúl Pardo Jimenez and Andrzej Wąsowski for supervising my thesis and helping me find my way. Neither would this endeavor have been possible without the help of Niels Steen Krogh and ZiteLab ApS who have been kind enough to provide our research data. Thanks should also go to the ScandRA consortium who have helped us with approvals and have contributed suggestions to my work. Lastly, I would like to thank those from outside of ITU I have had the fortune of having had as sparring partners just to make sure my work was not completely alien.

Contents

Declaration of Authorship iii

Abstract v

Acknowledgements vii

1 Introduction 1
1.1 Overview . . . 1
1.2 Proposed Thesis Contributions . . . 2
1.3 Description of Remaining Chapters . . . 2

2 Background 3
2.1 Background: Domain Knowledge . . . 3
2.1.1 Rheumatoid Arthritis . . . 3
2.1.2 Treating Rheumatoid Arthritis . . . 4
2.1.3 Data Privacy Concerns in RA Research . . . 5
2.2 Background: Differential Privacy . . . 6
2.2.1 An Overview of Differential Privacy . . . 6
ϵ-Differential Privacy . . . 6
(ϵ, δ)-Differential Privacy . . . 7
2.2.2 Composition of Differential Privacy . . . 7
2.2.3 Differential Privacy for Synthetic Data . . . 8
2.2.4 Marginal DP Synthetic Data Algorithms . . . 8
Multiplicative Weights Exponential Mechanism (MWEM) . . . 8
Maximum Spanning Tree (MST) . . . 9
Adaptive Iterative Mechanism (AIM) . . . 9
2.2.5 Remark: GAN-based DP Synthetic Data Algorithms . . . 9

3 Previous Work: Differential Privacy and its Application to Medical Research 11
3.1 Subgroup Importance: An overlooked risk with using differential privacy . . . 12

4 Data and Method 13
4.1 Data . . . 13
4.1.1 Variables . . . 13
4.1.2 Trends in Data . . . 14
4.1.3 Limitations . . . 15
4.2 Method . . . 16
4.2.1 Change to Subgroups . . . 19

5 Results: Evaluating Synthetic Data 21
5.1 Part 1: Synthetic Cohort Response . . . 21
5.2 Part 2: Synthetic Subgroup Responses . . . 24

6 Threats to Validity and Discussion **33**
6.1 Threats to Validity 33
6.2 Discussion 33

7 Conclusion **35**

A Synthesizer Parameters **37**

B Source Code **39**

Bibliography **41**

List of Figures

2.1 The Treat-to-Target Strategy 5

4.1 Illustration of Rate of Remission over Time on Original Data 15
4.2 Illustration of Rate of Remission over Time on Original Data for Male and Female Subgroups 16
4.3 Computation of Jensen-Shannon Distance between Original vs. Synthetic Data 17
4.4 Computation of Pearson Correlation between Original vs. Synthetic Data 18
4.5 Computation of Variances among Synthetic Data 19
4.6 Subgroup Split Strategy 20

5.1 NPJS Distance vs. Privacy Budget 21
5.2 Pearson Correlation vs. Privacy Budget 22
5.3 Variance vs. Privacy Budget 23
5.4 Subgroup Split Strategy (Re-used) 24
5.5 NPJS Distance vs. Privacy Budget (for subgroups) 25
5.6 Pearson Correlation vs. Privacy Budget (for subgroups) 28
5.7 Variance vs. Privacy Budget (for subgroups) 30

List of Tables

4.1 Data: Patient Variables 14
4.2 Rate of Remission Observed in Original Data 15
4.3 Rate of Remission Observed for Male/Female Subgroups in Original Data 16

5.1 NPJS Distance vs. Privacy Budget 22
5.2 Pearson Correlation vs. Privacy Budget 23
5.3 Variance vs. Privacy Budget 24
5.4 NPJS Distance vs. Privacy Budget (for subgroups) 27
5.5 Pearson Correlation vs. Privacy Budget (for subgroups) 29
5.6 Variance vs. Privacy Budget (for subgroups) 31

List of Abbreviations

ACR	The **A**merican **C**ollege of **R**heumatology
AIM	an **A**daptive and **I**terative **M**echanism for Differentially Private Synthetic Data
DP	**D**ifferential **P**rivacy
CRP	**C**-**R**eactive **P**rotein
DAS	**D**isease **A**ctivity **S**core
DMARD	**D**isease-**M**odifying **A**nti-**R**heumatic **D**rugs
GAN	**G**enerative **A**dversarial **N**etworks
GPA	**G**lobal **P**atient **A**ssessment
MST	**M**aximum **S**panning **T**ree
NPJS	**N**ormalized **P**airwise **J**ensen-**S**hannon distance
MWEM	**M**ultiplicative **W**eights **E**xponential **M**echanism
QoL	**Q**uality of **L**ife
RA	**R**heumatoid **A**rthritis
T2T	**T**reat **to** **T**arget

Dedicated to my mother Yeşim, to my father Nurullah and to my sister Irem.

Chapter 1

Introduction

1.1 Overview

Living in the Age of Data, we find ourselves generating and accumulating new information on the world, every day. Those systematic processes which govern the Data Age, have led to sectoral paradigm shifts. Hence in healthcare we are now used to seeing Smart Tablets being employed for patient surveys in clinics, slowly becoming as emblematic of healthcare as the older X-Ray Machines and CT-Scanners. The proliferation of clinical data in this manner has encouraged research, evidence-based decision-making and better understanding between patients and clinicians.

However, this widespread collection and exchange of clinical data has also raised serious concerns about privacy, as personal health information is highly sensitive and could be used to identify individuals. These growing concerns on patient privacy in healthcare have encouraged stricter control of data. Several countries have even developed data protection regulations and laws to ensure data privacy such as HIPPA (1996) in the US , the GDPR in the EU/EEA (2016) or the Cybersecurity Law of the People's Republic of China (2016). These regulations have trapped data in silos, hurting the transparency and granularity of research data; as well as the reproducibility of experimental results.

Rheumatoid arthritis (RA) is but one of many conditions whose research is hampered by the difficulty of publishing patient-level data. RA is an auto-immune disorder that affects the joints, and it is a chronic condition which may persist with patients for a lifetime. ***Differential privacy*** (DP), specifically differentially private synthetic data may offer a way to publish patient data with a trade-off between privacy and utility. The goal of applying differential privacy is to reduce patient-identifiability, such that it is no longer possible to link the data points to particular real-world individuals.

While DP has shown promise in preserving privacy, it is not a panacea. Implementing differential privacy requires careful consideration of the trade-offs between privacy, data utility, and computational efficiency. Moreover, differential privacy may not be equally effective for all types of data. In the context of medical data, where the stakes are high, there is a need to carefully evaluate the effectiveness of differential privacy. One serious concern whenever applying differential privacy is whether the noise introduced through differential privacy has a disruptive effect on the analysis of said data. For example, RA tends to affect women more than men (Aletaha and Smolen, 2018) and so if we were to partition the data into a male/female split we would expect to see a higher incidence or more drug-resistance from the female subgroup than males. The question of how differentially private synthetic data impacts subgroups is not well explored, therefore warranting investigation.

1.2 Proposed Thesis Contributions

In this thesis, I try to answer the question: '***Can differentially private synthetic data be used to gain realistic insights similar to clinical RA data?***'. To this end, I consider a number of state-of-the-art algorithms for differentially private synthetic data generation and evaluate these across an experiment comparing results to that of the original clinical data. Furthermore, I am interested in subgroup analyses, on the male/female subgroups for both DP and original data.

For differentially private data synthesis, I design an experimental set-up (See Sec. 4.2) where I compare the remission rate of a population over 12 month's time, between original clinical data and synthetically generated datasets. Given this experimental set-up I investigate the following research questions:

- **RQ1: Is there practical significance between algorithm selection along the privacy-utility trade-off?**
- **RQ2: How much of a privacy budget is required for differential privacy to preserve the characteristics of population subgroups?**

By identifying the opportunities and challenges of using differential privacy for publishing RA data, this thesis aims to provide insights into the effectiveness and limitations of differential privacy in the context of medical data publishing.

1.3 Description of Remaining Chapters

To further contextualize this thesis, I will cover background knowledge in Chapter 2. I will present background knowledge on Rheumatoid Arthritis in Section 2.1, differential privacy in Section 2.2. This will be followed by a brief literature review on previous work on DP for medical applications in Chapter 3.

Chapter 4 will detail my data and experimental set-up. Results are presented in Chapter 5 with discussion and threats to validity highlighted in 6. Finally, I will conclude with my closing remarks in Chapter 7.

Chapter 2

Background

2.1 Background: Domain Knowledge

In this section I go over the necessary domain knowledge to understand the challenges in RA Research.

2.1.1 Rheumatoid Arthritis

Rheumatoid arthritis (RA) is a chronic and progressive auto-immune disorder that affects the joints, leading to inflammation, pain, and stiffness. RA can significantly impact daily activities, such as work, household tasks, and leisure activities. The disease can also have a substantial negative impact on a person's quality of life (QoL) due to pain, fatigue, and decreased mobility. Studies have also shown that individuals with RA are more likely to have depression and anxiety, which can further exacerbate the negative impact on their QoL. (Aletaha and Smolen, 2018; Sokka et al., 2010). It is estimated that RA affects approximately 5 per 1000 adults worldwide, with women being more affected than men (Aletaha and Smolen, 2018).

Because RA is a chronic disease, a patient with RA should never expect a full recovery unlike a flu infection. Rather, the goal of RA treatment is 'remission', broadly defined as a state where the disease is essentially inactive or has very minimal activity. Remission from RA means that the patient can expect lessened inflammation, pain and stiffness in their joints. Another way we can conceptualize remission is as a fall in the bodily disease activity.

The American College of Rheumatology (ACR) has identified several disease activity measures as preferred options for assessing disease activity, and thereby remission, in RA(England et al., 2019):

- Disease Activity Score on 28 joints (DAS28, or DAS for short)
- Simplified Disease Activity Index (SDAI)
- Clinical Disease Activity Index (CDAI)
- Routine Assessment of Patient Index Data 3 (RAPID3)
- Patient Activity Scale (PAS) II (PAS II)

The choice between these measures is based on clinician preference. While RAPID3 and PAS-II are measures that rely on patient-reported data, they may be influenced by subjective patient factors such as mood or personality. On the other hand, measures such as DAS28, CDAI and SDAI, require both patient-reported and clinician-assessed data, for example the number of tender and swollen joints, levels of acute phase reactants such as C-Reactive Protein (CRP) and Erythrocyte Sedimentation

Rate (ESR), and the patient's global assessment of disease activity or general health. These measures are more comprehensive in nature and may provide a more objective and evidence-based assessment of disease activity and therefore are more widely used to monitor treatment response and guide treatment decisions based on each patient's individual disease activity level.

In this thesis, I have chosen to work with **DAS28-CRP** given that DAS28 is the disease activity metric that is most often recommended (Mian, Ibrahim, and Scott, 2019) and I have CRP measures available to me (more on this in Sec. 4.1). DAS28-CRP is computed according to Eq. 2.1. Remission for DAS28-CRP is defined as a score < 2.3 , so my interest is in seeing how many patients are able to bring their disease activity below this threshold.

$$\begin{aligned}\text{DAS28-CRP} = \\ 0.56 * \sqrt{(\text{Number of Tender Joints [0, 28]})} \\ + 0.28 * \sqrt{(\text{Number of Swollen Joints [0, 28]})} \\ + 0.014 * (\text{Patient Global Assessment Score [0, 100]}) \\ + 0.36 * \ln(\text{CRP (in mg/L)} + 1) + 0.96\end{aligned} \tag{2.1}$$

2.1.2 Treating Rheumatoid Arthritis

As stated above, RA is a chronic and progressive condition. Thus the treatment of RA is also a gradual process of reducing disease activity. The mainstay for RA treatment are the so-called Disease-Modifying Anti-Rheumatic Drugs (DMARDs). There are a variety of biologic and synthetic DMARDs in use and it is not always obvious which drug is optimal for which patient, seeing as patients can respond differently to drugs. In fact, patient responses are also gradual, therefore warranting a monitoring of the patient's condition over the course of their prescriptions. The popular treatment framework which describes how to monitor the development of a patient's condition is known as ***Treat-to-Target (T2T)*** (Smolen et al., 2016) .

The T2T strategy is illustrated in Fig. 2.1. T2T splits the responsibility of the patient's recovery between the healthcare professional and also the patient themselves. This strategy begins with an initial clinical assessment to determine the patient's current situation. Hence the name, the health professional then picks a **target** as a shared-decision with the patient that the pair want to be able to achieve. For example, a patient who is already in remission might seek to sustain it; a patient who has high disease activity will not anticipate immediate remission before reaching medium-low disease activity. Based on the chosen target, the health professional will develop a treatment plan where they prescribe drugs, often a DMARD, possibly paired with another drug, together with recommending lifestyle changes to help the patient's gradual improvement. The patient now enters a self-management phase where they, as best as they can, try to stick to the health professional's suggestions. At the next monitoring a few weeks later, the patient will be asked to assess how they feel their condition has changed. If the patient is seeing improvement, only minor changes are made to dosage and lifestyle and the patient goes home to come again in a few weeks. If the patient is not seeing improvement, then the health professional will devise a new treatment-plan based on a combination of clinical data

and **the patient's historical data**. Thus, the patient's own response to treatment is incorporated into the overall strategy.

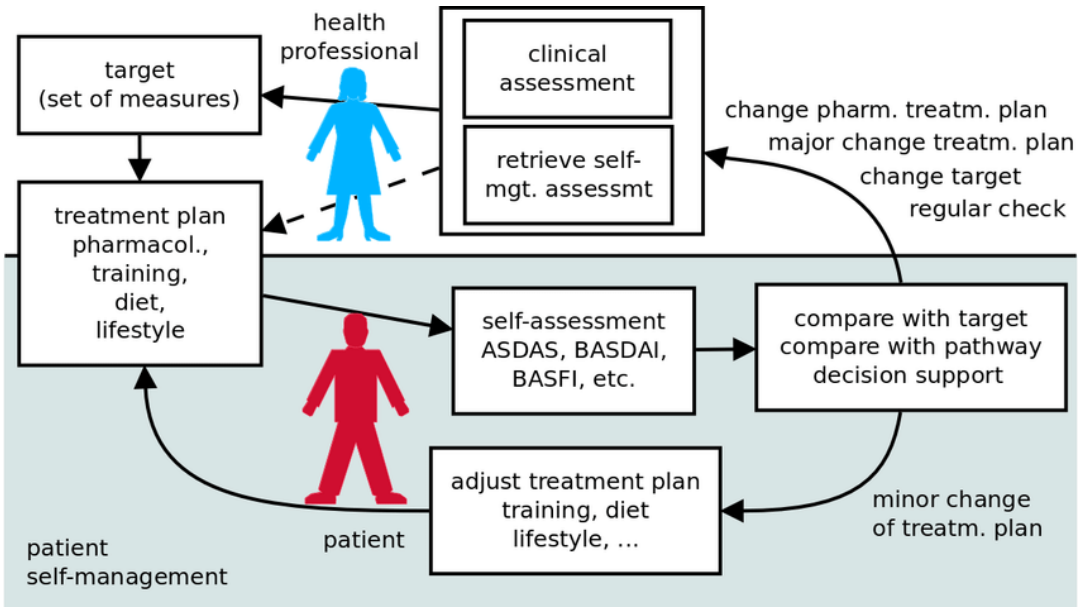


FIGURE 2.1: Diagram describing the Treat-to-Target strategy. Taken from Leister et al., 2016.

As T2T demonstrates, the treatment of RA requires routinely, personal data-collection. On one hand disease progression is clinically significant information, because it gives a clinician information on both the patient and the prescribed drug. Yet on the other hand, a value such as DAS28 over-time is a sensitive attribute from a privacy perspective. The patient's values directly tell us how advanced a stage their disease has reached. This dual nature of clinical data as both scientific evidence and personal condition puts clinical data into a rough spot within the realm of data privacy regulations.

2.1.3 Data Privacy Concerns in RA Research

RA Research is as burdened with patient-data vulnerabilities as any other field in medicine. Privacy risks for patients whose data are collected and retained for research range from honest-but-curious actors to unwanted exposure through the 3rd party breaching of centralised patient registries (Malin, Emam, and O'Keefe, 2013). As such, RA Research is also subject to many of the data regulations across the legal world.

Unfortunately, data-protection regulations to address these concerns have been destructive to the broader medical research ecosystem. In research secondary uses of data are common, where the same data is used for future research questions. The current frameworks, such as the active revision of the GDPR, which are based

around patient consent for each use-case are not compatible with extracting secondary uses in this manner (Peloquin et al., 2020) . Furthermore, there is some public opinion objecting to the status quo; a recently conducted anonymised survey in Ireland found that a majority both patients and doctors agreed that secondary use-cases of data should be implied by initial consent for patient-data processing (Davey et al., 2022).

The allure of synthetic data thus becomes obvious. Since differentially private synthetic records no longer correspond to particular individuals there is implicitly no need for consent for secondary, or even tertiary, uses of the data. Provided, of course, that the synthetic data is a sufficiently close enough of an approximation to some original real-world dataset. To this end, I have modeled this thesis to consider both a hypothetical primary and secondary use for the synthetic data generated. This is achieved by first observing the remission trend over the entire population (Sec. 4.2) and then subgroup trends for men and women (Sec. 4.2.1).

2.2 Background: Differential Privacy

2.2.1 An Overview of Differential Privacy

In computational security we often model privacy based on an assumed trade-off between privacy and utility. The maximum usefulness of data is only possible through absolute transparency, but this isn't ever feasible. On the other hand, the absolute state of privacy means that we get no information to work with at all. Thus one must find an appropriate balance between the two extremes of absolute utility and absolute privacy.

The framework of ***Differential Privacy*** (Dwork, 2006) is a tool to model this privacy-utility trade-off mathematically. Using DP we can obtain results based on some initial data, such that those results are adjusted for the right balance of privacy and utility to meet our needs.

ϵ-Differential Privacy

When talking about differential privacy, we assume the abstraction D of a particular dataset. In the most basic definition, dataset D is a collection of n records, where each record is a tuple of the form $\boldsymbol{d}_i$ representing the data of a single individual (indexed by $i = 1, \ldots, n$).

Recall that what we seek to accomplish with differential privacy is to disassociate the $\boldsymbol{d}_i$ individual-level tuples from the original individuals. Consider the domain $\boldsymbol{d}$ which spans all possible individual-feature tuples $\boldsymbol{d}_i$. Let us then generalize the domain $\boldsymbol{d}$ to datasets which have the same n number of records to describe the domain of all possible datasets shaped similar to D as $\mathcal{D}$. Then we describe two datasets over $\mathcal{D}$ which differ precisely only in the contents of some $\boldsymbol{d}_j$ as **adjacent** where j is between 1, n.

An algorithm A mapping from domain $\mathcal{D}$ to range $\mathcal{O}$ is considered ϵ-differentially private, if for any pair of adjacent datasets D_1, D_2 and any subset $S \subseteq \mathcal{O}$ the inequality in Eq. 2.2 holds. The probability that the output of A came from D_1, as opposed to the probability that it came from D_2 is bounded by e^{ϵ}.

$$Pr[A(D_1) \in S] \leq e^{\epsilon} Pr[A(D_2) \in S] \tag{2.2}$$

The ϵ parameter is a non-negative number called either the privacy loss parameter or privacy budget. ϵ is used to set a bound on how much minimum privacy we want to allow for. The closer ϵ is to zero the higher the uncertainty of the results. A small epsilon also thereby means more privacy for the resulting output data. Conversely, as ϵ approaches infinity the more certain our results will become. So for larger epsilon utility will increase.

A trivial algorithm to satisfy ϵ-differential privacy is the ***Laplace Mechanism*** (Dwork and Roth, 2013). In essence this is a mechanism which adds noise sampled from a Laplace distribution to some resulting numeric tuple. Consider a function f mapping from a dataset to a numeric tuple; or in other words $\mathbb{N}^{|D|}$ (i.e. $\mathcal{D}$) to $\mathbb{R}^k$. Now let Δf denote the sensitivity of this function f, that is the maximum magnitude the result of $f(D)$ can ever change by altering a single entry in dataset D. In other words, Δf is the maximum value $|f(D_1) - f(D_2)|$ can take for any pair of adjacent datasets (D_1, D_2). Now let Y_i denote a sample-tuple drawn from a Laplace distribution $Lap(\Delta f/\epsilon)$. The Laplace Mechanism as shown in Eq. 2.3 is ϵ-differentially private.

$$M_L(D, f, \epsilon) = f(D) + (Y_1, \ldots, Y_k) \tag{2.3}$$

Another fundamental ϵ-DP algorithm is the ***Exponential Mechanism***. The Exponential Mechanism is suited for situations where we want to release a result for arbitrary utility (Dwork and Roth, 2013) where some outputs may be 'better' than other outputs. The Exponential Mechanism works by assigning a utility or score to each possible output of a query. The scores represent how well each output aligns with the true query result. Higher scores indicate outputs that are more accurate or informative.

Let us denote this score function as u and the set of possible results as $\boldsymbol{R}$. Then $u : \boldsymbol{d} \times \mathcal{R} \mapsto \mathbb{R}$ and let Δu denote the sensitivity for u, as defined earlier for Δf in the Laplace Mechanism. The Exponential Mechanism $M_E(D, u, R, \epsilon)$ returns an element of $r \in \mathcal{R}$ with probability proportional to $exp(\frac{\epsilon u(D,r)}{2\Delta u})$ and it is ϵ-differentially private.

(ϵ, δ)-Differential Privacy

Designing an algorithm with too high a privacy guarantee can sometimes be infeasible. There is, therefore, also a relaxation of differential privacy (Dwork et al., 2006) called (ϵ, δ)-differential privacy. Here we introduce a δ parameter into the inequality in Eq. 2.2, sometimes called the approximation parameter. This gives us Eq. 2.4.

$$Pr[A(D_1) \in S] \leq e^{\epsilon} Pr[A(D_2) \in S] + \delta \tag{2.4}$$

This relaxation allows us to describe algorithms where even if strict ϵ-DP might not hold for all pairs of D_1 and D_2, the bound may at most be exceeded by δ. Because (ϵ, δ)-DP is more lenient, many algorithms can accommodate lower values for ϵ by allowing a δ approximation. Furthermore, any $\epsilon - DP$ procedure is equivalent to $(\epsilon, \delta = 0)$-DP; this can be demonstrated by substituting $\delta = 0$ into Equation 2.4.

2.2.2 Composition of Differential Privacy

While differential privacy is a very concise framework when describing how results relate to the original data, it comes with a certain pitfall when there are multiple operations of several DP algorithms. When applying differential privacy successively,

either multiple times to the original dataset or as part of a sequential composition, the total ϵ of the process accumulates (Dwork and Roth, 2013). Repeated querying can thus quickly exhaust whatever privacy budget ϵ was allocated on follow-up analyses.

If we want to be able to use the output of a differentially private algorithm we have to be clever about it. One approach is to use differential privacy to construct a synthetic data generator, to mimic the distribution of the original dataset. Since the synthetic data generator is the output of some algorithm A, it will only incur the initial ϵ cost to construct this generator.

2.2.3 Differential Privacy for Synthetic Data

Differentially private data synthesis is a compromise which allows for releasing a re-usable kind of result. This approach to differential privacy involves generating synthetic datasets that capture the statistical patterns and properties of some original dataset without revealing specific individual records. DP data synthesis broadly falls into two categories: marginal-based techniques and GAN-based techniques. I will explain these below, with a heavier emphasis on marginal-based techniques.

2.2.4 Marginal DP Synthetic Data Algorithms

A ***marginal*** is a central statistic which can capture low-dimensional structure from high-dimensional data. It is often thought of as analogous to a histogram, taking counts of values into bins (McKenna et al., 2022). A marginal considering some set r of features occurring under the $\boldsymbol{d}_i$-tuple is a frequency table of the counts for how many times a particular combination of values for the feature-set r is found in the original dataset D.

Crucially, the domain of possible values for $\boldsymbol{d}$ (the domain of the $\boldsymbol{d}_i$-tuple representations of individuals) must be finite for marginal techniques to be usable. Thus marginal techniques encourage binning variables and are not able to accommodate continuous variables without pre-processing.

Multiplicative Weights Exponential Mechanism (MWEM)

The ***Multiplicative Weights Exponential Mechanism (MWEM)*** algorithm (Hardt, Ligett, and McSherry, 2012) is hailed as the pioneering paper for marginal DP synthesis. The idea behind MWEM is to start with a random guess of what could pass as an imitation of the original dataset and then to iterate over it to construct a better imitation, while respecting differential privacy.

First, we start with a parallel dataset A_0 of size n with a histogram of weights. We use A_0 and its future iterations A_i to retain a kind of prior initially assumed as a uniform histogram. At each iteration we use the Exponential Mechanism to select an ideal query q (i.e. for the bin of some marginal) where the counts of our synthetic data and the original dataset differ. This query q may be a one-way or two-way marginal, considering multiple features at once. We then use the Laplace Mechanism to get an approximate count for q in the original dataset, m. Based on this count m, we update A_{i-1} to A_i as we try to match the expectation of the same count across our probabilistic dataset A_i. At the end of some T iterations we return a final dataset A which is based on an average of the previous iterations A_i.

MWEM ensures that the queries being checked and the counts being optimized for are differentially private. Thus MWEM allows us to complete the entire procedure of synthesizing data without peaking at the original dataset more than our privacy budget allows us.

Maximum Spanning Tree (MST)

The ***Maximum Spanning Tree (MST)*** algorithm (McKenna, Miklau, and Sheldon, 2021) is a more recent marginal-based technique that builds on MWEM. The MST algorithm comes with a few optimizations such as compressing the domain. Recall that the query choice in MWEM was rather vague in how many dataset features we wanted to consider at once. The algorithm gets its name from the fact that it constructs a maximum spanning tree based on mutual information between 2-way and 3-way marginals. This clever optimization ensures that important feature correlations in the original data are also preserved in the synthetic data. Otherwise, MST too follows a similar pattern of select q , measure m, update weights. It should also be noted that the MST algorithm is (ϵ, δ)-DP, allowing for an approximation term.

Adaptive Iterative Mechanism (AIM)

The ***Adaptive Iterative Mechanism (AIM)*** algorithm (McKenna et al., 2022) is a budget-adaptive and workload-adaptive and marginal data synthesis technique. Unlike its siblings, the AIM algorithm does not shy away from spending its budget; for example, it will attempt to expend some ϵ to start with a better prior instead of a uniform distribution. On the other hand, workload-adaptivity means that AIM can be configured to optimise specific marginals. However, because I want to simulate uses of the dataset that are unknown prior (as described in Sec. 4.2), I do not take advantage of this fact in my experiments. It should also be noted here that AIM too is an (ϵ, δ)-DP algorithm.

2.2.5 Remark: GAN-based DP Synthetic Data Algorithms

Some of the other novel approaches to DP-Synthetic Data are based on Generative Adversarial Networks (GAN)s (Goodfellow et al., 2014). GANs are machine learning architectures that consist of two sibling neural networks: a generator and a discriminator. The generator trains to generate imitations of some original training data to deceive the discriminator, where the discriminator trains to distinguish between forgeries made by the generator and the original data. The generator of a GAN can later be isolated to generate new synthetic data.

For a long time, there was a gap between the interests of researchers working with GANs and those working on DP. Even so, there have been very recent advances showing the potential of GANs respecting differential privacy. These include DP-GAN (Xie et al., 2018) and PATE-GAN (Jordon, Yoon, and Schaar, 2018). There have also been modifications optimised for handling tabular synthetic data in the form of PATE-CTGAN (Xu et al., 2019) and DP-CTGAN (Fang, Dhami, and Kersting, 2022) the latter of which was especially developed for synthesising medical data.

Although I will be referring back to GAN-based techniques later in this work, for the experiments in Chapters 4-5 I have chosen to specifically focus on marginal-based techniques. I have chosen to omit GAN-based techniques from my experiments, due to the parametric complexity and demanding need for the initial training sample size. Instead I will be working with marginal DP synthesis algorithms,

which as stated previously require data to be binned, but are better suited for accommodating my dataset as will later be described in Section 4.1.

Chapter 3

Previous Work: Differential Privacy and its Application to Medical Research

In this chapter, I will examine the progression of different approaches in differential privacy and how they have been linked to in medical research.

Since the framework of Differential Privacy (Dwork, 2006) was laid forth, there have been a number of attempts to incorporate DP into public health, mostly restricted to research. A recent literature survey on differential privacy (Ficek, 2021) found that the most common purposes for using differential privacy in health research were related to data release, predictive modelling and outputting a top-M ranking in statistics.

Data release does not necessarily mean releasing the whole of the data, but a slice or bit of it, typically within a clinic or between clinics. In fact, there has been much stigma around releasing public datasets which may contain private information, as the reverse-engineering of the 2006 Netlix Challenge has shown the possible risks this may entail (Narayanan and Shmatikov, 2006). Even so there has been some recent work reconsidering dataset release with differential privacy, such as by perturbation and constructing mixture-models (Wang et al., 2021).

More typically, DP for data release is intended for a query-based setting where a researcher makes specific requests. These queries may be to investigate counts (Vinterbo, Sarwate, and Boxwala, 2012) or generate contingency tables (Mohammed et al., 2013). Target domains for queries can be, for example, biomedical data (Cho et al., 2020) or genomic data (Almadhoun, Ayday, and Ulusoy, 2020).

Another common use-case for differential privacy in medical research is the aforementioned predictive modelling. Historical patient data can be used to construct predictive models, which can assist healthcare practitioners. However, data-privacy concerns (Sec. 2.1.3) prevent us from directly using sensitive patient data. As such, using differential privacy, has been a common workaround to supply training data for predictive modelling. This same pitfall has motivated alternative research outputs against publishing data. For example, a research group may choose to publish their predictive models trained with DP algorithms instead of the training data itself. Furthermore, releasing predictive models also acts as a counter-measure against the effects of composition (Section 2.2.2), because repeated queries come at no additional privacy cost.

In this context, DP for predictive modelling has matured so much as a niche that there are now differentially private neural-network architectures specifically meant for training differentially private predictive models such as DP-SGD (Abadi et al., 2016). Applications of DP predictive models range from drug sensitivity prediction (Niinimäki et al., 2019) to neuroimaging (Xiaoxiao et al., 2020).

Whenever the research objective is only concerned with modelling a predetermined outcome, predictive models are adequate. However, a predictive model is not easy to re-use when the research objective is subject to change (i.e. secondary uses of the data), or if we want apply techniques from classical statistics as opposed to machine learning. Here, differentially private data synthesis is a compromise which allows us to model a dataset, instead of only picking a specific outcome. Preliminary applications of DP synthetic data to medical research include privacy-preserving clinical data sharing for blood measures (Beaulieu-Jones et al., 2019) and generating sequential electronic health records for intensive-care patients (Lee et al., 2020). The research-emphasis on predictive modelling bleeds into the line of synthetic data as well, with the advent of GAN-based techniques for data synthesis such as DP-CTGAN (Fang, Dhami, and Kersting, 2022).

Nevertheless, DP Synthetic Data is not without its critics. Very recently, a team of security researchers (Stadler, Oprisanu, and Troncoso, 2022) have shown that differentialy private synthetic data is more difficult to achieve than it at first seems. Not only do the researchers show that popular implementations of some DP synthesis algorithms violate their differential privacy guarantees, but even that proper implementations fail at balancing utility and privacy. The researchers conclude that low-ϵ DP for data synthesis is too unusable (described at one point as 'unpredictable') and that high-ϵ DP degrades privacy rapidly.

Under the influence of differential privacy's common use-case for predictive modelling Stadler et al. and other researchers today tend to measure utility based on predictive accuracy. Unfortunately this definition of utility is only with respect to a predetermined outcome and does not so much tell us about the data itself. I have designed the experiment in this thesis to instead consider the scenario of also a secondary use of already published data.

3.1 Subgroup Importance: An overlooked risk with using differential privacy

A concern with DP is whether the noise introduced over differential privacy has a neutralising effect on data. The objective of differential privacy is to reduce patient-identifiability, but we don't want to lose feature-importance.

There is some research to suggest that differential privacy may make it more difficult for subgroups to remain distinct. One recent study (Santos-Lozada, Howard, and Verdery, 2020) explored how the error brought on through applying differential privacy was more severe in analyses of minority groups (US Hispanics and non-Hispanic Blacks) than in analyses of the majority (US Whites). Another recent study (Kurz et al., 2022) frames the issue of subgroups as a matter of granularity when investigating Medicaid participation in the US, as queried through DP. At the county-level they found errors as high as 10% for the counts of minority groups, whereas errors for state-level DP aggregations were negligible for all groups.

As stated above, RA is more common in women than men. Thus one more consideration of generating synthetic RA data should be on how the data for male-female subgroups are affected. Furthermore, the effects of DP on subgroups have been studied in query-based settings but not so much for synthetic data. Therefore, my aim in this thesis is to contribute to our understanding of how differential privacy through synthetic data affects subgroups, by examining the change in an overall population as well as its subgroups.

Chapter 4

Data and Method

My experiment for this thesis has me applying differentially private data synthesis to a cohort of Rheumatoid Arthritis patients, with the aim of producing a synthetic cohort where the population's response to treatment over time is comparable to that of the original cohort. Furthermore, because I am also interested in the question of how subgroups are affected by the choices made when synthesizing data, the second part of my experiment will examine how the male/female subgroups for synthetic patient data compare to the originals.

4.1 Data

The experimental data is part of a provisional dataset provided to me by my workplace, ZiteLab ApS[1], under the condition that it be used as part of research in my thesis. This dataset represents a cohort of Rheumatoid Arthritis patients from hospitals in Denmark. The entire dataset is comprised of patient-level data for $N = 1137$ patients who were recruited to follow a similar treatment strategy (Refer to Sec. 2.1.2 as needed). These patients had no prior DMARD history and were all initially prescribed an identical first-line DMARD: Methotrexate. Following a T2T strategy, during the treatment patients may be prescribed other drugs to supplement or replace Methotrexate. For 12 months the patients are then monitored to see whether their condition improves. Patient conditions are re-assessed at intervals of 3 months, 6 months and at the end of 12 months.

The data for my patient cohort is tabular with each row describing a patient and each column describing a particular variable. A complete-case analysis approach is taken, so only patients who had none of the values for the variables shown in Table 4.1 missing were included. These patient variables are further elaborated below.

4.1.1 Variables

In this section I describe the variables in the original data that synthetic data is to be generated from.

Sex: A string denoting the sex of the patient. Sex is a demographic predictor for remission. Recall that RA is more prevalent in women than in men (Aletaha and Smolen, 2018). Therefore, prior knowledge of the patient's sex can inform us about their prognosis. On the other hand; because sex is demographic information, taken together, sex and age can act as quasi-identifiers. Therefore the inclusion of sex as a dataset variable necessitates data privacy considerations.

Age: Age at the start of treatment. Age is a demographic predictor for remission. The older a patient is, the further RA disease activity is expected to have

[1] https://zitelab.eu/

TABLE 4.1: Data: Patient Variables

Variable	Range
Sex	$\{Male, Female\}$
Age	$(18, 85)$
CRP Level (Baseline)	$(0, 363)$
Patient Global Assessment (Baseline)	$[0, 100]$
DAS Remission (Baseline)	$\{0, 1\}$
DAS Remission (M3)	$\{0, 1\}$
DAS Remission (M6)	$\{0, 1\}$
DAS Remission (M12)	$\{0, 1\}$

progressed. Thus it is considerably more difficult for older patients to enter remission than younger ones. Moreover, older patients also tend to be less responsive to drugs due to age-related complications. Taken together, age and sex can act as quasi-identifiers. Therefore the inclusion of age as a dataset variable necessitates data privacy considerations.

CRP: C-Reactive Protein density in blood, at baseline. The level of CRP in the blood increases whenever there is inflammation in the body. While CRP is an objective measure, it can be confounded by inflammation caused by reasons besides RA such as bacterial infection. Thus in effect, CRP is one way to estimate the total inflammation across the body.

PGA: Patient Global Assessment Score at baseline. Abbreviated as PGA or sometimes PtGA, it is a patient-reported measure of their condition on a scale from 0-100. PGA at baseline is one way to identify how much the patient's condition has progressed prior to treatment. In addition, the PGA can also hint to distress and comorbidities (Nikiphorou et al., 2016), which may go undetected by the other clinical tests.

DAS Remission: A boolean value indicating remission as defined for DAS28-CRP (Sec. 2.1). The value at baseline (month 0) will register as 1 only if the patient is below the disease activity threshold to be considered in remission (< 2.3). Follow-up DAS28-CRP remission is also reported at months 3, 6 and 12.

4.1.2 Trends in Data

The primary objective here is to assess and track the remission rate among the entire (synthetic) patient population over a period of time. I denote this observational time-series I track, the 'remission curve'. Before moving onto synthetic data, the remission curve for the original dataset is illustrated in Figure 4.1 and shown numerically in Table 4.2. We can see that the Methotrexate treatment strategy initially had a good effect on the population up until month 3. Thereafter, flares in disease activity lowered the remission rate of the population through months 6 and 12. How I go about comparing this original remission curve to the remission curves for synthetic data is further described in Section 4.2.

Because my secondary interest is in the male-female subgroups, it is worth examining the original situation there as well. Table 4.3 show the remission rate for the male population $(= 328)$ and female population $(= 810)$ over time, respectively. Figure 4.2 illustrates how the subgroups relate to the overall population, with women being shown to be less likely to enter remission than men. Even so, the shared treatment strategy on both subgroups suffers from an issue of diminishing

returns as the population's condition is difficult to sustain past month 3. Therefore, when generating realistic synthetic data from this dataset, I seek to be able to preserve this expected distinction between male and female subpopulations.

TABLE 4.2: Rate of Remission Observed in Original Data

Population Average	Baseline	Month 3	Month 6	Month 12
	5.36%	19.16%	16.52%	13.44%

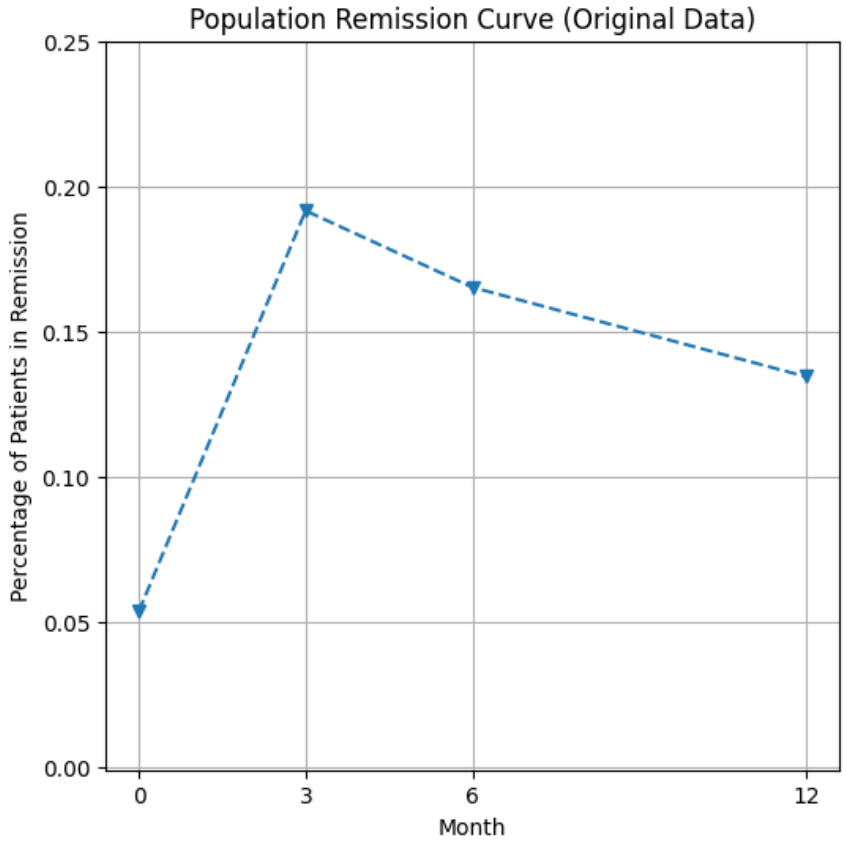


FIGURE 4.1: Remission curve for original data. An observational study on the unmodified data would have looked like this.

4.1.3 Limitations

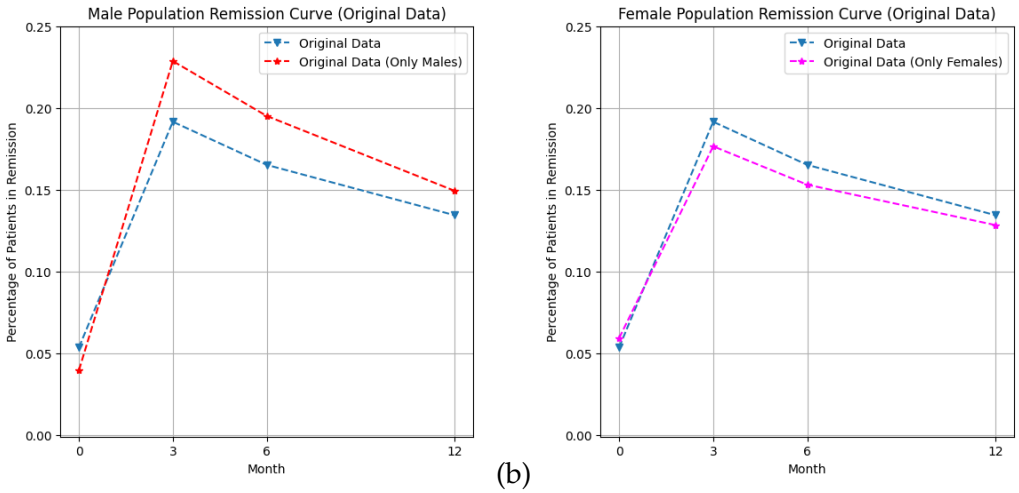
All of my patient data originates in hospitals from Denmark. Therefore, data trends are reflective of the Danish population and not necessarily global trends.

A complete-case analysis was taken because missing data is particularly common within the medical domain. Missing data, when not missing completely at random, can be quite revealing of the patients in the dataset and therefore warrants special consideration when applying DP (Das et al., 2022). While some implementations of DP algorithms can handle missing data, for the sake of consistency I have chosen to only work on patients who had all of their data available.

Finally, I ignore information between months 3-12, besides DAS remission. That means that intermediary indicators of the patient's condition such as CRP, PGA changes are ignored. Furthermore, drug and dosage changes as part of the T2T strategy are also abstracted away.

TABLE 4.3: Rate of Remission Observed for Male/Female Subgroups in Original Data

Population Average	Baseline	Month 3	Month 6	Month 12
Males	3.96%	22.87%	19.51%	14.94%
Females	5.93%	17.65%	15.31%	12.84%



(a) (b)

FIGURE 4.2: Remission curves for the Original Data for Males (a) and Females (b), as they compare to the whole population.

4.2 Method

I conduct an experiment where I simulate a series of observational studies on the unaltered dataset together with parallel patient datasets, synthesized through DP data synthesizers. In particular, the objective of these simulated observational studies is to measure the remission rate across the patient population over time, a 'remission curve' as described above in Sec. 4.1.2. The goal of this experiment is to see how the same remission curve would have come out differently, had an observational study been conducted on DP synthetic data of the same size, in the absence of the original.

The DP data synthesis algorithms covered in this experiment are all marginal-based. Namely, I use MWEM (Hardt, Ligett, and McSherry, 2012), MST (McKenna, Miklau, and Sheldon, 2021) and AIM (McKenna et al., 2022) as implemented in the SmartNoise Synthesizers Package v0.3.5 for Python 3.10.11 . Additional parameters can be found in Appendix A.

Because I am working with marginal-based DP synthetic data, all my variables needed to be either categorical or ordinal. While numeric information (i.e. continuous values) is available to me, I have binned these into discrete variables. Age was binned into 10 bins between $18 - 100$, PGA was binned into 4 quarters and CRP was binned non-linearly into four levels {Normal (< 1.0 dL), Marked ($10 > x \geq 1.0$ dL), Severe Elevation ($50 > x \geq 10$ dL), Extreme Elevation (≥ 50 dL)}.

For each of the DP data synthesis algorithms I generate ϵ-DP synthetic data, as I vary ϵ across a list of discrete points. To address the uncertainty brought on through random sampling I run $K = 20$ trials for generating different data synthesizers. To evaluate the performance of these trials I consider two utility metrics: Normalized Pairwise Jensen-Shannon (NPJS) Distance and Pearson's Correlation. I also measure

the variance for the $K = 20$ trials of a given (algorithm,ϵ) to understand how these choices govern the uncertainty of results.

The Normalized Pairwise Jensen-Shannon (NPJS) Distance is an ad hoc measure I ended up developing over the course of this project. An early suggestion I had was to use absolute difference as a utility metric. Unfortunately absolute difference is much better suited for counts rather than rates or probabilities. Thus I was led to look into the Jensen-Shannon distance which is an adequate way to relate the distance between two probability distributions. Along our remission curve in Fig. 4.1 we have four points representing a discrete probability for remission at a given month. For a particular synthetic dataset generated during a trial we can extract a similar synthetic remission curve with its own probabilities. With NPJS, I compute the pairwise Jensen-Shannon distance for the probabilities of the original remission curve and the synthetic curve and then normalise the result by the length of 4. The manner in which NPJS penalizes a hypothetical synthetic remission curve is illustrated in Figure 4.3. I then report the median of all trials, as well as the interquartile range. Just like any other distance, utility is maximised the more NPJS is minimised.

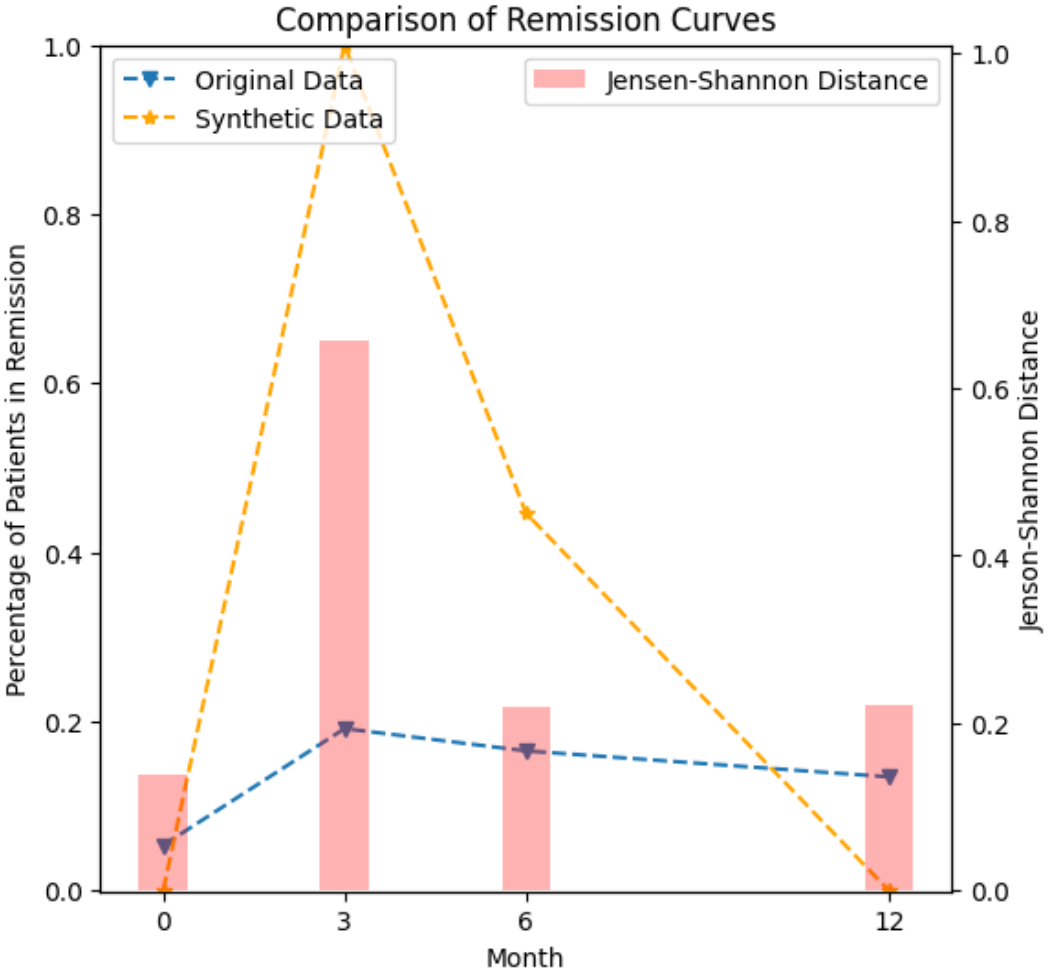


FIGURE 4.3: Comparing Jensen-Shannon Distance between Original and Synthetic Data. Notice how for this poor sample of synthetic data, the spike in month 3 is heavily penalised.

The second utility metric I consider is the Pearson correlation co-efficient. A safe

assumption to make is that if two time-series are close to one another, we would expect to see a strong positive correlation. Thus, we can assess the utility of synthetic data based on how correlated it appears to the original data. This is a way of considering the whole remission curve at once instead of breaking it apart. Figure 4.4 illustrates how this works in practice. Again, I then report the median of all trials, as well as the interquartile range. Only this time, utility is maximised when Pearson Correlation is closest to 1.

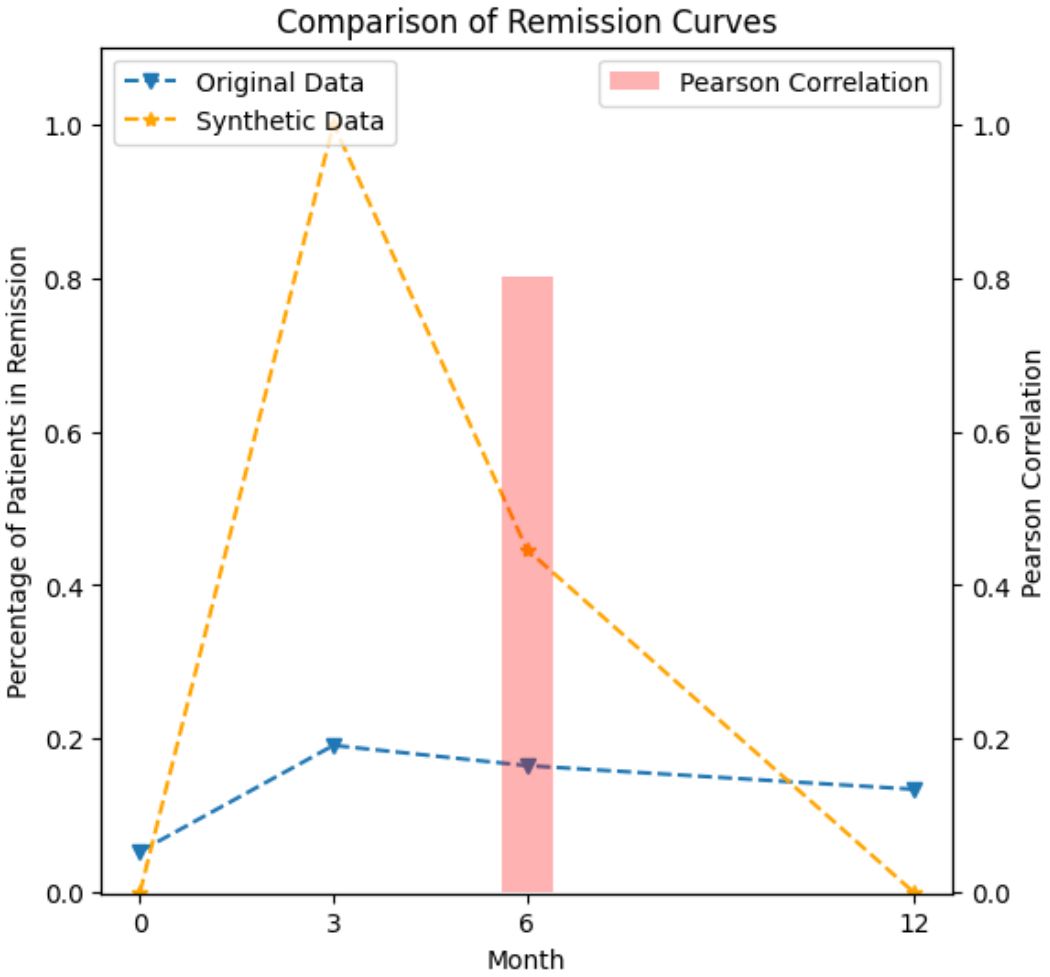


FIGURE 4.4: Computing the Pearson Correlation between Original and Synthetic Data. Notice how Pearson Correlation between the two remission curves is fairly strong, given that the structural shapes are similar.

Lastly, I measure the inconsistency of results by computing the variance for all remission curves for pairs of (algorithm, ϵ). This can show the uncertainty we ought to expect for a certain algorithm at that level of ϵ. In theory, this is like summing the mean of all distances to the hypothetical centroid of all the trials for this algorithm and level of ϵ, as depicted in Figure 4.5. However, please note that I also normalize the variance computed in this way by the length of the sequence 4. This is abbreviated to merely variance for sake of convenience in the later chapters.

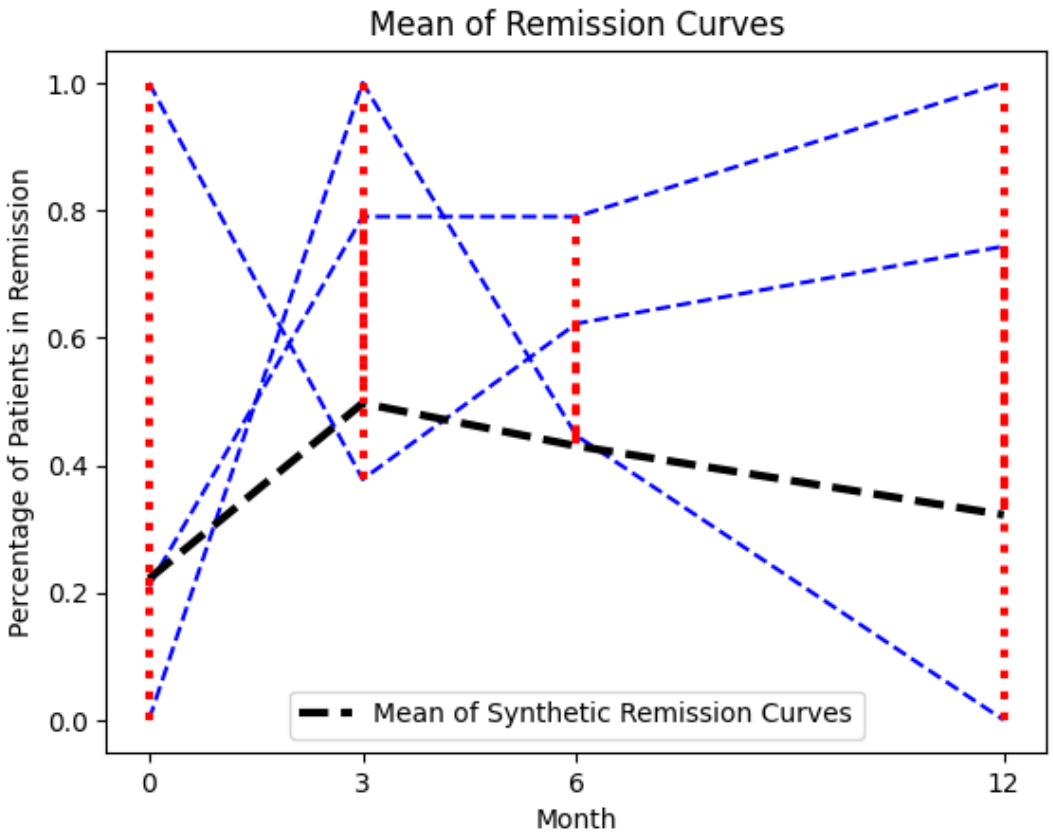


FIGURE 4.5: Computing the variance between the synthetic remission curves generated by identical parameters. Fixing the (algorithm, ϵ) parameters, I compute the variances for each time-point across the $K = 20$ trials to estimate how much variance is brought on by the choice made.

4.2.1 Change to Subgroups

The next step is to take the above experimental set-up and reuse the same synthetic data trials to see what we would get for a secondary use of the data. As stated previously, I also want to investigate if the trends for subgroups are preserved. To do this, I split the data into male/female subgroups based on the sex for patients as reported in the original and synthetic data. For clarity, I only split the data into male/female subgroups after we have synthesized new data. This is depicted in Figure 4.6. Thus we end up with 4 classes of subgroups: Males from the original data(1), females from the original data(2), males from any of the particular synthetic datasets(3) and finally females from any of the particular synthetic datasets (4).

I then pair these classes together and compute the same metrics that we had used for the overall population before. Specifically, I pair males from the original data with synthetic males (1-3). I pair females from the original data with synthetic females (2-4). Then finally I pair synthetic males with synthetic females (3-4) to compare the difference between their remission curves to that of the original males and females (1-2).

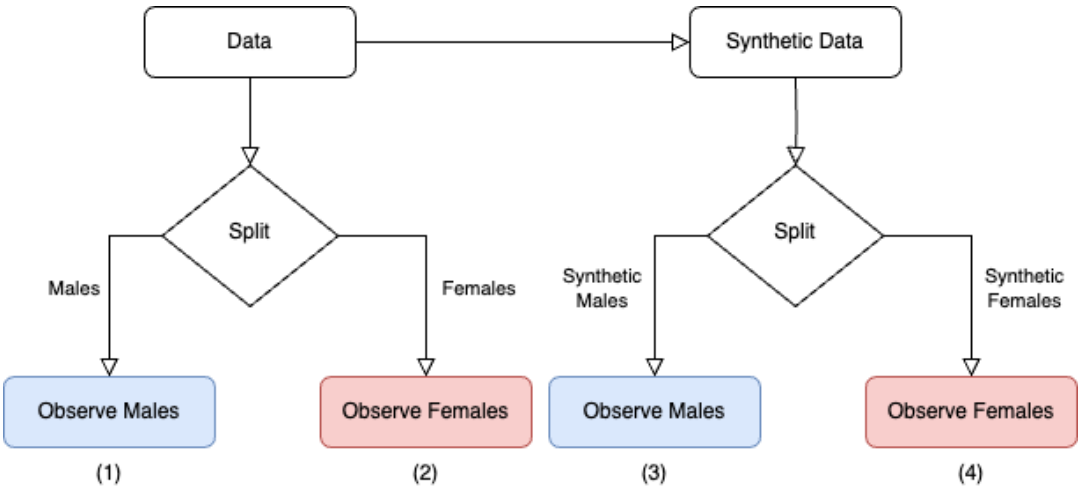


FIGURE 4.6: Subgroup Split Strategy.

Chapter 5

Results: Evaluating Synthetic Data

5.1 Part 1: Synthetic Cohort Response

In this section, I first report how well the remission curve for a whole synthetic dataset matched the remission curve for the whole original population. Figure 5.1 and Table 5.1 show a comparison of the privacy-utility trade-off when taking NPJS as the utility metric. Here, error bars are centered at the median utility of $K = 20$ trials, and span the 25th to 75th percentiles. Here, AIM clearly outperforms MWEM and MST as can be seen from the difference in magnitude on the y-axes of Figure 5.1.

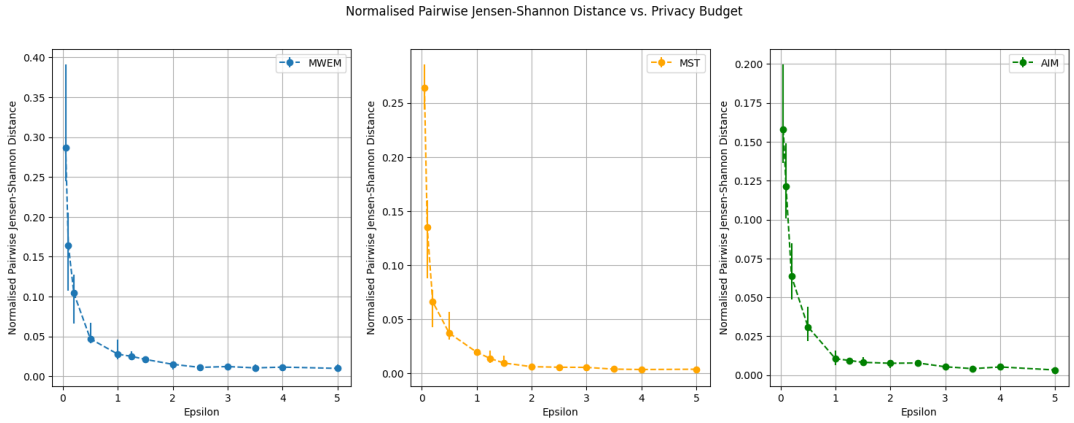


FIGURE 5.1: NPJS Distance vs. Privacy Budget

TABLE 5.1: NPJS Distance vs. Privacy Budget (Median with 25%, 75% IQR)

ϵ	MWEM	MST	AIM
0.05	0.2863 (0.2451, 0.3914)	0.2639 (0.2440, 0.2857)	0.1580 (0.1363, 0.1999)
0.10	0.1645 (0.1075, 0.2055)	0.1352 (0.0882, 0.1603)	0.1215 (0.1007, 0.1488)
0.20	0.1041 (0.0662, 0.1272)	0.0662 (0.0428, 0.0772)	0.0635 (0.0487, 0.0850)
0.50	0.0470 (0.0418, 0.0666)	0.0372 (0.0310, 0.0572)	0.0310 (0.0220, 0.0441)
1.00	0.0277 (0.0211, 0.0457)	0.0196 (0.0169, 0.0212)	0.0105 (0.0063, 0.0157)
1.25	0.0250 (0.0201, 0.0314)	0.0137 (0.0094, 0.0210)	0.0094 (0.0085, 0.0112)
1.50	0.0212 (0.0175, 0.0251)	0.0095 (0.0072, 0.0165)	0.0082 (0.0069, 0.0115)
2.00	0.0150 (0.0083, 0.0186)	0.0062 (0.0050, 0.0078)	0.0076 (0.0045, 0.0094)
2.50	0.0111 (0.0089, 0.0146)	0.0057 (0.0042, 0.0074)	0.0078 (0.0054, 0.0098)
3.00	0.0122 (0.0104, 0.0150)	0.0055 (0.0045, 0.0068)	0.0053 (0.0047, 0.0069)
3.50	0.0105 (0.0069, 0.0145)	0.0039 (0.0030, 0.0044)	0.0041 (0.0036, 0.0053)
4.00	0.0115 (0.0082, 0.0145)	0.0035 (0.0024, 0.0050)	0.0052 (0.0030, 0.0066)
5.00	0.0099 (0.0079, 0.0150)	0.0038 (0.0027, 0.0044)	0.0033 (0.0028, 0.0041)

The second metric I consider is Pearson's correlation between the original and synthetic remission curves, Figure 5.2 and Table 5.2 show a comparison of the privacy-utility trade-off when taking Pearson's correlation to be the utility metric. At $\epsilon \leq 0.1$, all three algorithms are seen to produce not only uncorrelated remission curves but even inversely correlated remission curves. There does not seem to a particularly good algorithm for maximising correlation, but by $\epsilon = 1.25$ all three algorithms are shown to be able to consistently generate strongly positively correlated (> 0.75) remission curves.

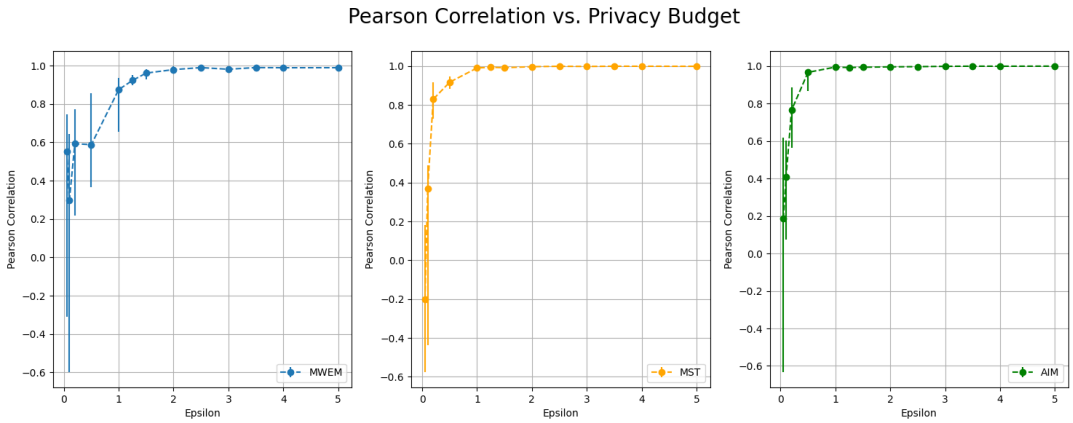


FIGURE 5.2: Pearson Correlation vs. Privacy Budget

TABLE 5.2: Pearson Correlation vs. Privacy Budget (Median with 25%, 75% IQR)

ϵ	MWEM	MST	AIM
0.05	0.5532 (-0.3117, 0.7453)	-0.2029 (-0.5755, 0.1797)	0.1861 (-0.6320, 0.6162)
0.10	0.2980 (-0.5984, 0.6442)	0.3691 (-0.4374, 0.4887)	0.4088 (0.0755, 0.6035)
0.20	0.5922 (0.2159, 0.7719)	0.8290 (0.7301, 0.9172)	0.7642 (0.5634, 0.8848)
0.50	0.5873 (0.3648, 0.8571)	0.9166 (0.8825, 0.9475)	0.9661 (0.8648, 0.9832)
1.00	0.8746 (0.6560, 0.9369)	0.9903 (0.9873, 0.9937)	0.9941 (0.9846, 0.9972)
1.25	0.9222 (0.8965, 0.9494)	0.9938 (0.9894, 0.9968)	0.9919 (0.9834, 0.9984)
1.50	0.9606 (0.9279, 0.9800)	0.9915 (0.9877, 0.9965)	0.9935 (0.9893, 0.9987)
2.00	0.9788 (0.9721, 0.9957)	0.9966 (0.9931, 0.9987)	0.9953 (0.9944, 0.9980)
2.50	0.9891 (0.9792, 0.9972)	0.9987 (0.9979, 0.9999)	0.9964 (0.9931, 0.9982)
3.00	0.9806 (0.9718, 0.9910)	0.9978 (0.9961, 0.9988)	0.9981 (0.9963, 0.9989)
3.50	0.9894 (0.9813, 0.9946)	0.9991 (0.9985, 0.9997)	0.9990 (0.9985, 0.9995)
4.00	0.9885 (0.9723, 0.9962)	0.9989 (0.9979, 0.9994)	0.9987 (0.9967, 0.9996)
5.00	0.9888 (0.9861, 0.9958)	0.9986 (0.9982, 0.9995)	0.9992 (0.9986, 0.9994)

Lastly, to better understand the uncertainty inherent in how these different algorithms synthesize data I consider the variance of trials for all (algorithm, ϵ) pairs. Figure 5.3 and Table 5.3 show a comparison of the uncertainty brought on by the choice of either the MWEM, MST or AIM algorithm and how this uncertainty varies with the privacy budget. MWEM has by far the highest variance of the three, with MST's and AIM's performance converging to very similar remission curves by $\epsilon = 1.0$.

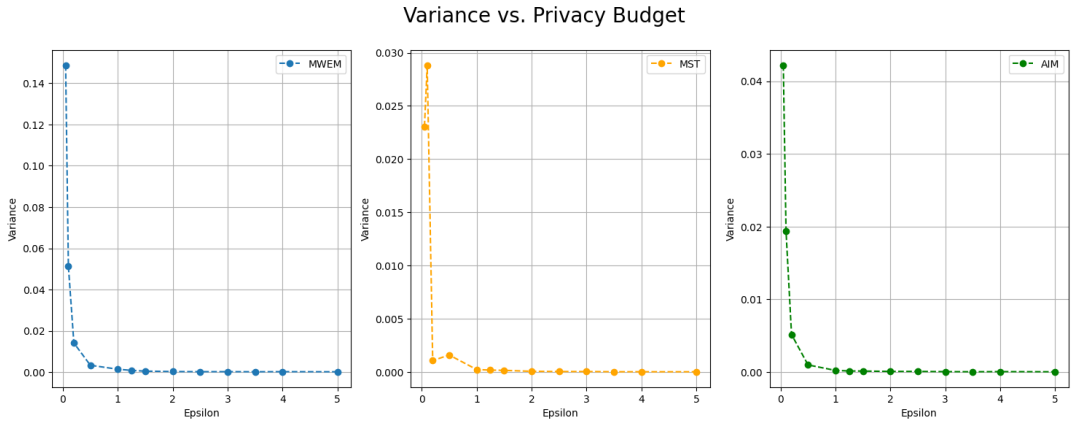


FIGURE 5.3: Variance vs. Privacy Budget

TABLE 5.3: Variance vs. Privacy Budget

ϵ	MWEM	MST	AIM
0.05	0.1487	0.0231	0.0422
0.10	0.0515	0.0288	0.0193
0.20	0.0141	0.0011	0.0051
0.50	0.0034	0.0016	0.0010
1.00	0.0014	0.0002	0.0002
1.25	0.0009	0.0002	0.0001
1.50	0.0005	0.0001	0.0001
2.00	0.0003	0.0001	0.0001
2.50	0.0002	0.0000	0.0001
3.00	0.0002	0.0000	0.0000
3.50	0.0002	0.0000	0.0000
4.00	0.0002	0.0000	0.0000
5.00	0.0002	0.0000	0.0000

5.2 Part 2: Synthetic Subgroup Responses

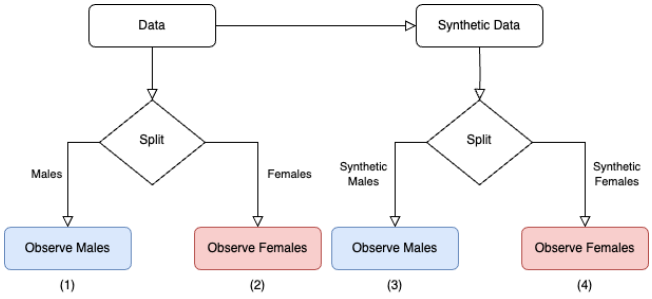


FIGURE 5.4: Diagram of subgroup splits in analyses for part 2

In this next section I present my results on comparing the remission curves for the male/female subgroups. As a refresher we have 4 subgroups as illustrated in Figure 5.4: (1) Males as observed in the original data, (2) females as observed in the original data, (3) males as observed in a particular synthetic dataset and finally (4) females as observed in a particular synthetic dataset.

Figure 5.5 and Table 5.4 show a comparison of the privacy-utility trade-off between the different subgroups when taking NPJS distance to be the utility metric. For clarity, here the higher plots show the distance as computed between the remission curve for the males of the original data (1) vs. that of the males of the synthetic data (3); the middle plots show the distance between the remission curve for the females of the original data (2) vs. that of the females of the synthetic data (4); finally, the lower plots show the distance between the remission curve for the males generated as synthetic patients (3) and that of the female synthetic patients (4). Notice the red line at the bottom intercepting the y-axis; this is the NPJS distance between the original male and female subgroups (1-2). Ideally, the synthetic data should precisely shoot for this level of similarity between the two groups, neither too above or below.

Comparing the top row with the middle row, it would seem that the algorithms are able to flatten their performance curves sooner for the middle row. In other words, synthetic data to match the female subgroup's remission curve can be achieved

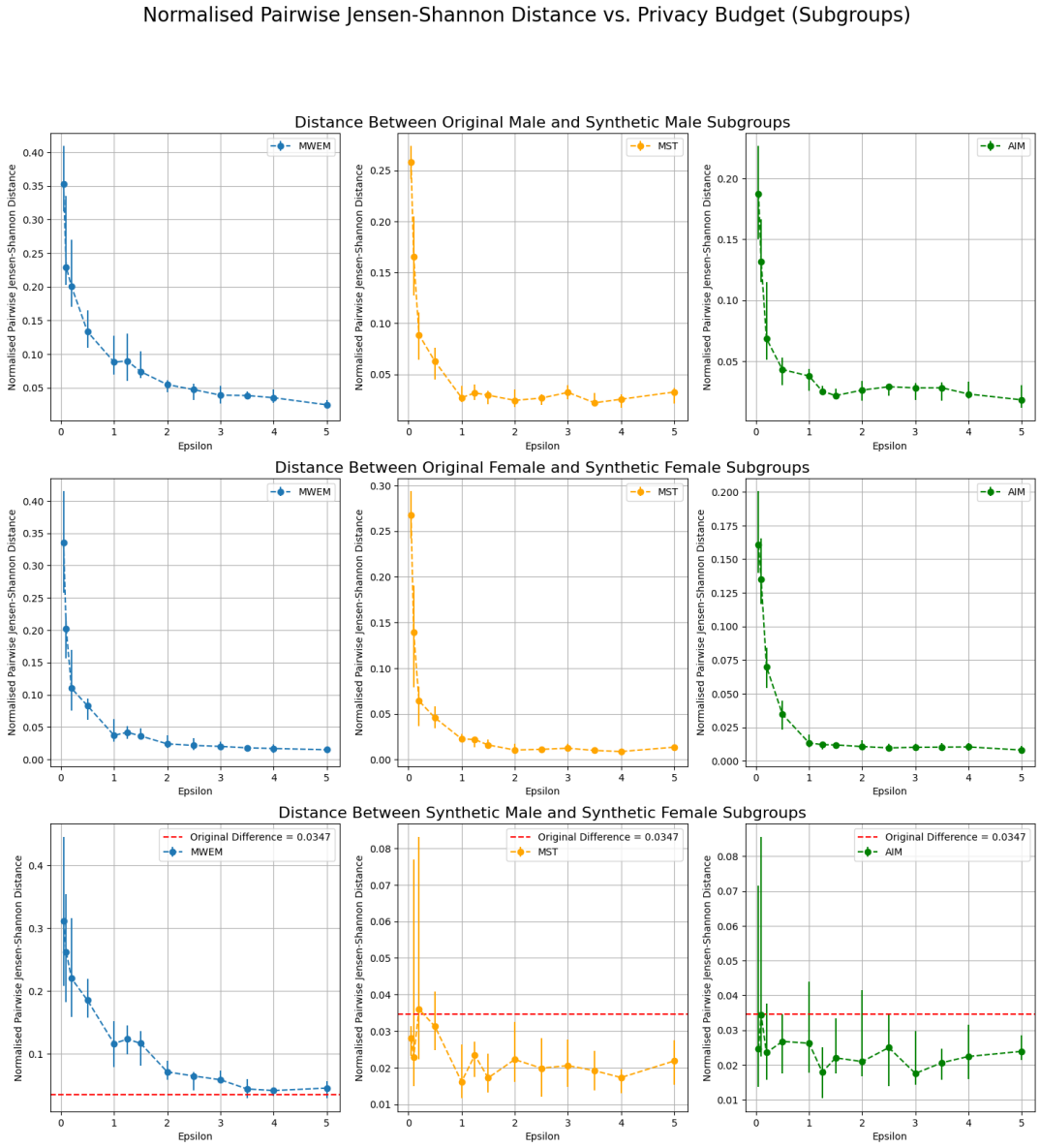


FIGURE 5.5: NPJS Distance vs. Privacy Budget for the different subgroups

at lower ϵ than for the male subgroup's remission curve. Once again, AIM is able to reach the maximum utility sooner than MST, MWEM. The lower row showing the distance between the synthetic male and female subgroup's remission curves has MWEM approaching the ideal level from above wheras MST and AIM remain below it.

An important discovery I made when computing these tables was the non-zero possibility that a dataset may be generated such that it is only composed of homogeneous males or females. This was the case for MWEM and MST at low values of epsilon. I chose to mitigate this by ignoring a homogeneous dataset for the computation of the opposite sex's remission curve (i.e. discarding some of the 20 trials to avoid NaN values).

TABLE 5.4: NPJS Distance vs. Privacy Budget for the different subgroups (Median with 25%, 75% IQR)

ϵ	MWEM	MST	AIM
	Original Males	vs.	Synthetic Males
0.05	0.3527 (0.3116, 0.4091)	0.2580 (0.2415, 0.2737)	0.1875 (0.1500, 0.2265)
0.10	0.2291 (0.2024, 0.3350)	0.1651 (0.1272, 0.2047)	0.1316 (0.1149, 0.1666)
0.20	0.2004 (0.1700, 0.2698)	0.0887 (0.0643, 0.1106)	0.0684 (0.0513, 0.1148)
0.50	0.1333 (0.1094, 0.1649)	0.0626 (0.0451, 0.0762)	0.0430 (0.0299, 0.0531)
1.00	0.0881 (0.0692, 0.1269)	0.0270 (0.0229, 0.0389)	0.0377 (0.0256, 0.0433)
1.25	0.0898 (0.0600, 0.1299)	0.0317 (0.0250, 0.0399)	0.0253 (0.0229, 0.0297)
1.50	0.0740 (0.0645, 0.1042)	0.0294 (0.0204, 0.0339)	0.0216 (0.0194, 0.0275)
2.00	0.0547 (0.0437, 0.0610)	0.0244 (0.0178, 0.0351)	0.0262 (0.0174, 0.0339)
2.50	0.0471 (0.0317, 0.0563)	0.0266 (0.0196, 0.0304)	0.0289 (0.0212, 0.0310)
3.00	0.0390 (0.0270, 0.0529)	0.0322 (0.0253, 0.0395)	0.0280 (0.0182, 0.0322)
3.50	0.0384 (0.0339, 0.0448)	0.0220 (0.0186, 0.0318)	0.0280 (0.0176, 0.0323)
4.00	0.0352 (0.0278, 0.0472)	0.0255 (0.0172, 0.0303)	0.0229 (0.0190, 0.0333)
5.00	0.0246 (0.0204, 0.0317)	0.0326 (0.0213, 0.0366)	0.0181 (0.0118, 0.0299)
	Original Females	vs.	Synthetic Females
0.05	0.3351 (0.2574, 0.4146)	0.2676 (0.2416, 0.2934)	0.1608 (0.1399, 0.2004)
0.10	0.2019 (0.1564, 0.2188)	0.1392 (0.0790, 0.1899)	0.1348 (0.1164, 0.1654)
0.20	0.1099 (0.0751, 0.1695)	0.0641 (0.0367, 0.0800)	0.0701 (0.0539, 0.0839)
0.50	0.0830 (0.0612, 0.0941)	0.0456 (0.0341, 0.0580)	0.0346 (0.0230, 0.0448)
1.00	0.0371 (0.0273, 0.0626)	0.0226 (0.0188, 0.0282)	0.0134 (0.0103, 0.0197)
1.25	0.0419 (0.0319, 0.0516)	0.0218 (0.0136, 0.0233)	0.0120 (0.0084, 0.0140)
1.50	0.0362 (0.0309, 0.0487)	0.0159 (0.0116, 0.0220)	0.0118 (0.0097, 0.0140)
2.00	0.0239 (0.0180, 0.0374)	0.0104 (0.0076, 0.0174)	0.0106 (0.0086, 0.0155)
2.50	0.0215 (0.0147, 0.0329)	0.0111 (0.0092, 0.0136)	0.0096 (0.0077, 0.0129)
3.00	0.0200 (0.0174, 0.0270)	0.0123 (0.0095, 0.0173)	0.0101 (0.0089, 0.0125)
3.50	0.0178 (0.0155, 0.0204)	0.0099 (0.0071, 0.0117)	0.0102 (0.0091, 0.0130)
4.00	0.0168 (0.0099, 0.0232)	0.0087 (0.0073, 0.0104)	0.0104 (0.0076, 0.0132)
5.00	0.0149 (0.0113, 0.0200)	0.0134 (0.0114, 0.0156)	0.0081 (0.0059, 0.0110)
	Synthetic Males	vs.	Synthetic Females
0.05	0.3109 (0.2081, 0.4455)	0.0281 (0.0233, 0.0314)	0.0246 (0.0138, 0.0715)
0.10	0.2625 (0.1824, 0.3538)	0.0229 (0.0150, 0.0769)	0.0343 (0.0223, 0.0856)
0.20	0.2199 (0.1581, 0.3163)	0.0360 (0.0223, 0.0832)	0.0235 (0.0156, 0.0376)
0.50	0.1852 (0.1570, 0.2189)	0.0313 (0.0248, 0.0409)	0.0267 (0.0176, 0.0346)
1.00	0.1164 (0.0788, 0.1519)	0.0161 (0.0116, 0.0264)	0.0262 (0.0177, 0.0438)
1.25	0.1234 (0.0987, 0.1449)	0.0234 (0.0175, 0.0272)	0.0179 (0.0104, 0.0250)
1.50	0.1166 (0.0810, 0.1366)	0.0172 (0.0133, 0.0238)	0.0220 (0.0176, 0.0334)
2.00	0.0712 (0.0591, 0.0890)	0.0222 (0.0160, 0.0325)	0.0209 (0.0167, 0.0415)
2.50	0.0648 (0.0417, 0.0715)	0.0199 (0.0120, 0.0281)	0.0250 (0.0139, 0.0346)
3.00	0.0588 (0.0537, 0.0729)	0.0206 (0.0148, 0.0276)	0.0175 (0.0142, 0.0298)
3.50	0.0442 (0.0293, 0.0599)	0.0192 (0.0138, 0.0246)	0.0206 (0.0158, 0.0247)
4.00	0.0416 (0.0332, 0.0455)	0.0173 (0.0130, 0.0196)	0.0224 (0.0160, 0.0315)
5.00	0.0456 (0.0295, 0.0560)	0.0218 (0.0153, 0.0274)	0.0239 (0.0213, 0.0284)

Moving onto Pearson correlations for the subgroups, these are shown in Figure

5.6 and Table 5.5. Once more notice the red-line in the lower plots, this indicates the ideal baseline for the Pearson correlation between the remission curves of the male and female subgroups. Serendipitously, in the original cohort the responses to treatment were almost proportional to one another at every interval and therefore the correlation comes out to ≈ 1. Again, we see that the performance curves for the female subgroup tends to flatten sooner than the male subgroup regardless of algorithm. The correlation between male and female subgroups' remission curves is maximised by AIM.

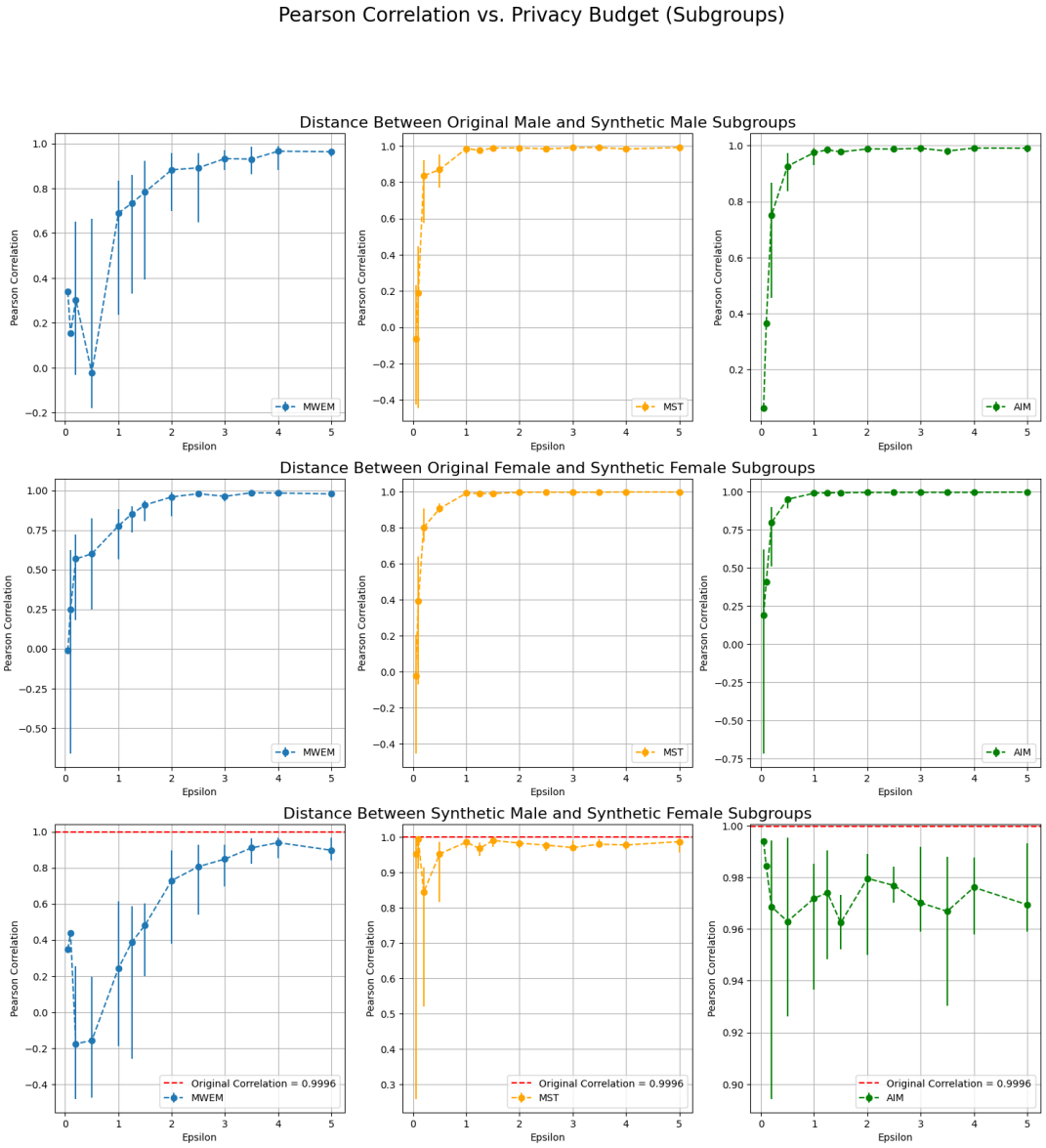


FIGURE 5.6: Pearson Correlation vs. Privacy Budget for the different subgroups

Finally, I examine the variance in the synthetic male (3) and female (4) subgroups. Figure 5.7 and Table 5.6 show a comparison of the uncertainty brought on by the choice of either the MWEM, MST or AIM algorithm for the two subgroups and how this uncertainty varies with the privacy budget. It can be explicitly seen that the subpopulation variances here are higher for all algorithms compared to the overall population. Furthermore, a higher privacy budget seems to be needed to bring

TABLE 5.5: Pearson Correlation vs. Privacy Budget for the different subgroups (Median with 25%, 75% IQR)

ϵ	MWEM	MST	AIM
	Original Males	vs.	Synthetic Males
0.05	0.3392 (0.0645, 0.6106)	-0.0632 (-0.4239, 0.2320)	0.0629 (-0.3156, 0.3994)
0.10	0.1542 (-0.5292, 0.7425)	0.1901 (-0.4444, 0.4450)	0.3641 (-0.0308, 0.7878)
0.20	0.3031 (-0.0308, 0.6506)	0.8351 (0.5754, 0.9206)	0.7523 (0.4555, 0.8667)
0.50	-0.0235 (-0.1797, 0.6647)	0.8692 (0.7687, 0.9518)	0.9260 (0.8380, 0.9719)
1.00	0.6896 (0.2348, 0.8338)	0.9858 (0.9561, 0.9924)	0.9748 (0.9313, 0.9900)
1.25	0.7334 (0.3315, 0.8593)	0.9751 (0.9637, 0.9863)	0.9846 (0.9704, 0.9887)
1.50	0.7834 (0.3927, 0.9220)	0.9878 (0.9803, 0.9919)	0.9770 (0.9653, 0.9880)
2.00	0.8829 (0.7002, 0.9559)	0.9891 (0.9773, 0.9965)	0.9878 (0.9745, 0.9908)
2.50	0.8917 (0.6475, 0.9565)	0.9832 (0.9787, 0.9941)	0.9872 (0.9750, 0.9923)
3.00	0.9326 (0.8831, 0.9703)	0.9899 (0.9749, 0.9939)	0.9901 (0.9745, 0.9952)
3.50	0.9304 (0.8625, 0.9866)	0.9911 (0.9848, 0.9957)	0.9801 (0.9652, 0.9896)
4.00	0.9657 (0.8817, 0.9882)	0.9824 (0.9759, 0.9922)	0.9909 (0.9862, 0.9946)
5.00	0.9633 (0.9405, 0.9735)	0.9915 (0.9700, 0.9985)	0.9905 (0.9754, 0.9976)
	Original Females	vs.	Synthetic Females
0.05	-0.0122 (-0.5637, 0.6212)	-0.0206 (-0.4533, 0.2045)	0.1909 (-0.7173, 0.6212)
0.10	0.2490 (-0.6616, 0.6212)	0.3942 (-0.0689, 0.6389)	0.4075 (0.1193, 0.6096)
0.20	0.5677 (0.1804, 0.7195)	0.8018 (0.7212, 0.9067)	0.7980 (0.5113, 0.8987)
0.50	0.6000 (0.2481, 0.8239)	0.9057 (0.8831, 0.9330)	0.9500 (0.8905, 0.9707)
1.00	0.7764 (0.5646, 0.8841)	0.9912 (0.9844, 0.9939)	0.9896 (0.9825, 0.9951)
1.25	0.8510 (0.7350, 0.9002)	0.9895 (0.9820, 0.9952)	0.9921 (0.9804, 0.9954)
1.50	0.9084 (0.8056, 0.9349)	0.9896 (0.9821, 0.9972)	0.9931 (0.9875, 0.9962)
2.00	0.9590 (0.8355, 0.9908)	0.9959 (0.9864, 0.9977)	0.9944 (0.9865, 0.9975)
2.50	0.9793 (0.9662, 0.9860)	0.9972 (0.9931, 0.9990)	0.9943 (0.9837, 0.9976)
3.00	0.9626 (0.9301, 0.9737)	0.9957 (0.9921, 0.9970)	0.9954 (0.9894, 0.9979)
3.50	0.9853 (0.9696, 0.9915)	0.9969 (0.9956, 0.9988)	0.9946 (0.9921, 0.9985)
4.00	0.9837 (0.9623, 0.9900)	0.9982 (0.9964, 0.9991)	0.9954 (0.9898, 0.9973)
5.00	0.9778 (0.9584, 0.9874)	0.9976 (0.9952, 0.9986)	0.9972 (0.9931, 0.9990)
	Synthetic Males	vs.	Synthetic Females
0.05	0.3470 (-0.3830, 0.6062)	0.9516 (0.2573, 0.9898)	0.9940 (0.9305, 0.9979)
0.10	0.4395 (-0.5889, 0.6763)	0.9931 (0.9099, 0.9985)	0.9844 (0.5441, 0.9964)
0.20	-0.1759 (-0.4831, 0.2539)	0.8431 (0.5203, 0.9146)	0.9686 (0.8942, 0.9943)
0.50	-0.1576 (-0.4730, 0.1965)	0.9520 (0.8163, 0.9860)	0.9629 (0.9261, 0.9954)
1.00	0.2424 (-0.1896, 0.6135)	0.9849 (0.9684, 0.9925)	0.9719 (0.9367, 0.9852)
1.25	0.3873 (-0.2586, 0.5886)	0.9669 (0.9467, 0.9830)	0.9740 (0.9484, 0.9904)
1.50	0.4819 (0.2010, 0.6043)	0.9899 (0.9721, 0.9946)	0.9624 (0.9521, 0.9733)
2.00	0.7290 (0.3794, 0.8951)	0.9823 (0.9683, 0.9913)	0.9796 (0.9500, 0.9889)
2.50	0.8059 (0.5398, 0.9290)	0.9766 (0.9598, 0.9855)	0.9768 (0.9701, 0.9842)
3.00	0.8485 (0.6961, 0.9277)	0.9697 (0.9599, 0.9899)	0.9701 (0.9591, 0.9918)
3.50	0.9101 (0.8237, 0.9611)	0.9795 (0.9695, 0.9946)	0.9668 (0.9304, 0.9878)
4.00	0.9394 (0.8535, 0.9672)	0.9770 (0.9636, 0.9887)	0.9761 (0.9579, 0.9875)
5.00	0.8970 (0.8415, 0.9683)	0.9865 (0.9557, 0.9960)	0.9694 (0.9589, 0.9932)

down the variance among the male subpopulation than the female subpopulation.

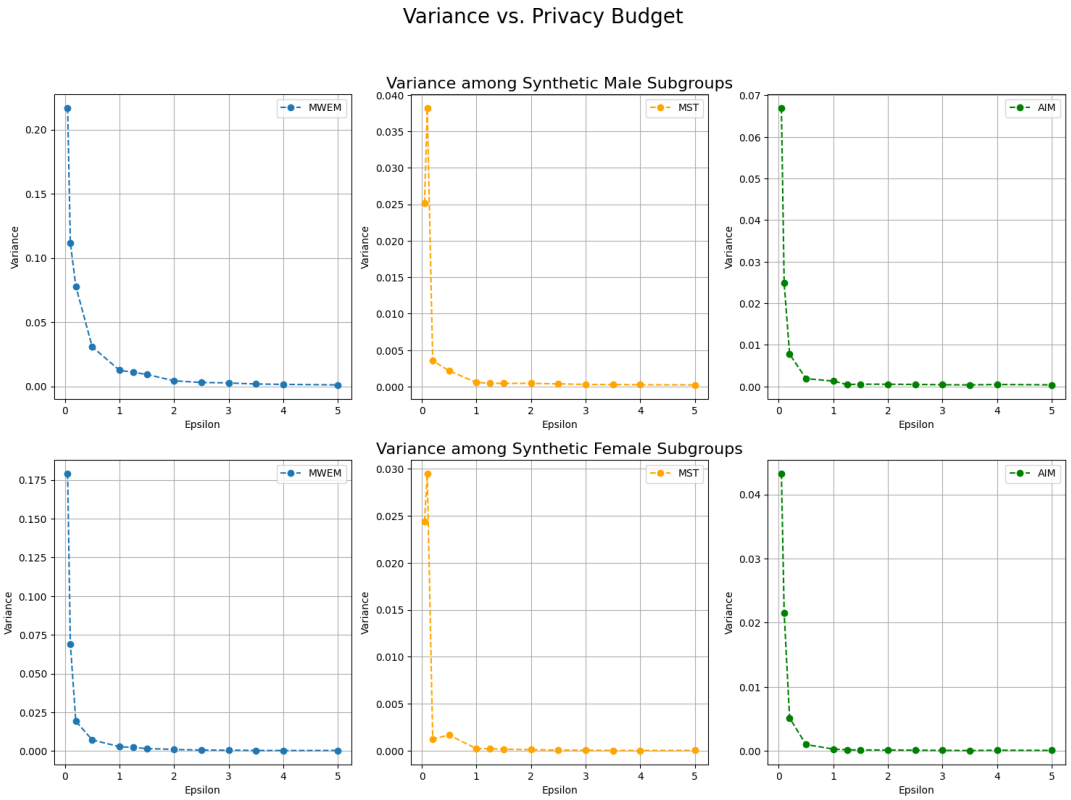


FIGURE 5.7: Variance of the Generated Male/Female Subgroups

TABLE 5.6: Variance of the Generated Male/Female Subgroups

ϵ	MWEM	MST	AIM
	Variance	among	Synthetic Males
0.05	0.2168	0.0252	0.0670
0.10	0.1117	0.0382	0.0249
0.20	0.0780	0.0036	0.0078
0.50	0.0308	0.0022	0.0019
1.00	0.0125	0.0006	0.0013
1.25	0.0110	0.0005	0.0005
1.50	0.0093	0.0005	0.0005
2.00	0.0043	0.0005	0.0005
2.50	0.0030	0.0004	0.0005
3.00	0.0027	0.0003	0.0004
3.50	0.0019	0.0003	0.0004
4.00	0.0016	0.0003	0.0005
5.00	0.0012	0.0002	0.0004
	Variance	among	Synthetic Females
0.05	0.1791	0.0244	0.0433
0.10	0.0690	0.0295	0.0215
0.20	0.0193	0.0013	0.0051
0.50	0.0072	0.0017	0.0010
1.00	0.0028	0.0002	0.0003
1.25	0.0023	0.0002	0.0002
1.50	0.0016	0.0002	0.0001
2.00	0.0010	0.0001	0.0001
2.50	0.0006	0.0001	0.0001
3.00	0.0006	0.0001	0.0001
3.50	0.0004	0.0000	0.0001
4.00	0.0003	0.0000	0.0001
5.00	0.0004	0.0001	0.0001

Chapter 6

Threats to Validity and Discussion

This section provides a discussion on my results. To further contextualise my findings, I open up with a list of threats to validity.

6.1 Threats to Validity

In this thesis, I have only focused on evaluating the performance of DP algorithms based on utility metrics. Sticking to the classical definition of differential privacy (Dwork, 2006) I define privacy as the opposite of utility. However, I have not considered engineered or targeted attacks (Stadler, Oprisanu, and Troncoso, 2022) and how those my exploit the algorithmic properties of the considered DP data synthesis algorithms.

Second, another weakness of my work is the small size. I synthesized differentially private datasets from an initial sample of $N = 1137$. However, most work on differentially private data synthesis deals with sample sizes in the tens of thousands(McKenna, Miklau, and Sheldon, 2021, Fang, Dhami, and Kersting, 2022). To draw a more strong generalization my findings for populations and sub-populations would warrant similar experiments on a larger dataset.

Lastly, I only run a limited number of $K = 20$ trials for all (algorithm,ϵ) pairs. Initially I aimed for higher, but because the runtime computation of AIM is very lengthy at higher levels of ϵ I settled on 20. This does bring some uncertainty to my results to await challenges from future research.

6.2 Discussion

My results have shown that working with low ϵ-DP for data synthesis might not be as unpredictable as some authors(Stadler, Oprisanu, and Troncoso, 2022) have described. While other researchers working on data synthesis ($\epsilon = 18$, Lee et al., 2020) or investigating the effects of DP on subpopulations ($\epsilon = 19.61$, Lee et al., 2020) have required privacy budgets of a much higher magnitude the AIM(McKenna et al., 2022) algorithm was able synthesize data approximating an real-world dataset for $\epsilon \leq 5$. AIM was able to synthesize a patient cohort whose remission curve was numerically close to and correlated with the original cohort's remission curve. Perhaps, research should consider highlighting input dimmensionality (and cardinality for marginal-based DP data synthesis) so that we can better relate ϵ privacy budgets.

On the other hand, I have also through this project seen the different things that can go wrong. For one, it's possible for scant ϵ to generate even inversely correlated results. Compared to the trend of improvement until month 3, and decline afterwards; DP data synthesis can possibly produce a mirror image with decline until

month 3 and improvement afterwards. Having a custodian, who has authorised access to the original data, also check basic use cases for a synthetic dataset before it is released is advisable, as I was able to do in my controlled experiments.

Moving on to the secondary objective of also generating data with feasible remission curves to match the subpopulations, this too was achievable but for a higher cost in more privacy loss. An interesting pitfall was the discovery that one could hypothetically encounter homogeneous populations that were all male or female. There was also the issue of imbalance between the female and male patient population, where women formed the majority. Similar to previous work I also found that DP obscures smaller groups rather than the majority (Santos-Lozada, Howard, and Verdery, 2020; Kurz et al., 2022). It was considerably more difficult for DP-synthetic data to match the remission curve trend for men, than women; as well as variance being higher for the generated synthetic male population's remission curve when ϵ is held constant. Given that AIM, MWEM and MST are all based on the Exponential Mechanism in one way or another; this suggests that some form of probabilistic regularization may be needed to prevent scores to favor a majority group.

Chapter 7

Conclusion

In this thesis I synthesized patient data with differential privacy, with the goal of it being re-usable to gain realistic insights on the population.

- **RQ1: Is there practical significance between algorithm selection along the privacy-utility trade-off?**
- Conclusion: There is practical significance relating to algorithm-selection, I found the AIM algorithm to outperform MWEM and MST for generating DP-Synthetic Data with higher utility at the same ϵ. AIM was able to generate less distant data, more correlated data and less varying data.
- **RQ2: How much of a privacy budget is required for differential privacy to preserve the characteristics of population subgroups?**
- Conclusion: Preserving the characteristics of the population subgroups costed more than preserving the characteristics of the whole population. AIM was able to achieve an optimal remission curve for the whole population around $\epsilon = 1$. On the other hand, the remission curve for the female subpopulation was well-preserved around $\epsilon = 1.5$ and the male remission curve for the male subpopulation reached its optimum at around $\epsilon = 5$.

Releasing synthetic data publicly can be used for observational studies, but at this stage it is recommendable to stick to population-wide trends rather than those of population subgroups. DP synthetic data would be welcome wherever some loss of precision is acceptable, in particular data visualization may benefit greatly from imprecise but approximate data.

Appendix A

Synthesizer Parameters

```
Synthesizer.create(
    'mwem', epsilon=eps[i],
    split_factor=10,
    q_count=None, iterations=None, splits=[],
    marginal_width=None, add_ranges=False,
    measure_only=False, max_retries_exp_mechanism=10,
    mult_weights_iterations=20,
    verbose=False
),

...

Synthesizer.create(
    'mst', epsilon=eps[i], delta=1e-09, verbose=False
),

...

Synthesizer.create(
    'aim', epsilon=eps[i], delta=1e-09,
    max_model_size=80, degree=2, num_marginals=None,
    max_cells=10000, rounds=None,
    verbose=False
)
```

Appendix B

Source Code

Some level of source-code will be made available at https://github.com/madprogramer/clinical-differential-privacy-project/, provided that there are no issues with leaking data.

Bibliography

Abadi, Martín et al. (Oct. 2016). "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. arXiv:1607.00133 [cs, stat], pp. 308–318. DOI: 10.1145/2976749.2978318. URL: http://arxiv.org/abs/1607.00133 (visited on 04/22/2023).

Aletaha, Daniel and Josef S. Smolen (Oct. 2018). "Diagnosis and Management of Rheumatoid Arthritis: A Review". eng. In: *JAMA* 320.13, pp. 1360–1372. ISSN: 1538-3598. DOI: 10.1001/jama.2018.13103.

Almadhoun, Nour, Erman Ayday, and Özgür Ulusoy (Mar. 2020). "Differential privacy under dependent tuples—the case of genomic privacy". In: *Bioinformatics* 36.6, pp. 1696–1703. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz837. URL: https://doi.org/10.1093/bioinformatics/btz837 (visited on 05/26/2023).

Beaulieu-Jones, Brett K. et al. (July 2019). "Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing". eng. In: *Circulation. Cardiovascular Quality and Outcomes* 12.7, e005122. ISSN: 1941-7705. DOI: 10.1161/CIRCOUTCOMES.118.005122.

Cho, Hyunghoon et al. (May 2020). "Privacy-Preserving Biomedical Database Queries with Optimal Privacy-Utility Trade-Offs". eng. In: *Cell Systems* 10.5, 408–416.e9. ISSN: 2405-4720. DOI: 10.1016/j.cels.2020.03.006.

Cybersecurity Law of the People's Republic of China (2016). Standing Committee of the National People's Congress. URL: http://www.lawinfochina.com/Display.aspx?LookType=3&Lib=law&Id=22826&SearchKeyword=&SearchCKeyword=&paycode= (visited on 05/01/2023).

Das, Soumojit et al. (July 2022). *Imputation under Differential Privacy*. arXiv:2206.15063 [cs]. DOI: 10.48550/arXiv.2206.15063. URL: http://arxiv.org/abs/2206.15063 (visited on 02/22/2023).

Davey, Matthew G. et al. (Aug. 2022). "General data protection regulations (2018) and clinical research: perspectives of patients and doctors in an Irish university teaching hospital". eng. In: *Irish Journal of Medical Science* 191.4, pp. 1513–1519. ISSN: 1863-4362. DOI: 10.1007/s11845-021-02789-8.

Dwork, Cynthia (July 2006). "Differential privacy". In: *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II*. ICALP'06. Berlin, Heidelberg: Springer-Verlag, pp. 1–12. ISBN: 978-3-540-35907-4. DOI: 10.1007/11787006_1. URL: https://doi.org/10.1007/11787006_1 (visited on 05/24/2023).

Dwork, Cynthia and Aaron Roth (2013). "The Algorithmic Foundations of Differential Privacy". en. In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4, pp. 211–407. ISSN: 1551-305X, 1551-3068. DOI: 10.1561/0400000042. URL: http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042 (visited on 04/22/2023).

Dwork, Cynthia et al. (2006). "Our Data, Ourselves: Privacy Via Distributed Noise Generation". en. In: *Advances in Cryptology - EUROCRYPT 2006*. Ed. by Serge Vaudenay. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 486–503. ISBN: 978-3-540-34547-3. DOI: 10.1007/11761679_29.

England, Bryant R. et al. (Dec. 2019). "2019 Update of the American College of Rheumatology Recommended Rheumatoid Arthritis Disease Activity Measures". eng. In: *Arthritis Care & Research* 71.12, pp. 1540–1555. ISSN: 2151-4658. DOI: 10.1002/acr.24042.

Fang, Mei Ling, Devendra Singh Dhami, and Kristian Kersting (2022). "DP-CTGAN: Differentially Private Medical Data Generation Using CTGANs". en. In: *Artificial Intelligence in Medicine*. Ed. by Martin Michalowski, Syed Sibte Raza Abidi, and Samina Abidi. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 178–188. ISBN: 978-3-031-09342-5. DOI: 10.1007/978-3-031-09342-5_17.

Ficek, Joseph (2021). "Differential Privacy for Regression Modeling in Health: An Evaluation of Algorithms". İngilizce. ISBN: 9798759958703. Ph.D. Ann Arbor, United States. URL: https://www.proquest.com/pqdtglobal/docview/2616265545/abstract/C32043580EA04F0EPQ/3 (visited on 04/28/2023).

Goodfellow, Ian J. et al. (June 2014). *Generative Adversarial Networks*. arXiv:1406.2661 [cs, stat]. DOI: 10.48550/arXiv.1406.2661. URL: http://arxiv.org/abs/1406.2661 (visited on 06/01/2023).

Hardt, Moritz, Katrina Ligett, and Frank McSherry (Dec. 2012). "A simple and practical algorithm for differentially private data release". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., pp. 2339–2347. (Visited on 05/23/2023).

Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC (June 2022). en-us. URL: https://www.cdc.gov/phlp/publications/topic/hipaa.html (visited on 05/01/2023).

Jordon, James, Jinsung Yoon, and Mihaela van der Schaar (Dec. 2018). "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees". en. In: URL: https://openreview.net/forum?id=S1zk9iRqF7 (visited on 06/01/2023).

Kurz, Christoph F. et al. (2022). "The effect of differential privacy on Medicaid participation among racial and ethnic minority groups". en. In: *Health Services Research* 57.S2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-6773.14000, pp. 207–213. ISSN: 1475-6773. DOI: 10.1111/1475-6773.14000. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.14000 (visited on 02/22/2023).

Lee, Dongha et al. (Sept. 2020). "Generating sequential electronic health records using dual adversarial autoencoder". In: *Journal of the American Medical Informatics Association : JAMIA* 27.9, pp. 1411–1419. ISSN: 1067-5027. DOI: 10.1093/jamia/ocaa119. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647348/ (visited on 05/26/2023).

Leister, Wolfgang et al. (Oct. 2016). "Towards Evidence-Based Self-Management for Spondyloarthritis Patients". In.

Malin, Bradley A, Khaled El Emam, and Christine M O'Keefe (Jan. 2013). "Biomedical data privacy: problems, perspectives, and recent advances". In: *Journal of the American Medical Informatics Association* 20.1, pp. 2–6. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2012-001509. URL: https://doi.org/10.1136/amiajnl-2012-001509 (visited on 05/31/2023).

McKenna, Ryan, Gerome Miklau, and Daniel Sheldon (Aug. 2021). *Winning the NIST Contest: A scalable and general approach to differentially private synthetic data*. arXiv:2108.04978 [cs]. DOI: 10.48550/arXiv.2108.04978. URL: http://arxiv.org/abs/2108.04978 (visited on 05/23/2023).

McKenna, Ryan et al. (Jan. 2022). *AIM: An Adaptive and Iterative Mechanism for Differentially Private Synthetic Data*. arXiv:2201.12677 [cs]. DOI: 10.48550/arXiv.2201.12677. URL: http://arxiv.org/abs/2201.12677 (visited on 05/23/2023).

Mian, Aneela, Fowzia Ibrahim, and David L. Scott (2019). "A systematic review of guidelines for managing rheumatoid arthritis". eng. In: *BMC rheumatology* 3, p. 42. ISSN: 2520-1026. DOI: 10.1186/s41927-019-0090-7.

Mohammed, Noman et al. (May 2013). "Privacy-preserving heterogeneous health data sharing". In: *Journal of the American Medical Informatics Association* 20.3, pp. 462–469. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2012-001027. URL: https://doi.org/10.1136/amiajnl-2012-001027 (visited on 05/01/2023).

Narayanan, Arvind and Vitaly Shmatikov (Oct. 2006). "How To Break Anonymity of the Netflix Prize Dataset". In: *ArXiv*. URL: https://www.semanticscholar.org/paper/How-To-Break-Anonymity-of-the-Netflix-Prize-Dataset-Narayanan-Shmatikov/c40e5c8b4957074644acdaf1f9f4332e63b5846b (visited on 05/15/2023).

Niinimäki, Teppo et al. (July 2019). "Representation transfer for differentially private drug sensitivity prediction". eng. In: *Bioinformatics (Oxford, England)* 35.14, pp. i218–i224. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btz373.

Nikiphorou, Elena et al. (Oct. 2016). "Patient global assessment in measuring disease activity in rheumatoid arthritis: a review of the literature". eng. In: *Arthritis Research & Therapy* 18.1, p. 251. ISSN: 1478-6362. DOI: 10.1186/s13075-016-1151-6.

Peloquin, David et al. (June 2020). "Disruptive and avoidable: GDPR challenges to secondary research uses of data". en. In: *European Journal of Human Genetics* 28.6. Number: 6 Publisher: Nature Publishing Group, pp. 697–705. ISSN: 1476-5438. DOI: 10.1038/s41431-020-0596-x. URL: https://www.nature.com/articles/s41431-020-0596-x (visited on 05/30/2023).

GDPR (Apr. 2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504 (visited on 05/31/2023).

Santos-Lozada, Alexis R., Jeffrey T. Howard, and Ashton M. Verdery (June 2020). "How differential privacy will affect our understanding of health disparities in the United States". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 117.24, pp. 13405–13412. ISSN: 1091-6490. DOI: 10.1073/pnas.2003714117.

Smolen, Josef S. et al. (Jan. 2016). "Treating rheumatoid arthritis to target: 2014 update of the recommendations of an international task force". en. In: *Annals of the Rheumatic Diseases* 75.1. Publisher: BMJ Publishing Group Ltd Section: Recommendation, pp. 3–15. ISSN: 0003-4967, 1468-2060. DOI: 10.1136/annrheumdis-2015-207524. URL: https://ard.bmj.com/content/75/1/3 (visited on 04/23/2023).

Sokka, Tuulikki et al. (Mar. 2010). "Work disability remains a major problem in rheumatoid arthritis in the 2000s: data from 32 countries in the QUEST-RA Study". In: *Arthritis Research & Therapy* 12.2, R42. ISSN: 1478-6354. DOI: 10.1186/ar2951. URL: https://doi.org/10.1186/ar2951 (visited on 04/23/2023).

Stadler, Theresa, Bristena Oprisanu, and Carmela Troncoso (Aug. 2022). "Synthetic Data – Anonymisation Groundhog Day". In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, pp. 1451–1468. ISBN: 978-1-939133-31-1. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.

Vinterbo, Staal A, Anand D Sarwate, and Aziz A Boxwala (2012). "Protecting count queries in study design". In: *Journal of the American Medical Informatics Association : JAMIA* 19.5, pp. 750–757. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2011-000459. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3422502/ (visited on 05/01/2023).

Wang, Hao et al. (Jan. 2021). "Why current differential privacy schemes are inapplicable for correlated data publishing?" In: *World Wide Web* 24, pp. 1–23. DOI: 10.1007/s11280-020-00825-8.

Xiaoxiao, Li et al. (Oct. 2020). "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results". en. In: *Medical image analysis* 65. Publisher: Med Image Anal. ISSN: 1361-8423. DOI: 10.1016/j.media.2020.101765. URL: https://pubmed.ncbi.nlm.nih.gov/32679533/ (visited on 05/27/2023).

Xie, Liyang et al. (Feb. 2018). *Differentially Private Generative Adversarial Network*. arXiv:1802.06739 [cs, stat]. DOI: 10.48550/arXiv.1802.06739. URL: http://arxiv.org/abs/1802.06739 (visited on 06/01/2023).

Xu, Lei et al. (Oct. 2019). *Modeling Tabular data using Conditional GAN*. arXiv:1907.00503 [cs, stat]. DOI: 10.48550/arXiv.1907.00503. URL: http://arxiv.org/abs/1907.00503 (visited on 06/01/2023).