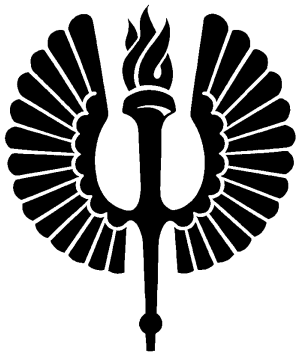


Generating Synthetic Longitudinal Patient Data with the PrivBayes Method

Katariina Perkonoja

Master's thesis
December 2020

DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF TURKU

UNIVERSITY OF TURKU
Department of Mathematics and Statistics

PERKONOJA, KATARIINA: Generating Synthetic Longitudinal Patient Data with the PrivBayes Method
Master's thesis, 45 pages, 38 appendix pages
Statistics
December 2020

In this thesis, the PrivBayes method is used to generate synthetic longitudinal patient data and the quality of the generated data is evaluated. In addition, this thesis briefly discusses the current situation of processing health data in Finland and proposes a simplistic definition of synthetic tabular data as well as presents different methods to evaluate the utility of generated synthetic data.

The PrivBayes method is based on approximating the association structure of a data set using a Bayesian network and generating synthetic data from the conditional distributions corresponding to the structure of the network. The method ensures the privacy of the data by applying differential privacy through the addition of noise in the data generation process in a specific way.

The method is applied to data collected from the database of Auria Clinical Informatics under permission number T152/2017. The data set consists of 2890 individual patients diagnosed with either type 1 or type 2 diabetes and seven different characteristics collected for each patient: age, body mass index, complications related to diabetes, gender, type of diabetes and two measurements for glycated hemoglobin that represent the repeated measurements in the data.

The PrivBayes method is evaluated by generating 27 different synthetic data sets, describing the structures of the Bayesian network of each data set and visually inspecting differences between the original data and each synthetic data set. Differences between data sets are considered in terms of similarity of univariate distributions, differences in Pearson's sample correlation coefficients and sample Cramer's V coefficients and the results of a linear mixed-effects model.

In conclusion, the PrivBayes method failed to produce synthetic longitudinal patient data of sufficient quality to be applicable as such in practice. However, this thesis revealed some shortcomings of the method and potential targets for further research and development.

Keywords: anonymity, Bayesian network, differential privacy, longitudinal data, the PrivBayes method, the Secondary Act, statistical disclosure control, synthetic data

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

PERKONOJA, KATARIINA: Synteettisen potilasseuranta-aineiston luominen PrivBayes-menetelmällä
Pro gradu, 45 s., 38 liites.
Tilastotiede
Joulukuu 2020

Tässä pro gradu -tutkielmassa käytetään PrivBayes-menetelmää synteettisen potilasseuranta-aineiston tuottamiseksi ja arvioidaan tuotetun aineiston laatua. Tämän lisäksi tutkielmassa kerrotaan lyhyesti terveystietojen käsittelyn nykytilanteesta Suomessa, minkä lisäksi ehdotetaan yksinkertaista määritelmää synteettiselle taulukkomuotoiselle aineistolle sekä esitellään menetelmiä tuotetun synteettisen aineiston käytettävyyden arvioimiseksi.

PrivBayes-menetelmä perustuu aineistossa esiintyvien assosiaatiorakenteiden mallintamiseen Bayes-verkon avulla ja synteettisen aineiston tuottamiseen ehdollisista jakaumista, jotka vastaavat verkon rakennetta. Menetelmä varmistaa aineiston tietosuojan soveltamalla differentiaalista yksityisyyttä, jossa aineiston tuotantoprosessiin lisätään tietyn tyyppistä kohinaa.

Menetelmää sovelletaan aineistoon, joka on kerätty Auria Tietopalveluiden tietokannasta tietolupanumerolla T152/2017. Aineisto koostuu 2890 yksittäisestä potilaasta, joilla on diagnosoitu joko tyypin 1 tai 2 diabetes, ja seitsemästä eri potilaita kuvaavasta muuttujasta: iästä, painoindeksistä, diabetekseen liittyvistä komplikaatiosta, sukupuolesta, diabeteksen tyypistä sekä kahdesta glykatoituneen hemoglobiinin mittauksesta, jotka edustavat seurantamittauksia aineistossa.

PrivBayes-menetelmää arvioidaan luomalla 27 erilaista synteettistä aineistoa, kuvailemalla kutakin aineistoa vastaava Bayes-verkon rakenne sekä arvioimalla visuaalisesti alkuperäisen aineiston ja synteettisen aineiston välisiä eroja yksiulotteisissa jakaumissa, Pearsonin otoskorrelaatio- ja Cramerin V-kertoimissa sekä lineaarisen sekamallin tuloksissa.

Tutkielman johtopäätöksenä voidaan todeta, että PrivBayes-menetelmä ei kyennyt tuottamaan riittävän laadukasta synteettistä potilasseuranta-aineistoa, jota voitaisiin sellaisenaan soveltaa käytännössä. Tutkielma kuitenkin paljasti joitakin menetelmän puutteita sekä mahdollisia kohteita jatkotutkimukselle ja -kehitykselle.

Avainsanat: anonymiteetti, Bayes-verkko, differentiaalinen yksityisyys, PrivBayes-menetelmä, seuranta-aineisto, synteettinen aineisto, tilastollinen tietosuoja, toisiolaki

Contents

1 Introduction

Patient data are generally classified as highly sensitive personal information. In Finland, the processing of such data is not only regulated by the EU General Data Protection Regulation (GDPR) [1] but also by several national laws, for example, the Act on the Secondary Use of Health and Social Data (the Secondary Act) [2]. Different regulatory schemes typically cause long permit processing times, and getting access to data may take months or even longer than a year [3]. Access to data becomes even more difficult if data are requested from several different data controllers. In order to facilitate the availability of Finnish social and health data, the Act on the Secondary Use of Health and Social Data was decreed. The purpose of the Secondary Act is also to ensure safe use of data specified in the Secondary Act and to protect personal information in such data.

Different Statistical Disclosure Control (SDC) methods have been developed to protect different types of data from disclosing confidential information. An inherent problem is that all SDC methods cause some degree of information loss, which affects statistical inference and learning. In other words, there is always a trade-off between addressing disclosure risk and safeguarding data utility [4]. The idea of generating synthetic data to preserve confidentiality was first introduced by Rubin in 1993 [5]. According to Rubin [5], synthetic data do not include any actual confidential data, that is, such data can be considered anonymous. This implies that synthetic data are no longer regulated by any personal data protection regulation, as anonymous data are not considered to be personal data [6]. According to Rubin [5], synthetic data should provide much better utility compared to anonymized data by imitating the original data not by only on its looks, but also valid inferences for legitimate estimands should be easily obtained. Ever since, different types of synthetic data sets have been generated with various methods.

This thesis was written while working at Auria Clinical Informatics (ACI), a department of the Turku University Hospital, as a part of the Finnish and Danish collaborative research project called Synthetic Health and Research Data (SHARED) funded by the Novo Nordisk Foundation (grant NNF19SA0059129). The objectives of the SHARED project are to evaluate the existing methods for synthetic data generation in the context of health data, develop and optimize new methods and assess the quality, safety and utility of the constructed synthetic data as well as explore the GDPR and the current national legislation in the context of synthetic data generation.

Researching the generation of synthetic data is also important from ACI's own perspective as one of the main tasks of ACI is to provide data sets to different types of scientific research projects and other purposes under the Secondary Act. Data sets are generated from patient data, originating from actual patient visits in the University Hospital, which means that they are highly sensitive and usage of such data is extremely regulated. However, if the data can be anonymized to protect the statistical units - in this context the patients - from being disclosed, such anonymized data could perhaps be provided, for example, for the development and innovation activities defined in the Secondary Act or be published in the context of scientific research. Finding suitable methods for producing such anonymous data is the motivation of this thesis. In addition, all views presented in the thesis are those of the author and not of ACI or the SHARED project.

The aim of this thesis is to describe the PrivBayes method [7], and use it to generate synthetic longitudinal patient data, and to evaluate the level of utility of the generated

data. The PrivBayes method is based on modeling the associations of the attributes in the data using a Bayesian network, utilizing the network in estimating the joint distribution of the attributes and generating synthetic data from the estimated distribution. In order to implement the algorithm, a preliminary R package is programmed. In addition, this thesis proposes a simplistic definition of synthetic data and presents general approaches to evaluate the similarity between synthetic and original data. The original data set used in this thesis is a data set collected by the author from the ACI's database (permission number T152/2017), where data retrieved from the operational systems of the Turku University Hospital are stored.

The structure of this thesis is as follows. Section 2 briefly discusses the history and the current state of the data protection legislation in Finland and introduces the two main concepts of this thesis: statistical disclosure control and synthetic data generation. Section 3 covers differential privacy, a method applied in PrivBayes to achieve private data. The PrivBayes method is introduced in Section 4 and different approaches to measure data utility are presented in Section 5. In Section 6, the PrivBayes method is applied to the original data set. The results, methods and limitations of this thesis are discussed in Section 7.

2 Protecting data privacy

This section briefly discusses the history of data protection as well as the current interpretation of the legislation in the context of anonymous and synthetic data in Finland. The section also introduces the main concepts of statistical disclosure control, a methodology of controlling the privacy of data, and synthetic data generation, an approach to achieve private data. These concepts serve as the basis for this thesis. For the most part, Subsection 2.1 is based on different legislation and online sources. Subsection 2.2 is based on the textbook *Statistical Disclosure Control* by Hundepool et al. [4] and Subsection 2.3 is based on a literature review and the textbook *Modeling Longitudinal Data* by Weiss [8].

2.1 Data protection legislation in Finland

Since the invention of computers and the World Wide Web, the amount of data has grown exponentially. The Global DataSphere [9] from the International Data Corporation (IDC) forecast that in 2020 more than 59 zettabytes of data will be created, captured, copied, and consumed. One zettabyte (ZB) is 1000^7 bytes, meaning one billion terabytes (TB), the size range currently used in leading hard drives. The fact that more and more data have emerged and become available has created a growing need for transnational and national data protection and security legislation. This thesis focuses on the Finnish legislation, and especially on the protection and use of health data. In general, different legislation and protection requirements apply to different types of data in different countries.

According to the time line created by the European Data Protection Supervisor on the history of European data protection legislation [10], the first EU directive on the protection of personal data was introduced in 1995 when Directive 95/46/EC was adopted. The directive defined that all data related to an identified or identifiable natural person are defined as *personal data*, a definition still in use today. The purpose of the directive was to protect individuals with regard to processing of personal data and the free movement of such data. As time went on and technologies evolved, the Directive 95/46/EC was considered insufficient and in 2012 The European Commission proposed a comprehensive reform of the EU's 1995 data protection rules. Four years later, Regulation (EU) 2016/679 was adopted, better known as the EU General Data Protection Regulation (GDPR). The GDPR was enforced on 25 May 2018 and is complied with in all EU Member States, including Finland. [10]

Based on the interview with P.-L. Heiliö [11], secretary of the preparatory working group and referendary of the Act, at the same time as the GDPR was being prepared, the Act on Secondary Use of Health and Social Data, later referred to as the Secondary Act, was being prepared in Finland. In Finland, as in many other Nordic countries, personal data have long been collected in various registers of different *data controllers*: a person, company, or other body that determines the purpose and means of personal data processing. However, obtaining such data for research purposes, for example, has been a laborious and long process. According to Heiliö [11], one of the purposes of creating the Secondary Act was to facilitate access to and processing of data and to enable these valuable data to be exploited throughout the EU. Several different parties, such as the Finnish Ministry of Social Affairs and Health, the Finnish Institute for Health and Welfare, Turku University Hospital, researchers and experts in various fields, clinicians and a representative of the

Data Protection Ombudsman, were involved in the preparation phase. The final version of the Secondary Act entered into force on 1 May 2019. In addition to the Secondary Act and the GDPR, Finland complies with the Data Protection Act, which is ordained to augment and clarify the protection of natural persons in the processing of personal data and the national application of the GDPR. These three laws must always be applied in parallel. [11]

According to the Finnish Ministry of Social Affairs and Health, "the purpose of the [Secondary] Act is to facilitate the effective and safe processing and access to the personal social and health data for steering, supervision, research, statistics and development in the health and social sector" [2]. The secondary uses referred to in the Secondary Act include scientific research, statistics, development and innovation activities, steering and supervision of authorities, planning and reporting duties by authorities, teaching and knowledge management. For example, the Secondary Act allows for a retrospective registry research without requiring a separate consent from the participants in the research, provided that the research meets the requirements set out in the Secondary Act, the Data Protection Act and the GDPR. The Secondary Act allows access to two different types of data, and the type of data depends on the intended use mentioned earlier. Aggregated statistical data, which are defined in the Secondary Act as statistical and reliably anonymous information, can be obtained with *data request* and confidential personal data specified in accordance with the Secondary Act can be obtained with *data permit*. [2]

Particularly interesting areas under the Secondary Act, from the perspective of patient data, are development and innovation activities, for which aggregated statistical information can be granted on request. Nevertheless, development and innovation activities are subject to at least one of the following objectives: the promotion of public health or social security, the development of social and health care services or the service system, the protection of the health or well-being of individuals or the safeguarding of their rights and freedoms in connection therewith [2]. Although development and innovation activities are limited to the above-mentioned uses, various health companies can still benefit from this opportunity provided by the law, as the internal processes for conducting scientific research in these companies are often heavier than those for development and innovation. Aggregated statistical data must also be primarily provided for educational activities, but if those activities cannot be performed with such data for reasons specified in the Secondary Act, personal data can also be provided for educational purposes.

In order to centralize, control and assure safe data processing, particularly in case of multi-controller research or when data are saved in the Kanta service (Finnish digital social welfare and health sector service) or when the data in question are register data from private social welfare and health care service providers, the Secondary Act designates a separate body, the data permit authority, to perform this task [2]. This data permit authority, known as Findata, became operational on January 1, 2020. Findata implements the Secondary Act by operating as a one-stop shop for the secondary use of health and social data. According to Findata's website [12], Findata's goals are to "improve data security and the data protection of individuals, speed up and streamline the utilisation of social welfare and health care data resources, decrease the duplication of work in permit processing and develop data descriptions for the social welfare and health care sector together with the controllers."

As the official data permit authority, Findata has the exclusive right to produce ag-

gregated statistical data in accordance with the data request and to ensure the anonymity of the results produced with the data, subject to the data permit, even in cases where the authorization has been granted by an individual data controller. For a justified reason, however, the data permit authority may grant the permittee the right to anonymize the results, provided that they are subsequently provided to Findata. In order to assist the data permit authority with anonymization, data protection and data security, the Secondary Act ordered the Finnish Ministry of Social Affairs and Health to set up a high-level expert group for Findata, with the task to establish guidelines for anonymization, data protection and data security. [2]

Since the Secondary Act is a new legislation and Findata has only started working at the time of writing this thesis, there exist no unambiguous views on the interpretation of the Secondary Act or on appropriate methods and procedures for data anonymization. For example, it is uncertain whether individual-level synthetic data can be considered as aggregated statistical data if the data generation method is based on aggregation of the original data. In addition to the Secondary Act, special caution is taken in the interpretation of the GDPR as well, and no official policy on what is considered as sufficiently anonymous data exists. For example, if the production of synthetic data is somehow based on simulation and random sampling and an observation is randomly generated but by chance resembles someone in the original data set, can the synthesized data be considered sufficiently anonymous? Although the interpretation of the Secondary Act is still open, reliable methods are needed to produce anonymous and synthetic data so that these data can be provided to those who need them within the framework of the Secondary Act.

2.2 Statistical disclosure control

A *microdata set* is a data set consisting of n records, also referred to as *statistical units* or *respondents*. Each record contains observations of p variables, denoted by X_{ij}, on an individual respondent i. According to Hundepool et al. [4], all other data formats are derived from microdata, which thus are the primary form that data are stored in. In this thesis, microdata and microdata set are generally referred to as data and data set if not otherwise mentioned. Furthermore, variables X_{ij} are also called *attributes* and statistical units ($i = 1, 2, \ldots, n$) are referred to as *subjects* in the context of health data. For convenience, we omit index i of the unit when not needed for clarity.

Hundepool et al. [4] coarsely divide attributes in the data into four categories depending on their level of sensitivity: identifiers, quasi-identifiers, confidential variables and non-confidential variables. *Identifiers* unambiguously identify the subject. These attributes include, for example, the social security number and the full name and they are usually removed or encrypted. Data from which all identifiers have been removed or encrypted are called *pseudonymized data*. A *quasi-identifier* is a set of attributes that, in combination, can be linked with external information to re-identify the subjects to whom the records refer to. Quasi-identifiers cannot be removed because any attribute in the data set potentially belongs to a quasi-identifier. *Confidential variables* contain sensitive information on the subject, e.g., health condition, religious view or membership in a union. *Non-confidential variables* contain non-sensitive information on the respondent, for example job and town of residence. However, these variables can be part of a quasi-identifier; in a small town it

is easy to resolve who the head of the police department is.

Statistical Disclosure Control (SDC) aims to protect data in such a way that they can be released without giving away confidential information that could be linked to specific individuals or entities. Therefore SDC is sometimes directly called anonymization. But what are anonymized or anonymous data? The following definition of anonymous data is based on the EU General Data Protection Regulation's 26th recital [13] and is used in this thesis.

Definition 2.1. *Anonymous data*
Any data D are considered anonymous if no statistical unit in D is identifiable, provided that:

(i) All the means reasonably likely to be used to directly or indirectly identify any unit, such as singling out, are considered.

(ii) To ascertain whether means are reasonably likely to be used to identify the statistical unit, account should be taken of all objective factors, such as the costs and the amount of time required for identification, taking into consideration the available technology at the time of processing and technological developments.

Anonymization of data can be performed by implementing SDC methods which minimize the risk of disclosure to an acceptable level while retaining as much information as possible, i.e., preserving data utility. Hundepool et al. [4] divide SDC methods into two categories: *perturbative* and *non-perturbative* methods. Perturbative methods falsify the data by purposely introducing an element of error, for instance by adding noise to the observations from a certain distribution. Non-perturbative methods reduce the amount of information released by suppression or generalization of data, e.g., the ages of those under 15 and those over 90 are rescaled within these limits. Note that although the methods are divided into the two groups, methods of both kind are perturbative in terms of information. The difference in designation is due to the fact that in one the perturbation is intentionally caused, while in the other it is mainly a consequence. SDC methods are highly dependent on the type of data and variables they are implemented on. As a result, a wide range of different methods have been developed to secure data. In addition, multiple methods can be implemented in one data set. The data set is usually a microdata set but methods for aggregated data exist as well.

2.2.1 Disclosure risk and data utility

According to Hundepool et al. [4], there exist different definitions of disclosure for different types of breaches. The following are only some of them. *Identity disclosure* occurs when a specific respondent can be recognised in a released data file. Identity disclosure is a risk especially for outliers, since they can be more easily identified based on their deviant values. In *attribute disclosure* sensitive information about a specific individual unit is revealed, this is especially a risk if no perturbation is applied. An example of an attribute disclosure could be that the released data file reveals that a subject has a rare disease. In case of identity disclosure, if confidential variables are present, it automatically leads to attribute disclosure.

When one can determine an attribute value or some characteristics of an individual unit more accurately than would have been possible before releasing the data, an *inferential disclosure* occurs. An example of inferential disclosure is if, based on the released data, one could infer that people living in a certain area have an increased risk of violent behaviour. This in turn could lead to discrimination against people living in the area. In this thesis, the focus is mainly on identity disclosure although also attribute disclosure is considered. Problems related to inferential disclosure are discussed in Section 7.

Hundepool et al. [4] point out that before applying any SDC methods, a disclosure scenario – or preferably multiple ones – should be formulated. This process includes mapping all quasi-identifiers and figuring out how they could be used in a breach. The disclosure scenario is usually based on the assumption that the adversary has access to other external data, which include identifiers and some other variables also included in the original data. These variables can then be used to match identifiers to records in the released data by finding overlapping values and patterns. Usually these variables are freely available demographics, such as gender, age and marital status.

After formulating the scenario, a maximum tolerated risk, i.e., the level of anonymity has to be decided. All SDC methods cause information loss of varying degree, so it is always a trade-off between disclosure risk and data utility. Therefore, there is no straightforward answer to which level or threshold should be used. The trade-off between the disclosure risk and utility can be illustrated by the risk-utility map presented in Figure 1. In addition, different anonymity measures protect against different breaches. Nevertheless, if the purpose of the data usage is known in advance, it usually helps to choose SDC methods and measures that preserve features relevant for that usage. [4]

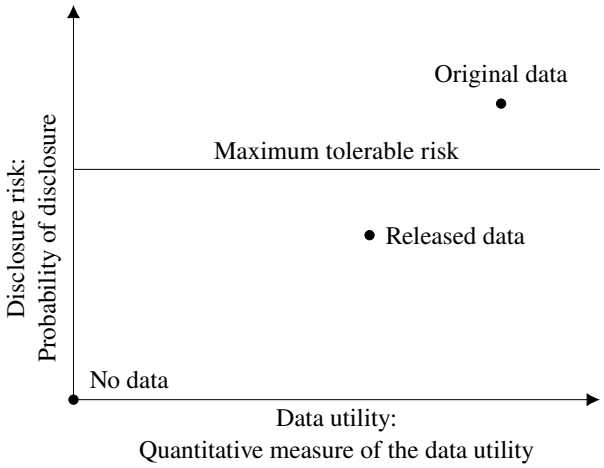


Figure 1: A risk-utility map illustrates the trade-off between data utility and disclosure risk. In the example, the released data is seen to achieve utility close to the original data while still staying under the threshold of maximum tolerable disclosure risk.

This thesis implements differential privacy, a relatively new measure of anonymity presented in Section 3. The main reason to choose this concept was that it is implemented in PrivBayes [7], which again is used to generate synthetic data in this thesis. However, differential privacy is one of the most frequently implemented privacy measures ever since its first introduction [14] and it is used by many major companies like Amazon, Apple and Google [15, 16, 17]. Utility measures, presented in Section 5, were selected on the basis of generalizability and their relevance for statistical inference considering longitudinal data.

2.3 Synthetic data generation

In 1993, Rubin introduced the idea of generating synthetic data via multiple imputation in order to overcome the shortcomings of SDC methods of that time [5]. He stated that synthetic data sets could be released for public use and that they not only preserve confidentiality but, moreover, preserve the ability to obtain valid statistical inferences. Synthetic data generation could solve privacy issues while at the same time allow wider and more free use of data. This thesis focuses on tabular data, more precisely on longitudinal tabular data, although synthetic data can also be generated for, e.g., images or texts and the same framework would apply to them.

If synthetic data are considered to be private, the generation of synthetic data can be thought of as a third SDC method in which new and anonymous data are created on the basis of existing data. The main difference between synthetic data generation and other SDC methods is that data synthesizing allows producing a larger number of observations than there were in the original data set. Since the current sizes of available data sets have increased, it is easier to generate synthetic data as it is easier to examine the processes that produce data. Although more data are available, they alone are not enough to guarantee the quality of synthetic data, as quality is affected by many other factors. However, the following two criteria can be considered as good starting points for ensuring the quality of synthetic data.

First, it must be ensured that the original data set itself is of high enough quality in order to draw valid statistical inferences. Second, the method by which the synthetic data set is produced must approximate as accurately as possible the probability distribution which generated the observed original data set. The first criterion can be ascertained, for example, by examining the sampling method and the theoretical representativeness of the data set. The assumption is that the empirical distribution of the original data set approximates the true underlying probability distribution. The second criterion is more complex and usually the most difficult part of generating synthetic data. In case of known multivariate distributions, like multivariate normal or multinomial distributions, the probability distribution can be presented in a closed form and is easier to estimate and use in synthesis. However, the assumptions of these distributions are rarely met in real-life data. Furthermore, the estimation of multivariate distributions becomes increasingly more difficult when the number of attributes in the original data set increases or if the attributes are correlated.

In general, data synthesization causes information loss although in some rare cases the data quality may improve. Synthetic data can be of better quality than the original ones, if the estimated distribution used in generation is actually closer to the true distribution than the empirical distribution of the original data. This can happen, for example, when noise is added to the synthetic distribution to achieve privacy and the added noise actually "corrects" the distribution. This, of course, can only be verified when more real-world data become available or if the generative model is somehow cross-validated. In other words, new data must be used to verify such improvements, as the reuse of the same data in distribution estimation may lead to under or overfitting, in which case the estimated distribution is unlikely to correspond to the actual distribution.

2.3.1 Definition of synthetic data

Based on a literature review and references cited in this thesis, no exhaustive and generally applied definition of synthetic data has been proposed. The most commonly used definition of synthetic data is that "synthetic data mimic original data". Some authors refine the definition by specifying how the synthetic data mimic the original data [5, 7, 18], but even then the level of detail varies. Some authors claim that any data generated from scratch qualify as synthetic data [19], while others point out that synthetic data can be either fully or partially synthetic [4]. In absence of a generally accepted definition of synthetic data, the following definition is proposed and used in this thesis. It is largely based on Rubin's [5] original presentation and a review of the use of synthetic data in the literature. The definition is also limited to tabular data as the original data $\boldsymbol{D}$ is assumed to be presented as a microdata set.

Definition 2.2. *Synthetic tabular data*
Let $\boldsymbol{D}$ be existing data, referred to as the original data, and assumed to be a realization of a random $n \times p$ matrix whose rows are independent random vectors $\boldsymbol{X}_i = [X_{i1}, X_{i2}, \ldots, X_{ip}]$ that each follow the same distribution F. Furthermore, the columns of the matrix are composed of the attributes desired for the synthetic data. Another data $\boldsymbol{D}^*$ is considered synthetic if the following two criteria are met:

(i) $\boldsymbol{D}^* = G'(\boldsymbol{D})$, where G' is a randomized algorithm that takes $\boldsymbol{D}$ as input, utilizes $\boldsymbol{D}$ in its operation, and outputs a random matrix of size $m \times p$, where m is an arbitrary number. The rows of $\boldsymbol{D}^*$ are independent and *anonymous* random vectors each following distribution F^*.

(ii) The two distributions satisfy $F^* \approx F$.

Criterion (i) of the definition rules out generating data from scratch without any background knowledge, since the algorithm used in the generation has to somehow utilize the original data that already exist. The prerequisite for utilizing the original data derives from the definition of the word *synthesis* – the composition or combination of parts or elements so as to form a whole [20] – and from the fact that this is usually how synthetic data have been created in the literature. Utilization of existing data is also intended to separate synthetic data from completely fictitious or simulated data and is expected to improve the realism and utility of synthetic data.

Criterion (i) also takes care of the deepest purpose of generating synthetic data, which is to avoid any restrictions on the use of data containing confidential information and thus enabling a more free and extensive use of data. This would not simply be the case if synthetic data were considered to contain confidential information, hence the synthetic data must be anonymous. This generally rules out resampling and permutation methods since they only reproduce the same observations and can thus lead to identity disclosure if similar combinations are present as in the original data. If these methods are to be used, and no perturbation methods are used in addition to those, special care must be taken to ensure that there are no similar observations as in the original data. Because if this permuted or re-sampled data were compared to any external data, a potential adversary could find out the correct personal information.

Criterion (ii) ensures that synthetic data are as realistic and usable as possible. However,

this is usually the most difficult part of generating synthetic data. In the simplest case, all attributes in the data set are independent from each other, therefore it is sufficient to estimate their univariate distributions and generate anonymous samples from each of them independently. However, even in this case finding good approximations to the underlying distributions can be difficult, let alone in case the attributes are not independent. The goodness of the approximation in criterion (ii) can be measured with different utility measures, some examples of which are discussed in Section 5.

Example 2.2.1.

Let the original data $\boldsymbol{D}$ be $n \times p$ random matrix and whose rows are random vectors $\boldsymbol{X}_i$ that follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, i.e., $F = N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Let G' be a randomized algorithm that takes $\boldsymbol{D}$ as input and outputs a matrix of random vectors $\boldsymbol{X}_i^*$ each of which follows distribution F^*. A good choice for F^* would be $N_p(\mathbf{0}, \boldsymbol{D}^T\boldsymbol{D}/n)$, where $\boldsymbol{D}^T$ denotes transpose of $\boldsymbol{D}$.

There are three aspects in Definition 2.2 that require special mention. First, the choice of the algorithm G' plays a key role, as it must produce both realistic and anonymous data. The algorithm must be a *randomized algorithm*, that is, an algorithm that exploits randomness as part of its logic, because the logic behind a deterministic algorithm can be solved and used to inverse the data back to the original. Computational efficiency as well as user-friendliness of G' are also features not to be forgotten.

Second, the desired number of attributes in the synthetic data set $\boldsymbol{D}^*$ is to be fixed before the synthesis but the number of generated records, i.e., random vectors is optional. The possibility to generate arbitrary number of records emphasizes the difference between synthetic and solely anonymized data since altering the values of the records in the original data set is generally not considered as generating synthetic data. Furthermore, if attributes are dropped from the original data before data synthesis, one must be aware of any dependencies among the attributes in the data and the potential consequences that the exclusion may have.

Finally, the definition is knowingly flexible to allow the use of different methods to both generate and evaluate synthesized data. The use of different methods is mandatory to estimate different joint distributions and to assess data quality. In a more general formulation of the definition, the original data $\boldsymbol{D}$ is assumed to be any kind of data, not necessarily tabular, and the distribution F is the distribution of the process that produced the data. Due to the flexibility, the generation of synthetic data requires careful design and development of methods.

2.3.2 Special features of longitudinal data

Weiss [8] defines *longitudinal data*, sometimes called panel data, as a special case of *repeated measures data*. In repeated measures data, measurements are collected repeatedly on each statistical unit and observations may or may not be distributed over one dimension. For example, the measurements may be spatially distributed in two or three dimensions. Repeated measures data may also be clustered or nested inside of larger units, e.g., households or schools. The defining feature of longitudinal data is that multiple observations within units are ordered across a single dimension that separates the measurements. The

dimension is usually time but alternative choices are equally valid as long as they separate the measurements and are linearly ordered.

According to Weiss [8], in longitudinal data, measurements are collected repeatedly over the dimension on each statistical unit and the units need not contribute the same number of measurements each. This type of data collection creates a temporal ordering of measurements, which is important because measurements closer in time within a unit are likely to be more similar than observations further apart in time. Time series is a special case of longitudinal data. In univariate time series, longitudinal data are observed over a long period of time on a single unit. Longitudinal data consist of many time series on a sample of statistical units and generally have fewer repeated measurements than traditional time series. Therefore, longitudinal data are usually multivariate, containing both numerical and categorical variables, and both univariate and multivariate analyses can be carried out.

Longitudinal data are widely encountered in health and medical research. They provide valuable information about change over time and allow many different and powerful statistical modelling methods, e.g., hierarchical and latent variable modelling. In addition, longitudinal data arise naturally from patient visits. Therefore, generating synthetic longitudinal patient data could rectify the current under-utilization of patient data without compromising patient privacy and safety.

Longitudinal data create challenges to generation of synthetic data, more specifically to criterion (ii) of Definition 2.2. As a consequence, generation of synthetic longitudinal data have not been studied as much as the generation of cross-sectional data, in which the dependence structure is generally assumed to be less complex. By definition, observations in longitudinal data are not independent and in order to preserve their temporal structure, special attention must be paid to the estimation of multivariate distributions. Furthermore, missing values are more common in longitudinal than in cross-sectional data, which can cause problems for synthetization as well as make the estimation of the joint distribution more difficult.

Some of the most commonly used methods to generate synthetic longitudinal data in the context of health data are Bayesian networks [18], hidden Markov models (HMMs) [21], neural networks, such as generative adversarial networks (GANs) [22] or variational autoencoders (VAEs) [23], as well as more complex simulation models like Synthea [24]. All of the above methods allow building hierarchical models, which have proven to be especially convenient if the data are correlated as in longitudinal data settings. A disadvantage of these methods is that they are all computationally heavy and require considerable expertise from the data processor.

3 Differential privacy

Several anonymity measures have been developed, each to fix the limitations of the previous ones [19, 25]. However, all these methods are based on identifying quasi-identifiers and their equivalence classes. This can sometimes be laborious, and the existing methods do not protect from every possible breach scenario [26]. In 2006, Cynthia Dwork postulated that "the risk to one's privacy, or in general, any type of risk, should not substantially increase as a result of participating in a statistical database" [14]. This is the foundation of *differential privacy*, a probability-based anonymization method with a mathematically proven privacy guarantee.

3.1 Definition of differential privacy

Various formulations, interpretations and extensions of differential privacy have been proposed in the literature. In this thesis, the definition follows closely the original definition introduced by Dwork et al. [27], in which the difference between data sets is defined through the *Hamming distance of the records*, as illustrated in Figure 2.

Definition 3.1. *Hamming distance of the records*
Let $\boldsymbol{D}_1$ be a data set consisting of records $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ that are independent row vectors of equal length of p. Let $\boldsymbol{D}_2$ be a similar data set but denote its row vectors by $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$, respectively. The Hamming distance between $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, denoted by $d_{\mathcal{H}}(\boldsymbol{D}_1, \boldsymbol{D}_2)$, is the number of indices where the records $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ differ. That is, $d_{\mathcal{H}}(\boldsymbol{D}_1, \boldsymbol{D}_2) = \sum_{i=1}^{n} \delta(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, where

$$\delta(\boldsymbol{X}_i, \boldsymbol{Y}_i) = \begin{cases} 0, & \boldsymbol{X}_i = \boldsymbol{Y}_i \\ 1, & \boldsymbol{X}_i \neq \boldsymbol{Y}_i. \end{cases}$$

Since $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ are already vectors, their distance is defined in a similar manner but by components, that is, $\boldsymbol{X}_i \neq \boldsymbol{Y}_i$ if and only if $X_{ij} \neq Y_{ij}$ for any $j = 1, \ldots, p$.

	A				**B**				**C**			
$i = 1$	0	1	1	0	0	1	1	0	0	1	1	0
$i = 2$	1	0	1	0	1	0	0	0	1	0	1	0
$i = 3$	0	1	0	0	0	1	0	0	0	0	1	0
$i = 4$	0	0	1	0	0	0	1	0	0	0	1	0

Figure 2: The Hamming distance of the records between data sets **A** and **B** is one because their records differ in only one pair of records when $i = 2$. Similarly $d_{\mathcal{H}}(\boldsymbol{A}, \boldsymbol{C}) = 1$ meaning that both **A** and **B** as well as **A** and **C** are neighboring data sets. Note the difference in Hamming distance between vectors and records: $d_{\mathcal{H}}(\boldsymbol{A}, \boldsymbol{C}) = 1$ but $d_{\mathcal{H}}(\boldsymbol{X}_3, \boldsymbol{Y}_3) = 2$, where $\boldsymbol{X}_3$ and $\boldsymbol{Y}_3$ are the third records of **A** and **C**, respectively. In addition, $d_{\mathcal{H}}(\boldsymbol{B}, \boldsymbol{C}) = 2$ as they differ by two records.

Based on Definition 3.1, the two data sets $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are said to be neighbors if and only if their Hamming distance of records satisfies $d_{\mathcal{H}}(\boldsymbol{D}_1, \boldsymbol{D}_2) = 1$. The definition of

neighboring data sets greatly influences the definition and implementation of differential privacy. As a consequence, the comparison of different data sets protected by differential privacy is not straightforward, since depending on the definition used for neighboring data sets, the methods may not be comparable. Premised on the above definition of neighboring data sets, differential privacy can be defined as follows:

Definition 3.2. *ε-differential privacy*
A randomized algorithm G satisfies ε-differential privacy if for any two neighboring data sets $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, and for any possible output O of G, we have

$$\mathbb{P}[G(\boldsymbol{D}_1) = O] \leq e^{\varepsilon} \cdot \mathbb{P}[G(\boldsymbol{D}_2) = O], \tag{1}$$

where $\mathbb{P}[\cdot]$ denotes the probability of an event and ε is called the *privacy budget*.

Although Definition 3.2 does not seem to be symmetric with respect to data sets $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, their positions can always be changed. Let $\mathbb{P}_1 = \mathbb{P}[G(\boldsymbol{D}_1) = O]$ and $\mathbb{P}_2 = \mathbb{P}[G(\boldsymbol{D}_2) = O]$. By changing the positions of the data sets, we get relations $\mathbb{P}_1 \leq e^{\varepsilon} \cdot \mathbb{P}_2$ and $\mathbb{P}_2 \leq e^{\varepsilon} \cdot \mathbb{P}_1$. By combining these inequalities we get

$$-\varepsilon \leq ln\left(\frac{\mathbb{P}_1}{\mathbb{P}_2}\right) \leq \varepsilon,$$

which illustrates how the probabilities $\mathbb{P}_1$ and $\mathbb{P}_2$ are desired to be close to each other.

The idea of differential privacy is that a hypothetical adversary can no longer draw disclosure-related conclusions based on differentially private data because the probability of observing any output O differs at most by factor e^{ε} regardless of whether any record is present or not. In other words, differential privacy mimics a situation in which conclusions drawn from a study would not significantly change even if any single participant would not have participated in the first place. As a result, the adversary cannot be certain of the presence of an individual – privacy stems from this uncertainty.

If $\varepsilon = 0$, the probabilities in (1) are equal, which means that the output O produced by the randomized algorithm G is independent of the input: no matter how we replace any record in the data set, the probability stays the same, meaning that the probability does not depend on the input. In practice this means that the randomized algorithm produces random noise, which is totally private but not useful. Increasing ε increases the potential maximum difference between the probabilities assigned to output O, meaning that the record is allowed to have more effect on the probability of the output – resulting in less randomness but at the same time making the record more identifiable.

One of the advantages of differential privacy is its *compositional property*, which results from the multiplicative formulation of (1). When multiple outputs O_i are published in a differentially private manner with privacy budgets ε_i and they are compared against each other, then based on the compositional property, the level of privacy is $\varepsilon = \sum \varepsilon_i$. In other words, the compositional property allows to determine an upper limit for the privacy budget. This is not the case, for example, for *k-anonymous* data sets, where anonymity is based on the fact that the information of each person in the data set cannot be distinguished from at least $k - 1$ other individuals whose information also appear in the data set. The compositional property is extremely convenient, since in order to apply differential privacy, it is not necessary to make any assumptions of the background

knowledge a hypothetical adversary may have – only the number of outputs to be released is needed to adjust the level of privacy. Differential privacy can also be applied without the need to map quasi-identifiers and it protects against all but inferential disclosure.

Example 3.2.1.

Suppose two researchers, John and Mary, both study independently how parental income is related to the distribution of children's health care visits in the private sector in Finland in 2018. In their respective studies, John uses the province declared for tax purposes while Mary uses the province where the family lived at the end of the year. John publishes his study in early 2020 and as part of the results states that "In families earning more than 1,000,000 euros a year in Southwest Finland ($n = 256$), children visited the private sector on average 14.5 times during 2018". Later that year, Mary publishes her own research and reports "In 2018, children from families earning more than 1,000,000 euros a year in Southwest Finland ($n = 255$) visited the private sector for care on average 14.2 times."

The results presented above are a very traditional and accepted way of reporting research results. In this situation, however, by combining the results of the above studies, it can be seen that one family either moved or otherwise disappeared from the sample of Mary's study. Calculation based on these reported results reveals that children in this family have been treated 91 times in the private sector in 2018. In Finland, some tax information is public, in addition to which address information can also be provided if it has not been made private. In this case, it is possible to find out which family is involved by using external sources and thus sensitive information has become public, even though it was not originally intended.

If the data on family visits used in the studies had been protected by an appropriate differential privacy method, no such reasoning could have been made. In addition, it would also have been possible to maintain the sensibleness of the study by ensuring that the probability of observed results does not change radically – up to a maximum of e^{ε} – as a result of the protection. Moreover, even if the hypothetical adversary had imagined that he had succeeded in revealing some sensitive information about the results, the conclusions would still have included uncertainty due to the differential privacy. Note that the figures used in the example are purely imaginary.

3.2 Methods to achieve differential privacy

Definition 3.2 requires that the algorithm G must be a randomized algorithm because the probability is taken over the randomness used by the algorithm. The algorithm G can be made ε-differentially private by different methods, but the two most commonly implemented ones that have been proven to achieve differential privacy are the *Laplace mechanism* [27] and the *exponential mechanism* [28]. The reason for using two different methods is that the Laplace mechanism works only for numerical outputs and the exponential mechanism is used for categorical outputs. Both mechanisms are applied as part of PrivBayes addressed in Section 4.

In the case of numerical outputs of G, the Laplace mechanism converts G into an

ε-differentially private algorithm by adding noise $\eta \overset{iid}{\sim} \text{Lap}(\mu, \lambda)$ into each output O. The location parameter μ is set to zero and the scale parameter $\lambda \geq S(G)/\varepsilon$. The quantity $S(G)$ is the *sensitivity* of G and measures the maximum possible change in G's outputs when one record is altered.

Definition 3.3. *Sensitivity*
The sensitivity of a function F that maps an input data set into a fixed size vector of real numbers is

$$S(F) = \max_{\boldsymbol{D}_1, \boldsymbol{D}_2} \|F(\boldsymbol{D}_1) - F(\boldsymbol{D}_2)\|_1, \tag{2}$$

where $\|\cdot\|_1$ denotes the L^1-norm and $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are any two neighboring data sets.

The exponential mechanism is used for categorical outputs of G. It releases differentially private version of G by sampling outputs ω from G's output sample space Ω with a probability proportional to $\exp(f_s(\boldsymbol{D}, \omega)/(2\Delta))$. The function f_s is a user-specific *score function* which takes any data set $\boldsymbol{D}$ and any element $\omega \in \Omega$ as input and for which a higher score indicates that ω is a more compatible output with respect to $\boldsymbol{D}$. The *scaling factor* $\Delta \geq S(f_s)/\varepsilon$ controls the degree of privacy and ensures differential privacy. For any two neighboring data sets $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ and any element $\omega' \in \Omega$, the function

$$S(f_s) = \max_{\boldsymbol{D}_1, \boldsymbol{D}_2, \omega'} |f_s(\boldsymbol{D}_1, \omega') - f_s(\boldsymbol{D}_2, \omega')|,$$

is referred to as the sensitivity of f_s since it is inherently of the same form as (2).

Both λ and Δ play key roles in achieving ε-differential privacy. In order to achieve differential privacy, λ must be at least $S(G)/\varepsilon$, i.e., a certain amount of noise must be accepted. Increasing λ increases the amount of noise, making the data more private but reducing the utility of the data. The same interpretation applies for Δ: increasing Δ reduces the effect of f_s so that with a sufficiently large Δ all outputs are approximately equally likely regardless of their score.

Since differential privacy is a feature of an algorithm rather than data, special attention must be paid to the design of the algorithm to avoid obscuring the signal in the data with excessive noise. For example, if the algorithm returns the average of its inputs and outliers are present, a considerable amount of noise should be added to achieve differential privacy. In addition, computational efficiency should also be considered so that the mechanism would be realistically applicable.

3.3 Limitations of differential privacy

Choosing the level of ε is not trivial. In general, it should be as small as possible to achieve sufficient level of privacy, but smaller values also mean more randomness. Dwork et al. [29] recommend that ε should not be larger than one, but in a more recent study [30], after interviewing different practitioners, the authors found no clear consensus how to approach the selection of ε. In this thesis, several different values of ε are applied and the effect of the parameter value on utility measurements is evaluated.

Applying differential privacy into small sample sizes may hide relevant information depending on the selected ε. Differential privacy ensures that the conclusions of the

research should be roughly equally likely to be reached with or without the contribution to data of any one individual. The problem with small sample sizes is that the change of one record can significantly affect the estimates of interest. This effect is hidden under differential privacy. The same applies to detecting and analyzing outliers since differential privacy generally hides their presence or absence.

Differential privacy is mainly applied via the Laplace or the exponential mechanism which are both dependent on the sensitivity of the used function. Selection of a highly sensitive function would cause too much noise and decrease the data utility. Thus, the variety of functions from which to select is limited to moderate or low sensitivity functions if a reasonable level of data utility is aimed at. Furthermore, finding such a function can be extremely difficult and requires skills from the data processor.

Even if the released data are protected by differential privacy, the data are not protected from inferential disclosure. For example, if the released data suggest that patients with cholesterol levels above a certain level and body mass index greater than 30 are more likely to die over a period of time, the information can be used to make decisions about individuals who meet these criteria, e.g., by raising the cost of health insurance. However, one could argue that the point of releasing data is to allow others to make valid inferences based on them and the removal of the inferential property would make data useless.

4 PrivBayes: method for generating synthetic longitudinal data

Differential privacy is increasingly applied in *privacy-preserving data publishing* (PPDP) methods. However, applying differential privacy to high-dimensional data is challenging. Many methods require a considerable amount of noise to be added to the data to achieve privacy, which in turn reduces data utility. To address this particular challenge, in 2014, Zhang et al. [31] presented the preliminary version of *PrivBayes*, a method which operates on low-dimensional marginal distributions instead of a high-dimensional joint distribution. In 2017, Zhang et al. [7] published a new, improved version of PrivBayes, which is applied in this thesis.

The operations of PrivBayes can be roughly divided into three phases:

1. *Network learning*: Construct a k-degree Bayesian network $\mathcal{N}$ over the attributes $X_1, X_2 \ldots, X_p$, in the data set $\boldsymbol{D}$, using an ε_1-differentially private method. The network provides a succinct model of the associations among the attributes.
2. *Distribution learning*: Use an ε_2-differentially private algorithm to generate a set of conditional distributions of $\boldsymbol{D}$, such that for each attribute-parent pair (X_j, Π_j) in $\mathcal{N}$, we have a noisy version, $\mathbb{P}^*\left[X_j|\Pi_j\right]$, of the conditional distribution.
3. *Data synthesis*: Use the Bayesian network $\mathcal{N}$ and the p noisy conditional distributions to derive an approximate distribution of the records in $\boldsymbol{D}$ and then sample records from the approximate distribution to generate a synthetic data set $\boldsymbol{D}^*$.

In summary, based on the compositional property of differential privacy, PrivBayes is an $(\varepsilon_1 + \varepsilon_2)$-private method constructing a Bayesian network $\mathcal{N}$ which is used to generate a private synthetic data set $\boldsymbol{D}^*$ approximating the original data set $\boldsymbol{D}$. Zhang et al. [7] point out that the formation of $\mathcal{N}$ requires careful selection of attribute-parent pairs and the value of k to obtain a close approximation of the original data set without violating differential privacy.

The reason why PrivBayes was selected as the method in this thesis is its way of modelling associations between variables using a hierarchical dependence structure. This is hypothesized to play a central role in modelling correlation structures in longitudinal data. In addition, Bayesian networks have been used in the past to generate synthetic time series [18] as well as correlated data [32]. PrivBayes is also a generic method that does not require prior knowledge of the workload and should therefore be applicable in a wide variety of analyses [7].

This chapter is limited to presenting the theory of the 2017 improved PrivBayes insofar as it is implemented in Section 6. The chapter is divided into three parts: Subsection 4.1 addresses the hierarchical encoding used for general domains, Subsection 4.2 focuses on the first phase of PrivBayes and Subsection 4.3 focuses on phases two and three.

4.1 Hierarchical encoding

The preliminary version of the PrivBayes method required that each attribute in the input data had to be transformed into a set of binary attributes. The authors stated that the transformation destroyed the semantics of natural attributes and degraded the utility of the output data [7]. Therefore, in 2017, Zhang et al. [7] introduced two encodings – *hierarchical* and *vanilla encoding* – and further developed the preliminary method to match with these encodings.

In hierarchical encoding, the original domain of each attribute is generalized using a taxonomy tree in order to reduce its *domain size*, that is, the size of the empirical sample space denoted by $|\mathrm{dom}(\cdot)|$ for a generic input. Each continuous attribute is divided into b bins from which a $\lceil log_2\, b \rceil$ high taxonomy tree is formed with the exception that the root is not taken into account in determining the height because it would generate only one class. For each categorical attribute, the taxonomy tree is built based on domain knowledge which can be, for example, some natural existing hierarchy. Vanilla encoding can be seen as a special case of hierarchical encoding, where each taxonomy tree consists of leaf vertices only, that is, the attribute is presented directly in as many classes as it has possible values. Examples of hierarchical encoding for continuous and categorical attributes are presented in Figures 3 and 4, respectively.

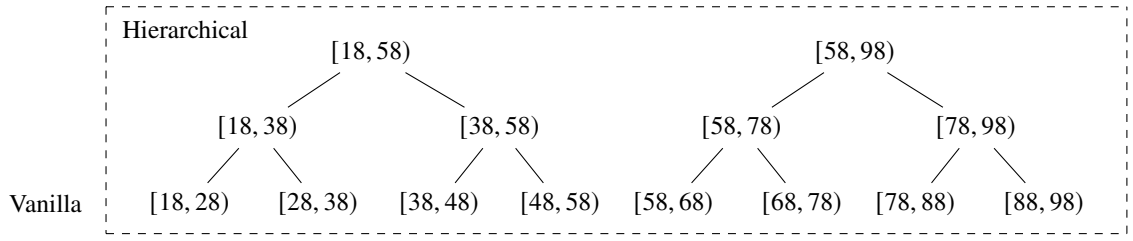


Figure 3: An example of hierarchical encoding of a continuous variable, age, using eight bins. In this example, for vanilla encoding, the age is initially assumed to be eight-class.

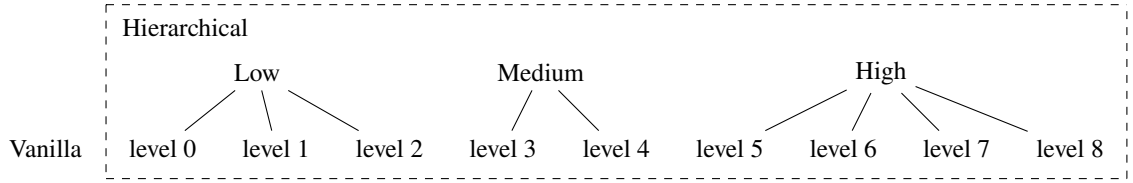


Figure 4: Hierarchical encoding of a categorical variable, International Standard Classification of Education (ISCED) 2011, based on a hierarchy defined by Eurostat [33]. Note that in the figure, levels $0, 1, \ldots, 8$, refer to the classification labels, not hierarchy levels.

Hierarchically encoded attributes, denoted by $X^{(i)}$, are also called *generalized attributes* and their domain size is defined by the number of nodes at level $i \in [0, \mathrm{height}(X))$, with leaf vertices at level 0. The larger i is, the more generalized the attribute is. Similarly, a subset of generalized attributes is called a *generalized subset*. The purpose of generalizing attributes is to reduce the dimension of the distribution so that the noise added in the distribution learning phase does not obscure the information, but at the same time the aim is to preserve the semantics of the attribute. Zhang et al. [7] recommended using hierarchical encoding as it provides more flexible encoding than vanilla encoding while preserving the semantics of the attributes. In addition, hierarchical encoding also performed best

in their encoding comparisons. Based on their recommendation and comparison results, hierarchical encoding is applied in this thesis and all presented algorithms are based on this method. Thus, in the algorithms presented in this thesis, $X = X^{(0)}$.

4.2 Private Bayesian networks

A *Bayesian network*, denoted by $\mathcal{N}$, is a probabilistic model that presents conditional independence between attributes using a *directed acyclic graph* (DAG). In a directed acyclic graph, each edge has a direction and there exists no cycles, i.e., non-empty directed paths from a node to itself. More specifically, the network illustrates three kinds of association between the attributes: *direct dependence*, *weak conditional independence* and *strong conditional independence*. Figure 5 illustrates a Bayesian network $\mathcal{N}_1$ over a set of five attributes: *age*, *education*, *hypertension*, *lifestyle* and *cardiovascular disease* (CVD).

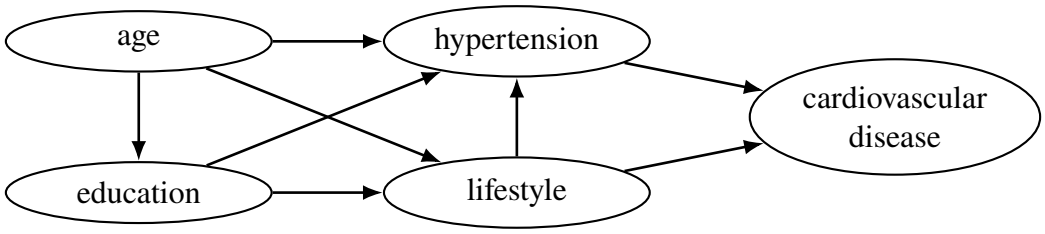


Figure 5: A Bayesian network $\mathcal{N}_1$ over five attributes. The network illustrates the conditional dependencies between the attributes.

In direct dependence, there is an edge between two nodes X_j and X_k, say, from X_j to X_k. The node X_j is called the *parent* of the node X_k and the set of all attributes that have an edge to X_k is called the *parent set* of X_k, denoted by Π_k. Together they form an *attribute-parent (AP) pair* (X_k, Π_k). For example, in Figure 5 there is an edge from hypertension to CVD which means that hypertension is a parent of CVD. However, there is also an edge from lifestyle to CVD, so the parent set of CVD is {hypertension, lifestyle}.

When there is a path from X_j to X_k but no direct edge, then X_j and X_k are conditionally independent given X_k's parent set Π_k. In Figure 5, there is a path from age to CVD but no direct edge, meaning that given the subject's lifestyle and occurrence of hypertension, her age and CVD are independent. This situation is called weak conditional independence. If there is no path between X_j and X_k, then X_j and X_k are conditionally independent given X_j's or X_k's parent sets and the situation is called strong conditional independence.

Definition 4.1. *Bayesian network*
Let $\mathcal{A}$ denote a set of p attributes. Formally, a Bayesian network over $\mathcal{A}$ is defined as a set of p AP pairs, $(X_1, \Pi_1), \ldots, \big(X_p, \Pi_p\big)$, such that

1. Each X_j is a unique attribute in $\mathcal{A}$;
2. Each Π_j is a subset of the attributes in $\mathcal{A} \setminus \{X_j\}$;
3. For any $1 \leq j < k \leq p$, we have $X_k \notin \Pi_j$, that is, there is no edge from X_k to X_j in $\mathcal{N}$.

The assumption that any X_j and any $X_k \notin \Pi_j$ are conditionally independent given Π_j allows the use of the chain rule to calculate the joint distribution:

$$\begin{aligned} F &= \mathbb{P}\big[X_1, X_2, \ldots, X_p\big] \\ &= \mathbb{P}[X_1] \cdot \mathbb{P}[X_2|X_1] \cdot \mathbb{P}[X_3|X_1, X_2] \cdot \ldots \cdot \mathbb{P}\big[X_p|X_1, \ldots, X_{p-1}\big] \\ &= \mathbb{P}[X_1|\Pi_1] \cdot \mathbb{P}[X_2|\Pi_2] \cdot \mathbb{P}[X_3|\Pi_3] \cdot \ldots \cdot \mathbb{P}\big[X_p|\Pi_p\big] \\ &= \prod_{j=1}^{p} \mathbb{P}\big[X_j|\Pi_j\big]. \end{aligned}$$

In addition, let $F_{\mathcal{N}}$ be the approximation of F defined by the estimated Bayesian network $\mathcal{N}$. Naturally, if the network captures precisely the conditional independence among the attributes in $\mathcal{A}$, then $F_{\mathcal{N}}$ would be a good approximation of the original distribution. This is the core of the PrivBayes method because to compute the joint distribution, it is enough to form only p low-dimensional conditional distributions $\mathbb{P}\big[X_j|\Pi_j\big]$.

The *degree of the network* $\mathcal{N}$, denoted by k, is the maximum size of any parent set Π_j in $\mathcal{N}$. The degree of $\mathcal{N}_1$ presented in Figure 5 is three, since the maximum size of any parent set in the network is three. All AP pairs in $\mathcal{N}_1$ are presented in Table 1. Notice how the third requirement of Definition 4.1 defines the order of the attributes.

Table 1: All attribute-parent (AP) pairs in $\mathcal{N}_1$ presented in Figure 5.

j	X_j	Π_j
1	age	$\emptyset$
2	education	{age}
3	lifestyle	{age, education}
4	hypertension	{age, education, lifestyle}
5	cardiovascular disease	{hypertension, lifestyle}

The concept of a Bayesian network may seem simple but the formation of the network in practice is highly non-trivial. First, what degree should be used? Second, how to select the parents of each attribute? Third, the construction of $\mathcal{N}$ can be modelled as an optimization problem, where the goal is to choose a parent set Π_j for each attribute $X_j \in \mathcal{A}$ to maximize a given score function, and this optimization problem has actually been proven to be NP-hard when $k > 1$. Finally, how to implement differential privacy to achieve private Bayesian networks? In PrivBayes, the solution is to use a greedy algorithm that selects a potential AP pair from among the candidates using the exponential mechanism with a carefully designed score function. These components are discussed in the following subsections but more specific details and proofs can be found in the 2017 article [7].

4.2.1 GreedyBayes

The most commonly used optimization algorithms for NP-hard optimization problems, such as hill-climbing or genetic algorithms, are not well suited to situations where differential privacy is applied. Such methods often lead to too noisy results, and as a consequence PrivBayes utilizes a greedy algorithm in the network learning phase. In PrivBayes, this algorithm is called *GreedyBayes*, described in Algorithm 1. In order to implement differential privacy, the privacy budget ε has to be fixed in advance. The privacy budget is used

in the PrivBayes method in two parts at different phases so, based on the compositional property, $\varepsilon = \varepsilon_1 + \varepsilon_2$. To recast these terms, Zhang et al. [7] proposed a new parameter $\beta \in (0, 1)$ to balance the quality of network learning and distribution learning phases by assigning $\varepsilon_1 = \beta\varepsilon$ and $\varepsilon_2 = (1 - \beta)\varepsilon$, respectively. In this thesis, several different values of β are considered.

Algorithm 1: GreedyBayes

Input: $\boldsymbol{D}, \theta, \varepsilon_1, \varepsilon_2$
Output: $\mathcal{N}$

1 $\mathcal{N} \leftarrow \emptyset$
2 $V \leftarrow \emptyset$
3 randomly select an attribute X_1 from $\mathcal{A}$
4 add $(X_1, \emptyset)$ to $\mathcal{N}$
5 add X_1 to V
6 **for** $j = 2$ **to** p **do**
7 $\Omega \leftarrow \emptyset$
8 **foreach** $X \in \mathcal{A} \setminus V$ **do**
9 $\mathrm{T}(X) \leftarrow \mathrm{MaximalParentSets}\left(V, \frac{n\varepsilon_2}{2p\theta|dom(X)|}\right)$
10 **if** $\mathrm{T}(X) = \emptyset$ **then**
11 add $(X, \emptyset)$ to Ω
12 **else**
13 **foreach** $\Pi \in \mathrm{T}(X)$ **do**
14 add (X, Π) to Ω
15 select (X_j, Π_j) from Ω using the exponential mechanism with a privacy budget of $\varepsilon_1/(p-1)$
16 add (X_j, Π_j) to $\mathcal{N}$
17 add X_j to V
18 **return** $\mathcal{N}$

In the PrivBayes method, AP pairs are selected from the output sample space Ω, generated by GreedyBayes, using the exponential mechanism (line 15 of Algorithm 1). As discussed earlier in Section 3, the use of the exponential mechanism requires a user-specific score function. Thus, the goal is to formulate a score function whose optimization leads to good approximation of the full joint distribution, i.e., $F_{\mathcal{N}} \approx F$, since these AP pairs are used in the second phase of PrivBayes to materialize $F_{\mathcal{N}}$. Zhang et al. [7] proposed the following score function to be used in the exponential mechanism:

$$R(X, \Pi) = \frac{1}{2}\left\|\mathbb{P}[X, \Pi] - \overline{\mathbb{P}}[X, \Pi]\right\|_1,$$

where $\mathbb{P}[X, \Pi]$ is the joint distribution $\mathbb{P}[X = x, \Pi = \pi]$ and $\overline{\mathbb{P}}[X, \Pi]$ is the product of the marginal distributions, i.e., $\mathbb{P}[X = x]\mathbb{P}[\Pi = \pi]$. Example 4.1.1 illustrates how $R(X, \Pi)$ can be calculated in the case where an AP pair consists of two discrete random variables. Since the purpose of the Bayesian network is to model the dependencies between the attributes in the data set, R measures the dependence between X and Π. For example, if X

and Π are independent of each other, such Π should not be chosen as the parent set of X as $R(X, \Pi) = 0$. Thus, a higher value of R indicates better compatibility between X and Π.

Zhang et al. [7] also proved that the sensitivity based on the above score function is $S(R) \leq 2/n^2 + 3/n$, where n is the number of records. In addition, GreedyBayes randomly selects X_1 and sets an empty set as its parent set, in which case only $(p-1)$ attributes with their parent sets are selected from Ω. As a consequence, the privacy budget in the exponential mechanism is $\varepsilon_1/(p-1)$. In order to achieve differential privacy, Δ in the exponential mechanism must be at least $S(R)/\varepsilon$. When Δ is set to its lower bound by using the upper bound of $S(R)$, the exponential mechanism samples each AP pair with a probability proportional to

$$\exp\left(\frac{\varepsilon_1 R(X, \Pi)}{2\left(\frac{2}{n^2} + \frac{3}{n}\right)(p-1)}\right).$$

Example 4.1.1.

Let X be a discrete random variable for which $x \in \{0, 1, 2, 3\}$ and let Π be a parent set of X that consists of only one other attribute with sample space $\pi \in \{0, 1, 2\}$. The following table presents the joint distribution $\mathbb{P}[X = x, \Pi = \pi]$ and the one-dimensional marginal distributions $\mathbb{P}[X = x]$ and $\mathbb{P}[\Pi = \pi]$, respectively.

$\mathbb{P}[X, \Pi]$	$x = 0$	$x = 1$	$x = 2$	$x = 3$	$\mathbb{P}[\Pi]$
$\pi = 0$	0.050	0.050	0.075	0.025	0.20
$\pi = 1$	0.10	0.15	0.025	0.025	0.30
$\pi = 2$	0.15	0.20	0.10	0.050	0.50
$\mathbb{P}[X]$	0.30	0.40	0.20	0.10	1.0

The score of the given AP pair is

$$R(X, \Pi) = \frac{1}{2}\Big[|(0.050 - 0.30 \cdot 0.20)| + \cdots + |(0.050 - 0.10 \cdot 0.50)|\Big].$$

4.2.2 θ-usefulness and MaximalParentSets

Choosing the degree k of the network in advance is not trivial, because while a larger k preserves more information about F by creating higher dimensional marginal distributions, a considerable amount of noise would have to be added to these distributions in the second phase of PrivBayes in order to achieve privacy, especially if the privacy budget is small. To address the trade-off between utility and privacy, Zhang et al. [7] introduced a new concept called *θ-usefulness*, which balances the informativeness of the Bayesian network and the robustness of distributions of AP pairs. Different values of θ are applied in this thesis and the effect of the parameter value on utility measurements is evaluated. For simplicity, the parameters θ, ε and β are later referred to as *hyperparameters* of the PrivBayes method.

Definition 4.2. *θ-usefulness*
A noisy distribution is θ-useful if the ratio of average scale of information to average scale of noise is at least θ.

As an example, a 3-useful noisy distribution is preferred over a 1.5-useful distribution, because its information to noise ratio is at least double that of the latter. In practice, a threshold for θ is set and the largest positive degree $\hbar$ of the network that guarantees θ-usefulness in distribution learning is chosen. If such a $\hbar$ does not exist, then $\hbar$ is set to zero, in which case no attribute has parents. In the case of non-binary attributes, a single $\hbar$ is not sufficient to guarantee θ-usefulness since the number of unique observed values of the different attributes may vary. As a result, each parent set can be of different size and thus another approach is required to determine the degree of the network. This approach will be discussed later in this section after the introduction of a maximal parent set and MaximalParentSets algorithm.

To form a Bayesian network, a parent set must be selected for each attribute from all possible combinations. However, any attribute cannot be selected as a parent, otherwise the network will not approximate the original data distribution well enough. As a solution, Zhang et al. [7] presented the concept of a *maximal parent set*.

Definition 4.3. *Maximal parent set*
For each attribute $X \in \mathcal{A} \setminus V$, its maximal parent set Π is a generalized subset of V that satisfies:

(i) $\mathbb{P}[X, \Pi]$ is θ-useful;

(ii) Π is maximal, that is, there is no generalized subset Π' of V such that $\mathbb{P}[X, \Pi']$ is θ-useful and Π' contains any extra attribute not in Π, or any shared attribute but with lower generalization level.

The motivation of criterion (i) in Definition 4.3 is to ensure that information in $\mathbb{P}[X, \Pi]$ will not be overpowered by the noise introduced in the distribution-learning phase. Zhang et al. [7] rationalized the requirement of maximality of Π with the monotonicity of the *mutual information*, that is, given two sets Π and Π' such that $\Pi' \subseteq \Pi$, the mutual information $I(X, \Pi') \leq I(X, \Pi)$ for any attribute X. In other words, we always choose the largest possible Π on the condition that the formed marginal distribution is θ-useful. The mutual information is defined as follows:

$$I(X, \Pi) = \sum_{x \in \mathrm{dom}(X)} \sum_{\pi \in \mathrm{dom}(\Pi)} \mathbb{P}[X = x, \Pi = \pi] \log \frac{\mathbb{P}[X = x, \Pi = \pi]}{\mathbb{P}[X = x]\mathbb{P}[\Pi = \pi]}.$$

Algorithm 2 describes a recursive algorithm called *MaximalParentSets*, used to find all suitable parent sets by recursively building two sets of maximal subsets, with and without a particular attribute $X \in V$, respectively, and then merging them to obtain the final result. Note that the MaximalParentSets algorithm is called within the GreedyBayes algorithm so it inherits its input parameters from there. The notation $\{\emptyset\}$ means a singleton set containing only an empty set.

Algorithm 2: MaximalParentSets

Input: V, τ
Output: $\mathcal{S}$

1 **if** $\tau < 1$ **then**
2 **return** $\emptyset$
3 **if** $V = \emptyset$ **then**
4 **return** $\{\emptyset\}$
5 pick an arbitrary attribute X from V
6 $\mathcal{S} \leftarrow \emptyset$
7 $\mathcal{U} \leftarrow \emptyset$
8 **for** $i = 0$ **to** height$(X) - 1$ **do**
9 **foreach** $Z \in$ MaximalParentSets$\left(V \setminus \{X\}, \frac{\tau}{|\mathrm{dom}(X^{(i)})|}\right)$ **do**
10 **if** $Z \in \mathcal{U}$ **then**
11 **continue**
12 add Z to $\mathcal{U}$
13 add $Z \cup \{X^{(i)}\}$ to $\mathcal{S}$
14 **foreach** $Z \in$ MaximalParentSets$(V \setminus \{X\}, \tau)$ **do**
15 **if** $Z \in \mathcal{U}$ **then**
16 **continue**
17 add Z to $\mathcal{S}$
18 **return** $\mathcal{S}$

In MaximalParentSets, V denotes the set of attributes from which parents can be chosen for any $X \in \mathcal{A} \setminus V$, defined earlier in GreedyBayes. Given an AP-pair (X, Π) with m cells in $\mathbb{P}[X, \Pi]$, the average scale of information in each cell is $1/m$ and the sensitivity of each one-dimensional marginal distribution $\mathbb{P}[X]$ is $2/n$. As a result, the average scale of noise added in the Laplace mechanism is $2p/(n\varepsilon_2)$. Consequently, $\mathbb{P}[X, \Pi]$ is θ-useful only if $m \leq n\varepsilon_2/(2p\theta)$, meaning that given an attribute X, it is sufficient to consider only those subsets of V as parents whose domain size is no greater than $\tau = n\varepsilon_2/(2p\theta|\mathrm{dom}(X)|)$. Therefore, the degree $\mathscr{k}$ of the network ultimately depends on how many attributes fulfill the condition $\frac{\tau}{|\mathrm{dom}(X^{(i)})|} \geq 1$, where $X^{(i)}$ is the generalized attribute on line 9 of MaximalParentSets algorithm, and which of the AP pairs are selected from Ω and in which order.

4.3 Noisy conditional distributions

The central idea of PrivBayes is to produce synthetic data by utilizing low-dimensional marginal distributions. This approach allows use of differential privacy without unduly compromising data utility. Moreover, in practice, the distributions are based on empirical data and are either discrete or are discretized, as explained in Subsection 4.1, in order to reduce the dimensionality and thus the amount of noise to be added. Algorithm 3 presents the pseudocode of *NoisyConditionals* that is used to materialize the distributions in PrivBayes.

NoisyConditionals forms conditional distributions from the noisy low-dimensional marginal distributions. The proper formulation of $\mathcal{N}$ ensures that each attribute is sampled before appearing in any parent set. To guarantee that the formed distributions are actual probability distributions, all negative numbers in $\mathbb{P}^*\left[X_j, \Pi_j\right]$ are set to zero and the distribution is normalized to maintain a total probability mass of 1. Example 4.3.1 shows how these distributions would be formed for $\mathcal{N}_1$ presented previously in Figure 5.

Algorithm 3: NoisyConditionals

Input: $D, \mathcal{N}, \varepsilon_2$
Output: $\mathcal{P}^*$

$\mathcal{P}^* \leftarrow \emptyset$
for $j = 1$ **to** p **do**
  materialize the joint distribution $\mathbb{P}\left[X_j, \Pi_j\right]$
  generate differentially private $\mathbb{P}^*\left[X_j, \Pi_j\right]$ by adding noise from $\text{Lap}\left(\frac{2p}{n\varepsilon_2}\right)$
  set negative values in $\mathbb{P}^*\left[X_j, \Pi_j\right]$ to 0 and normalize
  derive $\mathbb{P}^*\left[X_j|\Pi_j\right]$ from $\mathbb{P}^*\left[X_j, \Pi_j\right]$
  add $\mathbb{P}^*\left[X_j|\Pi_j\right]$ to $\mathcal{P}^*$
return $\mathcal{P}^*$

Example 4.3.1.

Given the Bayesian network $\mathcal{N}_1$ in Figure 5, whose AP pairs are presented in Table 1, the NoisyConditionals algorithm first estimates $\mathbb{P}[age, \emptyset] = \mathbb{P}[age]$, injects noise from $\text{Lap}\left(\frac{10}{n\varepsilon_2}\right)$ and then makes sure that the distribution is a valid probability distribution by setting negative values to 0 and normalizing the distribution to have a probability mass of 1. Since *age* did not have any parents, its conditional noisy distribution equals to $\mathbb{P}^*[age]$ and it's added to $\mathcal{P}^*$.

Then the algorithm moves to the next index and forms $\mathbb{P}[education, age]$, injects noise, standardizes the distribution and then derives $\mathbb{P}^*[education|age]$ and adds it to $\mathcal{P}^*$. In the next iteration, $\mathbb{P}[lifestyle, age, education]$ is formed, noise is injected, the distribution is standardized and the conditional distribution $\mathbb{P}^*[lifestyle|age, education]$ is formed and added to $\mathcal{P}^*$. This procedure is iterated for each attribute in the order of $\mathcal{N}_1$ until all conditional distributions have been materialized.

Once each of the noisy conditional distributions has been constructed, an arbitrary number of synthetic observations can be generated from them in the order indicated by the formulated Bayesian network.

5 Measuring utility

All SDC methods cause some degree of information loss. Generation of synthetic data is no exception since it is extremely unlikely to be able to estimate the true underlying distribution F perfectly. In addition, the generated records are required to be anonymous, meaning that records cannot simply be copied to preserve F. If the synthetic distribution F^* is too different from the original distribution F, the synthetic data set may lead to erroneous conclusions or indicate existence of phenomena that do not really exist. Thus, the veracity of any synthetic data set should always be assessed, although there are no general guidelines on when synthetic data are considered to be of sufficient quality to be utilized in practice nor how their veracity should be measured.

In this thesis, the examination of data utility is divided into two categories: methods that measure the overall similarity of the probability distributions and methods that evaluate the validity of statistical inference. In addition, since the definition of synthetic data requires the anonymity of data, it could be considered as one of the utility criteria. However, the examination of the anonymity of the synthetic data is excluded from this thesis, although it is an important aspect and is therefore discussed in Section 7. A very limited number of different methods have been chosen for this thesis, based on their importance in the analysis of longitudinal data and the ease of interpretation. In general, methods should always be appropriately selected based on the intended use and the properties of the synthetic data set in question.

5.1 Similarity of distributions

Probability theory forms the basis for the application of statistics, hence assessing the similarity of probability distributions is an important part of measuring the utility of synthetic data. There exist many methods to study probability distributions [34]. However, many of the methods have been developed for univariate distributions and comparing multivariate distributions – that are actually of interest – is challenging. With this in mind, the purpose of this chapter is to highlight some general considerations in comparing distributions and to present the methods used in this thesis.

The similarity of the sample spaces of synthetic and original data can be considered a good starting criterion. For example, a subject's height cannot be negative or his/her systolic blood pressure above 800. However, it is not enough to examine the sample space alone, but the event space must also be examined as it is possible that the method may generate combinations of observations that have a low probability of occurring or are not realistic, e.g., a 12-year-old girl having 5 children or a subject having follow-up measurements after death. The examination of the sample and event spaces is important since impossible or unlikely observations or combinations question the quality of the synthetic data very quickly.

The occurrence of missing observations or events should also be taken into account as systematic absence may lead to biased inference. In situations where the attributes in synthetic data are generated independently, the association structure that may be present in the original data is not taken into account at all. Correspondingly, the more repeated measures there are in the data, the more difficult it becomes to preserve the temporal structure. Thus, the probability of impossible observations and events increases if attributes are generated

independently or the data include a large number of repeated measures. However, the mere absence of impossible or missing values does not make the distributions sufficiently similar – the whole probability distributions should be similar.

5.1.1 Descriptive statistics

A summary statistic that quantitatively describes features of a collection of information is called a *descriptive statistic* and using and analyzing these statistics is called *descriptive statistics*. The difference between *inferential statistics* and descriptive statistics is the effort of descriptive statistics to summarize the sample while inferential statistics aims to learn something from the population that the sample is thought to represent. Nevertheless, the summary statistics used in descriptive statistics are often used in inferential statistics.

The summary statistics that are commonly used in descriptive statistics can be divided into two groups: measures of *central tendency* and measures of *variability* or *dispersion*. Central tendency is a typical or central value for a probability distribution. Examples of measures of central tendency are mean, median and mode. Variability or dispersion is the extent to which a probability distribution is stretched or squeezed. Examples of statistical measures measuring variability or dispersion are standard deviation or variance, range (maximum − minimum) and median absolute deviation. However, not all these summary statistics are suitable for describing categorical variables measured in nominal or ordinal level. Commonly used methods of descriptive statistics in case of categorical variables are different tables in which the distribution of the variable can be presented, for example, in terms of observed frequencies or proportions. Yet, tables can be difficult to interpret, especially if a variable can take a large amount of different values. As a consequence, the same information is often sought to be conveyed by different plots.

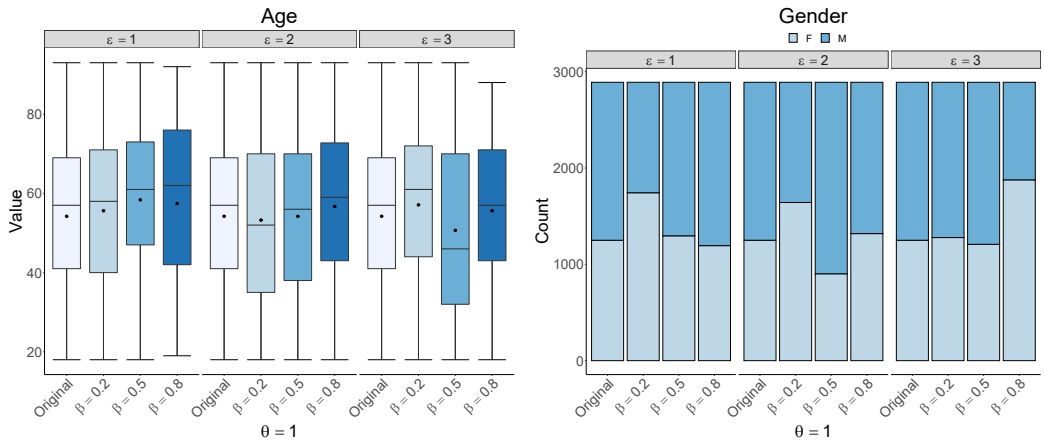
In generation of synthetic data, it would be desirable to maintain the associations between the variables. The preservation of associations between original and synthetic data can be examined, for example, by the difference between sample correlation coefficients, r, in the case of variables measured in ordinal, interval or ratio level and by the difference between Cramér's V coefficients, ϕ_c, in the case of variables measured in nominal level. In this thesis, the difference between the associations is calculated by subtracting the original value from the value of the synthetic data, in which case the sign of the difference is interpretatively meaningful. For example, if the sample correlation coefficient between two variables in the synthetic data is $r_s = -0.49$ and the corresponding value in the original data is $r_o = -0.50$, then $-0.49 - (-0.50) = 0.01$, which means that r_s is 0.01 higher than in the original data. The same interpretation of the sign of the difference also applies in case of ϕ_c: a positive coefficient indicates that the value in the synthetic data is greater than the original and vice versa.

Because descriptive statistics can be used effectively to summarize the properties of distributions, it is used in the comparison of the original and synthetic data. The summary statistics used in this thesis are presented using different visualization methods discussed in more detail in Subsection 5.1.2. The reason why tests designed to test hypotheses about summary statistics are not used in this thesis, although such tests exist, is that the visualizations are assumed to tell more about the nature of the data differences than test statistics.

5.1.2 Data visualization

Proper data visualizations provide insight into the descriptive statistics as well as into how the distributions differ. Some visualization techniques can reveal extreme values and abnormal combinations of values. Visualizations can be used for both univariate and multivariate distributions, although there exist fewer methods for the latter. Examples of different visualization methods for univariate distribution comparison are box plots, histograms, bar plots, strip charts, violin plots and plotting shift or difference asymmetry functions. Bivariate distributions can be visualized, for example, by forming a heat map from a correlation matrix, where each cell in the matrix presents the correlation between two variables, or by a scatter plot, where one variable is on the x-axis and the other on the y-axis, or a mosaic plot. In addition, grouping univariate visualization methods can also be used in bivariate comparisons. Correspondingly, multivariate distributions can be visualized, for example, using geometric projection, such as plotting the first two components of the principal component analysis as a scatter plot, or using iconographic methods such as glyphs.

In this thesis, the box plot is used as it can be used to present several of the previously mentioned descriptive statistics as well as outliers in one figure. Nevertheless, the box plot can only be used for continuous variables, and thus the bar plot, illustrated in Figure 6b, is used for discrete or categorical variables. Unlike the traditional box plot, the mean of the variable has been added to the plot. Figure 6a shows examples of modified box plots used in the thesis. To compare the original and synthetic distributions, a grouped plot can be formed with the distributions plotted side by side in the same plot.



(a) Box plot

(b) Bar plot

Figure 6: Panel (a) shows a collection of modified box plots with a bold line representing the median and a dot representing the mean. The upper and lower limits of the colored box equal to upper and lower quartiles, respectively. The length of the whiskers is 1.5 times the interquartile range (IQR), that is, the difference between the upper and lower quartile. The figure is stratified according to the hyperparameters used to generate the synthetic data so that one figure always represents one value of the usefulness θ indicated at the bottom of the figure. The distribution of the original data attribute is shown on the left under each value of ε for ease of comparison and is indicated on the x-axis. Panel (b) shows a bar plot of a two-class variable, in which the y-axis of the graph depicts the observed frequency for each class. A similar stratification of hyperparameters is used in the figure as for the box plots. Both visualization methods are used in this thesis for inter-distribution comparisons.

Differences between the correlation matrices in the synthetic and original data sets are studied by forming a heat map of the differences as defined in Subsection 5.1.1. A similar heat map is also generated for categorical variables by using the Cramér's V. Figure 7 shows an example of a heat map used in this thesis. Bivariate visualizations are recommended when examining associations between the response variable and explanatory variables if the model for which the synthetic data will be used is known. Such approaches are excluded from this thesis, however.

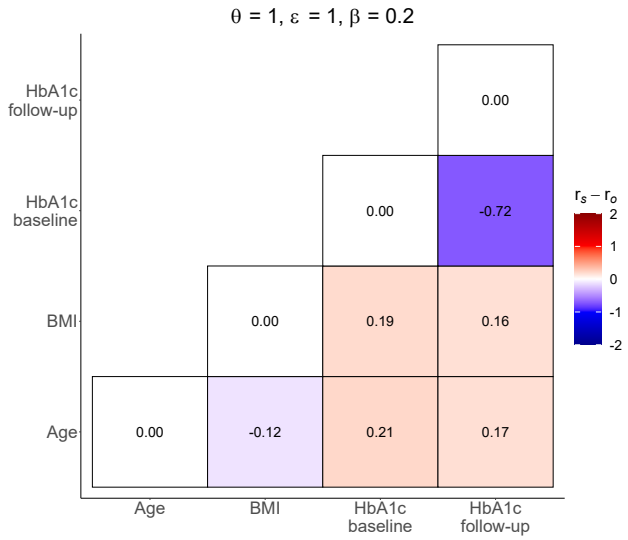


Figure 7: A lower triangle of a heat map of the differences between Pearson's sample correlation coefficients in two correlation matrices. If the sample correlation coefficient calculated from the synthetic data is higher, the difference is positive and vice versa. The values of the hyperparameters used to generate the synthetic data are shown in the title of the figure.

In case of longitudinal data, plotting individual trajectories is an effective way to find out whether trajectories deviate on average from the original trajectories or whether they contain impossible or unlikely events. However, the method becomes very laborious as the dimensions of the data increases, because simultaneous examination of many variables on an individual basis may require unit conversions and examination of many plots and, correspondingly, simultaneous examination of many subjects may hide important changes in the individual trajectories. Figure 8 shows an example of an individual trajectory plot for repeated measurements as used in this thesis.

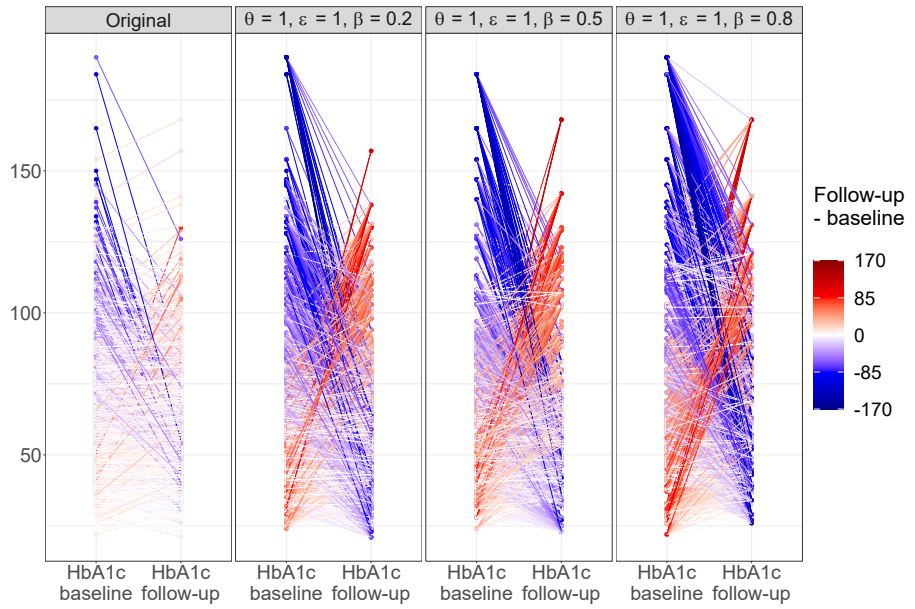


Figure 8: The individual trajectories are plotted in the figure so that the color of the line indicates whether the value has decreased (blue) or increased (red) from baseline to follow-up. The whiter the graph, the more similar the repeated measurements, which can also be interpreted as the correlation between the measurements. The headings in the sections of the figure indicate whether the data is original or synthetic, and in the latter case, the values of the hyperparameters used to generate the data are given in the heading.

5.2 Validity of statistical inference

Descriptive statistics is seldom able to answer all research questions and thus it is less often used alone in the conduct of research but rather as part of a broader analysis. Inferential statistics is used to deduce properties of an underlying probability distribution and to infer properties of a population. The data at hand are usually assumed to be a random sample from a larger population and are thus used to make inference about that population. Research questions that can be studied using statistical inference include, for example, whether a particular feature of an individual is associated with the onset of cardiovascular disease or how medication affects disease progression. In general, most studies aim to make some degree of statistical inference, thus the validity of statistical inference when using synthetic data should be studied.

Statistical inference can be approached from two different perspectives: *Bayesian inference* and *frequentist inference*. Without going into deeper details, the main difference between the approaches is that the Bayesian inference allows uncertainty about unknown parameters be expressed in terms of probability distributions whereas in the frequentist approach parameters are not described in terms of probability distributions and are treated as fixed. Depending on the selected inference approach, different statistical models are available with different assumptions. Of course, the research questions posed and the data available must always be taken into account when choosing the inferential approach and the methods of analysis.

This thesis applies the frequentist approach. The assessment of the validity of statistical inference focuses on differences in the estimated model parameters, overlaps in their

confidence intervals, and differences in statistical significance, as measured by p-values. These statistics play a key role in statistical inference, which is why they have been chosen as the evaluation criteria, although the need for the p-value has recently been questioned [35]. The results are presented in a table format, an example of which is given in Table 2.

5.2.1 Linear mixed-effects model

A *linear model*, or more specifically a linear regression model, is a commonly used statistical model in which the value of the continuous response variable Y is predicted by predictors $X_1, \ldots, X_k$, also called covariates. In a *mixed-effects model*, some of the covariates are fixed and some are random. A linear mixed-effects model (LMM) is a hierarchical model that can be used to account for the correlation structure in longitudinal data. It is generally applied in longitudinal data analysis and, as a result, the model is applied in this thesis to evaluate the validity of statistical inference between original and synthetic data.

Definition 5.1. *Linear mixed-effects model*
Let $\boldsymbol{Y}_i = \left[Y_{i1}, Y_{i2}, \ldots, Y_{in_i}\right]'$ be the vector of the response variable corresponding to the ith $(i = 1, \ldots, N)$ subject measured on n_i occasions. Thus, Y_{ij} denotes the jth measurement of the ith subject. The linear mixed-effects model is expressed as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown fixed effect parameters, including the intercept, $\boldsymbol{X}_i$ is a known $n_i \times k$ design matrix of the fixed effects, $\boldsymbol{Z}_i$ is a known $n_i \times q$ design matrix of the random effects, $\boldsymbol{b}_i$ is a $q \times 1$ vector of unknown random effects and $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ vector of random errors. It is assumed that $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$ are both independent and follow multivariate normal distributions with zero mean vectors, $\boldsymbol{0}$, and variance-covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}_i$, respectively. These matrices consist of unknown parameters and are assumed to have specific structured forms. Responses of different subjects are assumed to be independent.

Based on the above definition, the marginal distribution of $\boldsymbol{Y}_i$ is a multivariate normal distribution with the mean vector $\boldsymbol{X}_i\boldsymbol{\beta}$, and the variance-covariance matrix $\boldsymbol{V}_i = \mathbb{V}ar[\boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i] = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^T + \boldsymbol{R}_i$. This formulation now allows for correlations and unequal variances across the responses of a single subject. The diagonals of $\boldsymbol{G}$ and $\boldsymbol{R}_i$ form the between- and the within-individual variances, and thus their sum is the total variance. The ratio of the between-individual variance and the total variance is called the *intraclass correlation* (ICC). The greater the between-individual variance, the greater the ICC, and the greater the bias of a non-hierarchical model in the analysis of hierarchical data.

For the sake of simplicity, this thesis applies a linear mixed-effects model that can have an arbitrary number of fixed effects but includes only the random intercept term and therefore $\boldsymbol{G}$ suppresses to a scalar τ. Furthermore, $\boldsymbol{R}_i$ is assumed to equal $\sigma^2 I$, where I is the $n_i \times n_i$ identity matrix. Table 2 illustrates a table that is used to present the estimated models in this thesis.

Table 2: The table illustrates a results table of two different linear mixed-effects model for HbA1c: using the original and synthetic data, respectively. The table first describes fixed predictors and their coefficient estimates, 95% confidence intervals and p-values. The random part includes the estimate of the within-individual variance (σ^2) as well as the estimate for the between-individual variance (τ_{00}). The random effects part also includes the intraclass correlation coefficient (ICC) as well as the number of subjects in the data (N). The values of the hyperparameters used to make the synthetic data are indicated above the table.

synthetic: $\theta = 1, \varepsilon = 1, \beta = 0.8$

	HbA1c original			**HbA1c synthetic**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	74.99	71.36 – 78.62	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	0.00	-0.04 – 0.04	0.832
Gender [M]	1.31	0.07 – 2.55	**0.039**	-1.08	-3.11 – 0.95	0.297
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	0.54	-1.61 – 2.68	0.624
BMI	-0.09	-0.18 – -0.00	**0.041**	-0.02	-0.09 – 0.05	0.558
Random Effects						
σ^2	90.29			1107.51		
τ_{00}	230.76 $_{\text{Subject}}$			1.65 $_{\text{Subject}}$		
ICC	0.72			0.00		
N	2890 $_{\text{Subject}}$			2890 $_{\text{Subject}}$		
Observations	5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.001 / 0.002		

6 Generating synthetic longitudinal patient data with the PrivBayes method

In this section, the PrivBayes method is implemented to real-world longitudinal patient data and the results are presented as introduced in Section 5. This section also explains how the original data were generated and how the values of the hyperparameters of PrivBayes were selected. The results are presented as is and the results as well as the limitations of the method and the options for further development are discussed separately in Section 7.

In order to implement the method, a preliminary R-package called *SynthData* [36] was programmed. The generation of synthetic data sets and the utility comparisons were executed in ACI's own secure user environment with R version of 3.6.3 [37]. The source code of the SynthData package is available on the University of Turku's GitLab [36] and the source code used to generate and compare the synthetic data sets in this thesis can be found in Appendix A.

6.1 Original data and the selection of hyperparameters

The original data set was collected by the author from the database of Auria Clinical Informatics under the permission number T152/2017. The data set was collected from patients in the database who had an ICD-10 code, given in the parenthesis, corresponding to either type 1 (E10) or type 2 (E11) diabetes between the study period of January 1, 2017 – June 30, 2020. The gender of the patients, type of diabetes, and complications related to diabetes that occurred during the study period were included in the data set. In addition, the first and subsequent glycated hemoglobin (HbA1c) measurements during the study period, as well as the patient's age and body mass index (BMI) at the time point corresponding to or closest to the first measurement, were also included. To test the PrivBayes method, missing data were not allowed, and two patients were removed from the sample due to recording errors, after which 2890 patients were selected for the original data set.

The motivation behind the above choices was to simulate a reasonably realistic but simple study design that collects one repeated measurement of a selected response variable for each patient and allows studying the association between a few selected covariates and the response variable. In this thesis, HbA1c measurement was selected as the response variable because it is a standard measurement used to assess the blood glucose balance of diabetic patients in Finland. The domain size of HbA1c at baseline, that is, at the time of the first measurement, was 118 with the range of 22 to 190 mmol/mol and the domain size of the repeated measurement was 110, values ranging from 21 to 168.

The patient's age (18-93 years), BMI (13.3-90.7 kg/m^2), gender (female, male) and type of diabetes (type 1 or 2) were selected as covariates since they were thought to be possibly associated with the response variable. Their domain sizes were 75, 367, 2 and 2 respectively. Complication types were excluded from the covariates because its domain size was 16 and the inclusion of a multi-category attribute in the model would probably have caused problems in model fitting. However, the attribute was used to test the hierarchical encoding of a categorical attribute and to examine how such an attribute behaves when forming Bayesian networks. More comprehensive details on the distributions are presented in Subsection 6.2.

We next describe how the hyperparameters were chosen, a task that is highly non-trivial. Values for the usefulness θ were chosen to be $\theta = 1, 2, 3$. The value $\theta = 1$ was selected to study a situation where the ratio of average scale of information to average scale of noise is at least one. Values lower than $\theta = 1$ were not considered as they are less meaningful from a practical point of view, because in general the aim is to preserve as much information as possible. Therefore, values $\theta = 2, 3$ were chosen to investigate how the data behave if the ratio is increased, but the selection was limited to only two additional values to control the number of synthetic data sets produced.

Values of the privacy budget ε were initially planned to be at most one for all generated synthetic data sets based on a previous recommendation by Dwork et al. [29]. However, during the programming and testing of the PrivBayes method, this choice proved to be too restrictive when using generalized attributes and the values of ε were finally decided to be $\varepsilon = 1, 2, 3$. Although increasing the privacy budget undermines the privacy guarantees of differential privacy, the choice was justified by the fact that the generated synthetic data sets are not published in connection with this thesis and higher values of ε have been used in other studies [38, 39].

The value of β determines how the privacy budget ε is distributed between the exponential and the Laplace mechanism. A smaller value of β reduces the privacy budget of the exponential mechanism, resulting in very noisy Bayesian networks but in contrast, in the Laplace mechanism, less noise is added to marginal distributions. Correspondingly, a higher value of β helps to build more accurate, that is, less noisy, Bayesian networks but more noise needs to be added to the marginal distributions. Therefore, the values of β were selected to be 0.2, 0.5 and 0.8 to study how an emphasis on either mechanism or the even distribution between the mechanisms affects the quality of the synthetic data.

Based on the theory of the PrivBayes method, the following hypotheses were set:

- Data sets having the highest selected value of θ generally perform better in utility comparisons than data sets having a lower value since the former data sets should contain, on average, more information compared to the added noise.
- Data sets having the highest selected value of ε generally perform better in utility comparisons than data sets having a lower value because a larger privacy budget allows for a more accurate preservation of information (at the cost of privacy).
- Data sets having the lowest selected value of β generally perform better in utility comparisons than data sets having a higher value because a lower value restricts the formation of maximal parent sets less, in addition to which less noise is added in the Laplace mechanism.
- The fewer independent attributes have been used to generate the data set, the better the data set will perform in comparisons because the associations present in the longitudinal data are better preserved.

The validity of these hypotheses was tested by examining the results of the selected utility measurements.

Based on the choices of the hyperparameters, $3^3 = 27$ different synthetic data sets were generated in order to study how each choice affected the quality of the generated synthetic data set. In addition, since PrivBayes employs randomness, data sets were generated several times to visually assess the impact of randomness on the formed Bayesian

networks and distributions. Based on the visual inspection, a condition was added to the data generation that Bayesian networks were formed until the baseline measurement of HbA1c appeared as the parent of HbA1c follow-up measurement in at least one of the 27 networks. and this data were used in the main analysis. The results in the next subsection are based on a generated set of 27 data obeying this condition. Additionally, to evaluate the effect of the randomness on the results, we also generated several replicates of the same collection of 27 data but the plots and conclusions based on these (not shown here) were similar to the ones presented in the next section.

6.2 Results

The structures of all Bayesian networks corresponding to the synthetic data sets are described in Appendix B together with the hyperparameter values used to generate each set. The most common degree of the network was two with the frequency of 19, the minimum degree was one and the maximum degree was three. An examination of the network structures showed that attributes with smaller domain sizes were generally selected to the network first, medium-sized ones had more variation, and large-sized attributes were focused at the end of the network. This was despite the fact that network learning is subject to differential privacy, which could be expected to cause more variation in networks structures. Attributes with smaller domain sizes were also more frequently represented as parents. The observed frequencies corresponding to the placements in the Bayesian networks for each variable are described in Table 3.

Table 3: The table describes how many times each attribute appeared in position j in all formed Bayesian networks of the final round. The original domain size of each attribute is marked with a superscript and the attributes are arranged in the table in ascending order of domain sizes.

X_j	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	$j = 6$	$j = 7$
Gender^{2}	4	0	17	6	0	0	0
Type of diabetes^{2}	4	23	0	0	0	0	0
Complications^{16}	3	3	5	13	3	0	0
Age^{75}	2	1	4	3	8	3	6
HbA1c follow-up^{110}	8	0	0	0	3	9	7
HbA1c baseline^{118}	1	0	0	0	10	12	4
BMI^{367}	5	0	1	5	3	3	10

The effect of each hyperparameter value to the number of parents, AP pairs and the degree of the network is illustrated in Table 4. Increasing θ seemed to increase the number of independent attributes in the network when ε and β were kept constant, but only slightly. In contrast, a higher value of β seemed to have more impact: the hyperparameter value of 0.8 more often resulted in lower-degree networks and the only one-degree network, presented in Table 8.12 of Appendix B, was formed with the selection of $\beta = 0.8$. In addition, in this network, only one attribute had a parent. Furthermore, increasing ε seemed to reduce the number of independent attributes in the networks.

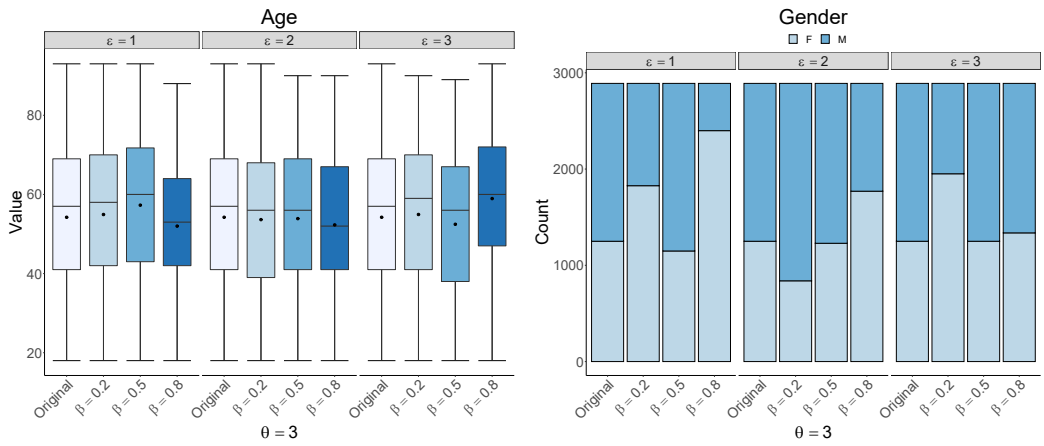
Despite the restriction set in data generation, only in one network, generated with hyperparameter values of $\theta = 1$, $\varepsilon = 3$ and $\beta = 0.2$, the HbA1c measurements formed

Table 4: The table describes the means of the number of AP pairs, the total number of parents as well as the degree of the network marginally with respect to each hyperparameter value.

Quantity	$\theta = 1$	$\theta = 2$	$\theta = 3$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\beta = 0.2$	$\beta = 0.5$	$\beta = 0.8$
Number of AP pairs	3.9	2.9	2.6	2.4	3.1	3.8	3.4	3.4	2.4
Number of parents	6.3	4.2	4.1	3.9	4.7	6.1	6.0	5.0	3.7
Degree of network	2.4	2.0	2.2	2.1	2.2	2.3	2.7	2.0	2.0

an AP pair. In the network, presented in Table 8.7 of Appendix B, the HbA1c baseline appeared as the parent of the HbA1c follow-up measurement, but it was generalized in a level with a domain size of four. Both the baseline and follow-up measurement were more often modelled as independent attributes, but in a few networks they had the type of diabetes as their parent. These networks are presented in Tables 8.4, 8.5, 8.7, 8.8 and 8.16 of Appendix B, respectively. The possible explanation why the baseline measurement of HbA1c did not occur more often as a parent of the follow-up measurement is discussed in Section 7. It is also worth mentioning that the HbA1c follow-up measurement was initialized eight times as the first attribute, i.e., modelled as independent attribute by default, whereas the HbA1c baseline measurement was only selected once.

The effect of the hyperparameter values was more evident when comparing the univariate distributions presented in Appendix C. For continuous attributes, the variation between the distributions generally appeared to decrease with increasing θ, as expected. However, this reduction in variation was less evident for the HbA1c measurements. In contrast, a similar reduction in variation – at least on the same scale – could not be observed for categorical attributes. Figure 9 illustrates univariate distributions of a continuous and a categorical attribute in the case of $\theta = 3$. In fact, the distributions of the categorical attributes seemed to be more homogeneous and similar to the original distribution when $\theta = 2$, illustrated in Figures 11.14, 11.17 and 11.20 of Appendix C.



(a) Continuous attribute

(b) Categorical attribute

Figure 9: Univariate distributions of age and gender generated with the value of $\theta = 3$. In the case of a continuous attribute, shown in panel (a), there is less variation between the distributions, whereas for a categorical attribute, shown in panel (b), the same is not observed.

Contrary to the hypothesis, for both continuous and categorical attributes, increasing the value of ε did not seem to increase the accuracy of the distribution when the synthetic equivalent that best matched the original distribution was selected for each hyperparameter

value. The frequencies of these selections with respect to hyperparameters ε and β are shown in Table 5. A lower value of β appeared to be associated with more accurate distributions for continuous attributes but not for categorical attributes. An interesting finding was that the distribution in which the baseline HbA1c measurement occurred as the parent of the follow-up measurement, illustrated in Figure 11.10 of Appendix C, differed significantly from the original distribution

Table 5: The table shows how many times the values of ε and β occurred for the best rated synthetic univariate distribution with respect to each continuous and categorical attribute. The assessment was based on the similarity of the synthetic and original distributions. The frequency of the continuous attributes is presented before the slash.

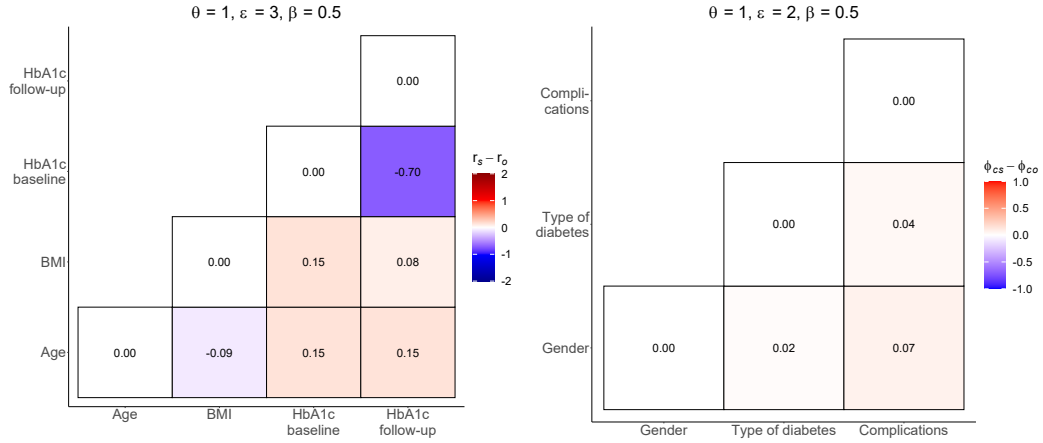
Hyperparameter	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$
$\beta = 0.2$	6/3	7/3	9/2
$\beta = 0.5$	6/3	5/4	1/2
$\beta = 0.8$	0/3	0/2	2/5

The differences in the sample correlation coefficients were quite similar in terms of both magnitudes and directions, and the values of the hyperparameters did not seem to have a large effect. In contrast, the differences in the ϕ_c coefficients were less stable, in addition to which the directions of their differences also varied. The heat maps of the differences between the sample correlation coefficients and Cramer's V coefficients are presented in Appendix D and Table 6 describes the means of the absolute values of the differences in r and ϕ_c coefficients marginally with respect to each hyperparameter value.

Table 6: The table describes the means of the absolute values of the differences in r and ϕ_c coefficients marginally for each hyperparameter value. The values on the diagonal were excluded from the calculation.

Coefficient	$\theta = 1$	$\theta = 2$	$\theta = 3$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\beta = 0.2$	$\beta = 0.5$	$\beta = 0.8$
r	0.25	0.25	0.26	0.26	0.25	0.25	0.26	0.25	0.26
ϕ_c	0.21	0.15	0.17	0.19	0.14	0.19	0.17	0.16	0.19

The difference $r_s - r_o$ was most severe for HbA1c measurements, with the mean difference of -0.75. In the network where the baseline HbA1c measurement occurred as the parent of the follow-up measurement, the difference was -0.78, which was the second largest difference in HbA1c measurements. In general, the degree of the networks or the number of independent attributes did not appear to be related to the coefficient differences, although the best overall results, illustrated in Figure 10, occurred for networks with at most two independent attributes.



(a) Correlation coefficients (b) Cramer's V coefficients

Figure 10: The figure illustrates the heat maps of the differences of the correlation (a) and Cramer's V (b) coefficients for which the coefficients of the synthetic data were closest to the original data. For computational and programming reasons, negative coefficients appear on the diagonal of Cramer's V coefficients, although the values on the diagonal are actually all zeros. The network structures corresponding to the heat maps are shown in Table 8.8 and Table 8.5, respectively.

The lack of correlation between the repeated measurements was particularly evident in the individual trajectory plots presented in Appendix E. In the graphs, the dark blue and red colors were overrepresented compared to the original graph, in addition to which there were values that occurred multiple times in the synthetic data while appearing significantly less often in the original data. Moreover, the synthetic data sets included individual changes that were clearly larger than those in the original data. Because the PrivBayes method operates with empirical sample space, there were no outlying observations in the individual trajectory plots that would have be outside the original set of values. The effect of hyperparameter values on individual trajectories was not evident from the results.

The lack of correlation was also reflected in the results of the linear mixed-effects models presented in Appendix F. None of the models was able to retain the results of the original model to the extent that the inferences from the analyzes would be similar. In a few models, some fixed parameter estimates were close to original estimates and confidence intervals overlapped, but then in turn, the variance between individuals was often 0, meaning that the variance in the data could be explained mostly by within-individual variance. This is also the reason why the program gave a warning of a singular fit, that is, the random effects structure was too complex to be supported by the data. The network structure or hyperparameter values did not appear to have a perceptible effect on the model results. Table 7 shows two models whose results were closest to those of the original model.

Table 7: Results of two different linear mixed-effects models that were closest to the original results. In the models of synthetic data sets 8 and 20, the predictors' estimates are close to the estimates of the original model and the confidence intervals cover the original estimates, except for gender and BMI in data set 8 and diabetes type in data set 20. For both data sets, the estimates of σ^2, τ_{00} and ICC differ significantly from the original estimates and the statistical significance is retained only for the diabetes type in the model of data set 8.

	HbA1c original			**HbA1c synthetic 8**			**HbA1c synthetic 20**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	73.45	70.34 – 76.56	**<0.001**	70.10	66.02 – 74.18	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	-0.03	-0.07 – 0.00	0.078	-0.04	-0.09 – 0.01	0.130
Gender [M]	1.31	0.07 – 2.55	**0.039**	0.09	-1.40 – 1.57	0.911	1.77	-0.07 – 3.60	0.059
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	-9.98	-11.50 – -8.45	**<0.001**	-0.98	-2.79 – 0.83	0.288
BMI	-0.09	-0.18 – -0.00	**0.041**	-0.00	-0.08 – 0.07	0.971	-0.02	-0.10 – 0.07	0.703
Random Effects									
σ^2	90.29			777.07			1192.69		
τ_{00}	230.76 $_{Subject}$			10.03 $_{Subject}$			0.00 $_{Subject}$		
ICC	0.72			0.01					
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.032 / 0.044			0.001 / NA		

7 Discussion

The main focus of this thesis was to study how synthetic longitudinal patient data could be generated by using the PrivBayes method and to measure the utility of the generated synthetic data. A variety of different utility comparison measures were considered in this thesis, extending from univariate distributions to individual trajectories and results of linear mixed-effects models. Based on the results, the PrivBayes method was unable to retain the correlation structure of the longitudinal data used in this thesis, and in only one network did the baseline level of the repeated measurement occur as the parent of the follow-up measurement. By looking at the algorithms of the PrivBayes method, the limiting factor may have been the MaximalParentSets algorithm and the sample size used in this thesis.

In MaximalParentSets, when $\tau = \frac{n\varepsilon_2}{2p\theta|\mathrm{dom}(X)|} < 1$, the attribute X is modelled as an independent variable by default. For the data set used in this thesis, this would mean that the ratio ε_2/θ should be at least 0.533 in order for the follow-up measurement of HbA1c to have parents at all. In addition, if the baseline measurement of HbA1c is desired to be the parent of the follow-up measurement, the value of τ is to be divided with the domain size of the corresponding generalization level of the baseline measurement. By combining these restrictions, in the case where the network learning phase is to be optimized with respect to repeated measurements so that the preceding measurement could appear as the parent of the subsequent measurement, the following equation can be used to determine the values of the hyperparameters as well as a sufficient sample size and a number of attributes in the original data set:

$$\frac{n\varepsilon_2}{2p\theta\left|\mathrm{dom}\left(X_{j,k}\right)\right|\left|\mathrm{dom}\left(X^{(i)}_{j,k-1}\right)\right|} \geq 1, \tag{3}$$

where $X_{j,k}$ is the kth repeated measurement of an attribute X_j at non-generalized level and $X^{(i)}_{j,k-1}$ the preceding measurement of that attribute at the desired generalization level i. In the case of more than one repeated measurement, the selection of hyperparameters can be performed by selecting the repeated measurement with the largest original domain size as the non-generalized measurement and its preceding measurement as $X^{(i)}_{j,k-1}$.

Based on the selected hyperparameter values and the restriction described in (3), the baseline measurement of HbA1c could occur as a parent of the follow-up measurement for only a few combinations of the selected hyperparameter values and only if the measurement was generalized to levels with domain sizes of two or four, respectively. Yet, only the AP pair in which the domain size of the baseline measurement was four occurred in the generated data sets, which could only be achieved in the case where $\theta = 1$, $\varepsilon = 3$ and $\beta = 0.2$. The network structure corresponding to this selection is presented in Table 8.7 of Appendix B. A possible explanation for this is that the value of the score function used in the exponential mechanism in the case where the baseline measurement can only have two different values is 0.02 while for four different values the value is 0.27. That is, the generalization of the attribute degrades the score which explains why the other AP pair was not selected in the exponential mechanism.

Nonetheless, the selected AP pair was clearly unable to retain the correlation between the measurements which is no wonder, since the four categories of the baseline measurement were $[22, 67.6]$, $(67.6, 113]$, $(113, 159]$ and $(159, 190]$. More specifically, the value

of the follow-up measurement was conditioned based on the baseline value within these categories and since the ranges in these categories are wide, the correlation between the measurements could not be retained. Because the method has the advantage of operating in the original sample space, for further development, it is sufficient to optimize only the modeling of the association structure between attributes. On the other hand, this advantage is also a weakness of the method, as it is not able to produce observations for a continuous variable from the whole range, but rather replicates the observations according to the original data.

There may also be other attributes in the data that are not repeated measurements, but whose association structure with other variables is to be preserved. In such a situation, X_j in (3) can be replaced by an attribute for which the network learning is to be optimized and one can select any another variable as the generalized attribute $X^{(i)}_{k\neq j}$ for which the association with X_j is to be optimized. For example, if the network learning phase had been optimized so that the association structures in the networks were as accurate as possible, BMI would have been chosen as X_j and non-generalized HbA1c baseline measurement as $X^{(i)}_k$. With these choices, and selecting $\theta = 1$, $\varepsilon = 3$, $\beta = 0.2$ and $p = 7$, the required sample size would have been $252,619$. Although the operation of the PrivBayes method was extended to general domains [7], looking at (3), it can be seen that either the sample size or the privacy budget of the Laplace mechanism (or both) need to increase significantly relative to the domain sizes of the attributes in the original data set in order to fully guarantee the usefulness of the method in practice. This is clearly a weakness of the PrivBayes method.

In the case of generalized attributes, if the sample size of the original data set is too small, the value of τ is also more often small, in which case the method has to form AP pairs from highly generalized attributes or model the attributes independently. This was reflected in the results by the fact that the network structures were quite similar despite the randomness created by the exponential mechanism. In addition, the PrivBayes method allows to initialize the repeated measurement as the first attribute, in which case the association with previous measurements may be lost if the attribute does not occur as a parent of any of these measurements. Zhang et al. [7] did not specify why AP pairs should be selected using differential privacy. One interpretation could be that a potential adversary cannot infer the value of an attribute with certainty based on the value of another attribute and thereby gain access to sensitive information. That is, the associations of the attributes of the original data cannot be inferred with certainty from the synthetic data. Nonetheless, it might be worthwhile to study how the quality of synthetic data would change if it were possible for a user to enter a ready-made network of hypothesized causal connections into the program or at least restrict the network learning by requiring certain connections to be in the Bayesian network. In addition, the method could be further developed to perform the optimization of hyperparameters as part of the operation of the program by applying, for example, cross-validation.

The results partially supported the hypotheses presented in Subsection 6.1. The main reason why not all hypotheses could be confirmed was the inconsistency of the results between the continuous and categorical attributes. One reason for this may be the inadequacy of the sample size used in the thesis as well as the values of the hyperparameters that forced the formation of Bayesian networks from a narrow set of attributes, possibly resulting in a deterioration in the quality of the synthetic data sets. Another reason may be

that the visual assessment used in this thesis is not accurate enough to detect differences and similarities, and quantitative methods might be more desirable. Therefore, it would be preferable to test the method in the future with quantitative criteria and data of sufficient quality, keeping also in mind that the method should work with real-world data, which may also contain missing observations, but whose existence was not allowed in this thesis. In addition, methods that measure the association between numerical and categorical attributes, with the exception of LMM, as well as multivariate methods were excluded from this thesis and their use should be considered in further studies.

The testing and verification of anonymity was excluded from this thesis, even though it is an essential part of the generation of synthetic data as the production of anonymous data is the basis of the whole generation. Verification of anonymity requires specific professional expertise that the author did not have enough at the time of writing this thesis and it would have required a broader literature review of de-identification methods that would have gone beyond the scope of this thesis. However, a quick look at the individual trajectory plots, presented in Appendix E, reveals that the observed values appear frequently in the data and thus cannot be distinguished from others and used to identify individuals. In an e-mail conversation with A. Bülow (lawyer at the research services of the University of Helsinki) and T. Shouterington (legal counsel at the Finnish Biobank Cooperative FINBB) [40], it was deliberated whether all synthetic data can be considered anonymous in principle if the data generation process is in some way based on random sampling and can therefore be considered as a special case of simulated data. If such an interpretation can be considered justified, for the PrivBayes method, it might be sufficient to apply differential privacy only on the distribution learning phase, or even to omit it altogether and directly generate new synthetic observations from the empirical conditional distributions.

There are also ethical and legal issues associated with the generation of synthetic data, but since these aspects do not belong to the field of mathematics or statistics, these topics were addressed only very superficially, even though they are important aspects and should also be studied in more detail in the future. Topics for further research in these areas include the development of more practical guidelines on what is considered sufficiently anonymous or what is a reasonable effort. It should be noted, however, that while it is possible to produce synthetic data, this procedure should be subject to the same ethical considerations as any other scientific activity. For example, data should not be synthesized and, in particular, published just because it is possible, but the benefits of publication or redistribution should be weighed against the potential disadvantages. This is especially true if it is possible to infer something from the data that, for example, may be clearly detrimental to a group of people, even if no single individual can be identified from the data. In other words, even if the inferential disclosure is not to be avoided in order to do research, its possible consequences must be assessed in advance and the potential risks of disclosing information that was not intended to be revealed should be acknowledged.

Although the PrivBayes method applied in this thesis did not produce synthetic data of sufficient quality to be applicable as such to the synthetic data generation of real-world longitudinal patient data, the results provided more detailed information on which areas require further research and development. In addition, the results of this thesis supported the previous conclusion as to why synthesizing longitudinal data is particularly challenging.

References

[1] European Parliament and Council. Regulation 2016/679/EU on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Official Journal. 2016 May;(L119):1–88.

[2] Finnish Ministry of Social Affairs and Health. Act 552/2019 on the Secondary Use of Health and Social Data [Internet]; 2019. Available from: https://stm.fi/en/secondary-use-of-health-and-social-data. Accessed 2020-03-25.

[3] Terveyden ja hyvinvoinnin laitos. Käyttölupien hakeminen [Authorisation application] [Internet]; 2020. Available from: https://thl.fi/fi/tilastot-ja-data/tutkimuskaytto/kayttoluvan-hakeminen. Accessed 2020-02-26.

[4] Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, et al. Statistical Disclosure Control. 1st ed. Chichester: Wiley; 2012.

[5] Rubin DB. Statistical Disclosure Limitation. Journal of Official Statistics. 1993 June;9(2):461–468.

[6] Office of the Data Protection Ombudsman. Pseudonymised and anonymised data [Internet]; 2020. Available from: https://tietosuoja.fi/en/pseudonymised-and-anonymised-data. Accessed 2020-08-10.

[7] Zhang J, Cormode G, Procopiuc C, Srivastava D, Xiao X. PrivBayes: Private Data Release via Bayesian Networks. ACM Transactions on Database Systems. 2017 October;42:1–41.

[8] Weiss RE. In: Modeling Longitudinal Data. 1st ed. New York: Springer; 2005. p. 1–4.

[9] IDC DataSphere. IDC's Global DataSphere Forecast Shows Continued Steady Growth in the Creation and Consumption of Data [Internet]; 2020. Available from: https://www.idc.com/getdoc.jsp?containerId=prUS46286020. Accessed 2020-09-19.

[10] European Data Protection Supervisor. The History of the General Data Protection Regulation [Internet]; 2017. Available from: https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en. Accessed 2020-09-19.

[11] Heiliö PL. "VS: Toisiolain historia" [RE: History of the Act on Secondary Use of Health and Social Data];. Received by Katariina Perkonoja. 21 September 2020. Email interview.

[12] Findata. About us [Internet]; 2020. Available from: https://www.findata.fi/en/about-us/. Accessed 2020-08-10.

[13] European Data Protection Party. Opinion 05/2014 on Anonymisation Techniques. Papers of the Article 29 Data Protection Working Party. 2014 March:5.

[14] Dwork C. Differential Privacy. In: Automata, Languages and Programming. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 1–12.

[15] Day one Staff. Protecting data privacy [Internet]; 2018. Available from: https://blog.aboutamazon.com/amazon-ai/protecting-data-privacy. Accessed 2020-03-25.

[16] Apple. Privacy [Internet]; 2020. Available from: https://www.apple.com/privacy/features/. Accessed 2020-03-25.

[17] Google. How Google anonymizes data [Internet]; 2020. Available from: https://policies.google.com/technologies/anonymization?hl=en-US. Accessed 2020-03-25.

[18] Christensen J, Pontoppidan NH, Rossing R, Anisetti M, Bamiou DE, Spanoudakis G, et al. Fully Synthetic Longitudinal Real-World Data From Hearing Aid Wearers for Public Health Policy Modeling. Frontiers in Neuroscience. 2019 August;13:850.

[19] Venkataramanan N, Shriram A. Data Privacy: Principles and Practice. 1st ed. Boca Ranton: CRC Press; 2017.

[20] Merriam-Webster. Synthesis [Internet]; 2020. Available from: https://www.merriam-webster.com/dictionary/synthesis. Accessed 2020-10-25.

[21] Dahmen J, Cook D. SynSys: A Synthetic Data Generation System for Healthcare Applications. Sensors. 2019 March;19:1181.

[22] Dash S, Dutta R, Guyon I, Pavao A, Yale A, Bennett KP. Synthetic Event Time Series Health Data Generation; 2019. ArXiv preprint arXiv:1911.06411.

[23] Baucum M, Khojandi A, vasudevan r. Improving Deep Reinforcement Learning with Transitional Variational Autoencoders: A Healthcare Application; 2020. ResearchGate preprint.

[24] Walonoski JA, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association. 2017 August;25:230–238.

[25] Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity. In: 2007 IEEE 23rd International Conference on Data Engineering; 2007. p. 106–115.

[26] Rajendran K, Jayabalan M, Rana ME. A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data. International Journal of Computer Science and Network Security. 2017 December;17:172–177.

[27] Dwork C, McSherry F, Nissim K, Smith A. Calibrating Noise to Sensitivity in Private Data Analysis. In: Theory of Cryptography. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 265–284.

[28] McSherry F, Talwar K. Mechanism Design via Differential Privacy. In: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. Washington DC, USA: IEEE Computer Society; 2007. p. 94–103.

[29] Dwork C, Smith A, Steinke T, Ullman J. Exposed! A Survey of Attacks on Private Data. Annual Review of Statistics and Its Application. 2017 March;4:61–84.

[30] Dwork C, Kohli N, Mulligan D. Differential Privacy in Practice: Expose your Epsilons! Journal of Privacy and Confidentiality. 2019 October;9(2).

[31] Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X. PrivBayes: Private Data Release via Bayesian Networks. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York, USA: Association for Computing Machinery; 2014. p. 1423–1434.

[32] Ping H, Stoyanovich J, Howe B. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management. New York, USA: Association for Computing Machinery; 2017. p. 1–5.

[33] Eurostat. International Standard Classification of Education (ISCED) [Internet]; 2020. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php/International_Standard_Classification_of_Education_(ISCED)#Implementation_of_ISCED_2011_.28levels_of_education.29. Accessed 2020-08-07.

[34] Miescke KJ, Liese F. Statistical Decision Theory. 1st ed. New York: Springer; 2008.

[35] Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". The American Statistician. 2019 March;73(sup1):1–19.

[36] Perkonoja K. SynthData: Tools for Generating Synthetic Data in R; 2020. R package version 1.0.1. Available from: https://gitlab.utu.fi/kakype/synthdata.

[37] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2020. Available from: https://www.R-project.org/.

[38] Xin B, Yang W, Geng Y, Chen S, Wang S, Huang L. Private FL-GAN: Differential Privacy Synthetic Data Generation Based on Federated Learning. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020. p. 2927–2931.

[39] Jälkö J, Lagerspetz E, Haukka J, Tarkoma S, Kaski S, Honkela A. Privacy-preserving data sharing via probabilistic modelling; 2020. ArXiv preprint arXiv:1912.04439v3.

[40] Bützow A, Southerington T. "VS: Synteettisen datan anonymiteetti, GDPR sekä laillisuus" [RE: Anonymity of synthetic data, GDPR and legitimacy]; 2020. Received by Katariina Perkonoja. 8 June 2020. Email interview.

Appendices

A R code

```
##############################################################################
## Generating Synthetic Longitudinal Patient Data with the PrivBayes Method ##
##############################################################################

## INFORMATION ##

# This is the source code of the empirical part of Katariina Perkonoja's
# master's thesis 'Generating Synthetic Longitudinal Patient Data with the
# PrivBayes Method'

## LIBRARIES ##

library(SynthData)
library(readr)
library(data.table)
library(ggplot2)
library(tidyr)
library(latex2exp)
library(pheatmap)
library(RColorBrewer)
library(lme4)
library(lmerTest)
library(sjPlot)



## ORIGINAL DATA ##

origdata <- data.table(read_csv("origdata.csv"))

# dropping id column = 1st column
origdata <- origdata[,-1]

# sample size of the original data set
orig_n <- nrow(origdata)


## HIERARCHICAL ENCODING ##

# numeric attributes to be encoded
numeric_attr <- c("age", "bmi", "hba1c_base", "hba1c_fol")

# categorical attribute to be encoded
categ_attr <- "comps"

# bin widths for numerical attributes
lapply(origdata[, numeric_attr, with = F], function (x) ceiling(length(unique(x))/2))

numeric_bins <- list("age" = 38, "bmi" = 184, "hba1c_base"= 59, "hba1c_fol" = 55)

# levels for categorical attribute to be encoded
categ_levels <- c("one", "two", "three", "four", "three", "two", "three", "two",
                  "one", "two", "three", "two", "zero", "one", "two", "one")

categ_levels <- list("comps" = list(categ_levels))


# applying hierarchical encoding
hierdata <- hierencoding(origdata, continvar = numeric_attr, categvar = categ_attr,
                         binwidths = numeric_bins, labellist = categ_levels)



## NETWORK LEARNING ##

# total privacy budgets
epsilon <- c(1, 2, 3)

# controls how the total privacy budgets are divided between
# the exponential mechanism and the Laplace mechanism
beta <- c(0.2, 0.5, 0.8)

# privacy budgets for the exponential mechanism
```

```r
ep1 <- epsilon %*% t(beta)

# privacy budgets for the Laplace mechanism
ep2 <- epsilon %*% t((1-beta))

# theta-usefulness
thetas <- c(1,2,3)

# greating several sets of synthetic data to inspect the randomness and
# simultaneously storing some interesting values, that is, the minimum and maximum
# correlation of HbAic and the maximum between-individual variance

mincor <- c(1,0)
maxcor <- c(-1,0)
maxtau00 <- c(0,0)


# function that returns tau00 from a list of data sets
calc_tau00 <- function (dataset){

  analysisdata <- data.table(dataset)

  analysisdata$Subject <- paste("Patient", 1:nrow(analysisdata))

  idcols <- colnames(analysisdata)[!(colnames(analysisdata) %in%
                                       c("hba1c_base", "hba1c_fol"))]

  analysisdata <- melt(analysisdata, id.vars = idcols)

  summodel <- summary(lmer(value ~ age + gender + dbtype + bmi + (1|Subject),
                           data = analysisdata))

  return(summodel$varcor$Subject[1])

}

# generating the set of 27 synthetic data sets for 10 times

for (iter in 1:10) {

  # initializing while requirement
  indep <- 0

  # requiring that at least in one network hba1c_fol has hba1c_base as parent

  while(indep < 1){

    # different networks
    networks <- list()

    for (i in thetas){

      for (j in 1:length(epsilon)){

        for (k in 1:length(beta)){


          network <- greedybayes(dataset = hierdata,
                                 theta = i,
                                 epsilon1 = ep1[j,k],
                                 epsilon2 = ep2[j,k],
                                 n = orig_n)

          networks <- append(networks, list(network))

          if("hba1c_base" %in% names(network[["hba1c_fol"]]$parents)) {

            indep <- indep + 1

          }

        }
      }
    }
  }

  ## DISTRIBUTION LEARNING ##

  noisydistrs <- list()

  # vectorizing epsilon 2 matrix by rows and multipying it by the number
```

```
  # of different thetas
  vecep2 <- rep(as.vector(t(ep2)), length(thetas))


  for (i in 1:length(networks)){

    noisydistrs[[i]] <- noisycond(network = networks[[i]], epsilon2 = vecep2[i])

  }



  ## DATA SYNTHETIZATION ##

  # all synthetic data sets have the same number of observations as the original so
  # that these data sets can be compared to each other

  synthsize <- orig_n

  synthdatasets <- list()

  for (i in 1:length(noisydistrs)){

    synthdataset <- genfromnoisy(noisyprobs = noisydistrs[[i]],
                                 size = synthsize,
                                 hierardata = hierdata,
                                 epsilon2 = vecep2[i])

    synthdatasets[[i]] <- synthdataset[colnames(origdata)]

    # changing numeric attributes back to numeric
    synthdatasets[[i]][,numeric_attr] <- lapply(synthdatasets[[i]][,numeric_attr],
                                                as.numeric)

  }

  # testing for correlations
  correlations <- unlist(lapply(synthdatasets, function (x) cor(x$hba1c_base,
                                                                x$hba1c_fol)))

  if (min(correlations) < mincor[1]){

    mincor <- c(min(correlations), which.min(correlations))

  }

  if (max(correlations) > maxcor[1]){

    maxcor <- c(max(correlations), which.max(correlations))

  }

  # testing for between-individual variance
  alltau00s <- unlist(lapply(synthdatasets, calc_tau00))

  if (max(alltau00s) > maxtau00[1]){

    maxtau00 <- c(max(alltau00s), which.max(alltau00s))

  }

  # Pause for-loop to inspect the structures of the Bayesian networks
  #and the conditional probabilities
  # Sys.sleep(120)

}

### FOR THE THESIS ###

## COMBINING DATA SETS ##

datasets <- data.table(origdata)
datasets$dataset <- "original"
datasets$epsilon <- NA
datasets$betas <- NA

# adding columns to all synthetic data sets and changing numeric attributes
# to numeric then combining the synthetic data set with the original and other
# synthetic data sets
for (i in 1:length(synthdatasets)){

```

```
  synthdatasets[[i]] <- cbind(synthdatasets[[i]], dataset = paste0("synth",i))

  synthdatasets[[i]]$epsilon <- paste0("epsilon == ",rep(epsilon, each = 3,
                                                         times = 3))[i]

  synthdatasets[[i]]$betas <- paste0("beta == ", rep(beta, times = 9))[i]

  datasets <- rbind(datasets, synthdatasets[[i]])

}

# adding values to variables
datasets[, epsilon := as.factor(epsilon)]
datasets[dataset == "original", betas := "Original"]
datasets[, betas := as.factor(betas)]
datasets$betas <- relevel(datasets$betas, "Original")

datasets[, gender := as.factor(gender)]
datasets[, dbtype := as.factor(dbtype)]
datasets[, comps := as.factor(comps)]
datasets$comps <- relevel(datasets$comps, "no")

# datasets needed for plotting
origep1 <- datasets[dataset == "original"]
origep2 <- datasets[dataset == "original"]
origep3 <- datasets[dataset == "original"]
origep1[, epsilon := "epsilon == 1"]
origep2[, epsilon := "epsilon == 2"]
origep3[, epsilon := "epsilon == 3"]

origepdata <- rbind(origep1, origep2, origep3)

# data sets stratified by the value of theta
theta1data <- datasets[dataset %in% paste0("synth",1:9)]
theta2data <- datasets[dataset %in% paste0("synth",10:18)]
theta3data <- datasets[dataset %in% paste0("synth",19:27)]

# Combining data sets
theta1data_o <- rbind(theta1data, origepdata)
theta2data_o <- rbind(theta2data, origepdata)
theta3data_o <- rbind(theta3data, origepdata)

## DISTRIBUTION COMPARISONS ##

colorpallette <- brewer.pal(4, "Blues")
colorpallette <- c(colorpallette, rep(colorpallette[2:4],2))
colorpallette2 <- colorRampPalette(brewer.pal(name = "Blues", n = 9))(16)

## boxplot ##

compare_boxplot <- function (num_attr, data, thetaval, maintitle) {

  num_attr <- enquo(num_attr)

  ggplot(rbind(data, origepdata), aes(x = dataset, y = !! num_attr, group = betas,
                                      fill = dataset)) +

    stat_boxplot(geom ='errorbar', linetype = 1, width = 0.75) +

    geom_boxplot(width = 0.75, outlier.shape=1, outlier.size = 1.5) +

    labs(x = parse(text = paste0(expression(theta), ' == ', thetaval)), y = "Value",
         title = maintitle) +

    scale_fill_manual(values = colorpallette) +

    scale_x_discrete(labels = parse(text = c("Original",
                                             rep(paste0(expression(beta),' == ',
                                                        c(0.2,0.5,0.8)),3)))) +

    theme_bw() +

    theme(legend.position = "none", plot.title = element_text(hjust = 0.5),
          panel.border = element_blank(), panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          axis.line = element_line(colour = "black"),
          axis.title.x = element_text(hjust = 0.5), text = element_text(size=34),
          axis.text.x = element_text(angle = 45,  size = 28, hjust = 1)) +

    facet_grid( ~ epsilon, scales = "free", labeller = label_parsed) +
```

```
317
318     stat_summary(fun = mean, geom="point", size = 2, aes(fill = dataset),
319                  position = position_dodge(0.75))
320
321 }
322
323 ## barplot ##
324
325 compare_barplot <- function (cat_attr, data, thetaval, maintitle, colorpal) {
326
327   cat_attr <- enquo(cat_attr)
328
329   ggplot(data, aes(x = betas, y = N, group = dataset, fill = !! cat_attr)) +
330
331     geom_bar(colour="black", stat = "identity") +
332
333     labs(x = parse(text = paste0(expression(theta), ' == ', thetaval)), y = "Count",
334          title = maintitle) +
335
336     scale_fill_manual(values = colorpal) +
337
338     scale_x_discrete(labels = function(l) parse(text=l)) +
339
340     theme_bw() +
341
342     theme(legend.position = "top", plot.title = element_text(hjust = 0.5),
343           panel.border = element_blank(), panel.grid.major = element_blank(),
344           panel.grid.minor = element_blank(),
345           axis.line = element_line(colour = "black"),
346           axis.title.x = element_text(hjust = 0.5), text = element_text(size=34),
347           axis.text.x = element_text(angle = 45,  size = 28, hjust = 1, vjust = 1),
348           legend.title = element_blank(), legend.text=element_text(size=20)) +
349
350     facet_grid( ~ epsilon, scales = "free", labeller = label_parsed)
351
352 }
353
354
355 somePDFPath = ".../Perkonoja_Katariina/SyntheticData/figures/univariate.pdf"
356 pdf(file=somePDFPath, width = 14, height = 12)
357
358 # Age
359 compare_boxplot(age, theta1data, 1, "Age")
360 compare_boxplot(age, theta2data, 2, "Age")
361 compare_boxplot(age, theta3data, 3, "Age")
362
363 # BMI
364 compare_boxplot(bmi, theta1data, 1, "BMI")
365 compare_boxplot(bmi, theta2data, 2, "BMI")
366 compare_boxplot(bmi, theta3data, 3, "BMI")
367
368 # hba1c_base
369 compare_boxplot(hba1c_base, theta1data, 1, "HbA1c baseline")
370 compare_boxplot(hba1c_base, theta2data, 2, "HbA1c baseline")
371 compare_boxplot(hba1c_base, theta3data, 3, "HbA1c baseline")
372
373 # hba1c_fol
374 compare_boxplot(hba1c_fol, theta1data, 1, "HbA1c follow-up")
375 compare_boxplot(hba1c_fol, theta2data, 2, "HbA1c follow-up")
376 compare_boxplot(hba1c_fol, theta3data, 3, "HbA1c follow-up")
377
378 # Gender
379 compare_barplot(gender,
380   theta1data_o[, .(.N), by = .(gender, betas, dataset, epsilon)][order(gender)],
381   1, "Gender", colorpallette[-1])
382 compare_barplot(gender,
383   theta2data_o[, .(.N), by = .(gender, betas, dataset, epsilon)][order(gender)],
384   2, "Gender", colorpallette[-1])
385 compare_barplot(gender,
386   theta3data_o[, .(.N), by = .(gender, betas, dataset, epsilon)][order(gender)],
387   3, "Gender", colorpallette[-1])
388
389 # Diabetes type
390 compare_barplot(dbtype,
391   theta1data_o[, .(.N), by = .(dbtype, betas, dataset, epsilon)][order(dbtype)],
392   1, "Type of diabetes", colorpallette[-1])
393 compare_barplot(dbtype,
394   theta2data_o[, .(.N), by = .(dbtype, betas, dataset, epsilon)][order(dbtype)],
395   2, "Type of diabetes", colorpallette[-1])
396 compare_barplot(dbtype,
```

```
  theta3data_o[, .(.N), by = .(dbtype, betas, dataset, epsilon)][order(dbtype)],
  3, "Type of diabetes", colorpallette[-1])

# Complication types
compare_barplot(comps,
  theta1data_o[, .(.N), by = .(comps, betas, dataset, epsilon)][order(comps)],
  1, "Type of complications", colorpallette2)
compare_barplot(comps,
  theta2data_o[, .(.N), by = .(comps, betas, dataset, epsilon)][order(comps)],
  2, "Type of complications", colorpallette2)
compare_barplot(comps,
  theta3data_o[, .(.N), by = .(comps, betas, dataset, epsilon)][order(comps)],
  3, "Type of complications", colorpallette2)


dev.off()


## heatmap ##

get_lower_tri<-function(cormat){

  cormat[upper.tri(cormat)] <- NA

  return(cormat)
}

corheatmap <- function(data1, data2) {

  res1 <- cor(data1)
  res2 <- cor(data2)

  heatdata <- res2-res1
  heatdata <- get_lower_tri( heatdata)
  heatdata <- reshape2::melt(heatdata, na.rm = T)


  ggplot(heatdata, aes(Var1, Var2)) +

    geom_tile(aes(fill = value), color = "black") +

    geom_text(aes(label = sprintf("%.2f", value)), size = 8) +

    scale_fill_gradientn(colours = c("darkblue", "blue", "white", "red", "darkred"),
                         values = scales::rescale(c(-2,-1,0,1,2)), guide="colorbar",
                         limits = c(-2,2),
                         name = expression(r[italic(s)] - r[italic(o)])) +

    theme_classic() +

    scale_x_discrete(labels = c("Age", "BMI", "HbA1c\nbaseline", "HbA1c\nfollow-up")) +

    scale_y_discrete(labels = c("Age", "BMI", "HbA1c\nbaseline", "HbA1c\nfollow-up")) +

    theme(axis.title = element_blank(), text = element_text(size=28),
          axis.text.x = element_text(size = 28), axis.text.y = element_text(size = 28),
          legend.key.size = unit(1.5, "cm"), plot.title = element_text(hjust = 0.5),
          legend.title = element_text(size = 28))

}
# original data
orig_num <- origdata[, .SD, .SDcols = numeric_attr]

# title parameters
titlepars <- cbind(theta = rep(1:3, each = 9), eps = rep(1:3, each = 3, times = 3),
                   beta = rep(c(0.2,0.5,0.8), times = 9))


# Cramer's V
cv.test = function(x,y) {

  CV = sqrt(chisq.test(x, y, correct=FALSE)$statistic /
              (length(x) * (min(length(unique(x)),length(unique(y))) - 1)))

  return(as.numeric(CV))

}

#  Cramers V for a data set of nominal variables
cv.mat <- function(dataset) {
```

```
  dataset <- as.data.frame(dataset)

  resultm <- matrix(nrow = ncol(dataset), ncol = ncol(dataset))
  colnames(resultm) <- colnames(dataset)
  rownames(resultm) <- colnames(dataset)

  for (var1 in 1:ncol(dataset)) {

    for (var2 in 1:ncol(dataset)) {

      resultm[var1,var2] <- cv.test(dataset[,var1], dataset[, var2])

    }
  }

  return(resultm)

}

# Cramer's V heatmap function
cvheatmap <- function(data1, data2) {

  res1 <- cv.mat(data1)
  res2 <- cv.mat(data2)

  heatdata <- res2-res1
  heatdata <- get_lower_tri(heatdata)
  heatdata <- reshape2::melt(heatdata, na.rm = T)


  ggplot(heatdata, aes(Var1, Var2)) +

    geom_tile(aes(fill = value), color = "black") +

    geom_text(aes(label = sprintf("%.2f", value)), size = 8) +

    scale_fill_gradientn(colours = c("blue", "white", "red"),
                         values = scales::rescale(c(-1,0,1)), guide="colorbar",
                         limits = c(-1,1),
                         name = expression(phi[italic(cs)] - phi[italic(co)])) +

    theme_classic() +

    scale_x_discrete(labels = c("Gender", "Type of diabetes", "Complications")) +

    scale_y_discrete(labels = c("Gender", "Type of\ndiabetes", "Compli-\ncations")) +

    theme(axis.title = element_blank(), text = element_text(size=28),
          axis.text.x = element_text(size = 28), axis.text.y = element_text(size = 28),
          legend.key.size = unit(1.5, "cm"), plot.title = element_text(hjust = 0.5),
          legend.title = element_text(size = 28))

}


somePDFPath = ".../Perkonoja_Katariina/SyntheticData/figures/heatmaps.pdf"
pdf(file=somePDFPath, width = 14, height = 12)

# correlation heatmaps
for (i in 1:27){

  plot <- corheatmap(orig_num, datasets[dataset == paste0("synth",i), .SD,
                                        .SDcols = numeric_attr])

  print(plot + ggtitle(label = bquote(paste(theta, ' = ', .(titlepars[i,"theta"]),', ',
                                            epsilon, ' = ', .(titlepars[i,"eps"]), ', ',
                                            beta, ' = ', .(titlepars[i,"beta"])))))

}


categ_attr <- colnames(origdata)[!(colnames(origdata) %in% numeric_attr)]
orig_cat <- datasets[dataset == "original", .SD, .SDcols = categ_attr]


# Cramer's V heatmaps
for (i in 1:27){

  plot <- cvheatmap(orig_cat, datasets[dataset == paste0("synth",i), .SD,
                                       .SDcols = categ_attr])
```

```
559
560   print(plot + ggtitle(label = bquote(paste(theta, ' = ', .(titlepars[i,"theta"]),', ',
561                                             epsilon, ' = ', .(titlepars[i,"eps"]), ', ',
562                                             beta, ' = ', .(titlepars[i,"beta"])))))
563 }
564
565 dev.off()
566
567
568 ## individual trajectories ##
569
570 # labeller for facet_wrap
571 new_labeller <- list()
572 new_labeller[["original"]] <- "Original"
573
574 for (i in 2:28) {
575
576
577   new_labeller[[i]] <-  bquote(paste(theta, ' = ', .(titlepars[i-1,"theta"]),', ',
578                                      epsilon, ' = ', .(titlepars[i-1,"eps"]), ', ',
579                                      beta, ' = ', .(titlepars[i-1,"beta"])))
580
581 }
582
583 names(new_labeller) <- c("original", paste0("synth",1:27))
584
585 # new labeller function
586 new_labellerf <- function(variable,value){
587
588   return(new_labeller[value])
589
590 }
591
592 # function which plots the individual trajectories
593
594 plotind <- function(dataset){
595
596   plotdata <- data.table(dataset)
597
598   plotdata$Subject <- rep(paste("Patient", 1:nrow(origdata)),
599                           times = length(unique(plotdata$dataset)))
600
601   plotdata[, Difference := hba1c_fol - hba1c_base, by = .(Subject, dataset)]
602
603   plotdata <- melt(plotdata, id.vars = c("dataset", "Subject", "Difference"))
604
605   ggplot(data = plotdata, aes(x = variable, y = value, group = Subject,
606                               colour = Difference)) +
607
608     geom_point(size = 2) +
609
610     theme_bw() +
611
612     geom_line() +
613
614     facet_wrap(~ dataset, ncol=4, labeller = new_labellerf) +
615
616     scale_color_gradientn(colours = c("darkblue","blue", "white", "red","darkred"),
617                           values = scales::rescale(c(-170,-85,0,85,170)),
618                           guide="colorbar", limits = c(-170,170),
619                           breaks = c(-170,-85,0,85,170),
620                           name = "Follow-up\n- baseline\n") +
621
622     theme(axis.title = element_blank(), text = element_text(size=34),
623           axis.text.x = element_text(size = 28), axis.text.y = element_text(size = 28),
624           legend.key.size = unit(1.5, "cm"), plot.title = element_text(hjust = 0.5),
625           legend.title = element_text(size = 30)) +
626
627     labs(x= 'Repeated measurement', y = 'Value') +
628
629     scale_x_discrete(labels = c("HbA1c\nbaseline", "HbA1c\nfollow-up"))
630
631 }
632
633
634 somePDFPath = ".../Perkonoja_Katariina/SyntheticData/figures/indiv.pdf"
635 pdf(file=somePDFPath, width = 18, height = 12)
636
637
638 for (i in seq(1,25,3)){
639
```

```r
  print(plotind(datasets[dataset %in% c("original", paste0("synth",i:(i+2))),
                         .(hba1c_base, hba1c_fol, dataset)]))

}


dev.off()

## LMM ##

# function which creates the result table for LMM

tableLMM <- function(dataset, ind = c()){

  analysisdata <- data.table(dataset)

  analysisdata$Subject <- rep(paste("Patient", 1:nrow(origdata)),
                              times = length(unique(analysisdata$dataset)))

  idcols <- colnames(analysisdata)[!(colnames(analysisdata) %in%
                                       c("hba1c_base", "hba1c_fol"))]

  analysisdata <- melt(analysisdata, id.vars = idcols)

  colnames(analysisdata) <- c("Age", "Gender", "Type of Diabetes", "BMI", "Dataset",
                              "Subject", "variable", "HbA1c")

  models <- list()

  for (data in unique(analysisdata$Dataset)){

    models[[data]] <- lmer(HbA1c ~ Age + Gender + `Type of Diabetes` + BMI + (1|Subject)
        ,
                           data = analysisdata[Dataset == data])

  }

  tableLMM <- tab_model(models[[1]], models[[2]], models[[3]], models[[4]],
                        dv.labels = c("HbA1c original", paste0("HbA1c synthetic ", ind))
                          )

  return(tableLMM)

}

# formatting data sets to create LMM tables
datasets2 <- data.table(datasets)
datasets2 <- datasets2[, c("epsilon", "betas", "comps") := NULL]

# printing out LMM tables
for (i in seq(1,25,3)){

  print(tableLMM(datasets2[dataset %in% c("original", paste0("synth",i:(i+2)))], i:(i+2)
      ))

}
```

B Constructed private Bayesian networks

Table 8: Tables present the structures of the Bayesian networks for the generated synthetic data sets. The domain size of the parent is indicated by a superscript.

(8.1) The structure of the Bayesian network for **synthetic data set 1**, produced with the following parameters: $\theta = 1$, $\varepsilon = 1$, $\beta = 0.2$. The degree of the network is three.

j	X_j	Π_j
1	Age	$\emptyset$
2	Type of diabetes	{Age^{75}}
3	Gender	{Type of diabetes^{2}, Age^{38}}
4	Complications	{Type of diabetes^{2}, Age^{3}, Gender^{2}}
5	HbA1c baseline	$\emptyset$
6	BMI	$\emptyset$
7	HbA1c follow-up	$\emptyset$

(8.2) The structure of the Bayesian network for **synthetic data set 2**, produced with the following parameters: $\theta = 1$, $\varepsilon = 1$, $\beta = 0.5$. The degree of the network is two.

j	X_j	Π_j
1	HbA1c follow-up	$\emptyset$
2	Type of diabetes	{HbA1c follow-up^{55}}
3	Gender	{HbA1c follow-up^{28}, Type of diabetes^{2}}
4	Complications	{Gender^{2}, Type of diabetes^{2}}
5	Age	$\emptyset$
6	HbA1c baseline	$\emptyset$
7	BMI	$\emptyset$

(8.3) The structure of the Bayesian network for **synthetic data set 3**, produced with the following parameters: $\theta = 1$, $\varepsilon = 1$, $\beta = 0.8$. The degree of the network is two.

j	X_j	Π_j
1	HbA1c baseline	$\emptyset$
2	Type of diabetes	{HbA1c baseline^{15}}
3	Gender	{Type of diabetes^{2}, HbA1c baseline^{8}}
4	Complications	{Type of diabetes^{2}}
5	Age	$\emptyset$
6	HbA1c follow-up	$\emptyset$
7	BMI	$\emptyset$

(8.4) The structure of the Bayesian network for **synthetic data set 4**, produced with the following parameters: $\theta = 1$, $\varepsilon = 2$, $\beta = 0.2$. The degree of the network is three.

j	X_j	Π_j
1	HbA1c follow-up	$\emptyset$
2	Type of diabetes	{HbA1c follow-up^{110}}
3	Gender	{Type of diabetes^2, HbA1c follow-up^{55}}
4	Complications	{Type of diabetes^2 , HbA1c follow-up^4, Gender^2}
5	Age	{Type of diabetes^2,Gender^2}
6	HbA1c baseline	{Type of diabetes^2}
7	BMI	$\emptyset$

(8.5) The structure of the Bayesian network for **synthetic data set 5**, produced with the following parameters: $\theta = 1$, $\varepsilon = 2$, $\beta = 0.5$. The degree of the network is two.

j	X_j	Π_j
1	Complications	$\emptyset$
2	Type of diabetes	{Complications^{16}}
3	Gender	{Type of diabetes^2, Complications^{16}}
4	Age	{Type of diabetes^2}
5	HbA1c baseline	{Type of diabetes^2}
6	HbA1c follow-up	{Type of diabetes^2}
7	BMI	$\emptyset$

(8.6) The structure of the Bayesian network for **synthetic data set 6**, produced with the following parameters: $\theta = 1$, $\varepsilon = 2$, $\beta = 0.8$. The degree of the network is two.

j	X_j	Π_j
1	Gender	$\emptyset$
2	Type of diabetes	{Gender^2}
3	Complications	{Type of diabetes^2, Gender^2}
4	Age	$\emptyset$
5	HbA1c baseline	$\emptyset$
6	HbA1c follow-up	$\emptyset$
7	BMI	$\emptyset$

(8.7) The structure of the Bayesian network for **synthetic data set 7**, produced with the following parameters: $\theta = 1$, $\varepsilon = 3$, $\beta = 0.2$. The degree of the network is three.

j	X_j	Π_j
1	Gender	$\emptyset$
2	Type of diabetes	{Gender^2}
3	Age	{Gender^2, Type of diabetes^2}
4	Complications	{Gender^2, Age^5, Type of diabetes^2}
5	HbA1c baseline	{Gender^2, Type of diabetes^2}
6	HbA1c follow-up	{HbA1c baseline^4}
7	BMI	$\emptyset$

(8.8) The structure of the Bayesian network for **synthetic data set 8**, produced with the following parameters: $\theta = 1$, $\varepsilon = 3$, $\beta = 0.5$. The degree of the network is two.

j	X_j	Π_j
1	BMI	$\emptyset$
2	Type of diabetes	{BMI^{92}}
3	Age	{Type of diabetes^2, BMI^2}
4	Gender	{Age^2, BMI^{46}}
5	Complications	{BMI^2, Age^{10}}
6	HbA1c follow-up	{Type of diabetes^2}
7	HbA1c baseline	{Type of diabetes^2}

(8.9) The structure of the Bayesian network for **synthetic data set 9**, produced with the following parameters: $\theta = 1$, $\varepsilon = 3$, $\beta = 0.8$. The degree of the network is three.

j	X_j	Π_j
1	BMI	$\emptyset$
2	Type of diabetes	{BMI^{46}}
3	Gender	{BMI^{23}, Type of diabetes^2}
4	Complications	{BMI^2, Gender^2, Type of diabetes^2}
5	HbA1c baseline	$\emptyset$
6	Age	$\emptyset$
7	HbA1c follow-up	$\emptyset$

(8.10) The structure of the Bayesian network for **synthetic data set 10**, produced with the following parameters: $\theta = 2$, $\varepsilon = 1$, $\beta = 0.2$. The degree of the network is two.

j	X_j	Π_j
1	Complications	$\emptyset$
2	Type of Diabetes	{Complications^{16}}
3	Gender	{Complications^{16}, Type of diabetes^{2}}
4	BMI	$\emptyset$
5	Age	$\emptyset$
6	HbA1c follow-up	$\emptyset$
7	HbA1c baseline	$\emptyset$

(8.11) The structure of the Bayesian network for **synthetic data set 11**, produced with the following parameters: $\theta = 2$, $\varepsilon = 1$, $\beta = 0.5$. The degree of the network is two.

j	X_j	Π_j
1	HbA1c follow-up	$\emptyset$
2	Type of diabetes	{HbA1c follow-up^{28}}
3	Gender	{Type of diabetes^{2}, HbA1c follow-up^{14}}
4	Complications	{Type of diabetes^{2}}
5	BMI	$\emptyset$
6	HbA1c baseline	$\emptyset$
7	Age	$\emptyset$

(8.12) The structure of the Bayesian network for **synthetic data set 12**, produced with the following parameters: $\theta = 2$, $\varepsilon = 1$, $\beta = 0.8$. The degree of the network is one.

j	X_j	Π_j
1	Gender	$\emptyset$
2	Type of diabetes	{Gender^{2}}
3	BMI	$\emptyset$
4	Complications	$\emptyset$
5	Age	$\emptyset$
6	HbA1c baseline	$\emptyset$
7	HbA1c follow-up	$\emptyset$

(8.13) The structure of the Bayesian network for **synthetic data set 13**, produced with the following parameters: $\theta = 2$, $\varepsilon = 2$, $\beta = 0.2$. The degree of the network is three.

j	X_j	Π_j
1	Complications	$\emptyset$
2	Type of diabetes	{Complications^{16}}
3	Age	{Type of diabetes^{2}}
4	Gender	{Age^{3}, Type of diabetes^{2}, Complications^{16}}
5	BMI	$\emptyset$
6	HbA1c baseline	$\emptyset$
7	HbA1c follow-up	$\emptyset$

(8.14) The structure of the Bayesian network for **synthetic data set 14**, produced with the following parameters: $\theta = 2$, $\varepsilon = 2$, $\beta = 0.5$. The degree of the network is two.

j	X_j	Π_j
1	Type of diabetes	$\emptyset$
2	Complications	{Type of diabetes^{2}}
3	Gender	{Complications^{16}, Type of diabetes^{2}}
4	Age	$\emptyset$
5	HbA1c baseline	$\emptyset$
6	HbA1c follow-up	$\emptyset$
7	BMI	$\emptyset$

(8.15) The structure of the Bayesian network for **synthetic data set 15**, produced with the following parameters: $\theta = 2$, $\varepsilon = 2$, $\beta = 0.8$. The degree of the network is two.

j	X_j	Π_j
1	BMI	$\emptyset$
2	Type of diabetes	{BMI^{23}}
3	Gender	{Type of diabetes^{2}, BMI^{12}}
4	Complications	{Type of diabetes^{2}}
5	HbA1c baseline	$\emptyset$
6	Age	$\emptyset$
7	HbA1c follow-up	$\emptyset$

(8.16) The structure of the Bayesian network for **synthetic data set 16**, produced with the following parameters: $\theta = 2$, $\varepsilon = 3$, $\beta = 0.2$. The degree of the network is two.

j	X_j	Π_j
1	HbA1c follow-up	$\emptyset$
2	Type of diabetes	{HbA1c follow-up^{110}}
3	Complications	{Type of diabetes^{2}, HbA1c follow-up^{7}}
4	Gender	{Complications^{5}, HbA1c follow-up^{28}}
5	Age	{Type of diabetes^{2}}
6	HbA1c baseline	{Type of diabetes^{2}}
7	BMI	$\emptyset$

(8.17) The structure of the Bayesian network for **synthetic data set 17**, produced with the following parameters: $\theta = 2$, $\varepsilon = 3$, $\beta = 0.5$. The degree of the network is two.

j	X_j	Π_j
1	HbA1c follow-up	$\emptyset$
2	Type of diabetes	{HbA1c follow-up^{55}}
3	Complications	{HbA1c follow-up^{4}, Type of diabetes^{2}}
4	Gender	{Type of diabetes^{2}, HbA1c follow-up^{28}}
5	Age	{Type of diabetes^{2}}
6	HbA1c baseline	$\emptyset$
7	BMI	$\emptyset$

(8.18) The structure of the Bayesian network for **synthetic data set 18**, produced with the following parameters: $\theta = 2$, $\varepsilon = 3$, $\beta = 0.8$. The degree of the network is two.

j	X_j	Π_j
1	Age	$\emptyset$
2	Type of diabetes	{Age^{19}}
3	Gender	{Type of diabetes^{2}, Age^{10}}
4	Complications	{Gender^{2}, Type of diabetes^{2}}
5	HbA1c baseline	$\emptyset$
6	BMI	$\emptyset$
7	HbA1c follow-up	$\emptyset$

(8.19) The structure of the Bayesian network for **synthetic data set 19**, produced with the following parameters: $\theta = 3$, $\varepsilon = 1$, $\beta = 0.2$. The degree of the network is three.

j	X_j	Π_j
1	Type of diabetes	$\emptyset$
2	Age	$\emptyset$
3	Complications	{Type of diabetes^2, Age^2}
4	Gender	{Age^3, Complications^5, Type of diabetes^2}
5	HbA1c baseline	$\emptyset$
6	BMI	$\emptyset$
7	HbA1c follow-up	$\emptyset$

(8.20) The structure of the Bayesian network for **synthetic data set 20**, produced with the following parameters: $\theta = 3$, $\varepsilon = 1$, $\beta = 0.5$. The degree of the network is two.

j	X_j	Π_j
1	BMI	$\emptyset$
2	Type of diabetes	{BMI^{12}}
3	Gender	{Type of diabetes^2, BMI^6}
4	Complications	{Type of diabetes^2}
5	HbA1c follow-up	$\emptyset$
6	HbA1c baseline	$\emptyset$
7	Age	$\emptyset$

(8.21) The structure of the Bayesian network for **synthetic data set 21**, produced with the following parameters: $\theta = 3$, $\varepsilon = 1$, $\beta = 0.8$. The degree of the network is two.

j	X_j	Π_j
1	HbA1c follow-up	$\emptyset$
2	Type of diabetes	{HbA1c follow-up^7}
3	Gender	{Type of diabetes^2, HbA1c follow-up^4}
4	BMI	$\emptyset$
5	Complications	$\emptyset$
6	Age	$\emptyset$
7	HbA1c baseline	$\emptyset$

(8.22) The structure of the Bayesian network for **synthetic data set 22**, produced with the following parameters: $\theta = 3$, $\varepsilon = 2$, $\beta = 0.2$. The degree of the network is two.

j	X_j	Π_j
1	Type of diabetes	$\emptyset$
2	Complications	{Type of diabetes^2}
3	Gender	{Complications^{16}, Type of diabetes^2}
4	BMI	$\emptyset$
5	HbA1c follow-up	$\emptyset$
6	HbA1c baseline	$\emptyset$
7	Age	$\emptyset$

(8.23) The structure of the Bayesian network for **synthetic data set 23**, produced with the following parameters: $\theta = 3$, $\varepsilon = 2$, $\beta = 0.5$. The degree of the network is two.

j	X_j	Π_j
1	HbA1c follow-up	$\emptyset$
2	Type of diabetes	{HbA1c follow-up^{28}}
3	Gender	{Type of diabetes^2, HbA1c follow-up^{14}}
4	Complications	{Type of diabetes^2, HbA1c follow-up^2}
5	BMI	$\emptyset$
6	HbA1c baseline	$\emptyset$
7	Age	$\emptyset$

(8.24) The structure of the Bayesian network for **synthetic data set 24**, produced with the following parameters: $\theta = 3$, $\varepsilon = 2$, $\beta = 0.8$. The degree of the network is two.

j	X_j	Π_j
1	BMI	$\emptyset$
2	Type of diabetes	{BMI^{12}}
3	Gender	{BMI^6, Type of diabetes^2}
4	Complications	{Type of diabetes^2}
5	Age	$\emptyset$
6	HbA1c follow-up	$\emptyset$
7	HbA1c baseline	$\emptyset$

(8.25) The structure of the Bayesian network for **synthetic data set 25**, produced with the following parameters: $\theta = 3$, $\varepsilon = 3$, $\beta = 0.2$. The degree of the network is three.

j	X_j	Π_j
1	HbA1c follow-up	$\emptyset$
2	Type of diabetes	{HbA1c follow-up^{55}}
3	Age	{Type of diabetes^{2}}
4	Gender	{HbA1c follow-up^{4}, Age^{19}}
5	Complications	{Age^{3}, Gender^{2}, Type of diabetes^{2}}
6	HbA1c baseline	$\emptyset$
7	BMI	$\emptyset$

(8.26) The structure of the Bayesian network for **synthetic data set 26**, produced with the following parameters: $\theta = 3$, $\varepsilon = 3$, $\beta = 0.5$. The degree of the network is two.

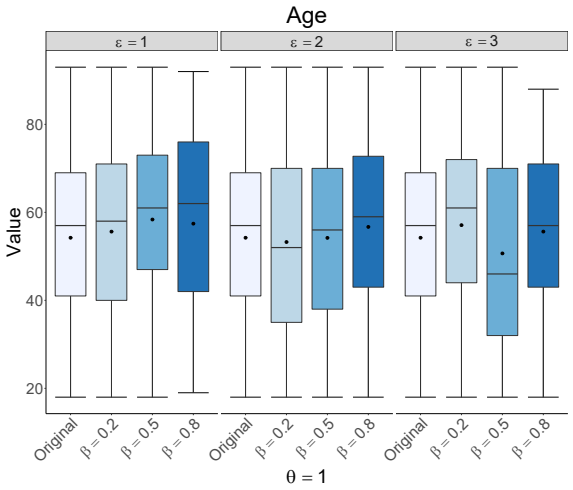
j	X_j	Π_j
1	Gender	$\emptyset$
2	Type of diabetes	{Gender^{2}}
3	Complications	{Type of diabetes^{2}, Gender^{2}}
4	BMI	$\emptyset$
5	HbA1c baseline	$\emptyset$
6	HbA1c follow-up	$\emptyset$
7	Age	$\emptyset$

(8.27) The structure of the Bayesian network for **synthetic data set 27**, produced with the following parameters: $\theta = 3$, $\varepsilon = 3$, $\beta = 0.8$. The degree of the network is two.

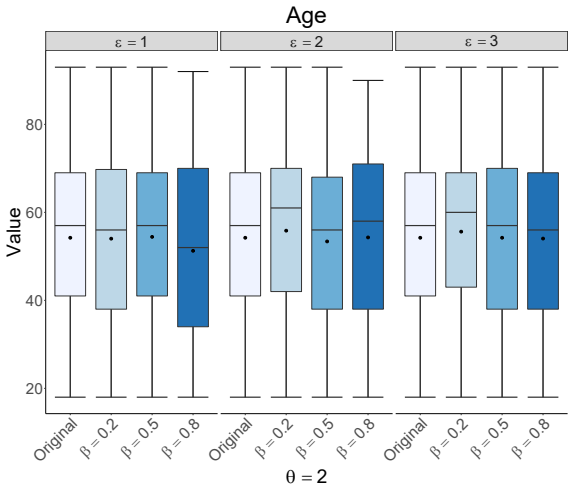
j	X_j	Π_j
1	Type of diabetes	$\emptyset$
2	Complications	{Type of diabetes^{2}}
3	Gender	{Complications^{5}, Type of diabetes^{2}}
4	BMI	$\emptyset$
5	HbA1c follow-up	$\emptyset$
6	HbA1c baseline	$\emptyset$
7	Age	$\emptyset$

C Univariate distributions

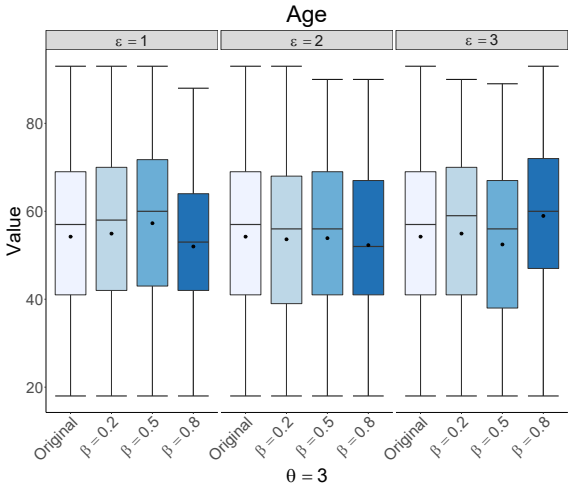
Figure 11



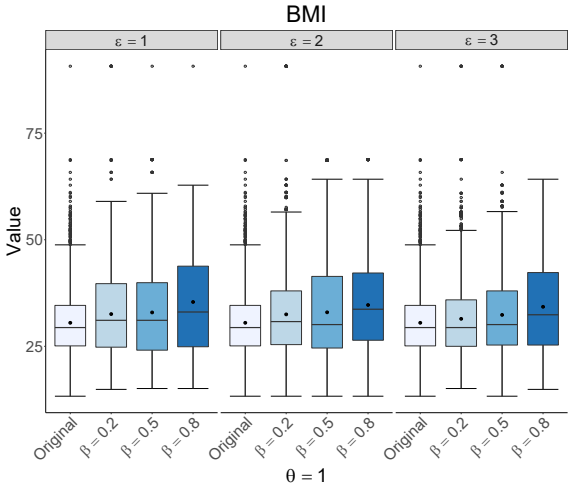
(11.1)



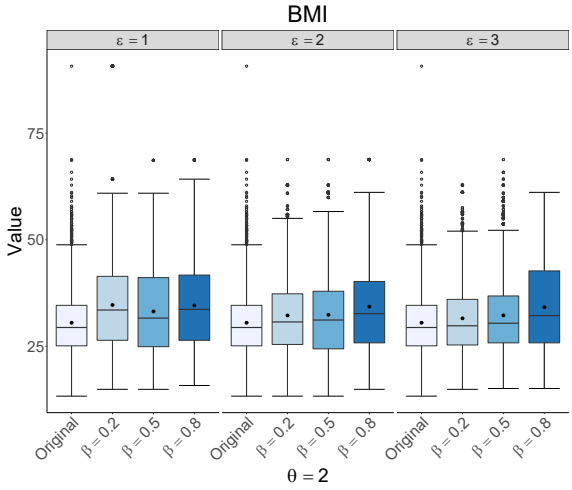
(11.2)



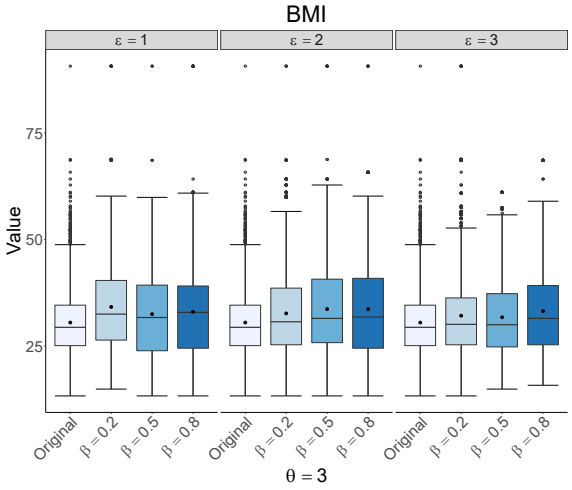
(11.3)



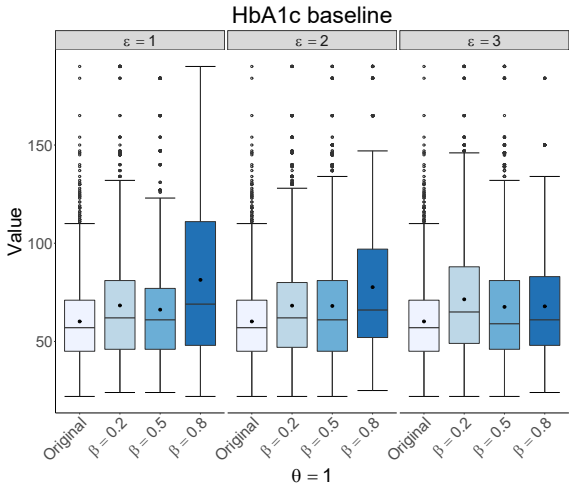
(11.4)



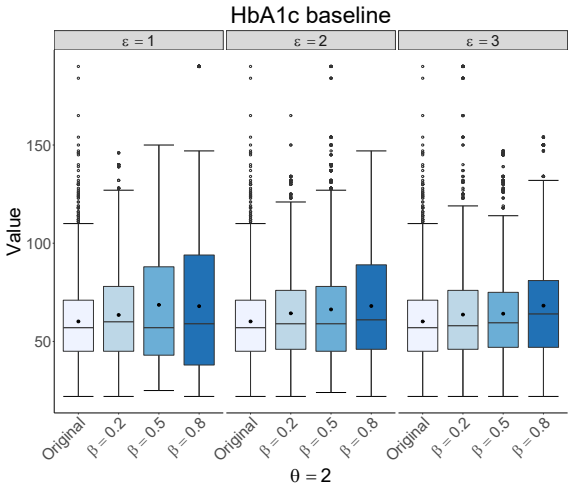
(11.5)



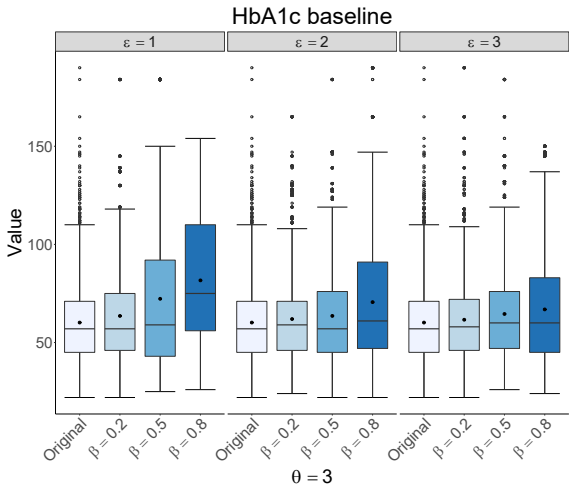
(11.6)



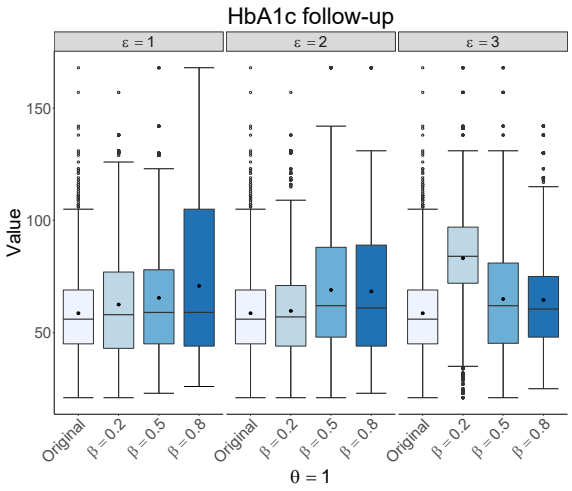
(11.7)



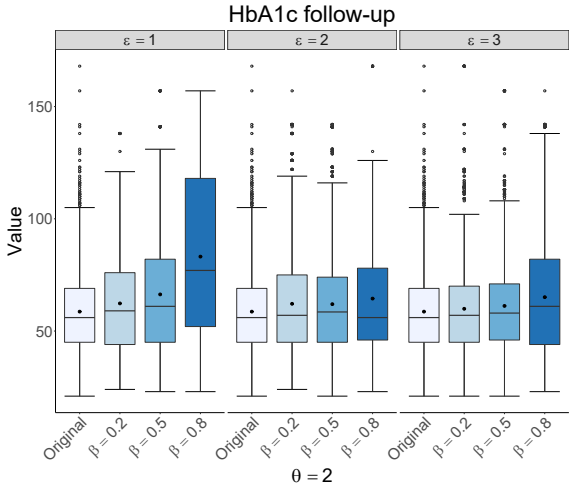
(11.8)



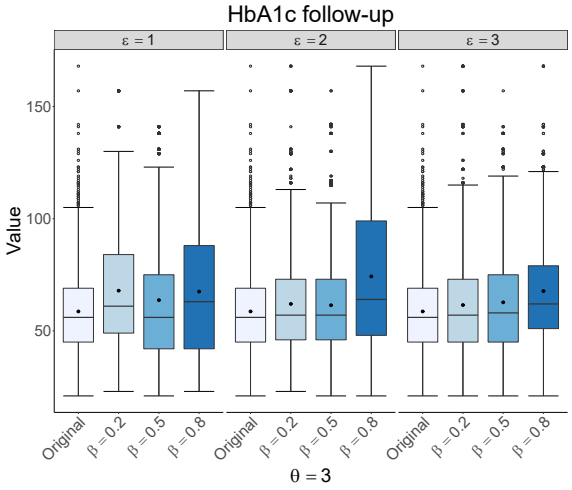
(11.9)



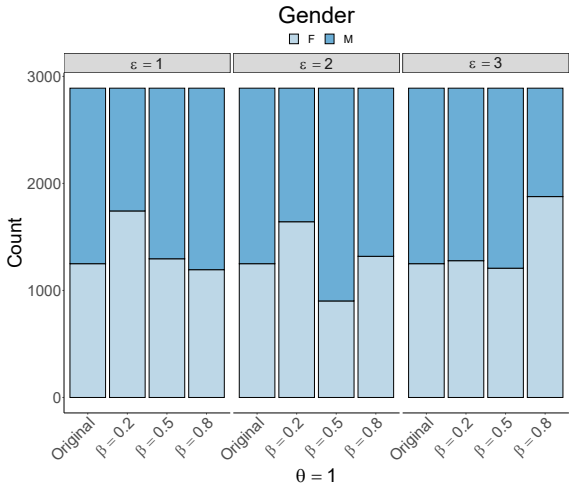
(11.10)



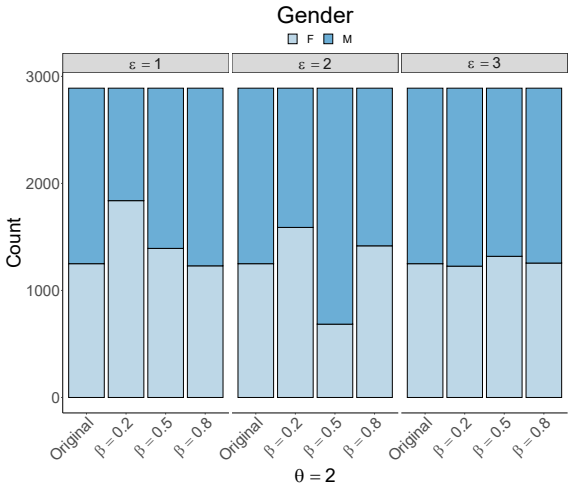
(11.11)



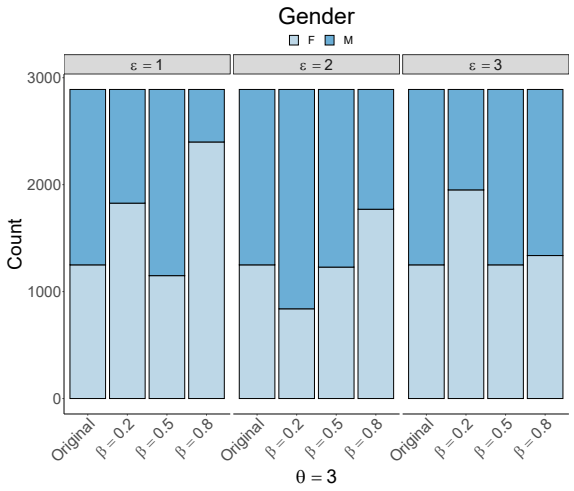
(11.12)



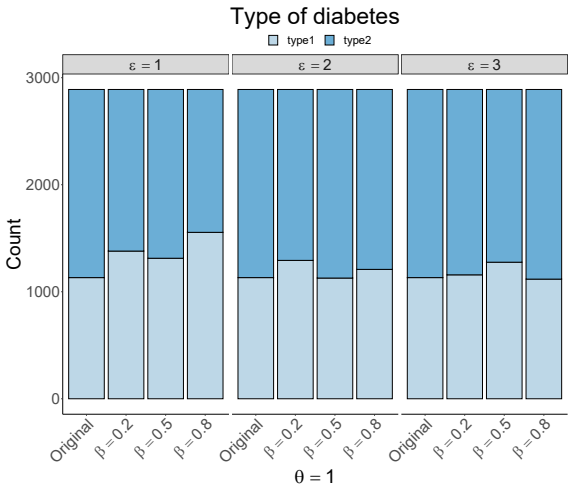
(11.13)



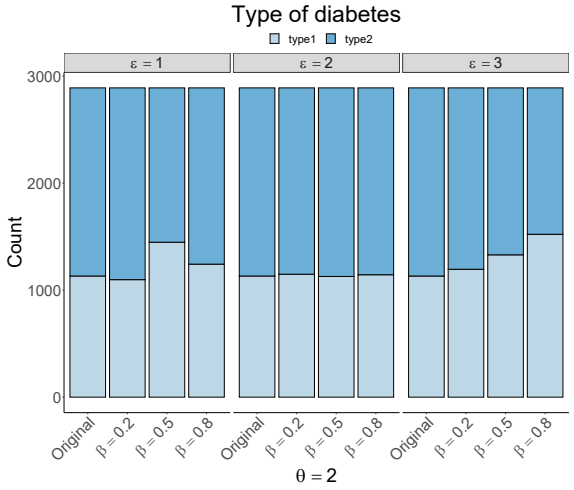
(11.14)



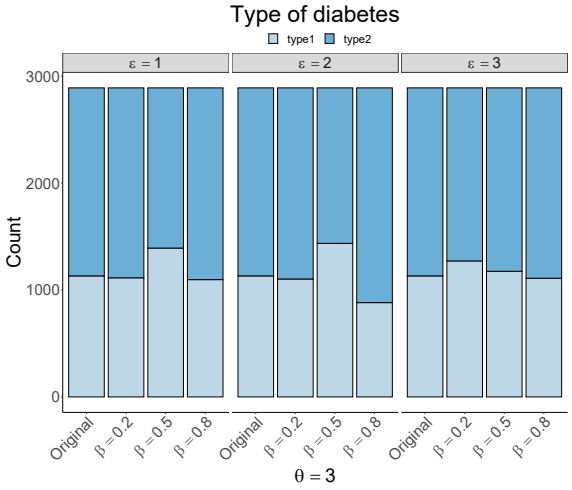
(11.15)



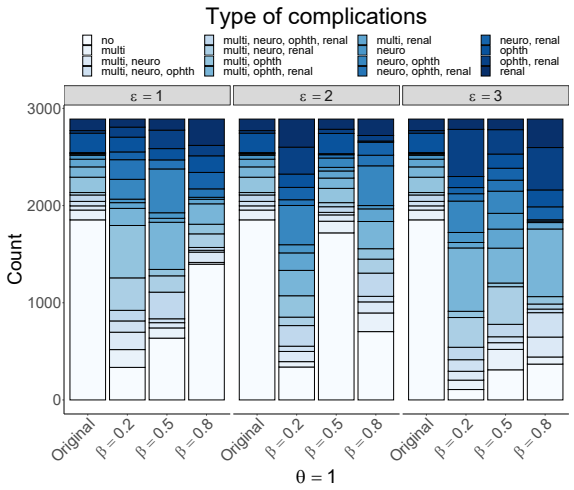
(11.16)



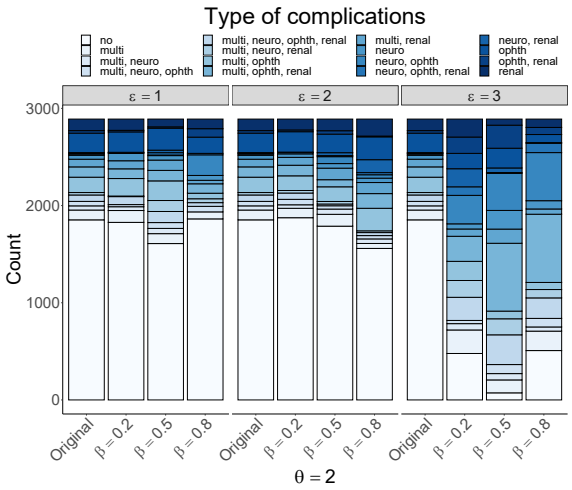
(11.17)



(11.18)



(11.19)



(11.20)

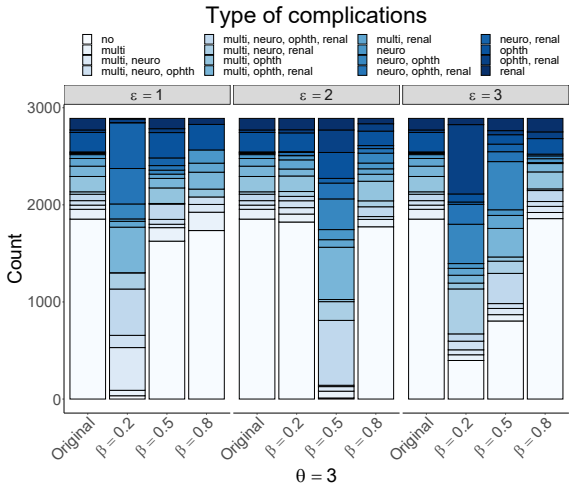
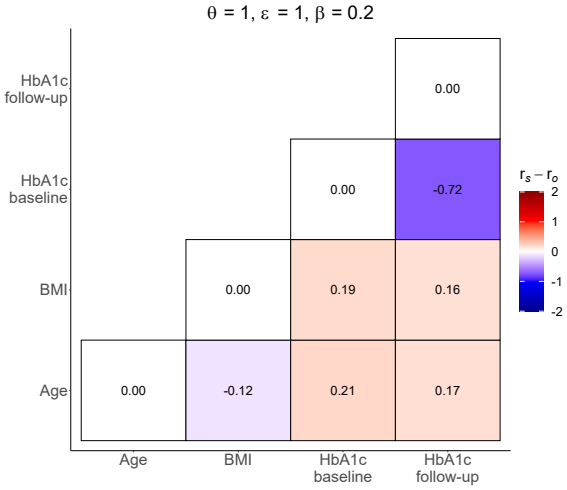


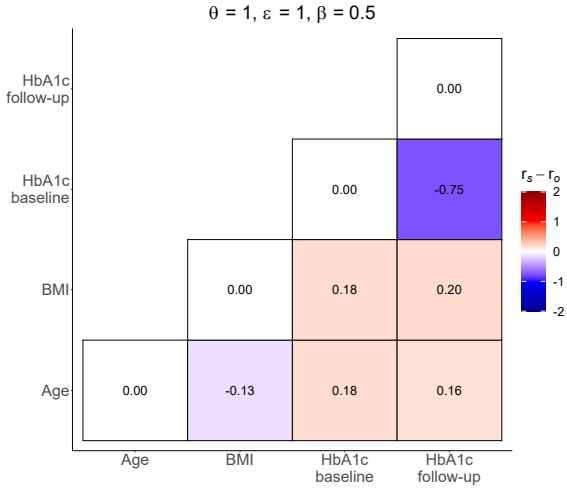
Figure 11

D Heat maps

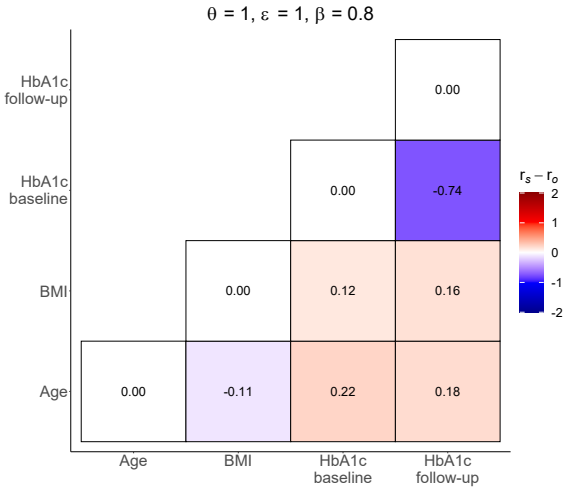
Figure 12: For computational and programming reasons, negative coefficients appear on the diagonal of Cramer's V coefficients, although the values on the diagonal are actually all zeros.



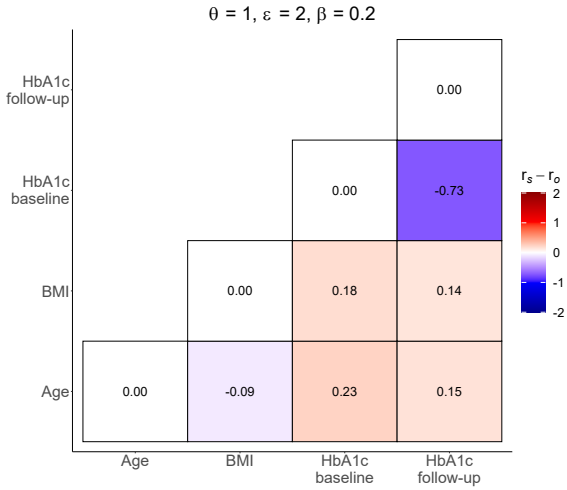
(12.1)



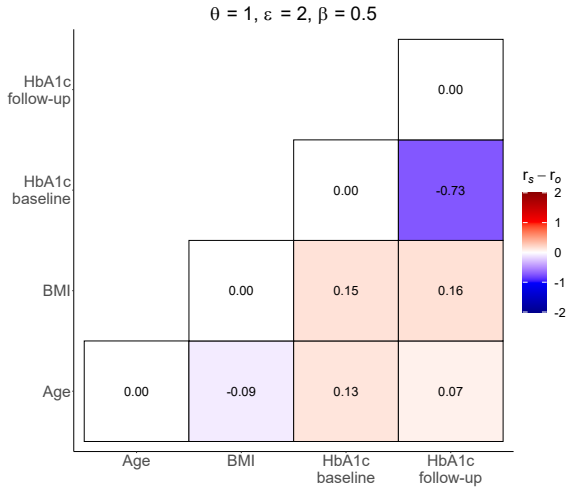
(12.2)



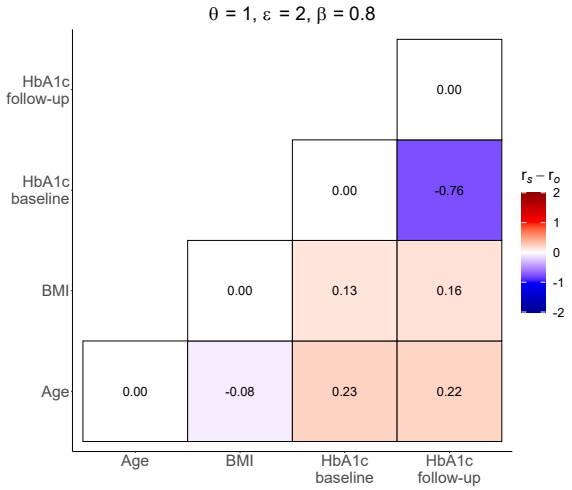
(12.3)



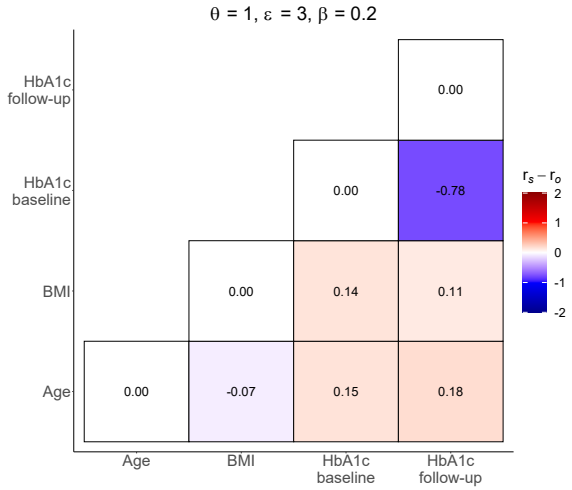
(12.4)



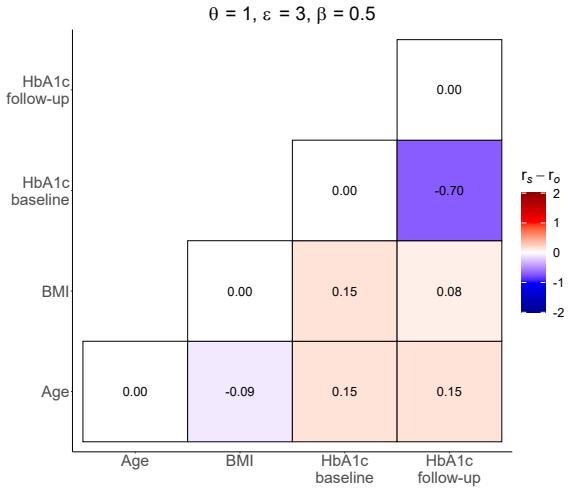
(12.5)



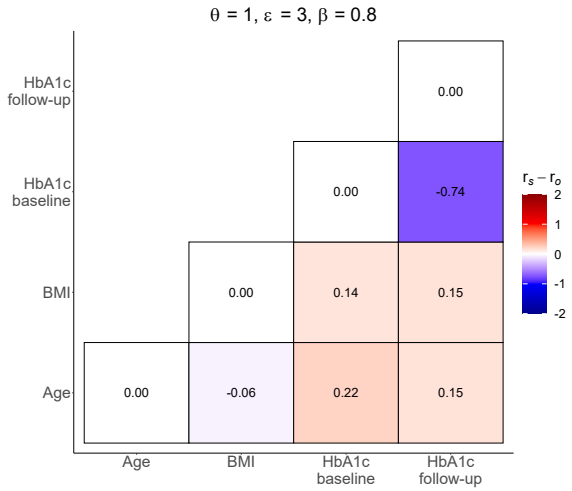
(12.6)



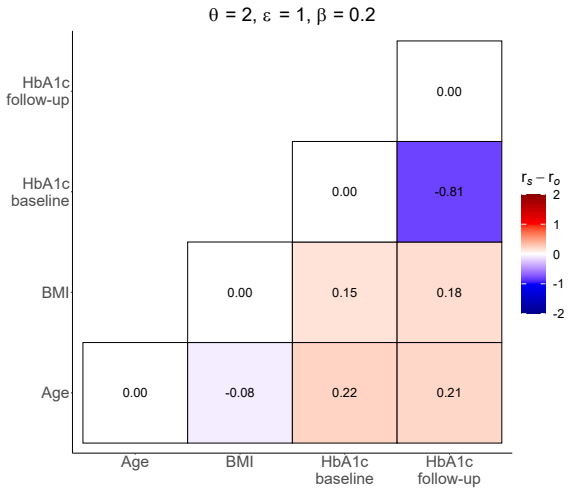
(12.7)



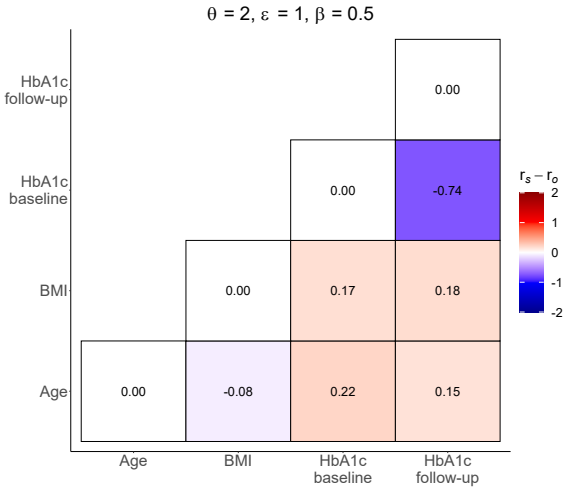
(12.8)



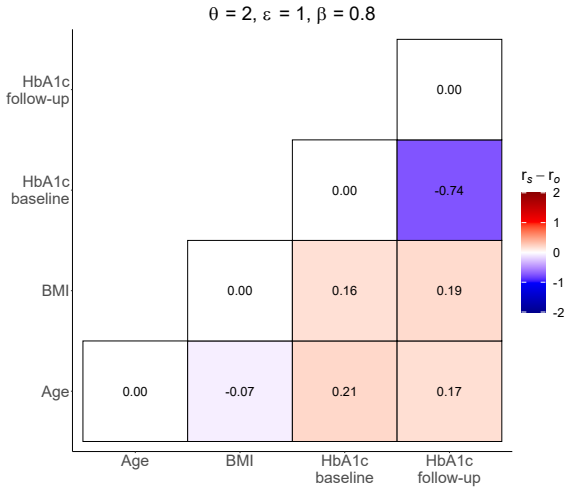
(12.9)



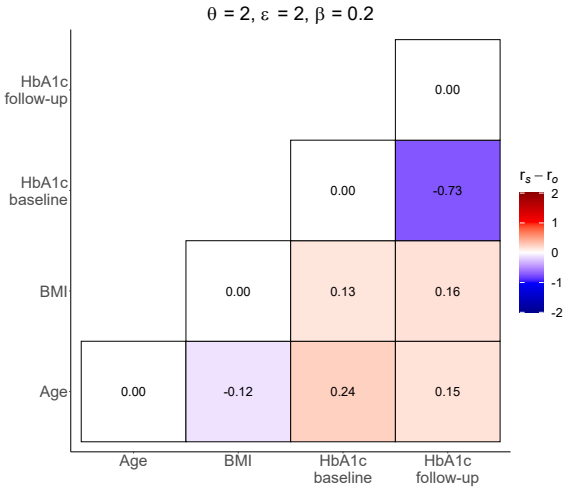
(12.10)



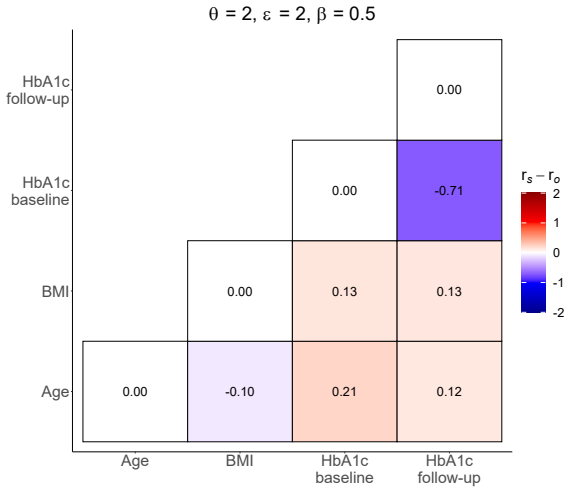
(12.11)



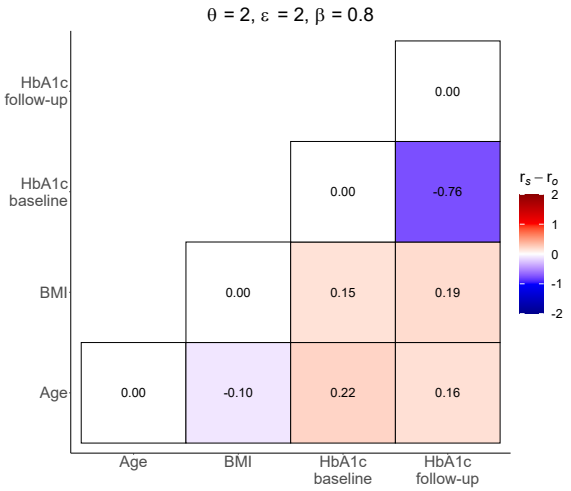
(12.12)



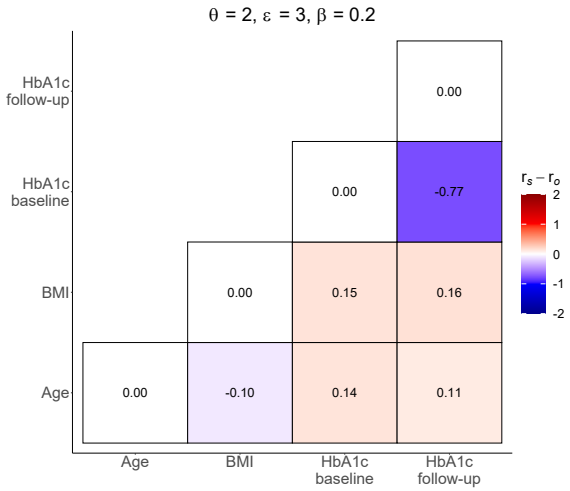
(12.13)



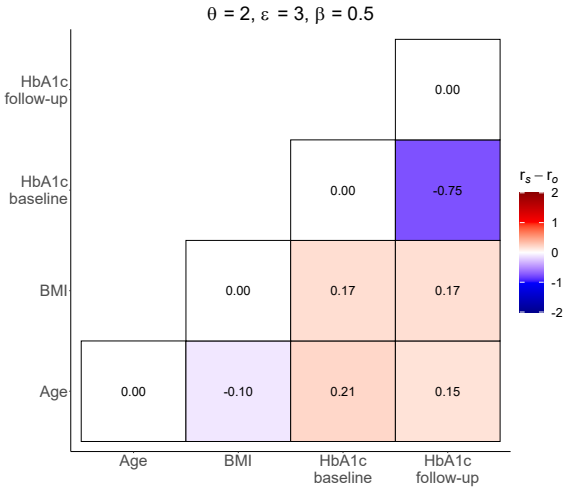
(12.14)



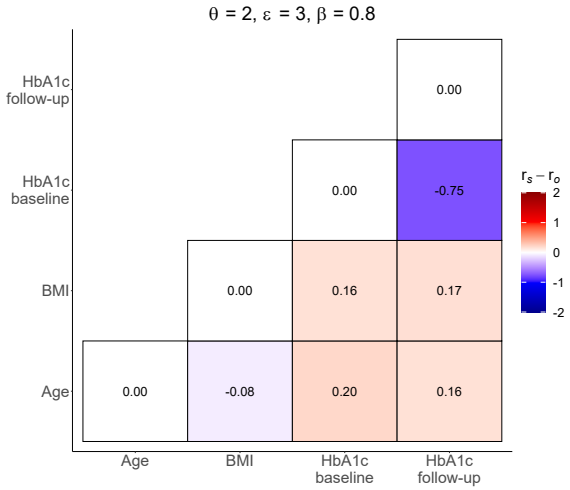
(12.15)



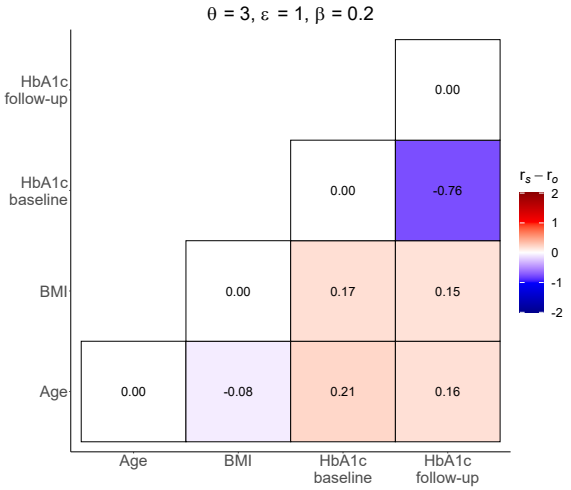
(12.16)



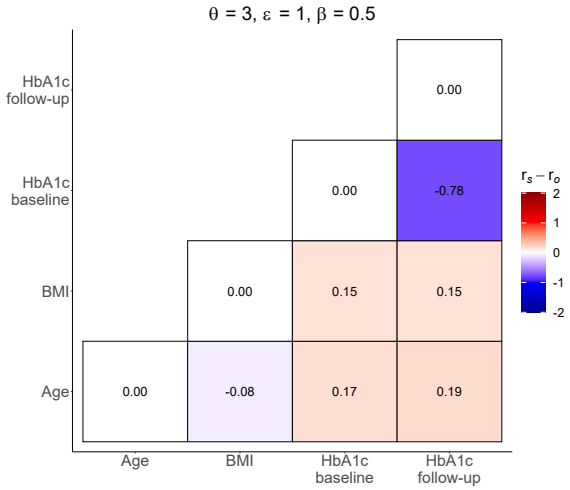
(12.17)



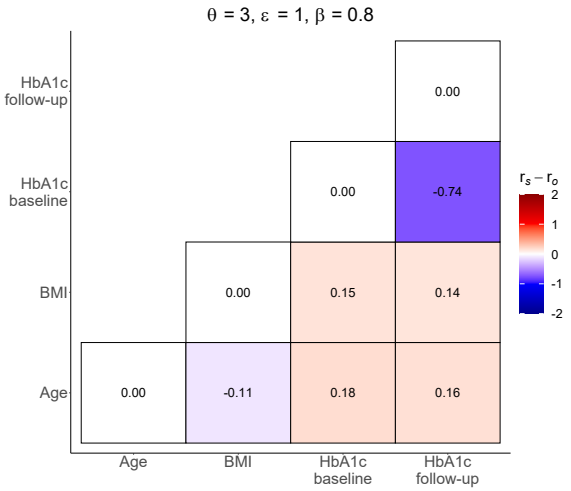
(12.18)



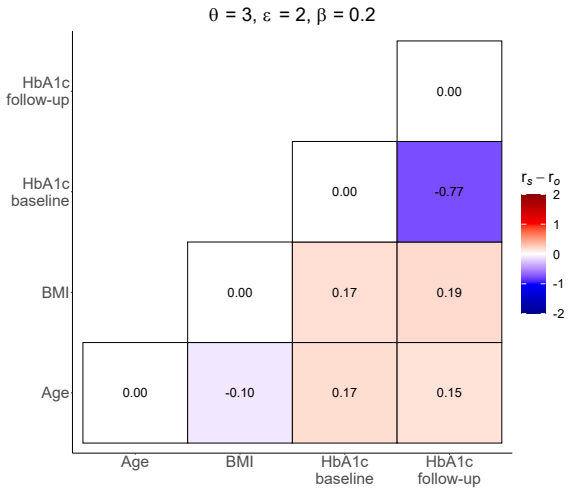
(12.19)



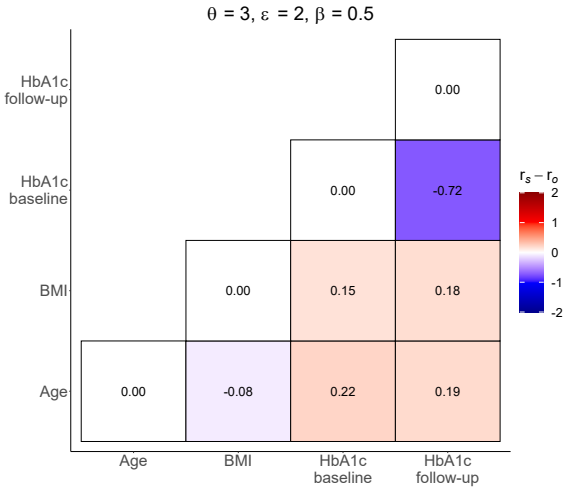
(12.20)



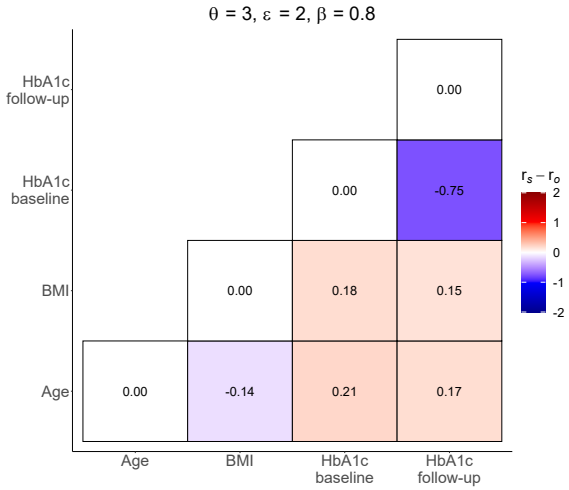
(12.21)



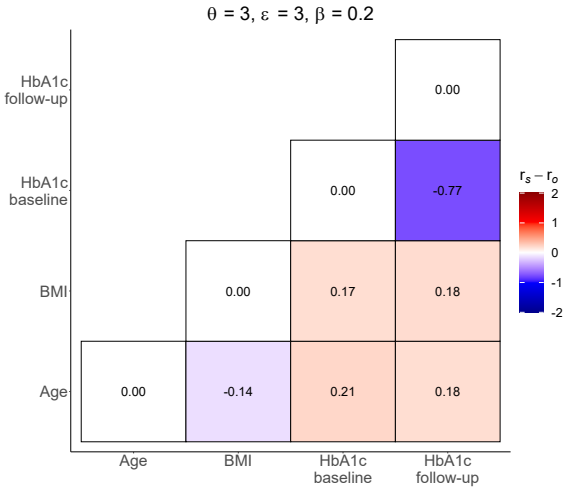
(12.22)



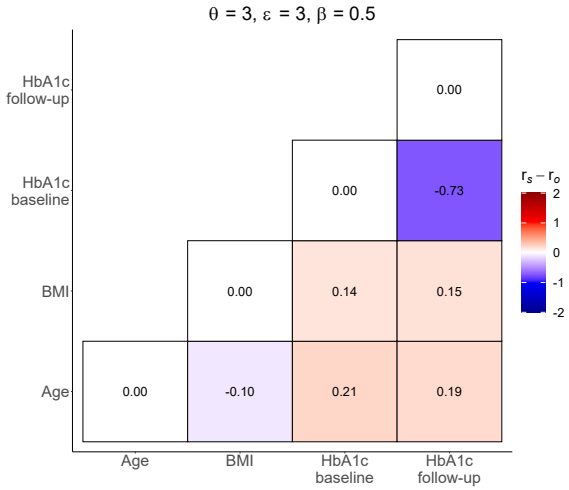
(12.23)



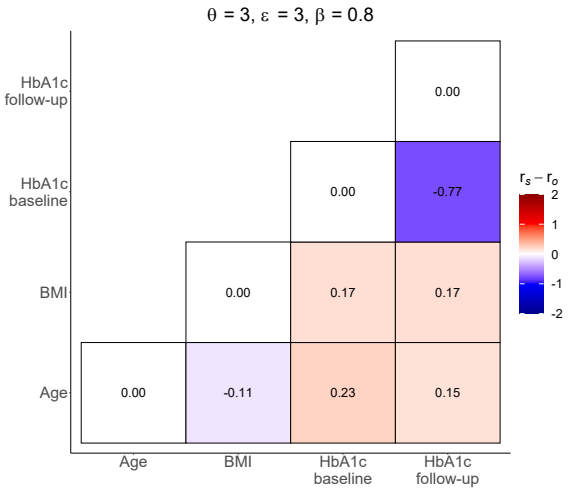
(12.24)



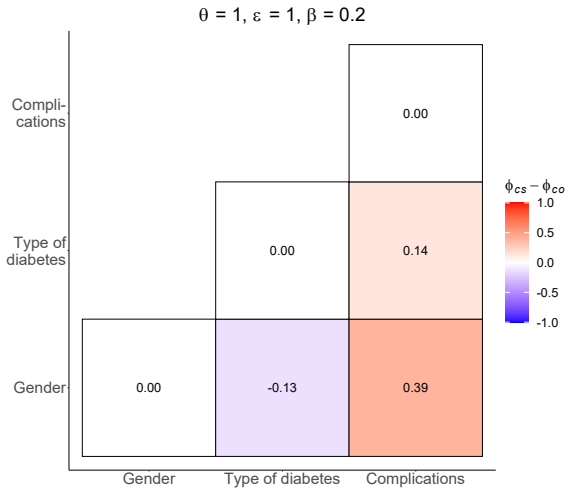
(12.25)



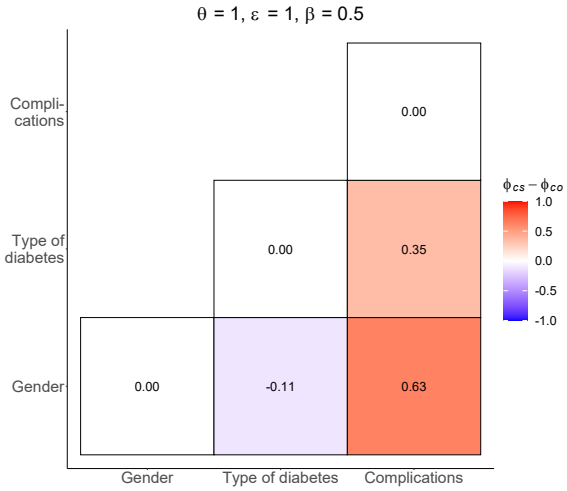
(12.26)



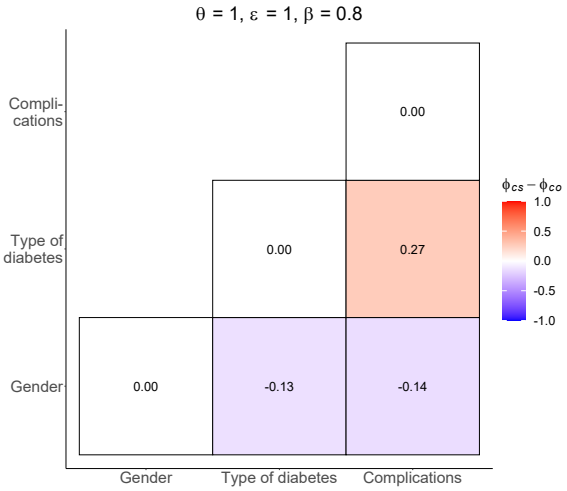
(12.27)



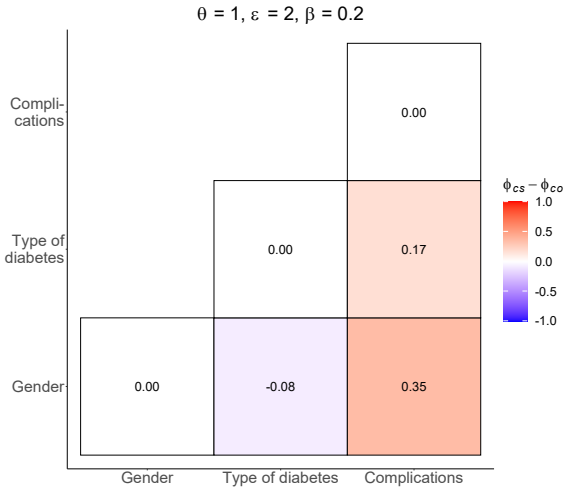
(12.28)



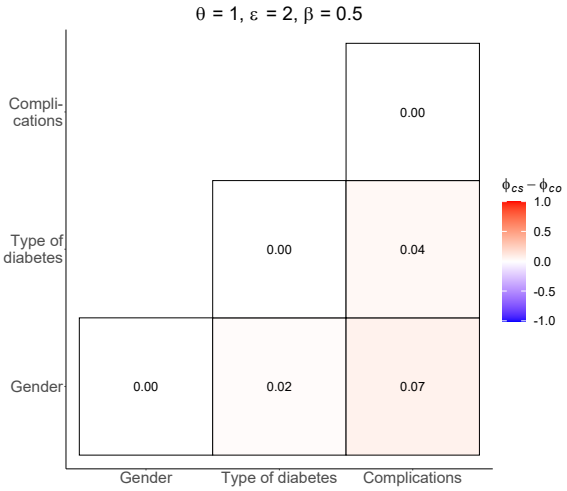
(12.29)



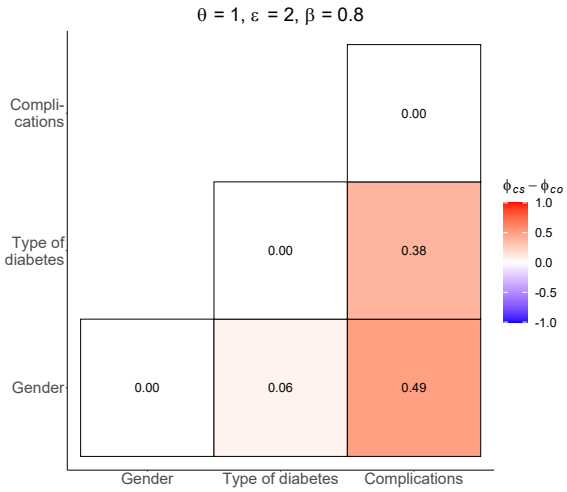
(12.30)



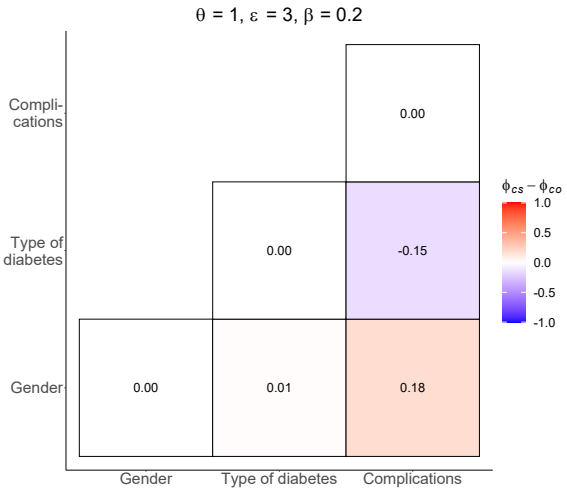
(12.31)



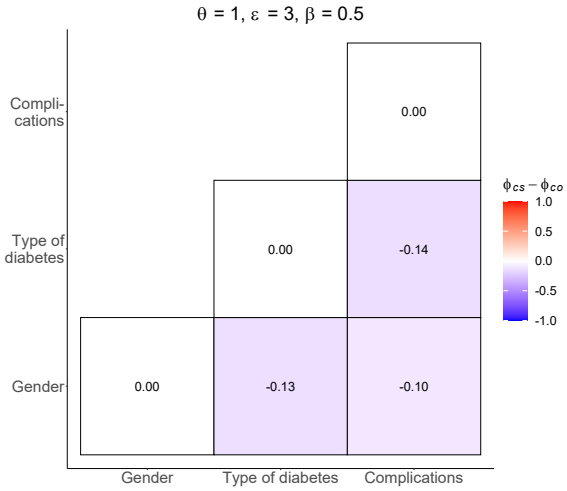
(12.32)



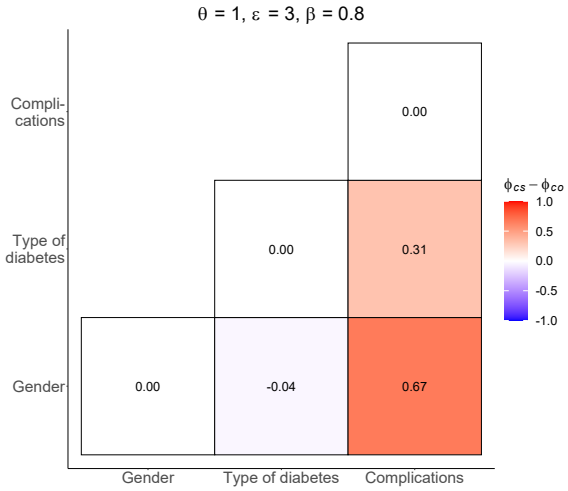
(12.33)



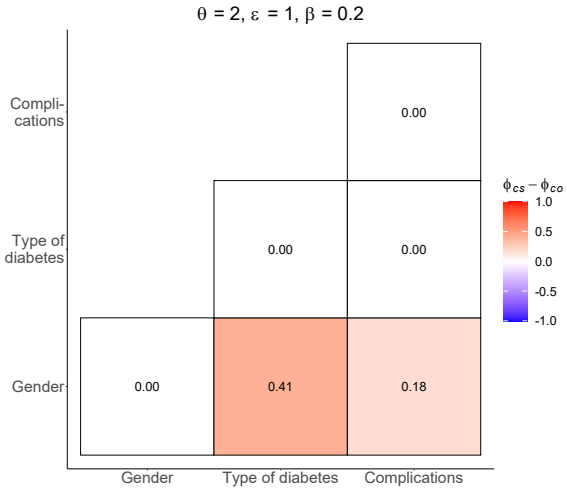
(12.34)



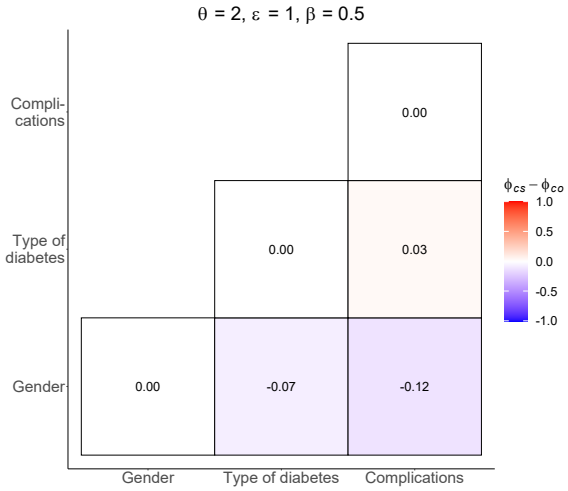
(12.35)



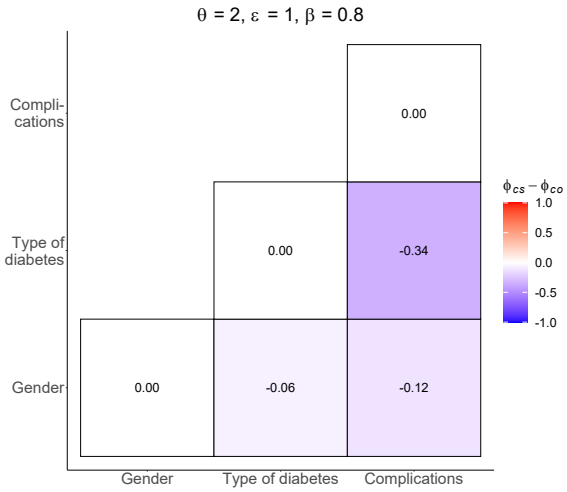
(12.36)



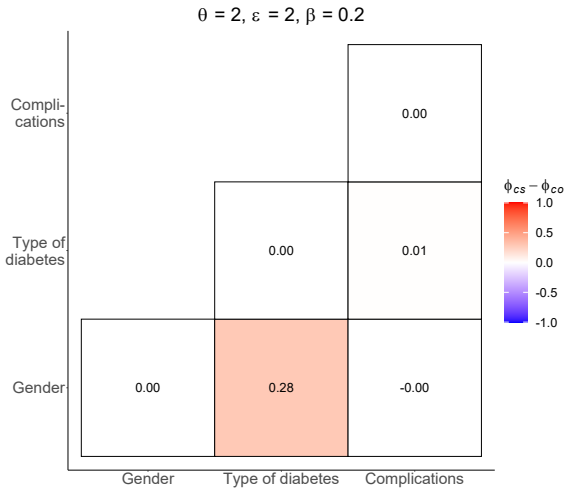
(12.37)



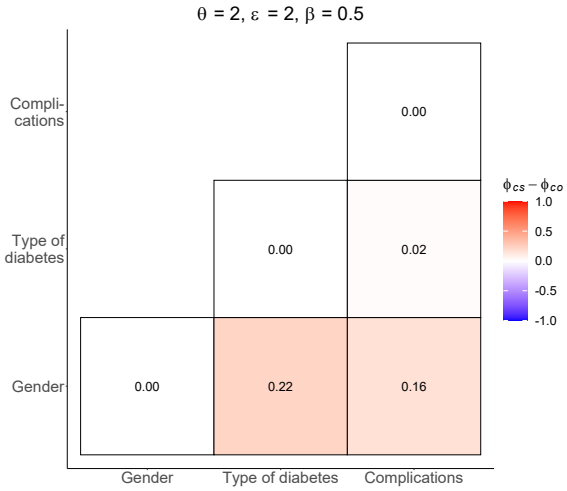
(12.38)



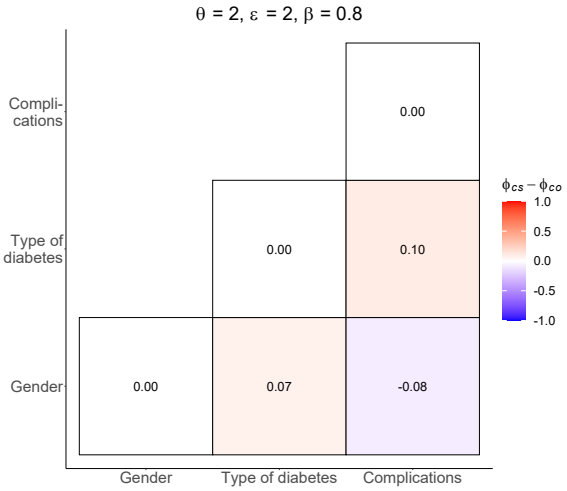
(12.39)



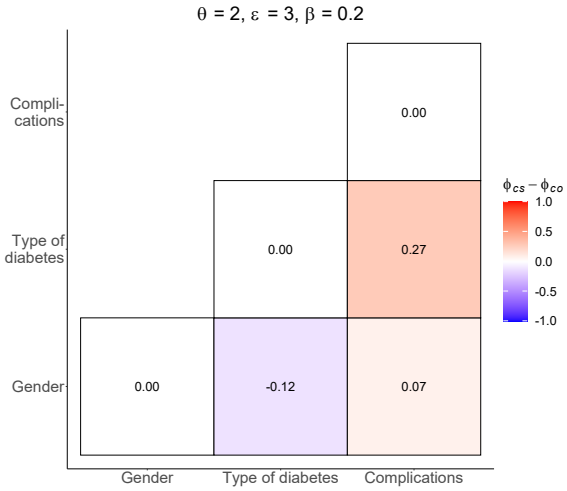
(12.40)



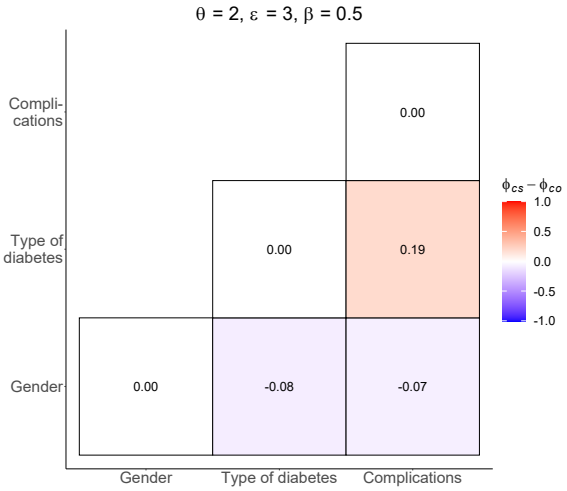
(12.41)



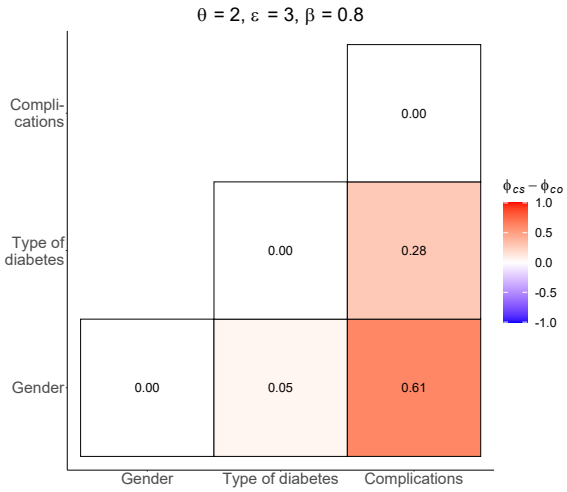
(12.42)



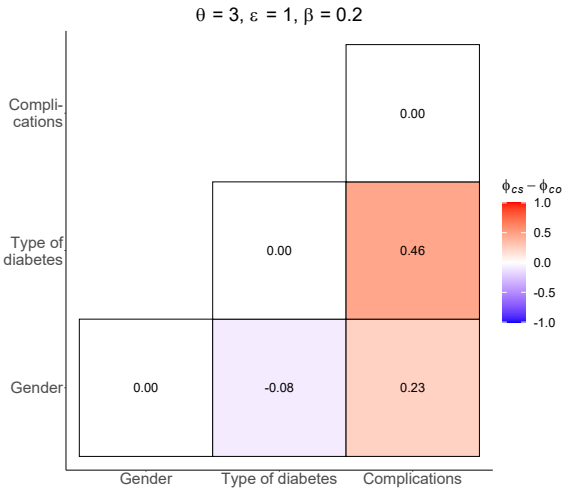
(12.43)



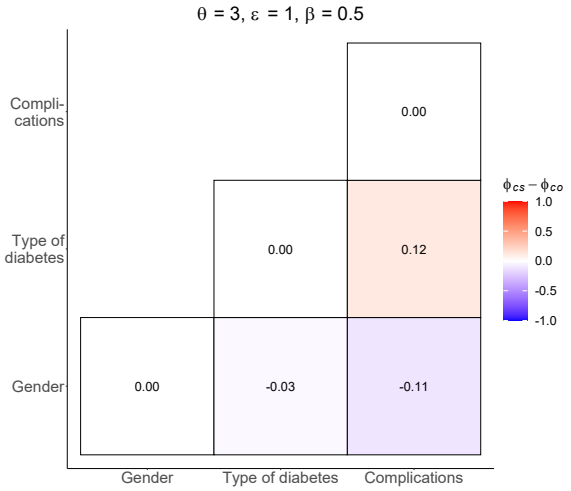
(12.44)



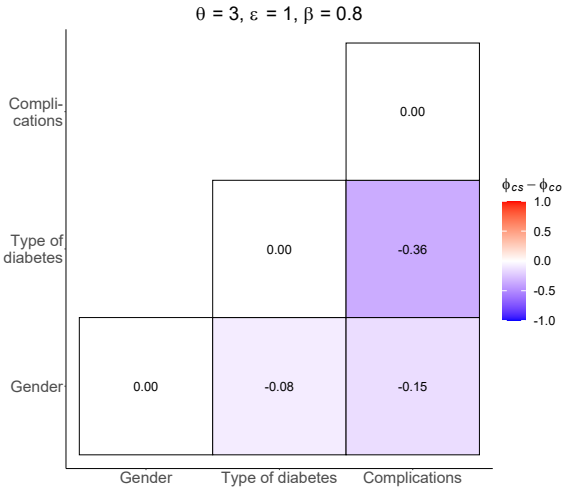
(12.45)



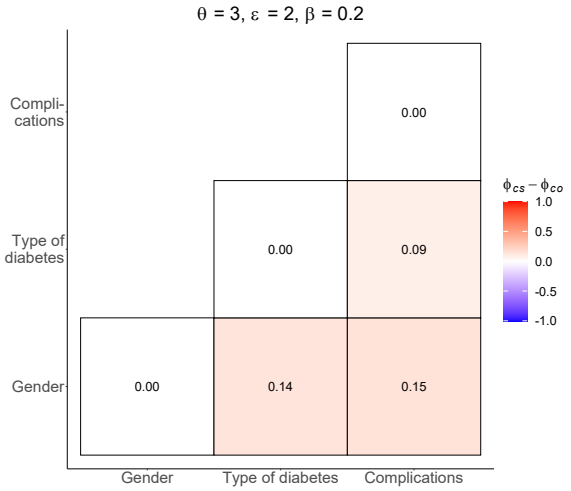
(12.46)



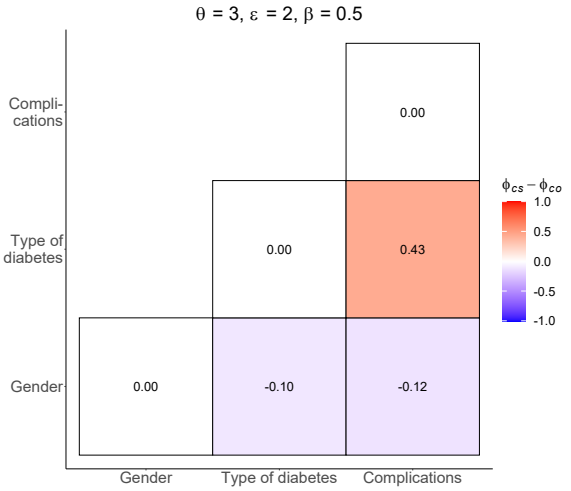
(12.47)



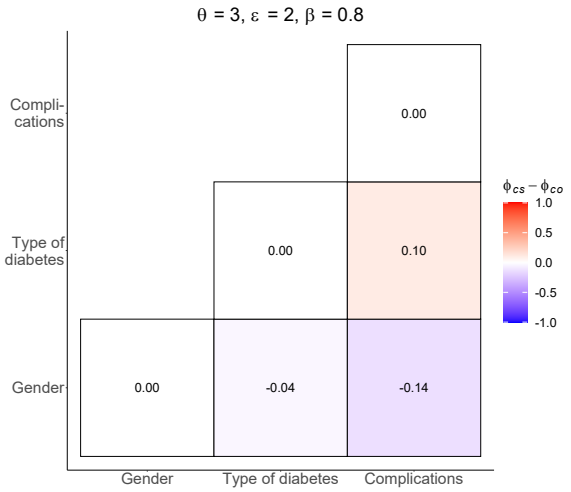
(12.48)



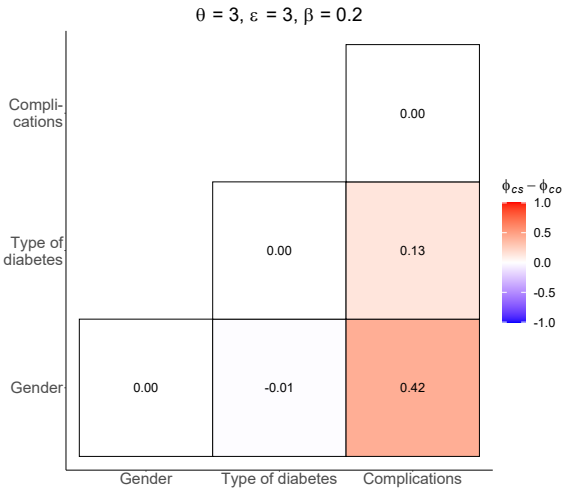
(12.49)



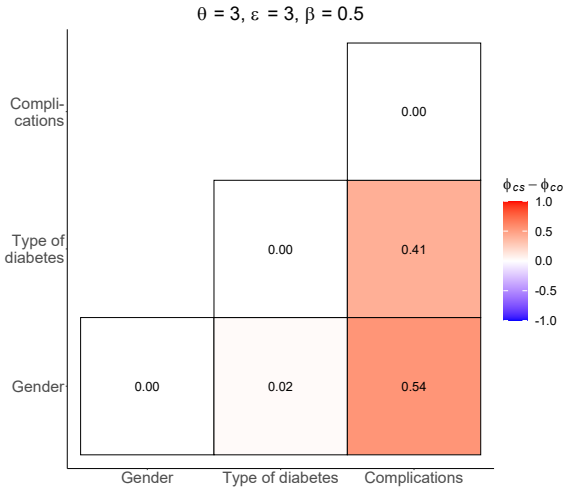
(12.50)



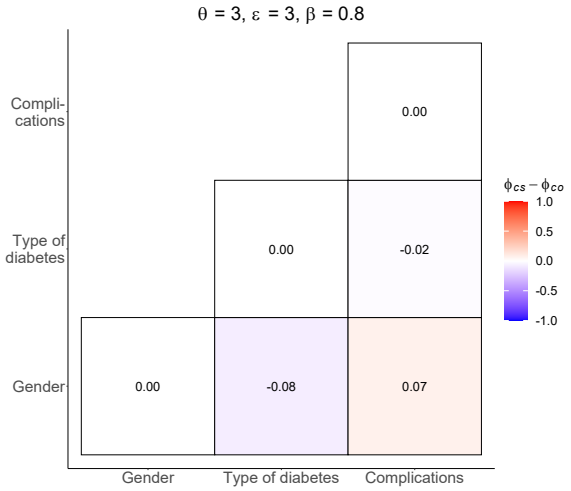
(12.51)



(12.52)



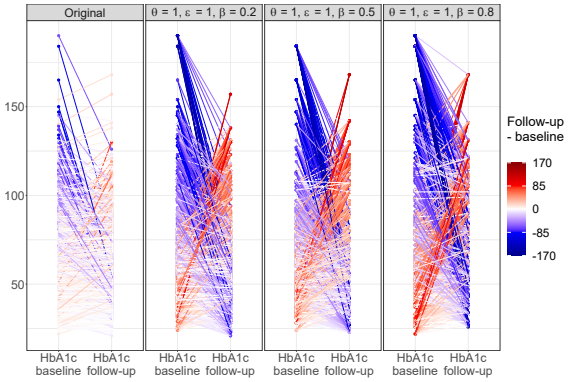
(12.53)



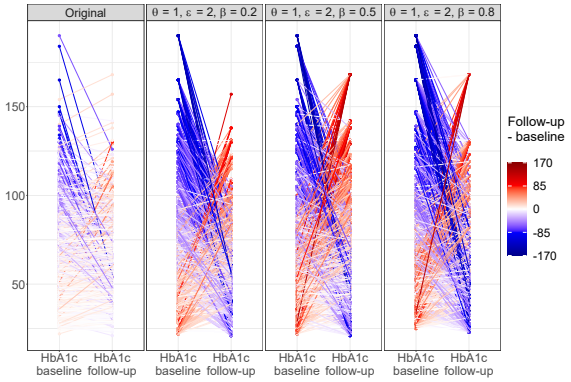
(12.54)

E Individual trajectories

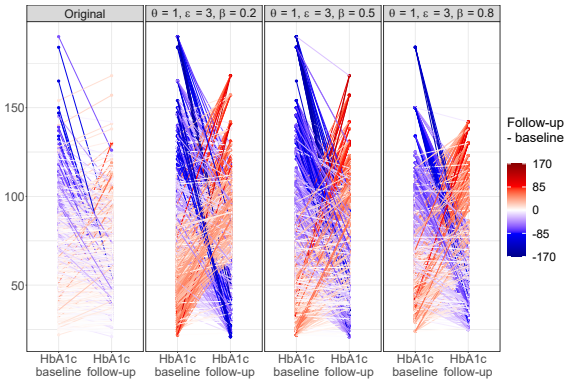
Figure 13



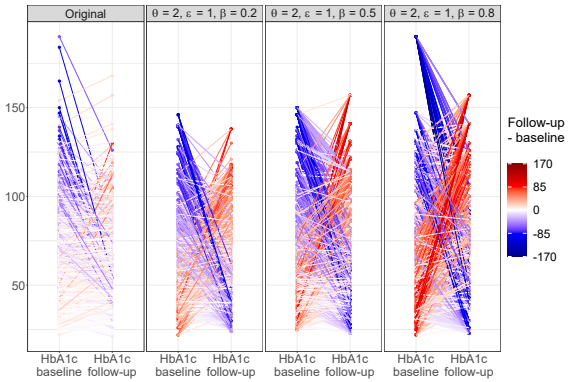
(13.1)



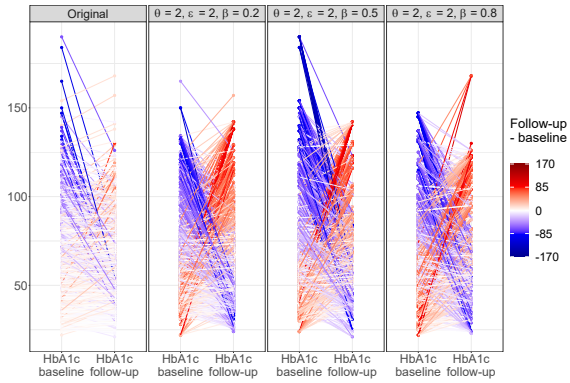
(13.2)



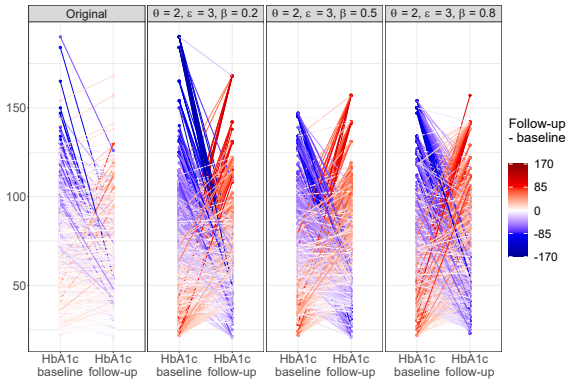
(13.3)



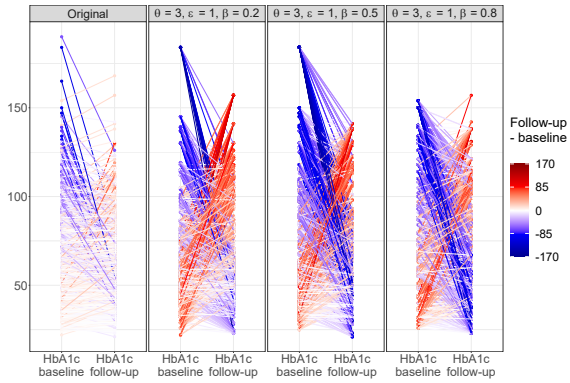
(13.4)



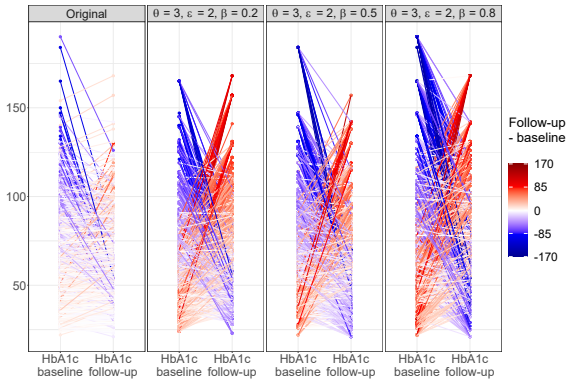
(13.5)



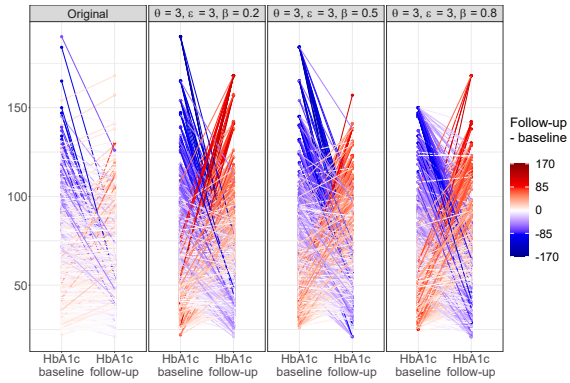
(13.6)



(13.7)



(13.8)



(13.9)

F Results of the linear mixed-effects models

synthetic 1: $\theta = 1, \varepsilon = 1, \beta = 0.2$, **synthetic 2**: $\theta = 1, \varepsilon = 1, \beta = 0.5$, **synthetic 3**: $\theta = 1, \varepsilon = 1, \beta = 0.8$, **synthetic 4**: $\theta = 1, \varepsilon = 2, \beta = 0.2$, **synthetic 5**: $\theta = 1, \varepsilon = 2, \beta = 0.5$, **synthetic 6**: $\theta = 1, \varepsilon = 2, \beta = 0.8$

	HbA1c original			**HbA1c synthetic 1**			**HbA1c synthetic 2**			**HbA1c synthetic 3**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	64.85	61.67 – 68.04	**<0.001**	65.56	62.20 – 68.93	**<0.001**	83.22	79.25 – 87.18	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	-0.01	-0.05 – 0.03	0.723	-0.03	-0.06 – 0.01	0.123	0.01	-0.03 – 0.05	0.556
Gender [M]	1.31	0.07 – 2.55	**0.039**	-0.52	-1.99 – 0.95	0.486	0.61	-0.86 – 2.08	0.415	4.04	2.15 – 5.93	**<0.001**
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	-0.52	-1.99 – 0.95	0.489	-0.87	-2.34 – 0.60	0.246	-17.39	-19.26 – -15.53	**<0.001**
BMI	-0.09	-0.18 – -0.00	**0.041**	0.04	-0.02 – 0.11	0.198	0.06	-0.00 – 0.13	0.060	-0.06	-0.14 – 0.02	0.115
Random Effects												
σ^2	90.29			748.76			793.65			1299.51		
τ_{00}	230.76 $_{Subject}$			13.66 $_{Subject}$			4.26 $_{Subject}$			0.00 $_{Subject}$		
ICC	0.72			0.02			0.01					
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.001 / 0.018			0.001 / 0.007			0.058 / NA		

	HbA1c original			**HbA1c synthetic 4**			**HbA1c synthetic 5**			**HbA1c synthetic 6**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	69.00	66.11 – 71.89	**<0.001**	77.02	73.57 – 80.47	**<0.001**	72.58	68.55 – 76.62	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	-0.01	-0.05 – 0.02	0.467	0.01	-0.03 – 0.05	0.672	0.05	0.00 – 0.09	**0.032**
Gender [M]	1.31	0.07 – 2.55	**0.039**	0.50	-0.89 – 1.89	0.483	-0.18	-1.82 – 1.46	0.826	-0.31	-2.10 – 1.48	0.733
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	-7.62	-8.95 – -6.30	**<0.001**	-14.30	-16.00 – -12.59	**<0.001**	-0.56	-2.37 – 1.25	0.545
BMI	-0.09	-0.18 – -0.00	**0.041**	-0.01	-0.07 – 0.05	0.723	-0.00	-0.07 – 0.06	0.904	-0.05	-0.13 – 0.03	0.189
Random Effects												
σ^2	90.29			644.49			843.64			1144.74		
τ_{00}	230.76 $_{Subject}$			0.00 $_{Subject}$			0.00 $_{Subject}$			0.00 $_{Subject}$		
ICC	0.72											
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.022 / NA			0.053 / NA			0.001 / NA		

synthetic 7: $\theta = 1, \varepsilon = 3, \beta = 0.2$, **synthetic 8**: $\theta = 1, \varepsilon = 3, \beta = 0.5$, **synthetic 9**: $\theta = 1, \varepsilon = 3, \beta = 0.8$, **synthetic 10**: $\theta = 2, \varepsilon = 1, \beta = 0.2$, **synthetic 11**: $\theta = 2, \varepsilon = 1, \beta = 0.5$, **synthetic 12**: $\theta = 2, \varepsilon = 1, \beta = 0.8$

	HbA1c original			**HbA1c synthetic 7**			**HbA1c synthetic 8**			**HbA1c synthetic 9**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	84.83	81.51 – 88.14	**<0.001**	73.45	70.34 – 76.56	**<0.001**	67.48	64.45 – 70.52	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	0.00	-0.04 – 0.04	0.989	-0.03	-0.07 – 0.00	0.078	-0.01	-0.04 – 0.03	0.717
Gender [M]	1.31	0.07 – 2.55	**0.039**	-7.14	-8.64 – -5.63	**<0.001**	0.09	-1.40 – 1.57	0.911	0.81	-0.58 – 2.21	0.253
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	-0.73	-2.21 – 0.74	0.329	-9.98	-11.50 – -8.45	**<0.001**	-0.18	-1.53 – 1.17	0.792
BMI	-0.09	-0.18 – -0.00	**0.041**	-0.10	-0.18 – -0.02	**0.011**	-0.00	-0.08 – 0.07	0.971	-0.03	-0.09 – 0.03	0.302
Random Effects												
σ^2	90.29			763.34			777.07			626.75		
τ_{00}	230.76 $_{Subject}$			0.00 $_{Subject}$			10.03 $_{Subject}$			5.05 $_{Subject}$		
ICC	0.72						0.01			0.01		
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.018 / NA			0.032 / 0.044			0.001 / 0.009		

	HbA1c original			**HbA1c synthetic 10**			**HbA1c synthetic 11**			**HbA1c synthetic 12**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	62.08	59.38 – 64.79	**<0.001**	64.49	60.79 – 68.20	**<0.001**	74.68	70.41 – 78.94	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	0.03	-0.00 – 0.06	0.079	-0.01	-0.05 – 0.03	0.512	-0.01	-0.06 – 0.04	0.691
Gender [M]	1.31	0.07 – 2.55	**0.039**	-0.73	-2.20 – 0.74	0.331	2.80	1.24 – 4.36	**<0.001**	0.04	-1.90 – 1.97	0.971
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	-0.76	-2.22 – 0.70	0.305	1.43	-0.13 – 2.98	0.072	-1.04	-2.98 – 0.89	0.291
BMI	-0.09	-0.18 – -0.00	**0.041**	0.00	-0.05 – 0.05	0.917	0.05	-0.03 – 0.13	0.239	0.06	-0.03 – 0.15	0.227
Random Effects												
σ^2	90.29			519.46			883.21			1367.83		
τ_{00}	230.76 $_{Subject}$			0.00 $_{Subject}$			10.09 $_{Subject}$			0.00 $_{Subject}$		
ICC	0.72						0.01					
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.001 / NA			0.003 / 0.014			0.000 / NA		

synthetic 13: $\theta = 2, \varepsilon = 2, \beta = 0.2$, **synthetic 14**: $\theta = 2, \varepsilon = 2, \beta = 0.5$, **synthetic 15**: $\theta = 2, \varepsilon = 2, \beta = 0.8$, **synthetic 16**: $\theta = 2, \varepsilon = 3, \beta = 0.2$, **synthetic 17**: $\theta = 2, \varepsilon = 3, \beta = 0.5$, **synthetic 18**: $\theta = 2, \varepsilon = 3, \beta = 0.8$

	HbA1c original			**HbA1c synthetic 13**			**HbA1c synthetic 14**			**HbA1c synthetic 15**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	65.34	62.25 – 68.43	**<0.001**	67.84	64.55 – 71.14	**<0.001**	65.67	62.32 – 69.01	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	0.01	-0.03 – 0.04	0.725	-0.03	-0.07 – 0.01	0.100	-0.00	-0.04 – 0.04	0.973
Gender [M]	1.31	0.07 – 2.55	**0.039**	-0.95	-2.31 – 0.41	0.170	0.44	-1.32 – 2.21	0.622	-0.31	-1.80 – 1.17	0.679
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	-1.07	-2.59 – 0.44	0.166	0.22	-1.32 – 1.76	0.780	0.34	-1.19 – 1.87	0.664
BMI	-0.09	-0.18 – -0.00	**0.041**	-0.04	-0.11 – 0.02	0.203	-0.08	-0.15 – -0.01	**0.023**	0.02	-0.05 – 0.08	0.632
Random Effects												
σ^2	90.29			537.51			689.76			778.60		
τ_{00}	230.76 $_{Subject}$			10.22 $_{Subject}$			21.89 $_{Subject}$			0.00 $_{Subject}$		
ICC	0.72			0.02			0.03					
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.001 / 0.019			0.001 / 0.032			0.000 / NA		

	HbA1c original			**HbA1c synthetic 16**			**HbA1c synthetic 17**			**HbA1c synthetic 18**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	67.64	64.62 – 70.65	**<0.001**	63.64	60.94 – 66.34	**<0.001**	67.50	64.37 – 70.64	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	-0.02	-0.05 – 0.02	0.432	0.00	-0.03 – 0.04	0.778	-0.02	-0.06 – 0.02	0.286
Gender [M]	1.31	0.07 – 2.55	**0.039**	-0.40	-1.62 – 0.81	0.515	-1.37	-2.55 – -0.18	**0.024**	-0.10	-1.58 – 1.39	0.900
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	-6.38	-7.74 – -5.02	**<0.001**	-2.43	-3.70 – -1.16	**<0.001**	0.27	-1.30 – 1.84	0.737
BMI	-0.09	-0.18 – -0.00	**0.041**	-0.03	-0.10 – 0.04	0.347	0.03	-0.03 – 0.08	0.402	0.01	-0.06 – 0.07	0.866
Random Effects												
σ^2	90.29			545.97			520.40			760.58		
τ_{00}	230.76 $_{Subject}$			0.00 $_{Subject}$			0.00 $_{Subject}$			0.00 $_{Subject}$		
ICC	0.72											
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.019 / NA			0.003 / NA			0.000 / NA		

synthetic 19: $\theta = 3, \varepsilon = 1, \beta = 0.2$, **synthetic 20**: $\theta = 3, \varepsilon = 1, \beta = 0.5$, **synthetic 21**: $\theta = 3, \varepsilon = 1, \beta = 0.8$, **synthetic 22**: $\theta = 3, \varepsilon = 2, \beta = 0.2$, **synthetic 23**: $\theta = 2, \varepsilon = 3, \beta = 0.5$, **synthetic 24**: $\theta = 3, \varepsilon = 2, \beta = 0.8$

	HbA1c original			**HbA1c synthetic 19**			**HbA1c synthetic 20**			**HbA1c synthetic 21**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	66.72	63.40 – 70.04	**<0.001**	70.10	66.02 – 74.18	**<0.001**	85.54	81.81 – 89.28	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	-0.02	-0.05 – 0.02	0.437	-0.04	-0.09 – 0.01	0.130	-0.03	-0.08 – 0.01	0.164
Gender [M]	1.31	0.07 – 2.55	**0.039**	-0.44	-1.91 – 1.03	0.557	1.77	-0.07 – 3.60	0.059	7.99	5.77 – 10.21	**<0.001**
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	0.23	-1.16 – 1.62	0.747	-0.98	-2.79 – 0.83	0.288	-15.16	-16.88 – -13.45	**<0.001**
BMI	-0.09	-0.18 – -0.00	**0.041**	-0.00	-0.07 – 0.06	0.900	-0.02	-0.10 – 0.07	0.703	-0.03	-0.11 – 0.04	0.354
Random Effects												
σ^2	90.29			687.10			1192.69			1041.20		
τ_{00}	230.76 $_{Subject}$			0.00 $_{Subject}$			0.00 $_{Subject}$			0.00 $_{Subject}$		
ICC	0.72											
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.000 / NA			0.001 / NA			0.061 / NA		

	HbA1c original			**HbA1c synthetic 22**			**HbA1c synthetic 23**			**HbA1c synthetic 24**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	63.42	60.65 – 66.20	**<0.001**	61.82	58.93 – 64.71	**<0.001**	72.37	68.65 – 76.10	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	-0.04	-0.07 – -0.01	**0.015**	0.01	-0.02 – 0.05	0.444	-0.01	-0.05 – 0.04	0.811
Gender [M]	1.31	0.07 – 2.55	**0.039**	-0.28	-1.62 – 1.05	0.678	1.72	0.44 – 2.99	**0.008**	-0.57	-2.36 – 1.23	0.536
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	-0.70	-1.95 – 0.55	0.273	-2.45	-3.72 – -1.19	**<0.001**	0.54	-1.65 – 2.72	0.631
BMI	-0.09	-0.18 – -0.00	**0.041**	0.04	-0.02 – 0.09	0.188	0.01	-0.05 – 0.06	0.846	0.01	-0.09 – 0.10	0.906
Random Effects												
σ^2	90.29			506.29			569.44			1021.31		
τ_{00}	230.76 $_{Subject}$			0.00 $_{Subject}$			13.94 $_{Subject}$			0.00 $_{Subject}$		
ICC	0.72						0.02					
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.002 / NA			0.004 / 0.028			0.000 / NA		

synthetic 25: $\theta = 3, \varepsilon = 3, \beta = 0.2$, **synthetic 26**: $\theta = 3, \varepsilon = 3, \beta = 0.5$, **synthetic 27**: $\theta = 3, \varepsilon = 3, \beta = 0.8$

	HbA1c original			**HbA1c synthetic 25**			**HbA1c synthetic 26**			**HbA1c synthetic 27**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	70.43	67.04 – 73.83	**<0.001**	58.53	55.78 – 61.28	**<0.001**	64.07	61.04 – 67.11	**<0.001**	67.05	63.67 – 70.43	**<0.001**
Age	-0.05	-0.09 – -0.01	**0.027**	0.01	-0.03 – 0.04	0.680	0.00	-0.03 – 0.04	0.776	-0.00	-0.04 – 0.03	0.873
Gender [M]	1.31	0.07 – 2.55	**0.039**	2.68	1.41 – 3.96	**<0.001**	-0.19	-1.45 – 1.08	0.773	-1.28	-2.65 – 0.10	0.069
Type of Diabetes [type2]	-10.48	-12.17 – -8.79	**<0.001**	1.07	-0.21 – 2.34	0.102	0.90	-0.37 – 2.18	0.166	0.40	-1.01 – 1.81	0.579
BMI	-0.09	-0.18 – -0.00	**0.041**	0.04	-0.02 – 0.09	0.202	-0.04	-0.10 – 0.03	0.304	0.03	-0.04 – 0.09	0.431
Random Effects												
σ^2	90.29			491.70			552.78			705.61		
τ_{00}	230.76 $_{Subject}$			0.00 $_{Subject}$			10.04 $_{Subject}$			0.00 $_{Subject}$		
ICC	0.72						0.02					
N	2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$			2890 $_{Subject}$		
Observations	5780			5780			5780			5780		
Marginal R^2 / Conditional R^2	0.098 / 0.746			0.004 / NA			0.001 / 0.018			0.001 / NA		