AI Models for Modeling and Simulation of Clinical Studies for Alzheimer's and Parkinson's Disease

DISSERTATION

ZUR

ERLANGUNG DES DOKTORGRADES (DR. RER. NAT.)

DER

MATHEMATISCH-NATURWISSENSCHAFTLICHEN FAKULTÄT

DER

RHEINISCHEN FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

vorgelegt von

MEEMANSA SOOD
aus

HIMACHAL PRADESH, INDIEN

Bonn, 2022

ANGEFERTIGT MIT GENEHMIGUNG DER
MATHEMATISCH-NATURWISSENSCHAFTLICHEN FAKULTÄT
DER RHEINISCHEN FRIEDRICH-WILHELMS-UNIVERSITÄT
BONN

1.GUTACHTER: PROF. DR. HOLGER FRÖHLICH
2.GUTACHTER: PROF. DR. THOMAS SCHULTZ
TAG DER PROMOTION: 4th July 2023
ERSCHEINUNGSJAHR: 2023

Abstract

Neurodegenerative diseases (NDDs) have a complex structure and most of them are untreatable that's why more research studies undertake translational paths for getting better insights into prevention, early detection, and better treatment options. A longitudinal understanding of disease development and progression across all biological scales is required for translational research of these diseases. However, due to the complexity underlying these diseases and their heterogeneous nature, there is a need for a comprehensive picture of a specific disease. For this purpose, multiple studies need to be compared and analyzed and several observational cohort studies and clinical trials are available for this purpose. Many of these clinical studies aim at early prognosis, drug development, and treatment of the disease. However, legal and ethical constraints typically do not allow for sharing of sensitive patient data. In consequence, there exist data silos, which slow down the overall scientific progress in translational research.

In our work, we suggest artificial intelligence (AI) based methods that are generative in nature and help to model and simulate the clinical studies for Alzheimer's disease (AD) and Parkinson's disease (PD). The key idea here is to describe a longitudinal patient cohort with the help of a Bayesian network (BN), in conjunction with deep learning methods. Our approach allows for incorporating arbitrary multi-scale, multi-modal data. As our method is generative in nature, we try to solve the problem of data sharing and data silos by generating synthetic data. We show that with the help of such a model, we can simulate subjects that are largely indistinguishable from real ones. Moreover, we demonstrate the possibility to simulate counterfactual interventions in a synthetic cohort. We also unravel the complexities underlying NDDs by disentangling and quantifying the connections between different clinical parameters.

Acknowledgments

Firstly I would like to express my deepest appreciation to my supervisor, Prof. Dr. Holger Fröhlich. I am very grateful that I got an opportunity to work in his group and contribute to the ongoing research in the field of Alzheimer's and Parkinson's disease. His constant scientific guidance, critical inputs and support has helped me to achieve the goals of my thesis. I would also like to thank Prof. Dr. Thomas Schultz for acceding to be the second reviewer of this thesis. I really appreciate the efforts of all my thesis committee members.

Next, I would like to extend my deepest gratitude to Prof. Dr. Martin Hofmann-Apitius who motivated me to take up challenges that really helped me to grow in my career and explore many opportunities. I am grateful to all my colleagues at Fraunhofer SCAI-BIO for being a part of this wonderful journey.

I would also like to extend my sincere thanks to the collaborators I worked with in RADAR-AD project, Dr.Andrew Owens, Prof. Dr. Dag Arsland and Neva Coello. It was a very stimulating project to work on and I believe I have learnt a lot through it. I am also grateful to Dr. Sebastian Heinzel from University of Kiel who supported me in my work during the final phase of my Ph.D.

I am extremely thankful to many friends and family (too many to name here but you know who you are!), their constant encouragement and emotional support was worth more than I can express on paper. I feel blessed and lucky to be surrounded by so many beautiful people.

Most importantly, I would like to express my profound gratitude to my parents, my brother and my sister. I thank them for always believing in me and I could not have undertaken this journey without their constant love and support. They have always encouraged me to chase my dreams.

My parents, just being a phone call away, were always there by my side and I thank them for being so patient and understanding. Their dedication towards their work and their positive attitude towards life has always been my source of inspiration. I am truly grateful for their unconditional and endless love.

Contents

List of Figures

List of Tables

When you learn a little, you feel you know a lot.
But when you learn a lot, you realize you know
very little.

Jay Shetty

1

INTRODUCTION

1.1 Key problems in translational neurological research

According to the World Health Organization, brain disorders are considered to be “one of the greatest threats to public health” with one in every four persons getting affected by neurological or mental health conditions at some point in their lives [1]. According to a study that was conducted by Global Burden of Disease, neurological diseases are the major cause of growing disability around the world [2]. A larger section of neurological diseases constitutes neurodegenerative disease (NDD)s which are characterized by progressive loss, degeneration, and death of nerve cells [3]. NDDs are age-related diseases, however, their onset might begin at an early age and they affect the life expectancy and quality of life of a person [4]. In the past years, there was a significant increase in the incidence of the diseases and as the population of the world ages, this increase is expected to continue [5]. NDDs affect up to 50 million people worldwide and roughly 10 million new cases are reported every year (https://www.who.int/news-room/fact-sheets/detail/dementia). NDDs pose a huge challenge to society and can become a burden as their cause is still unknown and no cure has been discovered [6]. AD and PD are the two most common types of NDDs.

1.1.1 Alzheimer's disease

AD, the most common cause of dementia [7], is a chronic NDD that causes problems with memory, thinking, and behavior. It is characterized by insidious onset and progressive impairment of behavioral and cognitive functions that includes memory, comprehension, language, attention, reasoning, and judgment [8]. It is hypothesized that neuritic plaques and neurofibrillary tangles (NFT) could characterize the pathophysiological mechanisms [9]. The NFT, also known as the amyloid plaques are the hallmark of AD. These are the extracellular deposits of amyloid beta (Abeta) protein present abundantly in the cortex of AD patients [10]. According to a report published by Alzheimer's Association [11], it is estimated that there were 6.2 million Americans aged 65 and older, affected by AD in 2021 and this number is projected to increase to 13.8 million by the year 2060. Some of the causes of the rise in the number of AD patients are an increasingly aging population, family history, stroke, and other underlying diseases occurring due to lifestyle choices such as cardiovascular diseases, diabetes, and high blood pressure.

1.1.2 Parkinson's disease

PD, the second most common cause of dementia is characterized by tremors and bradykinesia and is a progressive NDD [12]. It affects predominantly neurons that produce dopamine in a specific area of the brain called subastantia nigra (SN). It has been estimated that 6.2 million individuals in 2015 had the disease globally, compared to 2.5 million in 1990, and this number is expected to rise by more than double in the year 2040 [13]. PD is also known to affect 1-2% of individuals above 65 and its prevalence is increasing at a fast pace with the increase in the aging population [14]. Some of the causes of the exponential growth of PD cases are the increasing longevity, aging population above age 65, declining smoking population, and by-products of industrialization [15].

While there is a significant rise in the growth of NDDs, there is a dearth of treatments for these diseases. Moreover, due to several failed clinical trials around their established hypotheses, these diseases are regarded as complex multi-factorial diseases [16, 17]. These diseases are also accompanied by dysregulation at different biological scales ranging from mutations at the genetic level to structural and functional alteration of the brain at the clinical level [18]. Therefore, to understand the complexity of these diseases, a large amount of data is generated every day in hospitals, from medical devices, laboratories, clinical trials, and research studies such as observational cohorts, mobile devices, etc. This leads to the production of data with different modalities and varied scales, including imaging and non-imaging. In our work, we will focus on the data generated from observational longitudinal clinical studies.

As these diseases are untreatable and their cause is not clear, more research studies are dedicated towards a translational path for getting better insights into prevention, early detection, and better treatment options for these NDDs [6]. This translation path incorporates the field of translational neuroscience that translates basic clinical research into clinical applications and novel therapies for nervous system disorders [19]. This field aims to tackle these diseases by applying the "bench to bedside" concept and thereafter transforming the knowledge gained by basic research in science into interventions and applications for the treatment [20]. The hope is that this approach might change neuroscientific research and bridge the gap between clinical practice and neuroscience methods. It also helps to bring together neuroscience, neuroimaging, and clinicians for improving our understanding of symptoms and disorders and for better diagnostics and treatments [21]. Bridging the gap between several domains will also help us to refine and advance the application of discovery [22]. Nonetheless, there are also several challenges underlying bridging this gap in the field of translational neuroscience.

The translation of basic scientific findings into clinical practice is not a straightforward approach [23]. Some of the challenges include a lack of culture of translation in different institutions [24, 25], lack of adequate infrastructure [26, 25], inadequate training workforce [24, 25, 27] and facilities to conduct best practice clinical research [26, 25].

The other challenge is from the data-oriented perspective. As described above, there is a large amount of data generated from the observational longitudinal clinical cohorts for several features of these diseases. Hence, it is a challenging task to effectively translate all the features having different scales and modalities into one disease model or a clinical application. There is a lack of translation of adequate interoperable data from various stakeholders to a central orchestrator. A key requirement of translation is collaboration and this is often discouraged by the fact that the university system rewards individual research rather than joint clinical practices. This creates a culture divide between scientists and clinicians and compartmentalization of departments within universities and hospitals [24, 25, 27, 28]. Due to a lack of collaboration and more focus on individualized research, people often hesitate to share the data which leads to the formation of different data "silos". Furthermore, legal constraints such as the general data protection regulation (GDPR) [29] and ethical constraints especially in the United States and Europe in essence prohibit sharing of sensitive patient data across organizations. These aforementioned challenges also increase the time accounted for a particular research question and create a number of roadblocks.

In order to overcome these challenges, several initiatives have been established at the policy level, some of which include the creation of biomedical research centers that bring together the people working in a hospital setting and a university setting. However, the major roadblock of data sharing can be overcome by the simulation of a synthetic data cohort that is similar to the real cohort but not identical.

Considering all these challenges arising from the data perspective, it becomes essential to overcome these which brings us to the field of AI and Data Science. AI is a branch of computer science that deals with the simulation of human intelligence by a machine such as a computer [30]. Data Science uses scientific methods, algorithms based on AI, and tools to extract knowledge, interpret and understand the noisy structured and unstructured data [31, 32].

To give more context, we will discuss the need for AI in the field of healthcare specifically NDDs in the following section.

1.2 Need of Artificial intelligence (AI)

AI has grown dramatically in the 21st century [33]. It aims to enhance and expand the scope and efficiency of mankind in tasks focusing on remaking nature and governing society through intelligent machines, with the eventual goal of realizing a society where machines and people can coexist harmoniously [34]. AI was first described in medicine in the year 1976 when a computer algorithm was used to identify the causes of acute abdominal pain [35] and since then it has been rapidly growing and hugely impacting the field of healthcare. There can be several reasons for this growth, for e.g., it helps to relieve the manual workload of healthcare professionals [36].

In the medical field, the initial aim of data being stored digitally in the form of electronic health record (EHR) was meant to simplify and facilitate patient care, but doctors find it very difficult to navigate the technological systems. This has further been burdened by the bureaucracy and has resulted in burnout-related symptoms in many doctors [30]. AI could also synthesize the patient records and summarize the health concerns for the clinicians [37]; rather than manually analyzing the patient data; it could examine the available information much faster and highlight the core points [38]. It can also be used to automatically scan the diagnostic images and their interpretation and can work as an initial screening tool for the interpretation of scans and prioritizing those that are of concern. This also in turn reduces or relieves the workload of the physician. From an economic point of view, this could save time and resources [39, 40]. AI in the form of applications can also help to replace some tasks that are carried out by healthcare professionals that are repetitive and require little cognition [41, 42]. Primarily, AI has an enormous potential to augment clinical practice and patient care. It can assist in providing significant help in the field of NDDs by targeting diagnosis and monitoring e.g. detection of disease onset, classification of disease stages, improvement of the differential diagnosis, quantification of the disease progression, tracking of the effects of medication and treatment, etc. Considering all these target goals related to diagnosing and monitoring the disease, it has led to the collection of multi-scale and multi-level data for the diseases [43]. Due to the multi-scale and multi-level data and the presence of “data silos” in the community, the need for data sharing has increased in recent times. Therefore, data-sharing initiatives have also recently grown to contribute to the advancement of translational research. This vast amount of data has led to the accumulation of an enormous amount of information for diseases, which is beyond the human mind to comprehend. Therefore, AI plays a vital role in providing various methods to analyze large and complex data in order to understand and improve knowledge about diseases. To address the concept of data sharing, there are models based on AI that are generative in nature and help to simulate synthetic subjects. In the following section, we will discuss the existing AI approaches used to simulate synthetic data.

1.3 Existing AI approaches in the field of synthetic data generation

As described above, the need for synthetic data generation often arises from challenging legal situations and the extensive timelines needed to share real data. The thought is to provide researchers/data scientists a mechanism by which they could get insights into patterns of the real data while at the same time not having access to it. The most important and challenging part is to propagate AI methods for the generation of synthetic data. Synthetic data generation technology enables the research community to digitally generate the data they need in a given volume which is tailored to their specific needs. The data that is high in quality and realistic can be leveraged towards the advancement of methodological developments in the field of medicine. The data owners generally anonymize or de-identify the data to make sensitive patient data available to others. This could be done in several ways, including removing easily identifiable features (e.g. names and addresses), perturbing them (e.g. adding noise to birth dates), or grouping variables into broader categories to ensure more than one individual in each category [44]. While it might not be easy to re-identify the individuals based on the residual information contained in properly anonymized data. However, once this information is linked to the existing data sets e.g. social media platforms, they may contain enough information to identify specific individuals. There have been some efforts to determine the efficacy of de-identification methods but this remains inconclusive, specifically in the context of large datasets [45]. Therefore, it remains an extremely difficult task to guarantee that the re-identification of individual patients is not a possibility with the aforementioned approaches.

Synthetic data generation has been explored for roughly three decades [46] and it has been applied across several domains [47, 48] including patient data [49, 50, 51]. There has been a lot of focus on synthetic data because of its several advantages. They could either replace the entire real data, augment it, or be used as a proxy resource to accelerate research. It has the potential to impact patient care drastically enabling research on model development to move at a quicker pace. There have been a number of models for generation of synthetic data [47] but each model uses a different dataset and different evaluation metrics. Therefore, it becomes difficult to directly compare synthetic data generation methods.

Synthetic data generation can be divided into two categories: process-driven methods and data-driven methods. Synthetic data in process-driven methods is derived from computational or mathematical models of an underlying physical process. For e.g. numerical simulations, monte carlo simulations, discrete event simulations, etc. On the other hand, data-driven methods derive synthetic data from generative models that have been trained on observed data. Here we will mainly focus on data-driven methods because the true mechanism behind these diseases has not been understood so far. That means due to a lack of knowledge, it is impossible to write down a differential equation system, which would explain a disease and the observed symptoms. Some of the types of data-driven methods are imputation-based methods, full joint probability distribution methods, and function approximation methods. Rubin [52] and Little [53] first introduced imputation based methods for synthetic data generation in the context of statistical disclosure control (SDC) or statistical disclosure limitation (SDL) [47]. These methods are majorly concerned with reducing the risk of leakage of sensitive data when performing statistical analysis. Mathews and Harrel [54] released a general survey paper on data privacy methods related to SDL. The standard techniques are based on multiple imputation [55], where sensitive data is treated as missing data and randomly sampled imputed values are released instead of the sensitive data. Later, Raghunathan, Reiter and Rubin [56] extended these methods to the fully synthetic case. The limitation of the imputation-based methods is that while they are completely probabilistic, there is no guarantee that the resultant generative model is an estimate of a full joint probability distribution of the target/sampled population.

Patient data have several categorical features that must be handled carefully while generating synthetic data. This brings a huge challenge, specifically in high dimensions. Therefore, it becomes necessary to impose a certain kind of dependence structure on the data [57]. An example of this is the BN proposed by Zhang et al. [58], which approximates a joint distribution using a first-order dependence tree as a method for generating synthetic data with privacy constraints. Some of the more flexible methods that do not impose such dependence structures on the distributions are Bayesian non-parametric methods for multidimensional categorical data. Examples are latent Gaussian process methods [59] and Dirichlet mixture models [60, 61]. Several state-of-the-art machine learning methods are based on function approximation methods for e.g. deep neural network (DNN)s. A large number of parameters and a large amount of training data is required by these models. Generative adversarial network (GAN)s are a prominent class of DNNs for unsupervised learning tasks [62]. Specifically, two jointly trained networks are produced in these; one generates synthetic data that is intended to be similar to the training data, and the other aims to discriminate synthetic data from the actual training data. These methods have shown to be capable of learning high dimensional, continuous data [62, 63]. GANs for categorical data have also been proposed by Caminon, Hammerschmidt, and State [64] having specific applications to synthetic EHR data by Choi et al. [65].

We describe some of the existing synthetic data generation methods in detail in the following paragraphs. We also discuss the advantages and limitations of these approaches.

- Sampling from independent marginals: This method is known as the independent marginals (IM) method as it is based on sampling from empirical marginal distributions of each variable. This empirical marginal distribution is estimated from the observed data. Underlying are the key advantages and disadvantages of this approach:

    - It is a computationally efficient approach and a parallel estimation of marginal distributions for different variables is possible.
    - Limitation of this approach is that it does not capture statistical dependencies across variables and therefore the synthetic data generated by this method may fail to capture the underlying structure of the data.

- BNs: BNs are probabilistic graphical models where each node represents a random variable while the edges between the nodes represent the probabilistic dependencies between these random variables. When BN is used for synthetic data generation, the graph structure and the conditional probability distributions are typically inferred from the real data. There are primarily two steps underlying the learning process [44]; a) learning a directed acyclic graphs (DAG) from the data, which indicates all the pairwise conditional (in)dependencies among the random variables [45] and b) estimating the conditional probability tables (CPT) for each variable via maximum likelihood. The graph structure obtained from the real data shows the conditional dependencies among the variables. Sampling from the inferred BN can be a method to generate synthetic data. The key advantages and limitations of this approach are described as follows:
    - One of the advantages is that BN allows for integrating highly heterogeneous data within one modeling framework [66] while allowing to address one of the main challenging aspects of clinical study data, namely a large number of missing values.
    - The second advantage is, BNs belong to the family of generative models and hence can be used to simulate synthetic patient trajectories after model fitting [67].
    - One of the limitations is that graph structure learning is an NP-hard problem that might either be too costly to perform or impossible when the subjects are in small numbers but they have a large number of features. Also, they could only model DAG and not arbitrary dependency structures.

- Multiple imputation methods: These methods have been quite popular for synthetic data generation, specifically for the part where the data is considered to be sensitive [47]. Multivariate imputation by chained equations (MICE) [68] is one of the existing imputation methods and it has emerged as a vital method for masking sensitive content in datasets with privacy constraints. Here, the key idea is to treat sensitive data as missing data. Thereafter, this "missing" data is imputed with randomly sampled values generated from models trained on the non-sensitive variables. The multiple imputation software "mice" in R is used for generating and analyzing synthetic datasets [69]. The advantages and limitations of this approach are as follows:

  - MICE is computationally fast and is able to scale to massive datasets, both in the number of variables and samples.

  - It can easily deal with categorical and continuous values by precisely choosing either a Softmax or a Gaussian model for the conditional probability distribution for a given variable.

  - MICE is probabilistic, however, its limitation is that it does not guarantee the resulting generative model to be a good estimate of the underlying joint distribution of the data.

  - The other limitation is that it firmly relies on the flexibility of the model for the topological ordering of the DAG and also continuous probability distributions.

- GANs: GANs [62] have shown to be successful for generating complex synthetic data, such as medical images [70, 71, 72, 73, 74, 75] and text [76, 77, 78]. In this approach, two neural networks are jointly trained in a competitive manner: realistic synthetic data is generated by the first network while real and synthetic data are discriminated by the second one. One of the methods called deep de-Aliasing generative adversarial networks for fast compressed sensing mRI reconstruction (DAGAN) for Fast compressed sensing (CS) magnetic resonance imaging (MRI) Reconstruction, which is a conditional GAN. Combining this approach with existing MRI scanning sequences and parallel imaging, the simulation-based study could be translated to the real clinical environment. A well-known limitation of GANs is that it is not suited for generating categorical synthetic datasets. Another example we discuss here is medical generative adversarial network (medGAN) used for medical image-to-image translation [65]. Considering clinical patient data are largely categorical, works like medGAN have applied autoencoders that enable the transformation of categorical data to a continuous space. Due to this additional property to GANs, it can be applied for generating synthetic patient data. As medGAN was only applicable to binary and count data, an extension of it is called multi-categorical medGAN (MC-medGAN) [64] that fits multi-categorical data. The advantages and disadvantages of MC-medGAN are mentioned in the following paragraphs:

- Unlike BN, MC-medGAN is a generative approach that does not require rigorous probabilistic model assumptions. Hence it is more flexible compared to BN.
- The models based on GANs can be easily extended to deal with mixed data types. e.g. continuous and categorical variables.
- One of the limitations is as MC-medGAN is a deep learning model, it has a large number of parameters. The main issue with a large number of parameters and thus high model complexity is that the model is prone to overfitting. Hyperparameter tuning is yet another problem.
- The other limitation is that it is known to be very difficult to train GANs as the process of solving the min-max optimization problem can be very unstable. However, certain proposed variations of GANs such as Wasserstein GAN and its variants have significantly alleviated the problem underlying the stability of training GANs [79, 80].
- Another limitation of GANs is that they are prone to a so-called mode collapse, i.e. they tend to fit very well close to the statistical model of distribution, but deviate significantly from the training data in low-density regions. Hence, synthetic patient samples tend to look similar to the “average” patient. Once again, Wasserstein-GANs have been

proposed to address this issue.

Overall, the major limitation of GANs is that it cannot explicitly model time dependencies while accounting for missing and heterogeneous data.

- Autoencoders: Autoencoders are a specific type of neural network which are majorly designed to encode the input into a compressed and relevant representation, and then decode it back in a way that reconstructed output is as similar as possible to the input [81]. A special case of autoencoders that are deep generative models, called varitational autoencoder (VAE)s [82] [83] additionally employ variational inference to regularize the encoding distribution and ensure that the generation of new data is less prone to overfitting. The advantages and disadvantages of VAEs are mentioned as follows:
  - VAEs are generative because drawings from the latent distribution can be decoded again.
  - VAEs have recently been extended to deal with heterogeneous multimodal and missing data [84], which is the common situation in clinical studies, known as HI-VAE.
  - The limitation of VAE is that in a situation with comparably small data, a dense VAE model with several hidden layers could easily overfit.
  - Another limitation is, that interpretation of the neural network models is far more challenging than for BNs.

Additionally, several open-source packages also exist for synthetic data generation. Some of the examples include:

- R package synthpop: It focuses on synthesis of individual variables by sequential regression modeling [85].
- R package simPop [86]: It focuses on the simulation of synthetic populations based on household survey data and auxiliary information.
- Python package DataSynthesizer [48]: It uses BNs (with differential privacy respecting model training) or independent sampling from attributes, depend-

ing on the complexity and availability of data.

- Java-based simulator Synthea [50]: It is a rule-based approach for the synthesis of EHRs. It does not allow for simulating anything else.

Above, we have described various methods for synthetic data generation along with their advantages and their limitations. It makes it clear that the existing methods have several limitations that we need to overcome for generating synthetic data. These methods have limitations do deal with the following characteristics of clinical data; limited sample size while having an extensive number of features, inability to model time series, heterogeneous multi-modal and multi-scale data having missing values, inability to handle the computational complexity, not able to preserve the dependency structure of the real data in the synthetic data. Therefore, our work aims to overcome all these limitations. We have described the aims and summary of our approach in detail in the following section.

1.4 Aims and summary

The aim of this thesis is to use AI and specifically machine learning methods to model and simulate clinical studies. Considering the limitations of existing approaches described above, we develop a method for synthetic data generation that can handle these limitations. Our method can handle limited sample size, highly heterogeneous data with many variables having different distributions, multiple scales, and longitudinal data with many missing values. This thesis specifically aims to simulate the data from the patients and use this simulated data to further solve the issues related to data privacy and data sharing. As BN has several advantages, therefore our method takes BN as a base and further builds upon it. The advantages of BN that led us to primarily focus on it are mentioned below:

- BN doesn't need an enormous sample size.
- There is a straightforward mechanism to deal with missing values.
- They are white box models.
- They allow for simulating counterfactual scenarios.

- BN allows to create a synthetic representation of the original cohorts.

We also overcome the limitations underlying BN which are described in the next chapters.

The thesis is structured as follows:

- Chapter 2 provides the theoretical background of the various statistical and machine learning methods and approaches that have been used in our work.
- Chapter 3 comprises an evaluation of how far BNs allow to generate synthetic data from a longitudinal PD study with few variables. In particular, we introduce a concept to model missing values and mixed static and longitudinal data.
- Chapter 4 comprises the similar generative approach that was discussed in chapter 3 but with a modification. The main novelty is the extension of the BN concept such that high dimensional data can be incorporated by introducing the notion of a modular BN. The concept of sparse autoencoders was also introduced here as these were used for reducing dimensionality. This also includes the application of the approach to both AD and PD comprising well-established longitudinal observational cohort studies.
- Chapter 5 adds one more modification to the generative approach by replacing the sparse autoencoders with variational autoencoders. This approach is termed VAMBN. This decision was taken to overcome the limitations of sparse autoencoders. The main novelty of this method is that the approach avoids any discretization of data. VAMBN allows modeling of mixed static and longitudinal data of various heterogeneous data types and can deal with missing values. To make sure that the real data is not identified in the simulated set of data, differential privacy was also applied to the datasets.
- Chapter 6 talks about the application of VAMBN on AD studies and synthetic data generation in the context of data derived from sensor and device technologies. It discusses the generation of a global meta-cohort using two different studies. It also discusses the links between the data measures from the digital devices and already established questionnaire-based scores in AD.

- Conclusion brings the thesis to an end by summarizing the essence of the thesis. It primarily talks about the core message of using AI models for synthetic data generation and modeling of longitudinal data. It also discusses the scientific accomplishment of the thesis and its limitations along with the future outlook for research in patient simulation. longitudinal disease modeling and analysis of data derived from sensor and device technologies.

The work in this thesis is based on the following publications:

- “Bayesian network modeling of risk and prodromal markers of Parkinson’s disease, Preprint medRxiv, 2022”. The paper has been submitted to PLOS ONE and is currently under review.
- “Sood, M., Sahay, A., Karki, R., Emon, M. A., Vrooman, H., Hofmann-Apitius, M., Fröhlich, H. (2020). Realistic simulation of virtual multi-scale, multi-modal patient trajectories using Bayesian networks and sparse autoencoders. Scientific reports, 10(1), 10971. https://doi.org/10.1038/s41598-020-67398-4”
- “Gootjes-Dreesbach, L., Sood, M., Sahay, A., Hofmann-Apitius, M., Fröhlich, H. (2020). Variational autoencoder modular Bayesian networks for simulation of heterogeneous clinical study data. Frontiers in big Data, 3, 16.”
- “Evaluating Digital Device Technology in Alzheimer’s Disease via Artificial Intelligence, Preprint medRxiv, 2021.”

The smallest of actions is always better than the noblest of intentions.

Robin Sharma

2

THEORETICAL BACKGROUND

In this chapter, we will discuss the theoretical background underlying the algorithms that we have used in this thesis.

2.1 Bayesian networks (BNs)

BNs are probabilistic graphical models, where nodes represent variables and edges represent probabilistic stochastic dependencies between them [87]. These stochastic dependencies are characterized by conditional probability distributions (CPD), one for each variable. It encodes a joint probability distribution over a set of random variables $X = X_1, ..., X_n$. By definition, a BN is a pair G, Θ, where G is a DAG in which each node corresponds to one of the random variables [88] and Θ is a set of network parameters. The parents of X_i are represented by PA_i. Given its parents, X_i is independent of its non-descendants. The conditional probability distributions $P(X_i|PA_i)$ for each X_i are specified by network parameters, Θ. A crucial concept in BN, known as markov blanket (MB) [89] is important to understand. An MB of a node in a BN consists of its parents, children, and spouses (i.e. other parents of their common children). BNs follow an assumption called the markov assumption, meaning that every node in a BN, given its parents, is conditionally independent of its non-descendants. According to the markov chain rule; where every random variable X_i is directly dependent on its parents PA_i [90], the joint probability distribution for all the variables represented by a BN can be decomposed into a product of conditional probabilities using the graphical structure and the chain rule of probability calculus. This is represented by [91],

$$P(X|\theta) = \prod_{i=1}^{n} P(X_i|PA_i, \theta_i) \tag{2.1}$$

θ_i are the parameters of X_i.

To present the applicability of BN we illustrate an example of a hypothetical BN in Figure 2.1. Here the problem is to figure out the possible causes of cancer [91]. The possible causes of cancer could be exposure to UV radiation (R), smoking (S), unhealthy diet (D), obesity (O), alcohol (A), and physical inactivity (I). There is an edge from O to C indicating a higher risk of cancer for obese people. There are also direct edges from R and S to C indicating that cancer risk is dependent on UV radiation and smoking. D and I have edges pointing towards O suggesting that diet and physical activity are two of the reasons for obesity. Most importantly, there is no direct edge from D and I to C, this implies that knowing the state of O renders C independent of D and I. The joint distribution of all five variables can be factored by the following equation:

$$P(D, I, R, O, S, C) = P(C|R, O, S) \cdot P(R) \cdot P(S) \cdot P(O|D, I) \cdot P(D) \cdot P(I) \tag{2.2}$$

If each of the five variables in Figure 2.1 is assumed to be binary, the above factorization reduces the conditional probabilities (number of parameters) required to specify the full joint distribution from 64 to 14. This allows small datasets to parameterize large networks while still capturing all kinds of complex interactions.

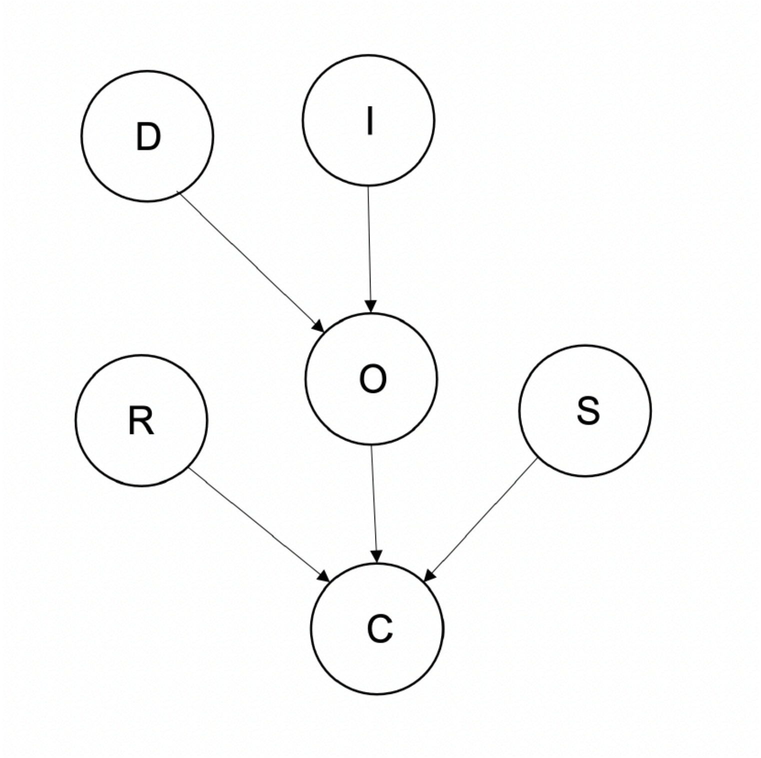


Figure 2.1: A BN representing the relationship between cancer incidence (C), UV radiations (R), obesity (O), physical inactivity (I), and unhealthy diet (D).

In most cases, the edges in the BN structure are unknown and therefore they need to be inferred from data. An important question is, how far the learned structure reflects existing causal relationships. In principle, it is not possible to uniquely identify the underlying causal DAG from observational data. Here comes the concept of markov equivalence. Markov equivalence is an equivalence graph structure present in BNs. Two graphs are known to be equivalent if and only if the set of markov properties of one graph is satisfied by the other graph [92]. Markov equivalence contains all DAGs encoding the equivalent conditional independencies and can be characterized by a completed partially directed acyclic graph (CPDAG) [93]. Indeed, if the BN is faithful to the underlying statistical distribution (i.e. models it correctly), then the true causal network is known to be part of a class of these equivalent graph structures. [67, 94].

There are several advantages of BNs some of which are listed below:

- They efficiently encode multivariate distributions and multinomial data.
- They are interpretable because the underlying graph structure can represent causal relationships.
- They exhibit a theoretical framework to simulate interventions via the "do" calculus. Judea Pearl developed a well-established theory for modeling and simulating interventions into BNs [95]: Assume we want to predict the intervention effect of $X_i = x$ on the remaining random variables in the BN, denoted as $P(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n | do(X_i = x))$. Pearl demonstrated in his work that this intervention effect can be computed by estimating the conditional probability distribution $P(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n | X_i = x)$ within a mutilated BN, in which all incoming edges into X_i have been deleted.

Despite these advantages, there comes a limitation as well. The most important constraint on the use of BN's is that the computation is NP-hard [96]. There are two NP-hard tasks in BNs, 1) Structure learning, and 2) Inference of the value of a specific random variable X_i: if we do not constrain conditional probability distributions to multinomial and Gaussian.

We need a set of parameters and a structure that can best encode the joint probability distribution. These are accounted in the BN learning process which is described in the following section.

2.1.1 Learning BN

The main aim of the BN learning process is to find the most suitable network that best encodes the joint probability distribution of a domain [97]. In general, BN learning can be divided into two parts, parameter learning, and structure learning; parameter learning involves describing the conditional probability distributions underlying a BN [98] and structure learning comprises finding the optimal DAG. These are described below:

Parameter Learning

It is the process of using data to learn distributions underlying a BN.

There are different types of parameter learning algorithms; maximum likelihood estimation (MLE), bayesian method, expectation-maximization algorithm, robust bayesian estimate, monte-carlo method, and gaussian approximation [90]. Given the scope of our work, here we will discuss the first two algorithms; MLE and Bayesian method. Parameters of a BN can be estimated from a data sample D. Here we assume, that the sample has no missing values.

We will describe the two methods as follows:

- MLE: It is a method of estimating the parameters of an (unknown) probability distribution by maximizing a likelihood function so that the observed data is most probable under the assumed statistical model. As mentioned above, a given set of observations, $(X_1, X_2, ..., X_n)$, from an unknown population is considered as a random sample [99]. The goal of MLE is to make inferences about the population that is most likely to have generated this sample, precisely the joint probability distribution of the random variables.
- Bayesian method: It is defined by the idea: given a distribution with unknown parameters and a complete set (C) observed data, Θ is a random variable with a prior distribution $p(\Theta)$, the changes of parameter Θ, namely

$p(\Theta|C)$, can be estimated according to the previous knowledge of the assumption of $p(\Theta)$. $p(\Theta|C)$ therefore represents the posterior probability of Θ. The aim of this method is to calculate the posterior probability which is then considered as a basis of parameter learning.

Considering the assumption of statistical independence between different parameters of a network, their estimation decomposes into various independent estimation problems, one for each variable and the potential configuration of parents. $p(\Theta|X_i, PA_i) = p(\Theta|X_i)$, because X_i and PA_i are statistically independent. The standard estimate for a parameter $\Theta|X_i$ is:

$$p(\Theta|X_i) = \frac{p(\Theta, X_i)}{p(X_i)} \tag{2.3}$$

An elementary difference between MLE and the Bayesian method is that point estimates of the parameters are provided by MLE method while Bayesian method maintains a constantly updated distribution over these parameters. Therefore, as new examples are provided, it enables the Bayesian method to continuously learn and improve new parameters [100].

Structure learning

For a dataset, $D = D_1, ..., D_N$, of the available variables in V, structure learning of BN is the problem of learning a network structure from dataset D.

The process of structure learning is NP hard [96] and a substantial amount of work in the research community has been dedicated to identifying heuristic-search techniques to identify good models. Here we will discuss six different types of structure learning algorithms. These algorithms are divided into score-based, constraint-based and hybrid algorithms. Greedy hill climbing (hc) and tabu search [101] are heuristic score-based optimization approaches, whereas MMPC [102] and SI-HITON-PC[103] are constraint-based approaches. Max-min hill-climbing (MMHC) [102] and restricted maximization (RSMAX2) [102] fall in the hybrid category as they used ideas from both score-based and constrained-based approaches.

Score-plus search-based algorithms: Score-based is a widely used approach for structure learning [87]. As the number of possible structures that can be learned from a graph is super-exponential with the number of nodes $|V|$, learning an optimal BN from D is considered to be an NP-hard problem [104]. Therefore, a lot of previous work focused on algorithms such as hc [105], and tabu search as they consider random restarts [106]. They limit the number of parents and parameters for each variable [107] and search the space representing equivalence classes of network structures [108]. These aforementioned algorithms use local search to search "good" networks; nonetheless, there is no guarantee offered by these algorithms to find the network that optimizes the scoring function. The main task is to find the most suited DAG based on certain score functions that measure its fitness to the data. A scoring function is used to measure the goodness of fit of a structure to the data in this approach. The goal of the learning problem here is then to search for the optimal scoring structure. Generally, the score approximates the probability of the structure given the data and represents a trade-off between how well the network fits the data and how complex the network is. Here, a decomposability assumption is made for the score [109]. That means, that we can calculate the score for a network structure as the sum of scores for the individual variables, and here the score for a variable is calculated entirely on the basis of a random variable and its parents [88]. Therefore,

$$Score(G|D) = \sum_{i=1}^{n} Score(X_i|PA_i, D), \tag{2.4}$$

Several scoring functions are represented in the form of a penalized log likelihood (LL) function. The likelihood of the data, given a structure, can be calculated as:

$$LL(D|G) = \sum_{j}^{N} logP(D_j|G) = \sum_{i}^{n} \sum_{j}^{N} logP(D_{ij}|PA_{ij}), \tag{2.5}$$

where $LL(D|G)$ is the log probability of D given G, D_{ij} is the instantiation of X_i in data point D_j and PA_{ij} represents the instantiation of X_i's parents in D_j. While adding an arc to the network, the arc could be ignored if it does not add any extra information to the network. The extra arcs could possibly lead to two problems, firstly they could lead to the overfitting of the training data and could result in lower performance on testing data. Secondly, the networks that are densely connected could increase the running time when they are used for inference and prediction type of analysis. This overfitting problem can be addressed by adding a penalty to the LL function and is represented by a penalized LL function. This helps to penalize the complex network. Hence, despite a very good LL score for complex networks, a high penalty term may help to reduce the score to fall below that of a less complex network. This is called decomposable penalized log likelihood (DPLL) scores and they are illustrated in the following form:

$$DPLL(G, D) = LL(D|G) - \sum_{i=1}^{n} Penalty(X_i, G, D) \tag{2.6}$$

Two of the most widely used DPLL scores for learning BN's are bayes dirichlet (likelihood) equivalent uniform (BDeu) [110] [105] [109], and Bayesian information criterion (BIC) [111]. The difference between these scoring functions is in their penalty terms.

- BDeu: It is a scoring function where 'e' stands for likelihood equivalence and 'u' stands for uniform distribution. Two assumptions called likelihood equivalence and structure possibility were defined by Heckermann, Geiger, and Chickering (1995) [109]. If two DAGs encode the same joint probability distributions, they are called equivalent. Bayesian dirichlet (BD) function computes the joint probability of a network for a given dataset [105]. The likelihood equivalence likelihood equivalence Bayesian drichlet (BDe) score was induced by heckermann, geiger and chickering (HGC95) theorem [109], and its expression is identical to BD expression. Buntine [110] proposed a particular case of BDe score, known as BDeu score. The score only depends on one assumption which is the equivalent sample size, referred to as N. N can be used to calculate all the required parameters and prior distribution over network structures. Under the assumption that all the network structures are equally likely, which means prior distribution over the network structures is uniform, N is the only input required for this scoring function. BDeu measures the posterior probability of a selected DAG given the available data while assuming uniform prior probability distributions on all the possible networks. BDeu is a DPLL scoring function and its penalty term is represented as follows:

$$PenaltyBDeu(X_i, G, D) = \sum_{j}^{q_i} \sum_{k}^{r_i} log \frac{P(D_{ijk}|D_{ij})}{P(D_{ijk}|D_{ij}, N_{ij})}, \tag{2.7}$$

where possible values of PA_i is represented by q_i, r_i is the number of possible values for X_i, D_{ijk} is the number of times $X_i = k$ and $PA_i = j$ in D, and the parameter determined based on the N specified by the user is represented by N_{ij}. The density of optimal network structure is learned with the scoring function BDeu, and it is correlated with N. Low N values give rise to sparser networks. Selecting an appropriate N could be difficult if the density of the network to be learned is unknown.

- BIC: BIC approximates BDeu. Both BDeu and BIC share the property of decomposability. It is represented by the formula:

$$BIC = -2 * LL + \Theta * log(N), \tag{2.8}$$

where N is the sample size of the training set and Θ is the total number of parameters. A good model requires a lower BIC score. The $\Theta * log(N)$ here is the penalty term and it grows with the number of parameters. The penalty term enables accurate estimation of the network on the given data set with size N, by filtering out over-complicated networks with too many parameters.

We will discuss primarily two types of score-based algorithms; hc [102] and tabu [101] search algorithm.

- hc: In these algorithms, network DAG is scored by the function f with regard to the training data, and a search method is used to explore a network with the best score [112]. This algorithm traverses the search space by beginning from an initial solution and executing a finite number of steps. At each step, the algorithm considers neighboring DAGs and chooses the one with the largest improvement in f. The algorithm terminates when there is no local change resulting in an improvement in f. Due to this greedy behavior, the execution terminates when the algorithm is trapped at a solution that maximizes f locally rather than globally. They explore the search space of DAG by single-arc addition, arc removal, and reversal, with random restarts to avoid local optima [113]. Due to the super-exponential cardinality of the search space [114], it seems a good objective to limit the areas of search space to be visited, particularly in domains with a substantial number of variables. The efficient evaluation of neighbors/DAGs is based on the scoring metrics which means decomposability in the presence of complete data. Hence, usage of the decomposability metric helps to efficiently evaluate the neighbor due to the change in only one arc at each move. This can further help to reuse the computation carried out in previous stages, and only the statistics related to the variables whose parents have been modified need to be recomputed. As hill climbing uses the operators of arc addition, deletion, and reversal, it takes advantage of the above-mentioned operation mode.

hc algorithms are prominent because of their trade-off between the quality of models learned and computational demands.

- Tabu search: Tabu search is a sub-heuristic algorithm that mimics the function of human memory [115]. It has fewer parameters and a simple structure and it can rely on a local search, akin to greedy hill climbing. Similar to the hill climbing approach, this algorithm also utilizes three operations; addition, removal, and reversal of edges, that generate neighborhoods without generating a network loop. It explores the neighbourhood for a local optimal solution and places it into a tabu table. After this step, in order to avoid duplication of the search process, tabu table serves as the tool to search the local optimal solution. This is done to make the next search as far as possible from the current one. Tabus are one of the distinctive features of tabu search when compared to hill climbing or other local search methods [116]. They also help to move the search away from the portions of the search space that have been visited before and thus, perform extensive exploration. As discussed above, while hill climbing can get stuck in local optima, tabu search maintains a tabu list to avoid the same. The tabu list holds objects that are recently used and are taboo to use for now. Consequently, moves that are comprised in tabu list are not accepted. Another aspect of tabu search is that it accepts a certain number of worsening moves if no improvement is possible. Hence tabu, despite still being a local search approach - has a higher chance to escape local optima than hc.

CONSTRAINT-BASED ALGORITHMS: Constraint-based algorithms can be used to learn causal graphical models under certain assumptions [117]. They identify conditional independent constraints with statistical tests and link nodes that are not independently observed. They provide a framework for learning the DAG of a BN using these tests taking the assumption that probabilistic independence and graphical separation imply each other [118]. The common tests that are used are, the mutual information test (for discrete BNs), the exact Student's t-test, and the Fisher's Z transformation for correlation (for gaussian Bayesian network (GBN)s). All constraint-based structure learning algorithms share a three-phase structure. The first phase is optional and consists of two steps. The first step is learning the MB of each node so that the number of potential DAGs can be reduced early on. Any algorithm that aims to learn MB can be plugged into step 1 and extended into a complete BN structure learning algorithm [119]. After all, MBs have been learned, they are checked for consistency (step 2) using their symmetry; by definition:

If any asymmetries are left, they are corrected by treating them as false positives and removing the violating nodes from each other's MBs. In the second phase, algorithm learns the skeleton of DAG, which is equivalent to learning the neighbors $N(X_i)$ of each node including their parents and children. If certain set of nodes $S_{(X_i,X_j)}$ separating a particular pair X_i, X_j are absent, it implies that either $X_i \rightarrow X_j$ or $X_j \rightarrow X_i$. Finally, arc directions are established [120] in the third phase. For some arcs, both directions are equivalent, which means they identify equivalent decomposition of the global distribution. This leads some arcs to be undirected and the algorithm will return a CPDAG that identifies an equivalent class containing multiple DAGs. In contrast to the score-based algorithms, the constraint-based approach mitigates the problem of heavy computational costs and extends the available network learning size, thus allowing the learning of larger BNs. A significant problem of these algorithms is their lower accuracy in comparison to the score-based approach. Here we discuss two types of constraint-based algorithms, MMPC and SI-HITON-PC.

- MMPC: Given that these algorithms consist of three stages, here first stage starts with an empty set and each node has all other nodes as potential parents. All possible variables for the parent-child set are added. The second stage removes several variables from this set via conditional independence tests. These first two stages result in all members of the parent-child set but it can also contain some false positives i.e. variables that are not included in the parent-child set. After the two stages, all the false positives are removed from the network. Finally, after stage three, the complete parent-child set of the target variable T is returned by the algorithm.
- SI-HITON-PC: It is a fast-forward selection technique for neighborhood detection that is designed to exclude the nodes early on, based on marginal association.

Hybrid algorithms: As the name suggests, hybrid algorithms are a mix of score-based and constraint-based algorithms. It uses both conditional independence to reduce the space of candidate DAGs and network scores to identify the optimal DAG amongst them. Here we will briefly explain two examples, MMHC and RSMAX2.

- MMHC: This algorithm is a combination of ideas from local learning, constraint-based, and search-and-score approaches. Firstly, it reconstructs the skeleton (i.e. the edges without their orientation) of a BN using MMPC and then orients the edges using greedy Bayesian scoring hill-climbing search.
- RSMAX2: It is a more general implementation of MMHC as it can use any combination of constraint-based and score-based algorithms [113].

2.2 Modular Bayesian networks (MBN)

As described above, learning DAG structure is NP hard [96], and it requires limiting the space of potential networks as much as possible. This is of distinct importance for datasets with many variables and a limited sample size. Therefore, module networks [121] have been introduced to address this situation. The key idea in module networks is to group variables into modules that share parameters [122]. During the BN structure learning process, only edges between modules are learned. The variables in the module have the same statistical behavior with the same set of parameters and local probabilistic model [121]. Enforcing this constraint on the network significantly helps to reduce the complexity of model space as well as the number of parameters. Thereafter, these reductions lead to more robust estimation and better generalization for unseen data. The explicit representation of the module network helps to gain insight into the domains that are often obscured by the intricate details of a large BN. The exact definition of variable groups differs between the datasets used. The key question is how to learn and encode a shared distribution for a module. In their original publication, Segal *et.al.* [121] assumed normally distributed data (such as gene expression) and employed decision trees to represent modules [121, 122]. The set of variables in a module shares the same set of parents and CPD. Here, each module is represented by a formal variable that is used as a placeholder for variables in the module. A module set C, consisting of K modules is a set of variables $M_1.....,M_K$. All the variables in a module should also have the same domain as they share the same CPD. $Val(M_j)$ is used to represent the set of probable values of the formal variable of the j^{th} module. Module component is defined by two components. The first component defines the template probabilistic model for each module in C and the model is shared by all the variables available in the module. Module network template, $T = (S, \Theta)$ for C, defines for each M_j:

- a set of parents $P_{a_{M_j}} \subset X$;
- a template of conditional probability distribution $P(M_j|P_{a_{M_j}})$ which specifies a distribution over $Val(M_j)$ for each assignment in $Val(P_{a_{M_j}})$

S is used to denote the dependency structure encoded by $\{P_{a_{M_j}} : M_j \in C\}$ and

Θ to denote the parameters that are required for CPD templates $\{P_{M_j}|P_{a_{M_j}}) : M_j \in C\}$.

The second component defines the function for module assignment that assigns each variable $X_i \in X$ to one of the K modules $M_1 ..., M_K$. A variable can only be assigned to the module having the same domain.

2.3 Logic sampling algorithm

This algorithm is used to generate synthetic data from BN. We want to investigate a new evidence E using the knowledge that is encoded in the BN. The posterior distribution can be investigated by $P(X|E, \text{BN}) = P(X|E, G, \Theta)$. The values for the root nodes are sampled from their (unconditional) distribution [113]. Furthermore, the values for the distribution of their children conditional on the respective sets of parents are generated. This process is performed iteratively until values for all nodes have been sampled.

2.4 Likelihood weighting algorithm

Likelihood weighting algorithm is an inference-based algorithm that helps to infer the value of node conditional on the observations for all the other nodes in the BN [113]. The nodes with known values are referred to as evidence variables, E. Therefore, we need to query the remaining nodes given the evidence nodes, to determine the state of the entire network. Here the sampling is not done from the original BN but from the BN where all nodes are in evidence E and are fixed. This network is known as a mutilated network. After the sampling is complete, a likelihood weight is assigned to the sample by multiplying the probabilities of each evidence variable given its parents. Thereafter, this result is stored in a map, represented by W here, that encompasses the association of all the variables in the network with its weight. To further explain it in more detail:

- A temporary variable w is assigned to 1 and it holds the calculated weight of the sample.
- A temporary variable x is set to empty and it holds the state of each node

for this sample.

- After this, each node in the network is examined and if the node is E node, the following calculation is performed, $w = w * p(\text{current node}|\text{parents of current node})$, where p is the probability of the E node given its parents. If the current node is not an E node, then its state is determined by sampling it. It does not add anything to the weight calculation. The state of the node is added to x whether it is an evidence node or whether it is discovered through sampling.

We will be left with x (state of the network) and w (likelihood weight associated with that state) after the entire network is examined for the particular sample. This is further added to the map, W where x is used as the key and w as the data value. If W already consists x, then w is added to the data value that is associated with x in W.

2.5 Missing data

One of the key challenges and common problems with longitudinal patient data is missing values. Overlooking missing data can lead to loss of statistical power, introduces bias in the estimation of parameters, can reduce the representativeness of the samples, cause incorrect estimation of variability in the data, and may complicate the analyses of the study [123]. In longitudinal clinical studies, many subjects could be present at baseline, or at one-time point and missing at another time point. This leads to non-monotone missing data patterns [124]. The missingness in the data can originate due to various reasons: (a) patients withdrawal from the study, e.g. because of worsening symptoms; (b) a certain diagnostic test is not measured at a particular visit (e.g. due to lack of patient agreement) (c) unclear further reasons like lack of availability of time, issues in the quality of data, etc. From a statistical point of view, these reasons divulge into three mechanisms of missing data which are explained as follows [46] [125]:

- Missing completely at random (MCAR): The probability underlying missing information is not related to either the specific value which is supposed to be obtained or other observed data. Hence, entire patient records could be skipped without introducing any bias. However, this type of missing data mechanism is probably rare in clinical studies.

- Missing at random (MAR): The probability of missing information depends on other observed data, but is not related to the specific missing value which is expected to be obtained. An example would be a patient drop out due to the worsening of certain symptoms, which are at the same time recorded during the study.
- Missing not at random (MNAR): Any reason for missing data, which is neither MCAR nor MAR. Here the missingness might depend on the unobserved data in addition to the observed data. MNAR is problematic because the only way to obtain unbiased estimates is to model missing data.

2.5.1 Dealing with missing data

In most of the longitudinal studies, the missing data is a combination of MAR and MNAR. Typically, multiple imputation methods have been proposed to handle missing data in longitudinal patient data [125]. One commonly used approach to handle missing data is complete case (CC) analysis [126]. CC analysis often assumes MCAR data and estimates the mean by the sample mean of the completers. In this case, if the patients who drop out are healthier or sicker than the patients who complete the study, the CC estimator will have a low or a high bias respectively. The other common approach to handling the missing data is the last observation carried forward (LOCF). LOCF just replaces any missing values of variable V after visit t by the value that V had at t. There are other types of single imputation approaches such as mean substitution, regression imputation, maxmin-likelihood, expectation maximization (EM), etc. These types of imputation approaches tend to underestimate the standard errors and thus overestimate the level of precision of the estimator. Therefore, “multiple imputation approach” have been established to deal with missing data [127]. It means multiple datasets can be created and the inference can be averaged over these datasets. They have several advantages compared to single imputation approaches. They have been shown to produce valid statistical inference that reflects the uncertainty associated with the estimation of missing data. Additionally, they have been proven to be robust to the violation of normality assumption and generate relevant results even for data set with high missingness and small sample size.

2.6 Autoencoders

Autoencoders are a specific type of neural network which are majorly designed to encode the input into a compressed and relevant representation and then decode it back in a way that the reconstructed output is as similar as possible to the input [81]. The main goal of autoencoders is to learn an "informative" representation of the data that can have various functions one of them being clustering in an unsupervised manner. Briefly, autoencoders perform non-linear dimensionality reduction [128]. An autoencoder takes a feature vector $x \in \mathbb{R}^{\mathrm{d}}$ as input and transforms/encodes it to a hidden representation $\tilde{x} \in \mathbb{R}^{\mathrm{q}}$ via

$$\tilde{x} = s(Wx + b) \tag{2.9}$$

where $s(\cdot)$ is a non-linear activation function e.g. sigmoid, rectified linear unit. Matrix W consists of weights and b is a bias vector. Several encoding steps can be performed sequentially, resulting in a deep autoencoder. The latent representation $\tilde{x}$ can be decoded/mapped back via

$$z = s^{'}(W^{'}\tilde{x} + b^{'}) \tag{2.10}$$

where $\mathrm{W}^{'}$, $\mathrm{b}^{'}$ are the parameters of the decoder that are not necessarily identical to the encoder. Moreover, $\mathrm{s}^{'}(\cdot)$ is a non-linear activation function, which may be different from $\mathrm{s}(\cdot)$. Autoencoders are trained to minimize the difference between reconstructions z and original inputs x. MSE is one of the ways to measure it.

2.6.1 Standard variational autoencoders (VAE)

Standard VAE [81] is considered to be the most popular form of autoencoders. VAEs were introduced by Kingma and Welling [82] and can be interpreted as a special type of BN, which has the form $Z \rightarrow X$, where Z is a latent, usually multivariate standard Gaussian, and X a multivariate random variable describing the input data. Moreover, for any sample (x, z), we have $p(x|z) = N(\mu(z), \sigma(z))$. One of the key ideas behind VAEs is to variationally approximate

$$log\, q(z|x) = log\, N(z|\mu(x), \sigma(x)) \tag{2.11}$$

This means that $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation of the approximate posterior and are outputs of a multilayer perceptron neural network that is trained to minimize for each data point, x the evidence lower bound (ELBO) criterion

$$log(x) \geq \frac{1}{2}\sum_{j=1}^{D}(1 + log\, \sigma_j(x)^2 - \mu_j(x)^2 - \sigma_j(x)^2) + \sum_{l} log\, p(x|z^{(l)}) \tag{2.12}$$

where $z = \mu(x) + \sigma(x) \odot \varepsilon^{(l)}$ with $\varepsilon^{(l)} \sim N(0, 1)$. Here, $\odot$ denotes element-wise multiplication.

2.7 Random forest (RF)

The RF is an "ensemble learning" technique constructed from an aggregation of a large number of decision trees, resulting in a reduction of variance compared to the single decision trees [129]. It is a significant modification of bagging that builds an extensive collection of de-correlated trees [130]. It helps to reduce the variance as compared to a single decision tree by averaging many noisy but approximately unbiased models.

The steps for generating an RF are explained below, for $b = 1$ tree to B number of trees:

1. Firstly, a bootstrap sample Z* of size N is drawn from the training data.
2. A RF tree, T_b is constructed for each bootstrapped sample, by recursively repeating the following steps for each leaf node of the tree. The following steps are repeated until the minimum node size n_{min} is reached.

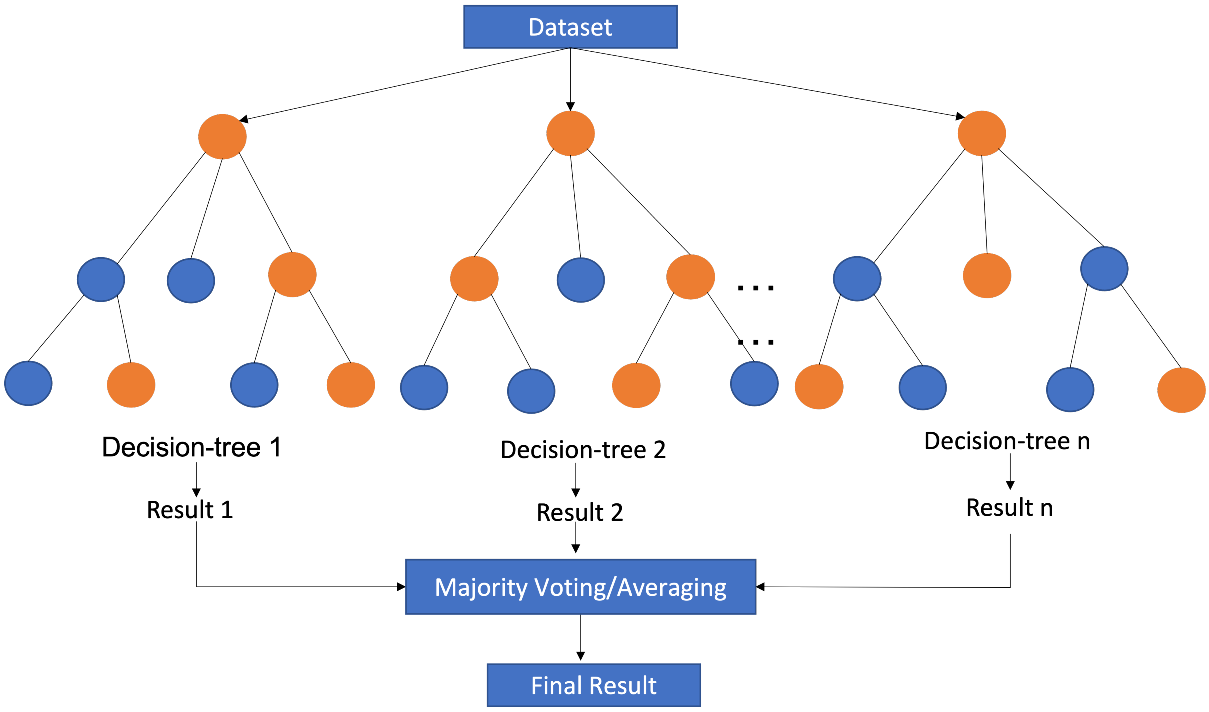


Figure 2.2: RF construction using multiple decision trees.

- Random m variables are selected from a list of p variables.
- Then, variables are picked according to the best split-point among m.
- Lastly, two daughter nodes are created from each node.

3. The ensemble of trees ${T_{b1}}^B$ are obtained as a result

An illustration of RF is provided in Figure 2.2: An average of independently and independently and identically distributed (i.i.d.) B random variables, each having variance σ^2, has total variance of $\frac{1}{B}\sigma^2$. For only identically distributed (i.d.) and not independent variables with having a positive pairwise correlation ρ, the variance of the average is,

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2.13}$$

In equation 2.13, if B increases, only the first term stays, and therefore the benefits of averaging are limited by the size of the correlation of pairs of bagged trees. The concept behind RFs is to reduce the correlation between the decision trees, without increasing the variance substantially and hence improve the variance reduction of bagging. The tree-growing process through random selection helps to achieve this.

Before each split, while growing a tree on a bootstrapped dataset, $m \leq p$ of the input variables is selected at random as a candidate for splitting.

Typically, values for m are $\sqrt{p}$ or as low as 1. The RF predictor is represented by the following equation after B such trees.

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x; \theta_b) \tag{2.14}$$

where θ_b depicts the bth random forest tree with regard to split variables, cutpoints at each node, and terminal node values. When m is reduced, it also assists in reducing the correlation between any pair of trees in the ensemble, and hence according to equation 2.14, it will reduce the variance of the average.

RF approach can be used both for classification and regression problems. In classification, an RF attains a class vote from each tree, and a majority vote is used for final classification. The default value here for m is $\sqrt{p}$ and the minimum node size is 1. On the other hand, in regression the predictions from each tree are averaged at a target point x. Here the default value for m is $p/3$ and the minimum node size is 5.

Out-of-bag (OOB) error estimate is used to calculate the performance of random forests. It is calculated by constructing a random forest predictor by averaging the trees corresponding to bootstrap samples considered as OOB or for which each observation $z_i = (x_i, y_i)$ did not appear.

2.8 Logistic regression

Logistic regression is a class of regression analysis that is used in classification when the dependent variable is categorical. Like all regression analyses, logistic regression is predictive and is used to explain the relationship between the dependent and independent variables. In a binary logistic model, the dependent variable has 2 possible values and is labeled as 0 or 1. The log-odds for the value labeled as 1 is a linear combination of the corresponding predictor variables, which is converted by the model into a probability varying between 0 and 1 using a logistic function. The input class is then determined by choosing a cutoff value, where inputs are classified in one class if their probability is higher than the cutoff or in another class if their probability is lower.

Consider a model with predictors $X = x_1, ..., x_i$ and a binary dependant variable $p = P(Y = 1|X)$, logistic regression can be expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 .... \beta_i x_i \tag{2.15}$$

where $\frac{p(X)}{1-p(X)}$ is called the odds.

and the probability of $Y = 1$:

$$p(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 x_1 ... \beta_i x_i}}{1 + e^{\beta_0 + \beta_1 x_1 ... \beta_i x_i}} \tag{2.16}$$

The only limit to our realization of tomorrow will be our doubts of today.

Franklin D. Roosevelt

3

Development of generative AI-based approach using Bayesian Networks

This chapter is an adaptation of our work "Bayesian network modeling of risk and prodromal markers of Parkinson's disease, Preprint medRxiv, 2022."

3.1 Introduction

PD is the second most common NDD [131], and there are several risks involved with the disease some of which are age, male sex, and environmental factors. Moreover, various genetic variants have been associated with the disease risk. The prime identifiers of the disease are a) cardinal motor symptoms such as tremor, rigidity, bradykinesia/akinesia, and postural instability b) clinical symptoms comprised by other motor symptoms such as postural abnormalities, gait disturbances, disturbances of speech, etc., and c) non-motor features such as dysphagia and diarrhea, autonomic, gastrointestinal, sleep, cognitive and neuropsychiatric disturbances. Pathologically, the loss of dopaminergic neurons in the substantia Nigra pars compacta (SNpc) and accumulation of misfolded α-synuclein are the hallmarks of PD. However, the etiology of the disease is still unclear in most of the patients and extensive research is going on to understand the disease. The disease is usually detected in its advanced stage and the latent phase can vary from 5 to 20 years [132]. This latent phase is termed as the prodromal phase of PD [133]. Some of the most recognized prodromal symptoms that might arise 10 years before the diagnosis of the disease are hyposmia, depression and anxiety, constipation and REM-sleep behavior disorder (RBD) [134], erectile dysfunction, somnolence [135], orthostatic hypotension, urinary dysfunction, possible subthreshold parkinsonism (MDS-UPDRS-III $>$ 3 excluding action tremor) / abnormal quantitative motor testing and abnormal dopaminergic Positron emission tomography/Single photon emission computed tomography positron emission tomography/single photon emission computed tomography (PET/SPECT) [136, 137, 135]. It is vital to identify the disease at its prodromal phase, in order to recognize the subjects at a higher risk of the disease much before the neurodegeneration occurs. Identifying the disease at the prodromal phase also provides an opportunity to slow or prevent the onset of motor symptoms when disease-modifying treatments become available [138, 139, 140]. Therefore, it becomes really important to identify the markers for the prodromal phase of the disease.

One of the studies focusing on identifying PD at prodromal phase is the TREND study. The overview of the study is provided in the next section. As we know that PD is a disease having multi-modal, multiscale, and heterogeneous data, it is important to have an understanding of the disease development and progression across all biologically relevant scales [122], that can further assist in early disease prediction. This can be achieved by modeling the health trajectory of the patient by developing disease risk [141, 142] and disease progression models [143, 144, 145, 146]. Accordingly, a compilation of a comprehensive overview of a specific disease requires comparison and analysis of multiple studies of the same disease. A large number of observational and clinical studies are being conducted in the context of early disease detection, drug development, and translational and pharmaceutical research. Nevertheless, there is a big problem of data silos in the medical field and it could be due to several reasons as discussed in Chapter 1. There are a number of legal and ethical constraints that restrict the organizations responsible for the study to share sensitive patient data. Moreover, there are a number of restrictions on terms of use, regulations imposed by the government, and an inability to get informed consent from marginalized populations. Organizations these days are keener on working across boundaries but that does not mean they have terminated their tendency to work with those most "like" them, continuing the rise in the "data silos" and thereafter the "silo effect" [147]. The problem of decentralized storage and strict data protection laws has also led to the generation of data silos [148]. The existence of data silos further slows down the overall scientific progress in translational research. Firstly, due to the highly sensitive nature of the health data, a lot of it is out of reach for researchers which could halt advancement in the research and slow down the development of new treatments. Therefore, it could curtail key findings that could lead to much-needed treatments and cures. Typically, siloed data is incompatible with the other data sets and is stored in a standalone system. This makes it difficult for users from other parts of the organization to access and use the data. They may also arise in any organization because separate business units might have different goals and want to operate independently. The data silos in healthcare also prevent pharmaceutical companies, physicians, and researchers from accessing and analyzing important data sets. Instead, it encourages each group to make conclusions based on the part of the information available to them. This further results in temporary fixes that are not sustainable in the long run and for patients, it results in delays in diagnosis, access to treatments, and proper care. Data silos create barriers for information sharing and collaboration amongst the research community.

In this work, we address this limitation of data sharing and data silos by building a generative model. The generative property of the model allows us to simulate subjects in the form of a SC that are sufficiently similar to the real ones and can be shared with the larger community. The idea behind our SC concept is to decipher the complex relationships between different biological scales and data modalities (clinical, genomic, etc.) within one modeling framework. This allows for generating synthetic patients that are highly similar to real ones with respect to relevant characteristics. Our work should at this point be discriminated from existing work on synthetic trial simulation, which mostly focuses on pharmacokinetic-pharmacodynamic (PKPD) modeling in clinical study design and typically involves mechanistic modeling of well-understood biological processes [149].

AI approaches, such as BNs [66] may offer possible solutions to these challenges. BNs can be used 1) to realistically simulate prospective cohorts, which could "at least partially" help to overcome restrictions posed by data privacy, and 2) to access such synthetic, comprehensive (population-based) cohort data 3) to model interdependencies of markers. Thereby, both the consideration of more generalizable evidence underlying PD prediction as well a more differentiated investigation and understanding of prodromal PD subtypes may be supported and possibly help to inform the design and recruitment for early intervention trials in prodromal PD.

The present study has two different aims: 1) to demonstrate the feasibility of generating a sufficiently realistic synthetic cohort, which shares statistical patterns of the original data and could allow researchers to gain a better understanding of the properties of the real data before formally applying for access to it and, 2) to model a BN with the interdependencies between longitudinal data of risk and prodromal markers of PD and incident PD status of a large prospective cohort (TREND).

3.2 Methodology

3.2.1 Overview of the data used

The data we use for this work comes from the TREND study. It started in 2009/2010 with the aim to collect early biomarkers for AD and PD. At present, the study consists of 1200 subjects, aged over 50 years with either one or several of the prodromal risk markers (hyposmia, RBD, depression) or none of these early markers that form the control group. These subjects are examined every two years and several examinations are performed including neurological, blood parameter collection, and medical history. Other tests include examining motor control, different tremor types, dexterity, slowing of distal movements, cognitive functions, etc. Questionnaires assessing sleep quality, mood, and activities of daily living are also recorded. For more information, visit `https://www.trend-studie.de`. We use the TREND study in our work in the context of PD.

The ten risk markers and ten prodromal markers of PD as assessed in the TREND study were defined as reported previously [150] and selected according to the recent International Parkinson and MDS Research Criteria for Prodromal PD [151].

Summary statistics of the variables are provided in Table 3.1. In this table, we present the descriptive statistics of longitudinal risk and prodromal marker data of PD-free individuals and incident PD cases. In total, data of 1178 PD-free was available and analyzed at their respective first visit (baseline). Of these individuals, 24 were diagnosed with PD (incident PD) at later visits (up to visit 6) of the TREND study and based on comprehensive neurological clinical diagnosis at the hospital. We provide the summary statistics of neurological, neuropsychiatric, environmental, lifestyle, and autonomic dysfunction biomarkers.

Variables	**At Baseline**	**Visit 2**		**Visit 3**		**Visit 4**	
	n-conv (1178)	n-conv (1172)	conv (6)	n-conv (1167)	conv (5)	n-conv (1162)	conv (5)
Age at baseline	63 (58, 68)	63 (58, 68)	74 (70, 74)	63 (58, 68)	75 (73, 76)	63 (58, 68)	70 (68, 71)
Sex (male/female)							
Male	599 (51%)	594 (51%)	5 (83%)	589 (50%)	5 (100%)	586 (50%)	3 (60%)
Female	579 (49%)	578 (49%)	1 (17%)	578 (50%)	0 (0%)	576 (50%)	2 (40%)
Sibling with PD (Yes/No)							
No	1010 (86%)	1007 (86%)	3 (50%)	1,003 (86%)	4 (80%)	999 (86%)	4 (80%)
Yes	168 (14%)	165 (14%)	3 (50%)	164 (14%)	1 (20%)	163 (14%)	1 (20%)
PRS							
Marker absent(Lowest quartile of PRS distribution)	247 (21%)	245 (21%)	2 (33%)	244 (21%)	1 (20%)	241 (21%)	3 (60%)
Borderline marker	500 (42%)	498 (42%)	2 (33%)	496 (43%)	2 (40%)	495 (43%)	1 (20%)
Marker present (highest PRS quartile)	252 (21%)	251 (21%)	1 (17%)	250 (21%)	1 (20%)	250 (22%)	0 (0%)
Missing	179 (15%)	178 (15%)	1 (17%)	177 (15%)	1 (20%)	176 (15%)	1 (20%)
GBA mutation carriers (Yes/No)							
No	1173 (99.6%)	1167(96.4%)	6 (100%)	1163 (99.7%)	4 (80%)	1158 (100%)	5 (100%)
Yes	5 (0.4%)	5 (0.4%)	0 (0%)	4(0.3%)	1 (20%)	4 (0.3%)	0 (0%)
SN							
SN-	839 (71%)	838 (72%)	1(17%)	835 (72%)	3(60%)	832 (72%)	3 (60%)
SN+	210 (18%)	205 (17%)	5 (83%)	203 (17%)	2(40%)	201 (17%)	2 (40%)
Missing	179 (15%)	129 (11%)	0 (0%)	129 (11%)	0(0%)	129 (11%)	0 (0%)
Pesticides (Yes/No)							
No	878 (75%)	855 (73%)	1(17%)	838 (72%)	1(20%)	811 (70%)	3 (60%)
Yes	19 (1.6%)	18 (1.5%)	0 (0%)	17 (1.5%)	0(0%)	17 (1.5%)	0 (0%)
Missing	281 (24%)	299 (26%)	5 (83%)	312 (27%)	4(80%)	343 (29%)	2 (40%)
Solvents (Yes/No)							
No	774 (66%)	753 (64%)	1 (17%)	738 (63%)	1(20%)	712 (61%)	2 (40%)
Yes	129 (11%)	125 (11%)	0 (0%)	122 (10%)	0(0%)	117 (10%)	1 (20%)
Missing	275 (23%)	294 (17%)	5 (83%)	307 (26%)	4(80%)	333 (29%)	2 (40%)
Smoking							
Marker absent	535 (45%)	478 (41%)	2 (33%)	437 (37%)	3(60%)	393 (34%)	0 (0%)
Borderline marker	533 (45%)	509 (43%)	4 (67%)	467 (40%)	2(40%)	418 (36%)	4 (80%)
Marker present	109 (9.3%)	83 (7.1%)	0 (0%)	72 (6.2%)	0(0%)	61 (5.2%)	1 (20%)
Missing	1 (<0.1%)	102 (8.7%)	0 (0%)	191 (16%)	0(0%)	290 (25%)	0 (0%)

Variables	At Baseline	Visit 2		Visit 3		Visit 4	
	n-conv (1178)	n-conv (1172)	conv (6)	n-conv (1167)	conv (5)	n-conv (1162)	conv (5)
Diabetes_II (Yes/No)							
No	1131 (96%)	1015 (87%)	5 (83%)	919 (79%)	5 (100%)	820 (71%)	4 (80%)
Yes	47 (4%)	55 (4.7%)	1 (17%)	58 (5%)	0 (0%)	53 (4.6%)	1 (20%)
Missing	1 (<0.1%)	102 (8.7%)	0 (0%)	190 (16%)	0 (0%)	289 (25%)	0 (0%)
Physical Inactivity Code (Yes/No)							
No	542 (46%)	343 (29%)	1 (17%)	752 (64%)	4 (80%)	695 (60%)	3 (60%)
Yes	142 (12%)	98(8.4%)	0(0%)	221 (19%)	1 (20%)	175 (15%)	2 (40%)
Missing	494 (42%)	731 (62%)	5 (83%)	194 (16%)	0 (0%)	292 (25%)	0 (0%)
RBD (Yes/No)							
No	1116 (95%)	1041 (89%)	6 (100%)	956 (82%)	4 (80%)	846 (73%)	4 (80%)
Yes	54 (4.6%)	28 (2.4%)	0 (0%)	21 (1.8%)	1 (20%)	27 (2.3%)	1 (20%)
Missing	8 (0.7%)	103 (8.8%)	0 (0%)	190 (16%)	0 (0%)	289 (25%)	0 (0%)
MDS-UPDRS-III							
No motor deficit	1006 (85%)	980 (84%)	0 (0%)	912 (78%)	1 (20%)	784 (67%)	1 (20%)
Borderline motor deficit	120 (10%)	66 (5.6%)	1 (17%)	39 (3.3%)	2 (40%)	63 (5.4%)	0 (0%)
Subthreshold parkinsonism	52 (4.4%)	24 (2%)	5 (83%)	26 (2.2%)	2 (40%)	26 (2.2%)	4 (80%)
Missing	0 (0%)	102 (8.7%)	0 (0%)	190 (16%)	0 (0%)	289 (25%)	0 (0%)
Hyposmia							
Marker absent	913 (78%)	860 (73%)	1 (17%)	746 (64%)	0 (0%)	664 (57%)	1 (20%)
Borderline marker	244 (21%)	186 (16%)	4 (67%)	192 (16%)	5 (100%)	159 (14%)	4 (80%)
Marker present	17 (1.4%)	4 (0.3%)	0 (0%)	24 (2.1%)	0 (0%)	42 (3.6%)	0 (0%)
Missing	4 (0.3%)	122 (10%)	1 (17%)	205 (18%)	0 (0%)	297 (26%)	0 (0%)
Constipation							
Marker absent	1015 (86%)	891 (76%)	4 (67%)	810 (69%)	2 (40%)	754 (65%)	5 (100%)
Borderline marker	139 (12%)	135 (12%)	2 (33%)	110 (9.4%)	2 (40%)	89 (7.7%)	0 (0%)
Marker present	15 (1.3%)	23 (2%)	0 (0%)	28 (2.4%)	1 (20%)	26 (2.2%)	0 (0%)
Missing	9 (0.8%)	123 (10%)	0 (0%)	219 (19%)	0 (0%)	293 (25%)	0 (0%)
Excessive Daytime Somnolence (Yes/No)							
No	0(0%)	32 (2.7%)	1 (17%)	383 (33%)	1 (20%)	833 (72%)	5 (100%)
Yes	0(0%)	1 (<0.1%)	0 (0%)	12 (1%)	0 (0%)	38 (3.3%)	0 (0%)
Missing	1178 (100%)	1139 (97%)	5 (83%)	772 (66%)	4 (80%)	291 (25%)	0 (0%)
Symptomatic Hypotension							
Marker absent	918 (78%)	793 (68%)	5 (83%)	746 (64%)	5 (100%)	736 (63%)	3 (60%)
Borderline marker	230 (20%)	218 (19%)	0 (0%)	198 (17%)	0 (0%)	93 (8%)	1 (20%)
Marker present	28 (2.4%)	52 (4.4%)	1 (17%)	27 (2.3%)	0 (0%)	41 (3.5%)	1 (20%)
Missing	2 (0.2%)	109 (9.3%)	0 (0%)	196 (17%)	0 (0%)	292 (25%)	0 (0%)
Urinary Dysfunction							
Marker absent	733 (62%)	691 (59%)	4 (67%)	634 (54%)	2 (40%)	625 (54%)	3 (60%)
Borderline marker	391 (33%)	286 (24%)	1 (17%)	269 (23%)	2 (40%)	192 (17%)	1 (20%)
Marker present	51 (4.3%)	82 (7%)	1 (17%)	64 (5.5%)	1 (20%)	52 (4.5%)	1 (20%)
Missing	3 (0.3%)	113 (9.6%)	0 (0%)	200 (17%)	0 (0%)	293 (25%)	0 (0%)
Depression (Yes/No)							
No	830 (70%)	735 (63%)	3 (50%)	664 (57%)	4 (80%)	594 (51%)	4 (80%)
Yes	348 (30%)	335 (29%)	3 (50%)	313 (27%)	1 (20%)	279 (24%)	1 (20%)
Missing	0 (0%)	102 (8.7%)	0 (0%)	190 (16%)	0 (0%)	289 (25%)	0 (0%)
Global Cognitive Deficits (Yes/No)							
No	971 (82%)	911 (78%)	6 (100%)	852 (73%)	4 (80%)	782 (67%)	4 (80%)
Yes	194 (16%)	142 (12%)	0 (0%)	111 (9.5%)	1 (20%)	80 (6.9%)	1 (20%)
Missing	13 (1.1%)	119 (10%)	0 (0%)	190 (16%)	0 (0%)	300 (26%)	0 (0%)

Table 3.1: Summary statistics of age, risk and prodromal markers of PD-free individuals and incident PD cases at different time points, absolute and relative (%) frequencies of marker presence or median (IQR in brackets) are given unless specified otherwise.

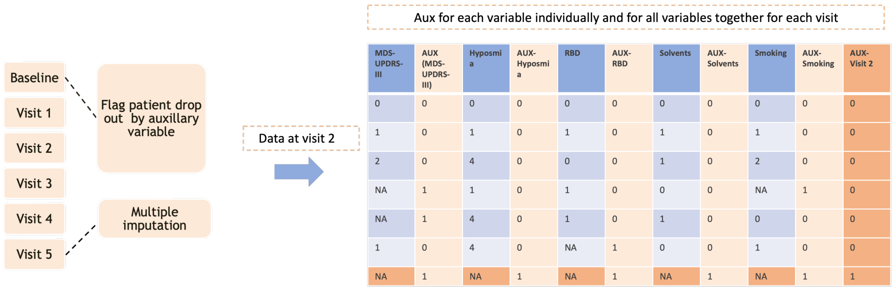
3.2.2 BN based approach

We propose a BN-based approach to simulate a realistic SC, to understand the links between different features that were measured in the study, and describe the longitudinal patient trajectories in a multi-modal, multi-scale manner. As explained in chapter 2, BNs are probabilistic graphical models, where nodes represent variables and edges represent probabilistic stochastic dependencies between them [87] characterized by a CPT for each variable. Please refer to Chapter 2 for a detailed description of BN.

In our case, there exists a subset $\tilde{X} \subset X$ (X is a set of random variables in the BN), such that measurements are time-dependent, i.e. $\tilde{x} = (\tilde{x}(1)), ..., \tilde{x}(T))$ with T being the number of visits. Dynamic BNs [152] usually deal with this situation by implicitly unfolding the BN structure over time, i.e. introducing for each visit t a separate copy $\tilde{X}(t)$ of $\tilde{X}$ while requiring that edges always point from time slice t to time slice $t + 1$ (corresponding to a first-order Markov process). This implicit unfolding assumes a stationary Markov process, i.e. parameters "Θ" do not change with time. In our setting this assumption is most likely wrong, because patients change in their disease outcome during the course of a study, i.e. $p(\tilde{X}(t)|\tilde{X}(t{-}1)) \neq (\tilde{X}(t+1)|\tilde{X}(t))$. Hence, we here use an unfolding strategy, in which we explicitly use different copies $\tilde{X}(t)$ for each time point.

One of the key challenges with longitudinal data as described in chapter 2 is dealing with the missing data and especially the data that is MNAR. We mitigated this limitation by defining auxiliary variables for each variable and visit. These auxiliary variables are fixed parents of all nodes, which contain missing values in a "systematic" way i.e. following an MNAR pattern, and we do not learn any more edges for auxiliary variables. Furthermore, for the features that are assessed at different visits, we enforced auxiliary variables to point from one visit to the next visit. An advantage of using auxiliary variables is that they make parameter estimates conditionally dependent on the missingness information, therefore accounting for the potential biases of the multiple imputations (due to hidden confounding factors). The missing data approach is illustrated in Figure 3.1.

To learn a BN corresponding to the MDS research criteria for prodromal PD, prospective data of established risk and prodromal markers of PD that have been collected in the TREND study [153] entered the analyses. In this work, we compiled a BN with 10 risk markers and 10 prodromal markers as well as age. These markers were assigned to different domains including: autonomic dysfunction


Baseline
Visit 1
Visit 2
Visit 3
Visit 4
Visit 5
Flag patient drop out by auxillary variable
Multiple imputation
Data at visit 2
Aux for each variable individually and for all variables together for each visit

MDS-UPDRS-III	AUX (MDS-UPDRS-III)	Hyposmia	AUX-Hyposmia	RBD	AUX-RBD	Solvents	AUX-Solvents	Smoking	AUX-Smoking	AUX-Visit 2
0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	1	0	1	0	1	0	0
2	0	4	0	0	0	1	0	2	0	0
NA	1	1	0	1	0	0	0	NA	1	0
NA	1	4	0	1	0	1	0	0	0	0
1	0	4	0	NA	1	0	0	1	0	0
NA	1	NA	1	NA	1	NA	1	NA	1	1

Figure 3.1: Aux/indicator variables defined for each variable and visit, "AUX Visit 2" is a parent of other AUX variables at visit 2.

(constipation, symptomatic orthostatic hypotension, erectile and urinary dysfunction; based on a self-report questionnaire), lifestyle features and related diseases (physical inactivity, non-smoking (self-report questionnaire), diabetes type II (self-reported medical diagnosis)), environmental features (occupational pesticide and solvent exposure; self-report questionnaire), neuropsychiatric features (depression (life-time diagnosis or acute depression based on international classification of diseases 10th revision (ICD-10) criteria), global cognitive deficit (based on comprehensive neuropsychological testing)), neurological features (incident PD diagnosis (PD conversion based on neurological diagnosis), sub-threshold Parkinsonism (based on MDS-UPDRS-III), (pRBD; based on a self-report questionnaire), hyposmia (Sniffin Sticks), SN hyperechogenicity based on transcranial ultrasnd), genetic factors (first-degree family history of PD (self-reported), polygenic risk scores (PRS) of PD, pathogenic glucocerebrosidase (GBA) mutations based on genetic testing) and demographic factors (age, sex). Since erectile dysfunction was only assessed in males, this prodromal marker was not included in the final BN to avoid biases in the model.

Notably, the TREND data of risk and prodromal markers of PD has been discretized such that all variables (except for age) indicate the presence or absence (or borderline status) of a marker in an individual TREND participant, as published previously for the TREND cohort and suggested by the MDS research criteria for prodromal PD [151, 136]. For each variable, a CPT was determined as the overall BN was learned.

Constraints on the network structure

In most cases, the edges in BN are not known and they need to be inferred from the data. We need to answer an important question how far the learned BN structure reflects existing causal relationships in the data? If the BN is able to learn the existing statistical distributions in the data and faithfully represent the underlying statistical distribution, then the true causal network is known to be a part of a class of equivalent graph structures, called CPDAG [67, 94]. Considering this assumption, the CPDAG follows the same skeleton as the true causal graph, and might also have some undirected edges. Therefore, it is important to restrict the CPDAG equivalence class as much as possible by prior knowledge to allow the correct orientation of as many edges as possible. We imposed the following constraints for the BN structure:

- No node can affect environmental factors except the factors themselves.
- Genetic factors like GBA mutation and PRS cannot be dependent on any other feature except themselves, PD family history, and sex.
- PD family history cannot be dependent on any other features except genetic factors, age, and sex.
- No other feature can be affected by the conversion feature (the conversion node is the node that represents if the subject has converted to PD or not).
- Age and sex cannot be dependent on any other node.
- Auxiliary variable that was created based on the missingness of a certain feature can only have an influence on their corresponding feature and the auxiliary variable for the same feature at the next time point.

BN structure and parameter learning

We learned the network on four different algorithms from R-package bnlearn [113] hc, tabu search, RSMAX2, and MMHC (described in detail in Chapter 2). Cross-validation was used to assess the generalization ability of a BN model and compare different structure learning algorithms. Selection between different BN structure

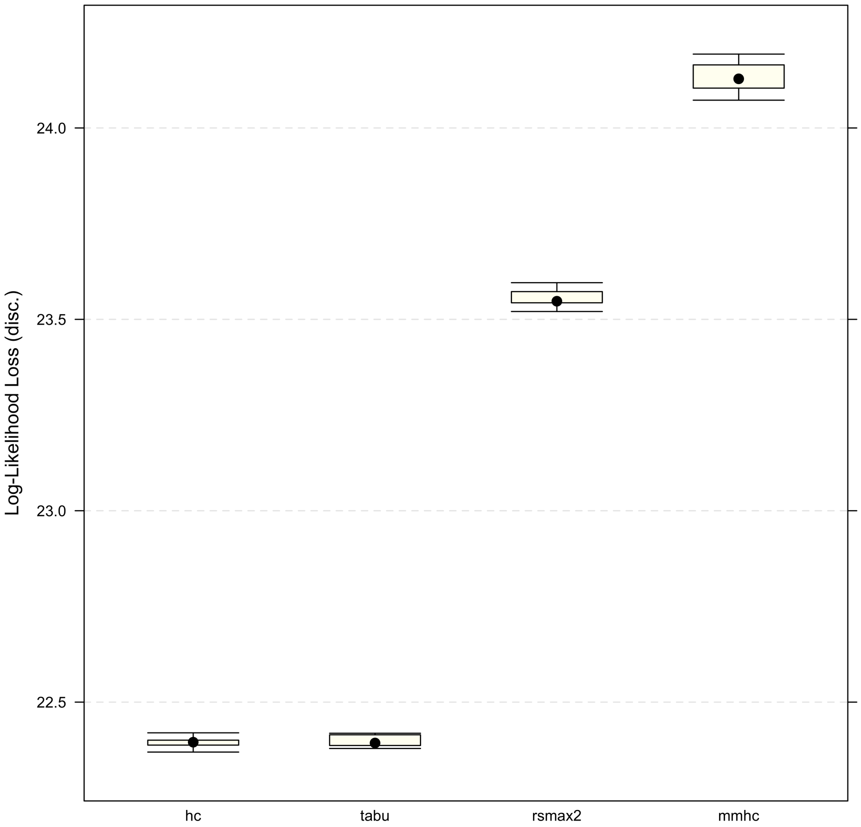


Figure 3.2: Comparison of different BN structure learning algorithms via 10-fold cross-validation for the data. The y-axis depicts the negative log-likelihood of the test data.

algorithms was done via k-fold cross-validation, meaning overall data were randomly split into k (k = 10) folds, and BN structure together with its parameters was learned using k-1 folds. If the overall population is correctly modeled by fitting the BN (and not just the training data), the data in the left out fold with high probability should fall into the same statistical distribution that is described by the BN. Negated expected log-likelihood of the test data is used to quantify this. The algorithm with the least negative log-likelihood is considered to be the best performing and in this case, hc satisfied the criteria. The comparison between different algorithms is illustrated in Figure 3.2.

We trained a BN based on the data of all 1178 subjects using a non-parametric bootstrap [154] by randomly selecting n = 1178 for 1,000 times, with replacement, and for each of these 1,000 bootstrap samples, we learned a complete BN structure. The relative frequency of observing a particular edge among 1,000 boot-straps was determined (see BN edges in Figure 3.3) and served as an indicator of the level of statistical confidence, i.e., a higher value means stronger support by the data for the existence of the respective connection. A value of 1.0 indicates two specific nodes were interdependent in all of the 1,000 learned BNs, and a value of 0.5 indicates in 50% of the BNs an interdependency was observed.

Once a BN topology is defined, parameters were then inferred using a Dirichlet prior to account for parent-child node configurations that were not observed [109].

3.2.3 Simulation of a synthetic TREND study cohort

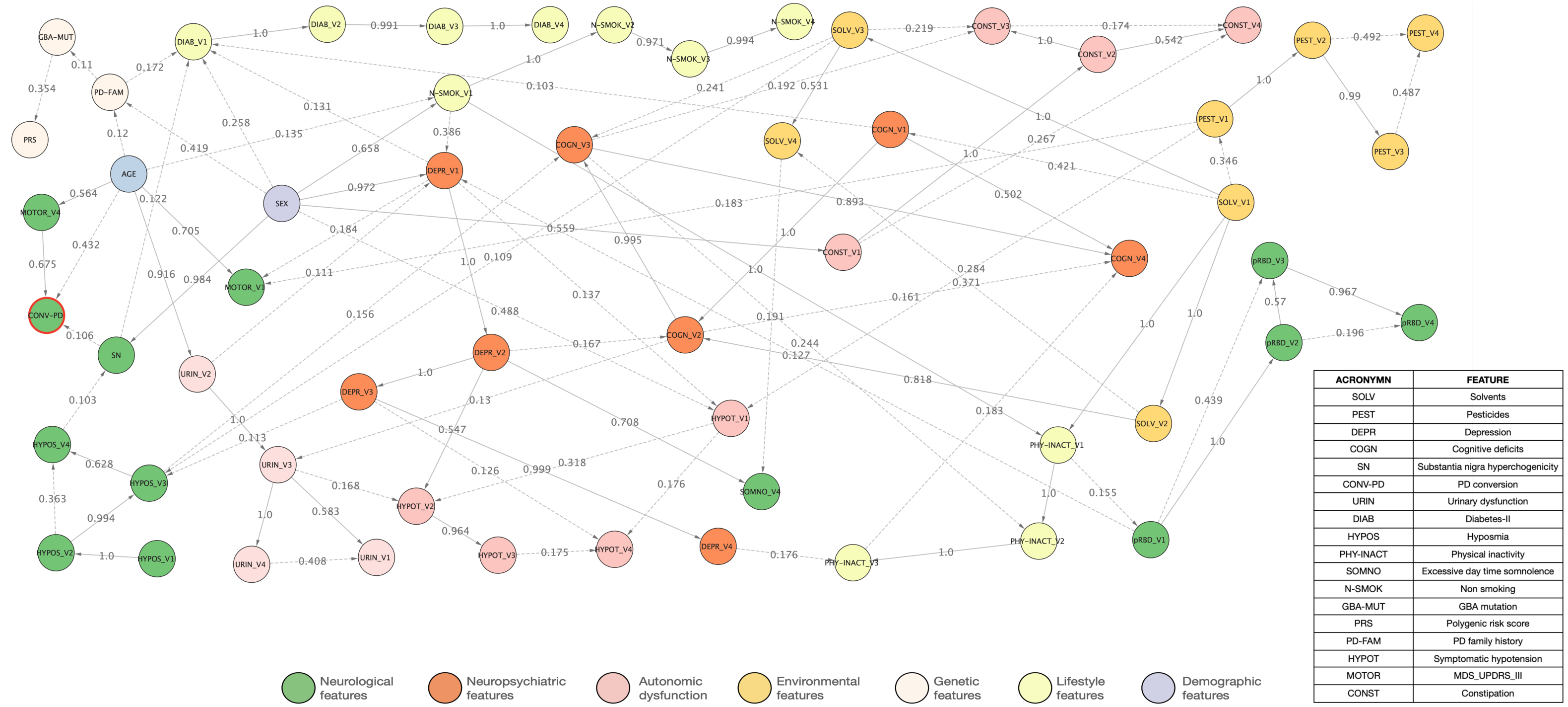
As discussed earlier, we know that BNs belong to the class of generative machine learning models. It means they learn the multivariate statistical distribution underlying the observed data. Therefore, random samples drawn from the model correspond to synthetic subjects. This was done by first drawing random values from a node's distributions and subsequently from the distributions of the children of that node while conditioning on the values of the parents. We also calculated the KLD [155] between real and synthetic data distributions. KLD is used to quantify the distributions of a random variable and measure the similarity between the two probability distributions of the same variable. Hence, it measures the divergence of one distribution from another. If the distributions are an exact match, the KL divergence is 0, otherwise, it can lie between 0 and ∞.

3.3 Results

3.3.1 The BN of risk and prodromal markers of PD in the TREND study

Our analysis using BN of the longitudinal TREND data resulted in a quantitative network between different variables, which is depicted in Figure 3.3.

A wide range in the level of confidence regarding the interconnectedness, i.e., statistical interdependence, was observed between several nodes and domain clus-



ACRONYMN	FEATURE
SOLV	Solvents
PEST	Pesticides
DEPR	Depression
COGN	Cognitive deficits
SN	Substantia nigra hyperchogenicity
CONV-PD	PD conversion
URIN	Urinary dysfunction
DIAB	Diabetes-II
HYPOS	Hyposmia
PHY-INACT	Physical inactivity
SOMNO	Excessive day time somnolence
N-SMOK	Non smoking
GBA-MUT	GBA mutation
PRS	Polygenic risk score
PD-FAM	PD family history
HYPOT	Symptomatic hypotension
MOTOR	MDS_UPDRS_III
CONST	Constipation

Figure 3.3: Interdependencies between different risk markers and prodromal markers of PD. The depicted BN represents interdependencies between variables learned from prospective TREND data. Domains of marker nodes are indicated by colored circles. The node of the "Conversion to PD" is indicated by a red circled outline. Numbers on edges indicate the level of statistical confidence (bootstrap probability), and dashed edge lines indicate confidence <0.5 while solid lines indicate a confidence ≥ 0.5. A higher value indicates a higher confidence in the existence of a connection. Nodes isolated from the rest of the network are not shown. V represents each visit.

ters of nodes. High probabilistic confidence (>0.5) of edges between different markers in the BN was found for edges between age to sub-threshold Parkinsonism (MDS-UPDRS-III) and urinary dysfunction, sex to SN hyperechogenicity, depression, non-smoking to constipation; depression to symptomatic hypotension and excessive daytime somnolence; solvent exposure to cognitive deficits and to physical inactivity; and non-smoking to physical inactivity. Pairwise co-occurrences of different markers showing edges with probabilistic certainties of >0.2 in the BN were shown and statistically tested for significance in Table 3.2. P-values have been calculated based on a Chi-square test and corrected for multiple testing using Holm's method. These findings remain significant in logistic regressions (additionally accounting for age and sex). All of these edges also showed statistically significant co-occurrences between markers, except for sex and PD family history, sex, and diabetes type-II (visit 1), occupational solvent exposure (visit 3), and constipation (visit 3), as well as GBA mutation carriers and PRS. These associations were no longer significant after accounting for multiple testing using Holm's method.

The BN revealed both expected as well as novel connections between risk and prodromal markers and the phenoconversion to PD. Plausibly, the nodes with edges directed to the conversion to PD comprised (prior) subthreshold parkinsonism indicated by MDS-UPDRS-III scores, age, and (with lower statistical confidence), SN hyperechogenicity. Further expected marker interdependencies were observed for edges pointing from depression and solvent exposure to global cognitive deficits, which itself was linked to physical inactivity while non-smoking was linked to physical inactivity. Edges pointing from depression to excessive daytime somnolence, pointing from solvent exposure and depression to hyposmia, or pointing from hyposmia to global cognitive deficits and to SN hyperechogenicity demonstrated further expected interdependencies. Unexpected interdependencies were observed from depression to non-smoking; pesticide exposure to symptomatic hypotension; physical inactivity to urinary dysfunction; and edges with directionality from SN hyperechogenicity, global cognitive deficits, sex, and PD family history to diabetes. Interestingly, constipation was dependent on sex, global cognition, and occupational solvent exposure. Surprisingly, little interdependencies were observed for pRBD, which was only linked to depression and received an edge from physical inactivity. Nodes with genetic features were not dependent on other markers except for sex being linked to PD family history, which itself was linked to diabetes.

Nodes of the same marker assessed at different time points were largely highly interdependent, except for subthreshold parkinsonism (MDS-UPDRS-III) for which visit 2 and visit 3, were not linked to other nodes of the BN. MDS-UPDRS-III at visit 1 showed no edge with the corresponding nodes of other visits, but instead only received edges from depression and pesticide exposure at visit 1.

Risk/prodromal marker	**Risk/prodromal marker**	**Participants (n)**			**p-value**
		No	Borderline	Yes	
Male sex	Depression at visit 1	471		128	<0.0001*
Female sex		359		220	
Male sex	Non-smoker at visit 1	228	320	51	<0.0001*
Female sex		308	213	58	
Male sex	SN hyperechogenicity	453		146	<0.0001*
Female sex		515		64	
Male sex	Constipation at visit 1	549	45	5	<0.0001*
Female sex		472	96	11	
Male sex	PD family history	529		70	0.013
Female sex		481		98	
Male sex	Symptomatic hypotension at visit 1	508	83	8	<0.0001*
Female sex		412	147	20	
Male sex	Diabetes type II at visit 1	567		32	0.024
Female sex		564		15	
Exposure to solvents at visit 1	Cognitive deficits at visit 1	302		86	<0.0001*
No exposure to solvents at visit 1		678		112	
Exposure to solvents at visit 2	Cognitive deficits at visit 2	246		176	<0.0001*
No exposure to solvents at visit 2		682		74	
Exposure to solvents at visit 3	Cognitive deficits at visit 3	201		237	<0.0001*
No exposure to solvents at visit 3		668		72	
Exposure to solvents at visit 3	Constipation at visit 3	394	32	12	0.029
No exposure to solvents at visit 3		626	88	26	
Exposure to pesticides at visit 1	Symptomatic hypotension at visit 1	13	20	1	<0.0001*
No exposure to pesticides at visit 1		907	210	27	
Presence of depression at visit 2	Day time somnolence at visit 4	414		26	<0.0001*
Absence of depression at visit 2		725		13	
Presence of depression at visit 2	Symptomatic hypotension at visit 2	213	200	27	<0.0001*
Absence of depression at visit 2		588	122	28	
Non-smoker at visit at visit 1	Physically active at visit 1	107		429	<0.0001*
Borderline smoker at visit 1		82		451	
Smokers at visit 1		63		46	
Non-smoker at visit at visit 1	Depression at visit 1	389		147	<0.0001*
Borderline smoker at visit 1		382		151	
Smokers at visit 1		59		50	
Presence of Global cognitive deficits at visit 2	Physically active at visit 3	224		85	<0.0001*
Absence of Global cognitive deficits at visit 2		193		676	
GBA mutation carries	Polygenic risk score	21	24	7	0.002
GBA mutation non-carriers		226	655	245	
Age (Older than 65 years)	Conversion to PD	500		23	<0.0001*
Age (Younger than/ equal to 65 years)		654		1	

Risk/prodromal marker	**Risk/prodromal marker**	**Participants (n)**			**p-value**
		No	Borderline	Yes	
Age (Older than 65 years)	Subthreshold parkinsonism at visit 4	462	31	30	<0.0001*
Age (Younger than/ equal to 65 years)		616	32	7	
Age (Older than 65 years)	Subthreshold parkinsonism at visit 1	418	71	34	<0.0001*
Age (Younger than/ equal to 65 years)		588	49	18	
Age (Older than 65 years)	Urinary Dysfunction at visit 1	269	220	34	<0.0001*
Age (Younger than/ equal to 65 years)		465	173	17	
Age (Older than 65 years)	Non-smoking at visit 1	260	236	27	<0.0001*
Age (Younger than/ equal to 65 years)		276	297	82	

Table 3.2: Statistical testing of the co-occurrence of markers in the TREND data as suggested by edges in the TREND BN of real data. P-values have been calculated based on a Chi-square test and corrected for multiple testing using Holm's method. Significant findings (after correction for multiple testing) are indicated by an asterisk. Findings remain significant in logistic regressions additionally accounting for age and sex.

3.3.2 Comparison of real and synthetic data

We sampled synthetic subjects as real participants and then compared the distributions of each variable visually at each visit (Figure 3.4-3.7).

As seen in Figures 3.4-3.7, the KLD value is close to 0 which means we have a good match between the distribution of real and synthetic data, the lower the value the better the match. We also measured spearman rank correlation for real and synthetic data, and we observed that the correlation between the variables remains preserved. The plot is illustrated in Figure 3.8. We used the relative error (*re*) of the "Frobenius norm" (norm) of the matrix to compare the correlation matrices for real (*RD*) and synthetic data (*SD*), which is 0.36. It was calculated using the following formula (in Equation 3.1), where corr.matrix refers to the correlation matrix:

$$re = \|(corr.matrix(RD) - corr.matrix(SD))\| \;/\; \|(corr.matrix(RD))\| \qquad (3.1)$$

The generative property of the BN allowed the simulation of synthetic versions of the prospective data of the TREND study and to extract individual synthetic participant profiles including age and the risk and prodromal markers of PD. Table 3.3 shows five arbitrary examples of synthetic subjects (from the synthetic cohort with the same sample size) and three real subjects together with their individual data (at visit 4) on age, sex, MDS-UPDRS-III, pRBD, depression, global cognitive deficits, and PD conversion status.

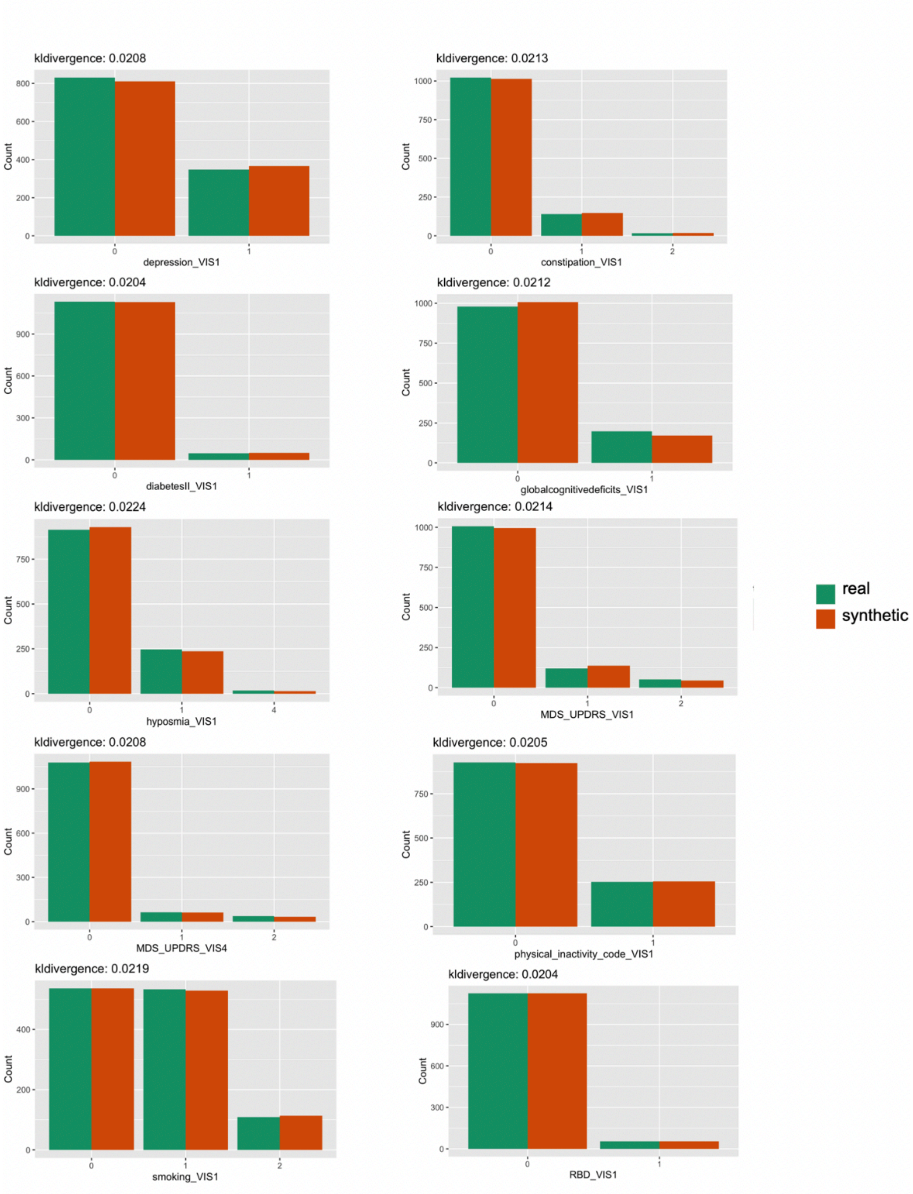


Figure 3.4: Examples of real and simulated subjects at visit 1 generated via the BN model trained on TREND study data. The Figure compares the distributions of features at visit 1 for real subjects (green) and synthetic/simulated subjects (red). KLD between the real and synthetic subjects is mentioned at the top of each plot.

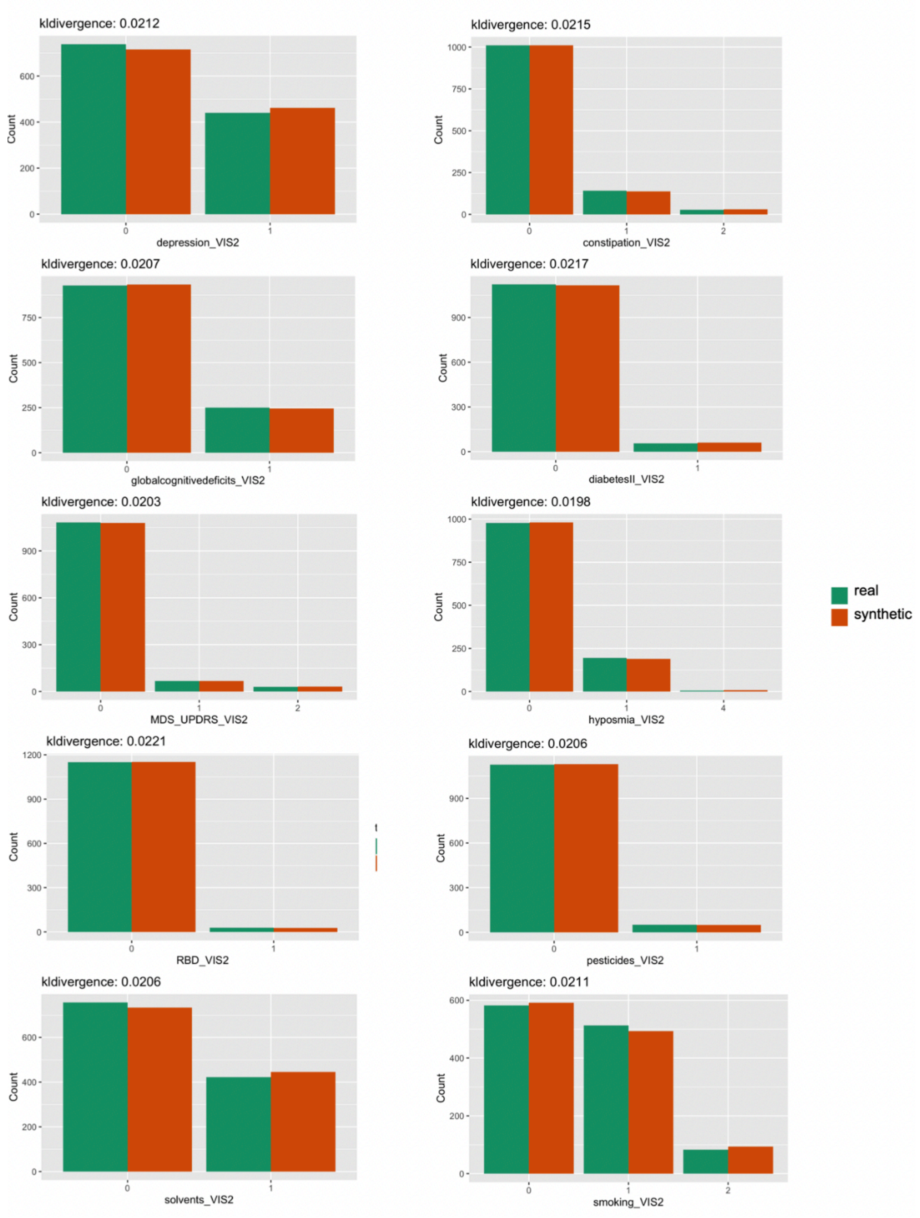


Figure 3.5: Examples of real and simulated subjects at visit 2 generated via the BN model trained on TREND data. The Figure compares the distributions of features at visit 2 for real subjects (green) and synthetic/simulated subjects (red). KLD between the real and synthetic subjects is mentioned at the top of each plot.

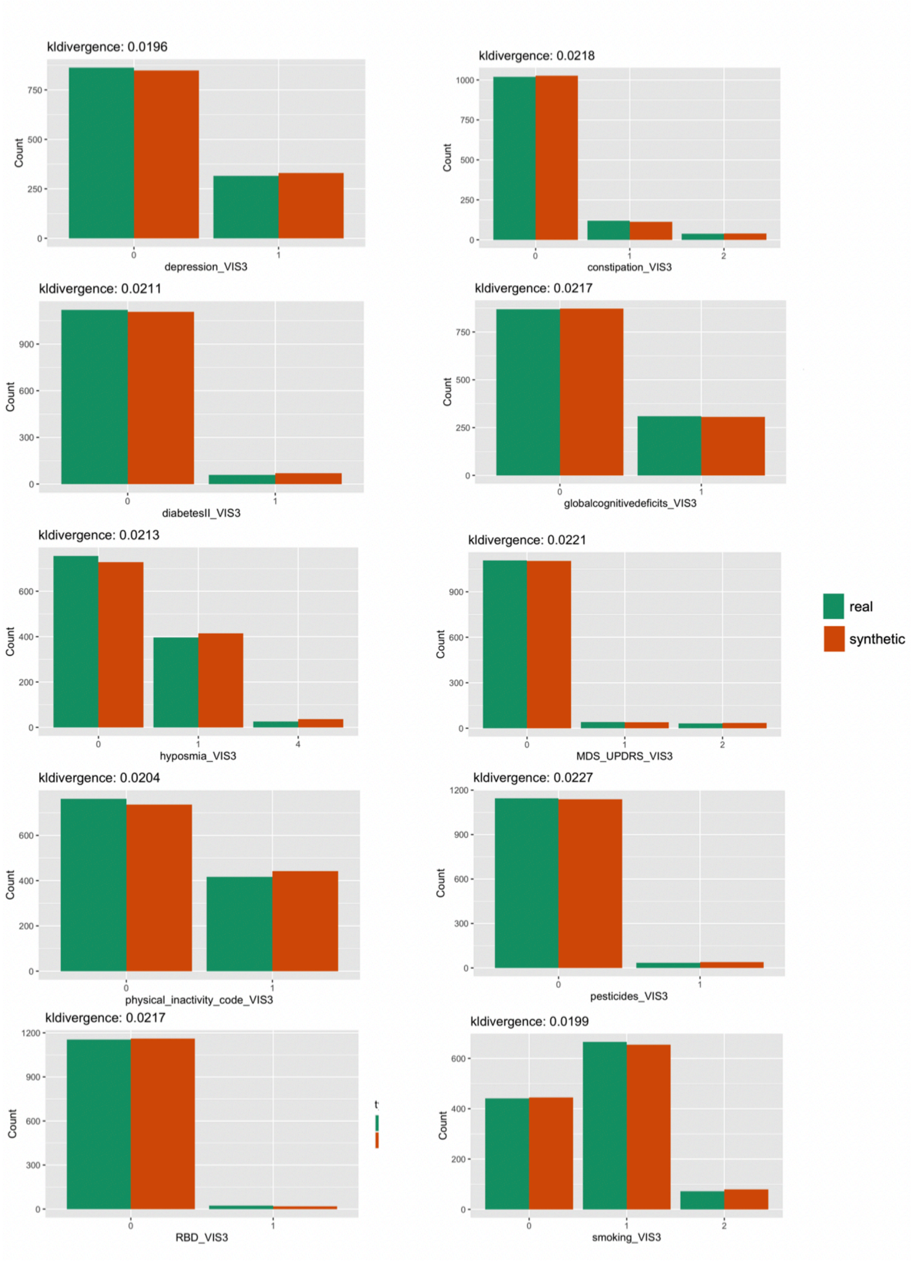


Figure 3.6: Examples of real and simulated subjects at visit 3 generated via the BN model trained on TREND data. The Figure compares the distributions of features at visit 3 for real subjects (green) and synthetic/simulated subjects (red). KLD between the real and synthetic subjects is mentioned at the top of each plot.

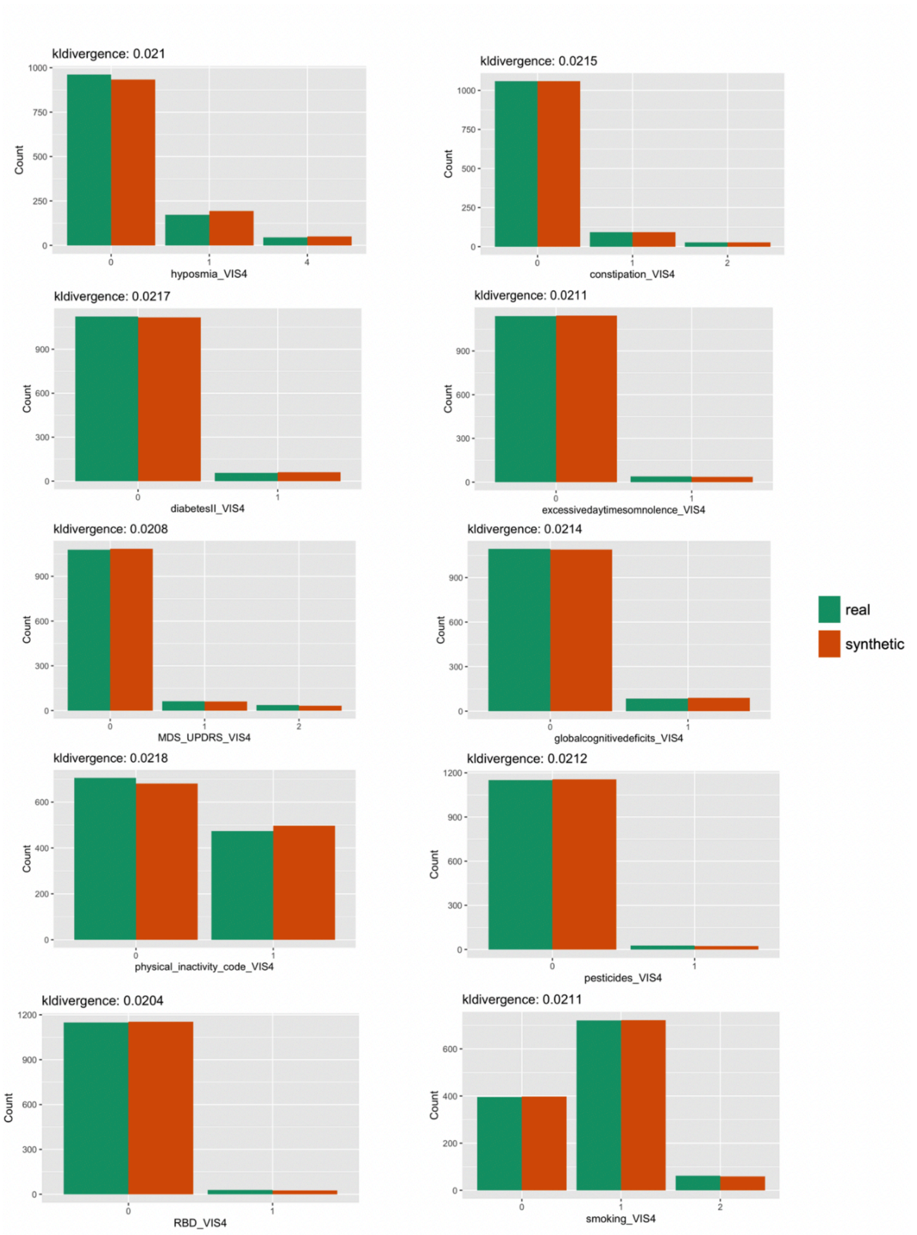


Figure 3.7: Examples of real and simulated subjects at visit 4 generated via the BN model trained on TREND data. The Figure compares the distributions of features at visit 4 for real subjects (green) and synthetic/simulated subjects (red). KLD between the real and synthetic subjects is mentioned on the top of each plot.

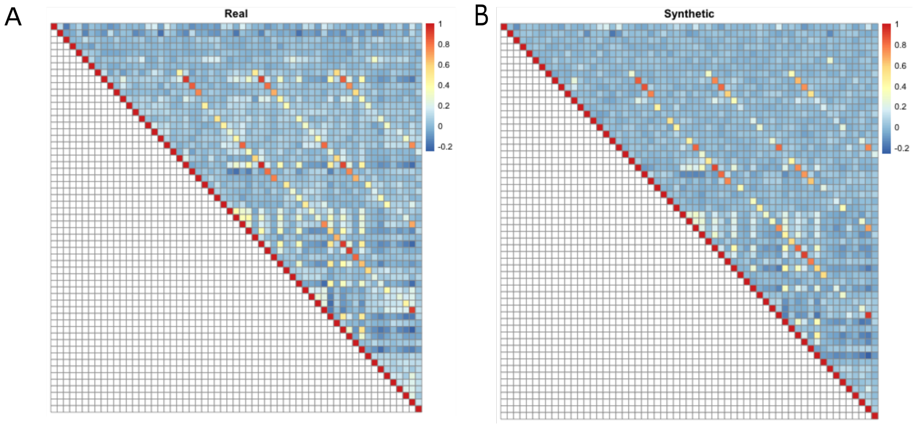


Figure 3.8: Spearman rank correlation for (A) real and (B) synthetic subjects. Legend indicates the strength of correlation, the darker the color (red or blue) the higher the correlation (positive (red) or negative(blue)), relative error for correlation matrices is 0.36.

Subjects	Age	Sex	MDS-UPDRS-III at visit 4	pRBD at visit 4	Depression at visit 4	Global cognitive deficits at visit 4	Conversion to PD
Synthetic subject # 1	68	Male	No motor deficit	No	No	No	No
Synthetic subject #2	67	Female	No motor deficit	No	No	No	No
Synthetic subject #3	**68**	**Male**	**Subthreshold parkinsonism**	**No**	**No**	**Yes**	**Yes**
Synthetic subject #4	73	Female	No motor deficit	No	No	No	No
Synthetic subject #5	**69**	**Male**	**Borderline motor deficit**	**No**	**No**	**No**	**Yes**
Real subject #1	63	Female	No motor deficit	No	No	No	No
Real subject #2	**68**	**Male**	**Subthreshold parkinsonism**	**No**	**No**	**No**	**Yes**
Real subject #3	**70**	**Male**	**Borderline motor deficit**	**No**	**No**	**No**	**Yes**

Table 3.3: Examples of synthetic and real subjects and their demographics, selected prodromal markers, subthreshold parkinsonism (MDS-UPDRS-III) and PD conversion status at visit 4. The rows in bold represent the similarity between the real and synthetic subjects data for incident PD cases.

3.3.3 Evaluating the utility of synthetic TREND subjects

To evaluate the utility of synthetic subjects generated by the BN model we performed different tests:

- We generated the same number of synthetic individuals as real individuals for the data and then tested whether a conventional RF classifier was able to separate between synthetic and real subjects within 10 times repeated 10-fold cross-validation scheme. That means we sequentially left out 1/10 of the subjects and trained an RF on the remaining subjects to learn the discrimination between real and synthetic subjects. We used the left-out portion of the data to assess the prediction performance of the RF. We used the pAUC at a pre-specified true positive rate of 99% for real subjects as a measure of the

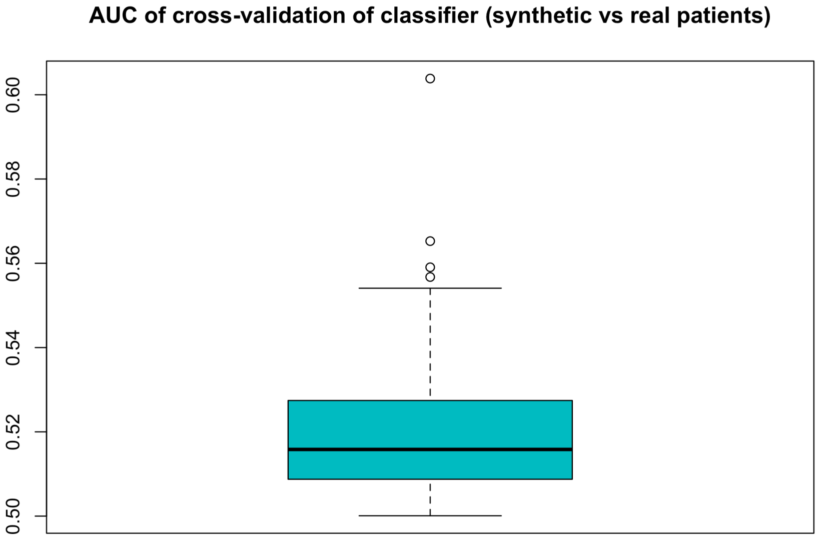


Figure 3.9: Performance of a RF classifier to correctly identify a given number of real TREND participants among synthetic subjects. The performance was measured via the pAUC at a pre-specified detection rate of $\geq$ 99% for real participants. The pAUC was assessed on test sets within 10 repeats of a 10-fold cross-validation procedure. Accordingly, boxplots show the distribution of the tenfold cross-validated pAUC that was obtained from 10 repeats of the cross-validation procedure.

prediction performance. The pAUC has been normalized between 0% and 100%. The area under the receiver operator characteristic (ROC) curve at which the detection rate for real subjects was between 99% and 100% served as an indicator of the validity of the synthetic TREND participants. This was done to account for the fact that misclassification of a synthetic TREND participant as real would be far less relevant as the other way around. In our case, a pAUC slightly above the chance level was achieved (Figure 3.9, indicating that synthetic subjects cannot reliably be discriminated from real ones by machine learning). The multiple correspondence analysis (MCA) [156] plot shown in Figure 3.10 indicates the similarity of synthetic subjects in relation to real ones.

- As a third test, we trained and evaluated the prediction performance of different machine learning models on real as well as synthetic data. More

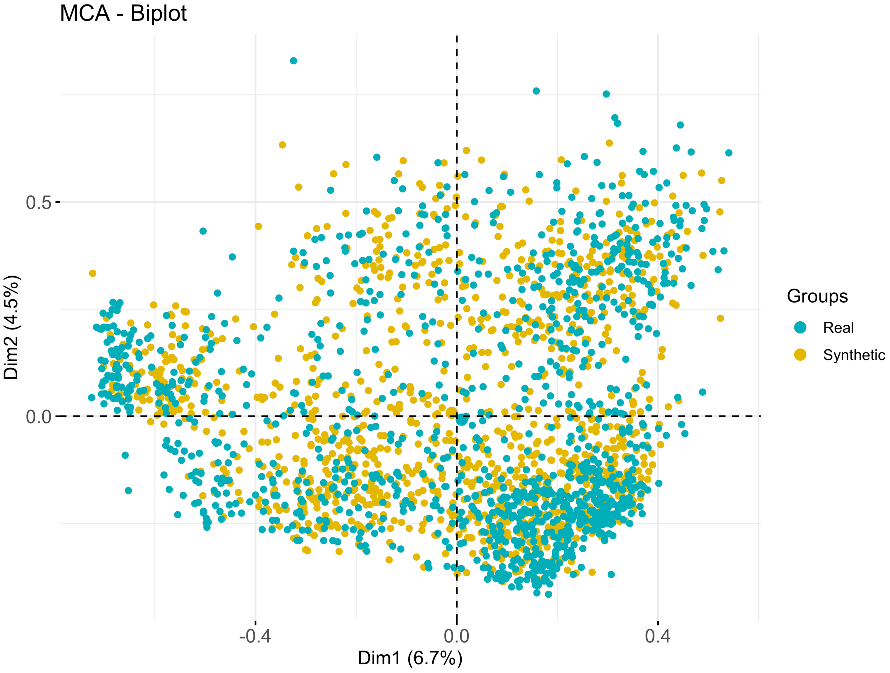


Figure 3.10: Multiple correspondence analysis (MCA) analysis plot of prospective data of real (in blue) and synthetic (in yellow) TREND participants.

specifically we here focused on the prodromal markers; pRBD, hyposmia, and depression. We trained a machine learning model (a RF classifier) to test the prediction ability of several variables to predict these prodromal markers at multiple visits. Outcomes at a subsequent visit were predicted by training the classifier on variables from the previous visit. For example, to predict the prodromal marker at visit 2, the classifier was trained on all the markers (measured longitudinally in the study) at visit 1. We either trained and tested the classifier on real subjects or trained the classifier on simulated/synthetic subjects generated by the BN and subsequently tested the classifier on real subjects. We evaluated the prediction performance of machine learning models using 10-fold cross-validation repeated 10 times. The overall dataset was randomly split into 10 folds, of which sequentially one of the folds was left out for testing the model, while the rest of the data was used for training. The prediction ability was measured via the AUC [157]. Despite synthetic data generally showing high similarity to real data, our results indicate a loss of ~10% AUC when training on synthetic subjects compared to training on real subjects (Figure 3.11). This could be due to slight differences between real and synthetic data regarding the distribution of individual variables (e.g. hyposmia, physical inactivity in Figure 3.6) as well as correlation structure (Figure 3.8). Notably, RFs are a comparably complex machine learning method, which allows for modeling highly nonlinear structures.

Altogether these results highlight that synthetic data share many patterns of real patient data, but they are not identical and hence do not necessarily allow for coming to identical statistical conclusions.

3.4 Conclusion

This work demonstrates a BN-based method by which we can unravel the complexities underlying a disease, by establishing connections between different clinical features of the disease. The model brings together heterogeneous multi-scale and multi-modal data together accounting for missing data patterns. We devised a method to identify MNAR patterns and deal with them using auxiliary variables. The present study shows the feasibility of generating a BN on prospective data on established risk and prodromal markers of PD in the TREND cohort of elderly PD-free individuals and incident PD cases. (1) The BN model showed several expected as well as non-obvious interdependencies of these markers, which may be explained by co-occurrences of markers in the TREND cohort and/or by methodological aspects of marker assessment. (2) The BN allowed the creation of a synthetic representation of the TREND cohort regarding the risk and prodromal marker interdependencies and to derive marker profiles of individual synthetic participants. The multitude of marker interdependencies as revealed through the AI-supported BN modeling approach may have important methodological implications for evidence-based PD prediction approaches as well as for the understanding of the interplay of different markers in the prodrome of PD and in potential prodromal PD subtypes. Based on our findings from a BN model of established PD markers in the prospective TREND cohort, we could show that for many markers independence might not be met.

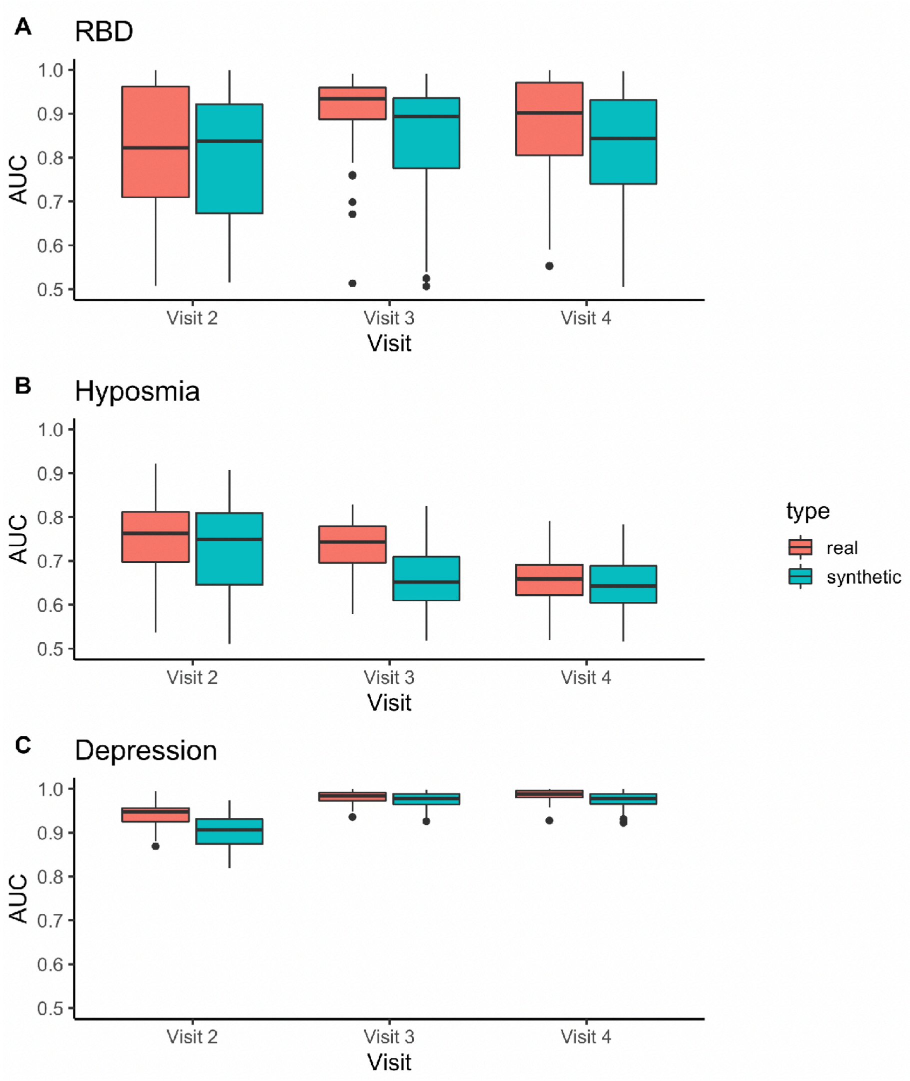


Figure 3.11: Performance of an RF classifier with subjects trained on real and tested on real data (red) and trained on synthetic and tested on synthetic data (blue) for (A) pRBD (B) hyposmia and (C) depression as clinical endpoints of the prospective data. The classification accuracy is indicated as AUC-ROC curves. The boxplots show the distribution of the AUC derived from the real/synthetic training data within 10 times repeated 10-fold cross-validation. In the case of training on synthetic data, results have been averaged over 50 repeated samplings of the same number of synthetic as real subjects.

Our approach also tries to solve the problem of data sharing and data silos by a realistic simulation of virtual clinical subject trajectories across multiple biological scales and data modalities outside the area of mechanistically well-understood biological processes. This was achieved by modeling the data with the help of BN which serves as a generative model, from which patient trajectories can be drawn. Using our method, we showed that synthetic and real patient trajectories are highly similar, but not identical. Hence, our proposed approach opens the possibility to build synthetic patients and at the same time realistic versions of clinical studies across multiple disease areas in the future. These synthetic studies could then be shared with the larger research community, even, if the raw data cannot be because of legal or ethical constraints. Hence, our method could help to unlock one of the key bottlenecks in biomedical research in data-scarce disease areas. Our proposed approach is not without limitations: BN structure and parameter learning require sufficiently large datasets that are representative of the disease population. The BN model makes the re-identification of real patients from the training data relatively unlikely. However, in its current implementation, our approach does not provide strict theoretical guarantees for this situation. But, we like to point out that privacy-preserving training of neural network models is possible in principle [158] which we have demonstrated in chapter 5 in the advancement of this work. Overall, this work demonstrates the potential of modern AI approaches to advance our understanding of prodromal PD and facilitate data sharing.

Greatness starts with being grateful.

Vex King

4

Modification of generative AI-Based approach with Sparse Autoencoders

This chapter is based on our work published in "Meemansa Sood, Akrishta Sahay, Reagon Karki, Mohammad Asif Emon, Henri Vrooman, Martin Hofmann-Apitius, and Holger Fröhlich. Realistic simulation of synthetic multi-scale, multi-modal patient trajectories using bayesian networks and sparse auto-encoders. Scientific reports, 10(1):1–14, 2020."

4.1 Introduction

As described earlier, BN is a generative model that helps us to simulate patients that are realistic in nature. It is also useful for longitudinal understanding of disease development and progression across all biologically relevant scales. Here, we applied BN to two datasets ADNI (`https://adni.loni.usc.edu/`) and

PPMI)(https://www.ppmiinfo.org/). The motivation behind the application of BN to these studies was to generate synthetic data subjects for them. In contrast to the previous data, the challenge is that there are many features and a comparably small number of subjects. However, many features are semantically related to each other, as they describe related aspects such as demographic information, clinical characteristics, or molecular mechanisms. The features have similar properties that can be combined together in the form of modules. Therefore, in this work, we describe a longitudinal patient cohort with the help of BN in conjunction with a deep learning technique called, sparse autoencoder. Altogether, this allows for simulating subjects that are sufficiently similar to real ones. Researchers could then develop models and generate hypotheses based on SCs that can, later on, be tested with the help of real data within their own organization. Moreover, we also demonstrate that SCs open the opportunity to simulate scenarios, which have not been observed in reality (e.g. a certain shift towards a more healthy population).

4.2 Motivation behind sparse autoencoders

The idea of sparse autoencoders was introduced in order to aggregate data on the level of variable groups, which do not make any assumptions about the underlying statistical distribution. Some variables in the datasets are continuous, and others are discrete. We discretize all the variables to enable efficient BN structure and parameter learning for arbitrary statistical distributions and non-linear dependencies between variable groups. The concept of sparse autoencoders is described in more detail in Chapter 2.

4.3 Methodology

The workflow of our approach is illustrated in Figure 4.1.

4.3.1 Datasets used

We used two longitudinal observational cohorts here, ADNI for AD and PPMI for PD. The overview of these datasets is given in the following paragraphs.

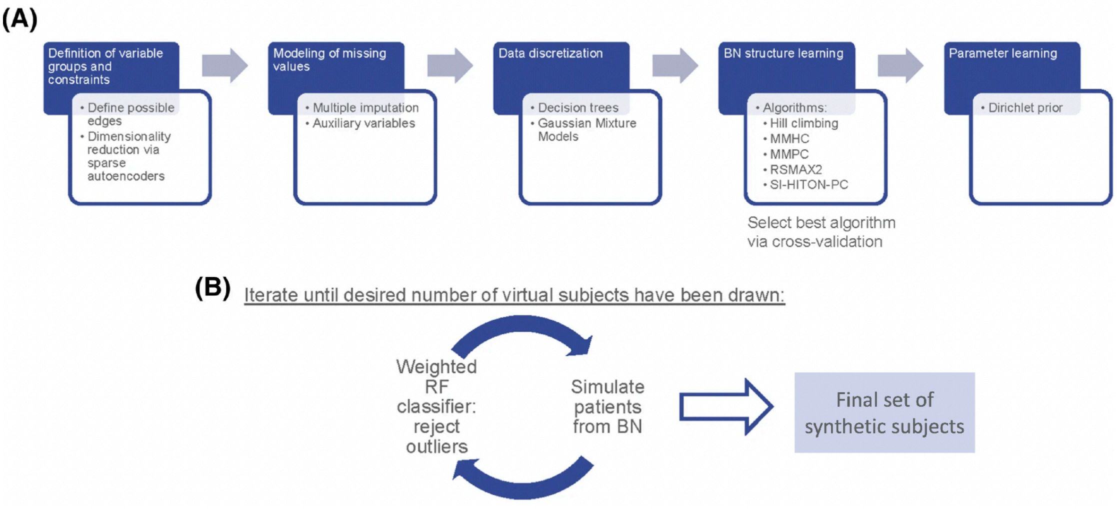

(A)
Definition of variable groups and constraints
• Define possible edges
• Dimensionality reduction via sparse autoencoders
Modeling of missing values
• Multiple imputation
• Auxiliary variables
Data discretization
• Decision trees
• Gaussian Mixture Models
BN structure learning
• Algorithms:
• Hill climbing
• MMHC
• MMPC
• RSMAX2
• SI-HITON-PC
Select best algorithm via cross-validation
Parameter learning
• Dirichlet prior
(B)
Iterate until desired number of virtual subjects have been drawn:
Weighted RF classifier: reject outliers
Simulate patients from BN
Final set of synthetic subjects

Figure 4.1: Overview about our modeling approach for longitudinal patient cohorts: (A) Approach to estimate BN, including dimensionality reduction via sparse autoencoders and modeling of missing data. (B) Conservative approach to simulate synthetic subjects.

ADNI

ADNI (adni.loni.usc.edu) was launched in 2004 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. It unites the expertise and efforts of scientists from varied disciplines and organizations who help us to gain a better understanding in the field of AD [159]. It is a longitudinal study and is undergoing at multiple sites throughout the world. It assesses clinical, imaging, genetic and bio-specimen biomarkers through the process of normal aging to early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), dementia, or AD. There are diverse data sets available in ADNI related to clinical data, genetic data, MRI data, positron emission tomography (PET) image data, etc. Therefore, it produces an accumulation of data that is heterogeneous, complex, and large. The fundamental goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be integrated to measure the progression of MCI and early AD. The other global goal is to validate the biomarkers for use in AD clinical treatment trials. ADNI is divided into 4 subsets; ADNI-1 which was the initial five-year study and was further extended by ADNI-GO, ADNI-2, and ADNI-3. There are 2600 subjects as of September 2021, out of which 837 are CN, 672 are MCI, 115 have significant memory concern (SMC), 340 are EMCI, 185 are LMCI, and 451 are in the AD stage. For up-to-date information on the study, visit `www.adni-info.org`.

In the following text we give an overview of the subjects and variables we use for our work for ADNI data and provide summary statistics for the same in Table 4.1.

Overview of ADNI data

This study includes 417 CN patients, 106 subjects with SMC, 310 subjects with EMCI, 562 subjects with LMCI, and 342 subjects, which were diagnosed with AD at the beginning of the study. In this work, we used longitudinal data from 689 subjects that were initially either diagnosed with AD (n = 342) or converted into AD patients during the study. In our work, ADNI data includes single nucleotide polymorphism (SNP) based genotype, APOE4 status, cerebrospinal fluid (CSF) biomarkers, volume measurements of seven brain regions as well as different clinical and neuropsychological test results. In addition to the 7 brain volume measurements provided in the original ADNIMERGE dataset, we calculated 68 cortical brain region volumes from raw images using Desikan parcellation which we explain in the next section. Out of more than 300,000 SNPs that have been commonly measured in the ADNI1 and ADNI2/GO phases of the ADNI study we focused on 110, which have previously been implicated as relevant in the transition of a normal/cognitively impaired state to AD [142]. We grouped all features measured in ADNI into brain volumes, cortical brain regions, cognition tests, CSF markers, genotype (SNPs + APOE4 status), demographic features, and baseline diagnosis (see exact definitions in Table 4.2). We generally discarded features with more than 50% missing values, which reduced the number of visits modeled by our approach to baseline, month 6, month 12, and month 24 (see Table 4.3).

Variable	**Dementia, N = 3411**	**MCI, N = 3221**	**NL, N = 261**
Age	76 (71, 80)	74 (70, 79)	76 (72, 78)
Gender			
Male	189 (55%)	196 (61%)	12 (46%)
Female	152 (45%)	126 (39%)	14 (54%)
yearsOfEducation	16.00 (13.00, 18.00)	16.00 (14.00, 18.00)	16.00 (14.00, 18.00)
Marital Status			
Married	285 (84%)	262 (81%)	19 (73%)
Widowed	34 (10.0%)	34 (11%)	6 (23%)
Divorced	14 (4.1%)	20 (6.2%)	1 (3.8%)
Never married	8 (2.3%)	6 (1.9%)	0 (0%)
Unknown	0 (0%)	0 (0%)	0 (0%)
Ethnicity			
Unknown	3 (0.9%)	1 (0.3%)	0 (0%)
Not Hisp/Latino	327 (96%)	312 (97%)	26 (100%)
Hisp/Latino	11 (3.2%)	9 (2.8%)	0 (0%)
apoe4			
0	114 (33%)	110 (34%)	13 (50%)
1	162 (48%)	162 (50%)	13 (50%)
2	65 (19%)	50 (16%)	0 (0%)

Table 4.1: Summary statistics of ADNI demographic data, and APOE4 status grouped by diagnostic stages, data are n (%), or median (Intra Quartile Range (IQR) in brackets), unless specified otherwise

Group	Contained Features	#Bins	MSE of autoencoder	p.adjust (real vs simulated data)
brain*	brain regions: Hippocampus, Entorhinal, Ventricles, Fusiform, Intracranial volume (ICV),Mid temporal lobe	Baseline: 4, Month 6: 12, Month 12: 4, Month 24: 2	Baseline: 0.016, Month 6: 0.013, Month 12: 0.013, Month 24: 0.013	Baseline - 1, Month 6 - 1, Month 12 - 1, Month 24 - 1
Cog.*	cognition scores: MMSE, MOCA, CDRSB, ADAS11, ADAS13, RAVLT, FAQ	Baseline: 6, Month 6: 3, Month 12: 3, Month 24: 2	Baseline: 0.022, Month 6: 0.013, Month 12: 0.016, Month 24: 0.017	Baseline - 1, Month 6 - 0.2616, Month 12 - 0.4262, Month 24 - 0.1625
CSF	ABETA, TAU, PTAU	Baseline: 2	Baseline: 0.015	Baseline - 1
SNP.bl	APOE status + 110 SNPs	Baseline: 2	Baseline: 0.06	Baseline - 1
brain68.bl	68 cortical brain regions	Baseline: 2	Baseline: 0.016	Baseline - 1
FDG	PET imaging diagnostics	Baseline: 4	NA	FDG - 1
Demographic features (treated separately	Age, Gender, Education, Race, Ethnicity, Marital status	Age: 11, Gender: 2, Education: 16, Race: 4, Ethnicity: 3, Marital status: 4	NA	Age - 1, Gender - 1, Education - 1, Race - 1, Ethnicity - 1, Marital status - 1
DX	Diagnosis at baseline and subsequent time points	Baseline: 3, Month 6, 12, 24: 4	NA	Baseline - 0.619, Month 6 - < 0.001, Month 12 - < 0.001, Month 24 - < 0.001

Table 4.2: Feature groups defined for ADNI dataset, number of bins for each feature, and MSE for each autoencoded feature. P-values correspond to a Chi-square test (null hypothesis: synthetic and real patient samples come from the same distribution). P-values were corrected for multiple testing using Bonferroni Holm's method. Note that p-values tend to become smaller the more samples are tested. $*$ Features considered as time-dependent. NA (not applicable).

Annotation for clinical visit	Months
bl	Baseline
m06	Month 6
m12	Month 12
m24	Month 24

Table 4.3: Description of suffixes of variable names in ADNI

Calculation of cortical brain region volumes in ADNI

All available MRI scans (T1-weighted scans) from the ADNI database were quantified by an open-source, automated segmentation pipeline at the Erasmus University Medical Center, The Netherlands. The number of slices of the T1w scans varied from 160 to 196 and the in-plane resolution was 256 × 256 on average, yielding an overall voxel size of 1.2 × 1.0 × 1.0 mm. From the 1715 baseline ADNI scans, the volumes of 34 bilateral cortical brain regions, 68 structures in total, were calculated using a model and surface-based automated image segmentation procedure, incorporated in the FreeSurfer Package (v.6.0, `https://surfer.nmr.mgh.harvard.edu/`). Segmentation in Freesurfer was performed by rigid-body registration and nonlinear normalization of images to a probabilistic brain atlas. In the segmentation process, each voxel of the MRI volumes was labeled automatically as a corresponding brain region based on a cortex parcellation (subdivision) guide. In this case, the cortical parcellation method, implemented by Desikan and Killiany in 2006 [160], was used for brain segmentation. For the subdivision of the human cerebral cortex into gyral-based regions, Desikan and Killiany manually identified the 34 cortical regions in the individual hemispheres. This information was encoded into an atlas that was utilized to automatically label region of interest (ROI)s. Desikan and Killiany showed that compared to manual segmentation, their automated method reached an intra-class correlation coefficient (ICC) of 0.835 across all of the ROIs. The mean distance error was less than 1 mm.

PPMI

PPMI was launched in the year 2010 by Michael J.Fox Foundation and a core group of scientists and industry partners with a mission to identify biomarkers of PD, onset, and progression (`https://www.michaeljfox.org/ppmi-clinical-study`). It is a landmark study in the field of PD that brings together collaborators from around the world to create a robust open-access data set and biosample library (`https://www.ppmi-info.org/about-ppmi`). The aim of this collaborative work is to speed up scientific breakthroughs and new treatments. As this data has open-access data sharing, it also helps to provide a deeper understanding of the disease and informed design of many therapeutic trials. PPMI is a multi-modal, longitudinal observational, multi-center cohort and it assesses the progression of clinical features, imaging outcomes, biological and genetic markers, and digital markers across all stages of PD from prodromal to moderate disease. The study aims to identify markers related to disease progression that can help:

- To accelerate therapeutic trials, further reducing the progression of disability caused due to the disease.

- To develop quantitative measures via which an optimal interval change is established between different stages of the disease (prodromal to PD).

The PPMI cohort additionally consists of multiple cohorts from a network of clinical sites. Specifically, it comprises eight cohorts with different clinical and genetic characteristics. As of February 2021, 1683 subjects were enrolled in this study. For up-to-date information on the study, visit `www.ppmi-info.org`.

In the following text, we give an overview of the subjects and variables we use for our work for PPMI data and provide summary statistics for the same in Table 4.4.

Overview of PPMI data

Variable	**N = 362**
Age	62 (55, 69)
Gender	
Female	122 (34%)
Male	240 (66%)
yearsOfEducation	16.00 (14.00, 18.00)

Table 4.4: Summary statistics of PPMI demographic data are n (%), or median (IQR in brackets), unless specified otherwise

Here we used data from 362 de-novo PD patients. These untreated subjects were diagnosed with PD for 2 years or less and showed signs of resting tremor, bradykinesia, and rigidity during the last 2 years. The dataset contains 831 clinical variables, which we categorized into 12 groups, such as patient demographics, patient PD history, imaging, non-motor symptoms, CSF markers, and UPDRS (see the complete list and exact definition in Table 4.5). PPMI assesses clinical variables at baseline and 11 follow-up visits. Noteworthy, some variables were assessed irregularly and not for all patients, yielding missing values. We generally discarded features with more than 50% missing values for modeling purposes. Accordingly, there were 12-time points included in our model, but not all variables were available at each time point (see Table 4.6).

Group	Contained Feature	#Bins	MSE of autoencoder	p.adjust (real vs simulated data)
Patient_ENROL_AGE	Age at enrollment	4	NA	1
Patient_Simplified_Gender	Child bearing capacity	Male, Female	NA	1
Patient_Gender	Gender	Child bearing capacity - yes, no	NA	1
Patient demographic	Gender, Ethnicity- Is subject Hispanic/Latino, Identify self as Am Indian/Alaska Native, Identify self as Asian, Identify self as Black/African American, Identify self as Hawaiian/Other Pacific, Identify self as White, Race not specified, Origin population	7	NA	1
Patient PD history	Biological Mother, Biological Mother with PD, Biological Father, Biological Father with PD, Full Siblings, Full Siblings with PD, Half Siblings, Half Siblings with PD, Maternal Grandparents, Maternal Grandparents with PD, Paternal Grandparents, Paternal Grandparents with PD, Maternal Aunts and Uncles, Maternal Aunts and Uncles with PD, Paternal Aunts and Uncles, Paternal Aunts and Uncles with PD, How many children do you have, How many children with PD, PD Family History, Duration of the disease at enrollment, Handedness, Number of years of education	7	NA	1
UPDRS	UPDRS- Total Unified Parkinson's Disease Rating Scale score	Baseline-2, visit 1- 3, visit 2 -5, visit 3- 2, visit 4- 4, visit 5-2, visit 6-2, visit 7- 2, visit 8- 1, visit 9- 2, visit 10- 2, visit 11- 6	NA	Baseline - 1, visit 1 - 1, visit 2 - 1, visit 3 - 0.34, visit 4 - 0.01, visit 5 - 1, visit 6 - 1, visit 7 - 1, visit 8 - 1, visit 9 - 1, visit 10 - 1, visit 11 - 1
	UPDRS1-Non-motor experiences of daily living	Baseline-3, visit 1- 3, visit 2 -4, visit 3- 5, visit 4- 4, visit 5-2, visit 6-4, visit 7- 2, visit 8- 2, visit 9- 3, visit 10- 4, visit 11- 3	NA	Baseline - 1, visit 1 - 1, visit 2 - 1, visit 3 - 1, visit 4 - 1, visit 5 - 1, visit 6 - 1, visit 7 - 1, visit 8 - 1, visit 9 - 1, visit 10 - 1, visit 11 - 1
	UPDRS2-Motor experiences of daily living	Baseline-3, visit 1- 5, visit 2 -2, visit 3- 3, visit 4- 2, visit 5-6, visit 6-2, visit 7- 2, visit 8 - 5, visit 9- 3, visit 10- 4, visit 11- 5	NA	Baseline - 1, visit 1 - 1, visit 2 - 1, visit 3 - 1, visit 4 - 1, visit 5 - 1, visit 6 - 1, visit 7 - 1, visit 8 - 1, visit 9 - 1, visit 10 - 1, visit 11 - 1
	UPDRS3-Motor examination	Baseline-2, visit 1- 2, visit 2 -5, visit 3- 4, visit 4- 4, visit 5-2, visit 6-2, visit 7- 2, visit 8- 1, visit 9- 5, visit 10- 3, visit 11- 3	NA	Baseline - 1, visit 1 - 1, visit 2 - 1, visit 3 - 0.51, visit 4 - 0.02, visit 5 - 1, visit 6 - 1, visit 7 - 1, visit 8 - 1, visit 9 - 1, visit 10 - 1, visit 11 - 1

Group	Contained Feature	#Bins	MSE of autoencoder	p.adjust (real vs simulated data)
Medical history	WGTKG - Weight (in Kilograms), HTCM - Height (in Centimeters), TEMPC- Temperature (in Celsius), SYSSUP -Supine BP – systolic, DIASUP - Supine BP – diastolic, HRSUP-Supine heart rate, SYSSTND-Standing BP – systolic, DIASTND-Standing BP – diastolic, HRSTND- Standing heart rate	Baseline -2, visit 1 - 6, visit 2- 7, visit 3 - 7, visit 4- 4, visit 5 - 4, visit 6 - 4, visit 7 - 3, visit 8 - 4, visit 9- 5, visit 10 - 4, visit 11- 8	Baseline: 0.019, Visit 1: 0.015, Visit 2: 0.013, Visit 3: 0.016, Visit 4: 0.018, Visit 5: 0.015, Visit 6: 0.018, Visit 7: 0.016, Visit 8: 0.019, Visit 9: 0.017, Visit 10: 0.016, Visit 11: 0.010	Baseline - 1, visit 1 - 1, visit 2 - 1, visit 3 - 0.51, visit 4 - <0.01, visit 5 - 1, visit 6 - 1, visit 7 - 1, visit 8 - 1, visit 9 - 1, visit 10 - 1, visit 11 - 1
Non-motor	DVT_TOTAL_RECALL-Derived-Total Recall T-Score, DVS_LNS-Derived-LNS Scaled Score, ESS-Epworth sleepiness scale, QUIP-Questionnaire for Impulsive-Compulsive Disorders in PD, SCOPA-Scales for outcomes in Parkinson's disease-autonomic, STA-State Trait Anxiety Total Score	Baseline- 2, Visit 2- 6, visit 4- 4, visit 6- 5, visit 8- 4, visit 10- 3	Baseline: 0.027, Visit 2: 0.018, Visit 4: 0.024, Visit 6: 0.023, Visit 8: 0.029, Visit 10: 0.023	Baseline - 1, visit 2 - 1, visit 4 - 0.07, visit 6 - 1, visit 8 - 1, visit 10 - 1
RBD	REM Sleep Behavior disorder (RBD)	Baseline- 2, Visit 2- 3, visit 4- 3, visit 6- 3, visit 8- 3, visit 10- 4	NA	Baseline - 1, visit 2 - 1, visit 4 - 1, visit 6 - 1, visit 8 - 1, visit 10 - 1
CSF	Abeta 42 (pg/ml)	Baseline - 2	NA	Baseline - 1
	CSF Alpha-synuclein (pg/ml)	Baseline -2, visit 2 - 3, visit4 - 4, visit6 - 4, visit 8 - 2	NA	Baseline - 1, visit 2 - 1, visit 4 - 0.52, visit 6 - 1, visit 8 - 1
	p-Tau181P (pg/ml)	Baseline - 3	NA	Baseline - 1
	Total tau (pg/ml)	Baseline - 2	NA	Baseline - 1
	t-tau/Abeta 1-42	Baseline - 3	NA	Baseline - 1
	p-tau/Abeta 1-42	Baseline - 4	NA	Baseline - 1
	p-tau/t-tau	Baseline - 3	NA	Baseline - 1
Biological	ALDH1A1 (rep 1),ALDH1A1 (rep 2), GAPDH (rep 1), GAPDH (rep 2), HSPA8 (rep 1), HSPA8 (rep 2), LAMB2 (rep 1), LAMB2 (rep 2), PGK1 (rep 1), PGK1 (rep 2), PSMC4 (rep 1), PSMC4 (rep 2), SKP1 (rep 1), SKP1 (rep 2), UBE2K (rep 1), UBE2K (rep 2)	Baseline - 4, visit 8 - 6	Baseline: 0.011, Visit 8: 0.007	Baseline - 1, visit 8 - 1
Imaging	MRI results	3	NA	Baseline - 1

Table 4.5: Feature groups defined for PPMI dataset. = treated as individual variables. For UPDRS "off-medication scores" were used. P-values correspond to a Chi-square test (null hypothesis: synthetic and real patient samples come from the same distribution). P-values were corrected for multiple testing using Bonferroni Holm's method. Note that p-values tend to become smaller the more samples are tested.

The annotations for clinical visits for PPMI data are described in Table 4.6.

Annotation for clinical visit	**Months**
V00	Baseline
V01	Visit 01 (Month 3)
V02	Visit 02 (Month 6)
V03	Visit 03 (Month 9)
V04	Visit 04 (Month 12)
V05	Visit 05 (Month 18)
V06	Visit 06 (Month 24)
V07	Visit 07 (Month 30)
V08	Visit 08 (Month 36)
V09	Visit 09 (Month 42)
V10	Visit 10 (Month 48)
V11	Visit 11 (Month 54)

Table 4.6: Description of suffixes of variable names in PPMI

4.3.2 Dimensionality reduction via MBNs using sparse autoencoders

Dimensionality reduction is performed using sparse autoencoders. For this purpose, the group of features to which each feature belongs was defined as illustrated in Table 4.2 for ADNI and Table 4.5 for PPMI.

As described earlier, learning the true CPDAG structure is NP hard [96]. Longitudinal observational cohorts like ADNI and PPMI have many variables and limited sample sizes. These types of data structures require the CPDAG structure to limit the space of potential networks substantially. For this purpose, the type of networks known as module networks has been introduced [121]. These networks have been described in detail in Chapter 2. The key idea in Module Networks is to group variables into modules, which share parameters. During the BN structure learning process, only edges between modules are learned. In our case, modules comprised e.g. imaging related features, plasma biomarkers, SNPs, medical history, cognition scores, etc (Table 4.2 and 4.5). The key question is, how to learn and encode a shared distribution for a module. In their original publication, Segal *et al.* relied on the assumption of normally distributed data (such as gene expression) and employed decision trees to represent modules [121]. In this work, we used sparse autoencoders, which can weigh the influence of different variables on the aggregate module score. Furthermore, autoencoders do not make any distribution assumption. We enforced sparsity in the network by introducing drop-out units in the input layer. Furthermore, we used an l_2 penalty for all weights. We constructed a separate sparse autoencoder for every variable group and different combinations of activation functions were tested via a grid search. Grid search also involved the tuning of l_2 penalty and drop-out ratio of input units. Autoencoders were trained for brain volumes, cognitive scores, CSF features, SNP features, and cortical brain regions for ADNI data (Table 4.3); and medical history, non-motor, and biological features for PPMI data (Table 4.5). The loss function optimized by the autoencoder networks was the MSE. Tuned hyper-parameters of autoencoder networks included the activation function (rectified linear unit or hyperbolic tangent), the input dropout ratio (0%, 5%, 20%, 50%), l_2 penalty (10^{-4},..., 10^{4}), and the network architecture. More specifically, we tested the following architectures:

- One hidden layer with one hidden unit.
- Two hidden layers: first layer with n/2 units, second with one hidden unit.
- Three hidden layers: first layer with n/2 units, second with n/4 units, and third with one hidden unit.

For each combination of hyper-parameters, a separate autoencoder training was performed for at most 500 epochs, but stopped earlier, if the MSE did not improve for 5 rounds. The best autoencoder model was selected according to the MSE criterion. We here relied on the h2o autoencoder implementation (`http://docs.h2o.ai/`). Tables 4.3 and 4.5 show the MSE obtained for each autoencoded module. To understand the influence of individual features on each of the autoencoder networks, we applied the method by Gedeon et al. [161], which is based on the idea that the relative contribution of the ith input to the jth output of a neuron can be estimated by:

$$P_{ij} = \frac{|W_{ij}|}{\sum p|W_{pj}|} \tag{4.1}$$

where the sum runs over all inputs of the neuron. P_{ij} can be regarded as the weight of the edge $i \rightarrow j$ in the neural network graph. Now assume that j itself feeds into a further neuron k. Gedeon [161] suggests to estimate the overall impact of i on k by:

$$P_{ik} = \sum_r P_{ir} P_{rk} \tag{4.2}$$

That means we take the product of edge weights along a path connecting i and k and sum over all alternative paths. The definition can directly be extended to deeper networks.

4.3.3 Dealing with missing data

As ADNI and PPMI are most likely a combination of MAR and MNAR mechanisms, missing data was dealt in a similar way as described in chapter 3. The difference here is that instead of individual variables, we grouped the variables in the form of modules, and therefore, we introduced one auxiliary variable for each variable group/module and visit to account for patient drop-out, i.e. MNAR. Moreover, in the case of features that are assessed at different visits, we enforced auxiliary variables to point from one to the next visit. For example, in ADNI dataset we introduced auxiliary variables for brain volume measurements at baseline, visit at the 6th month, visit at the 12th month and visit at the 24th month. Accordingly, the auxiliary variable for the feature “brain.bl” (brain volume at baseline) was also a parent of the auxiliary variable for the feature “brain.m06” (brain volume at month 6) (Figure 4.2). Details about the precise definition of auxiliary variables used in our work can be found in Tables 4.7 and 4.8 for ADNI and PPMI respectively.

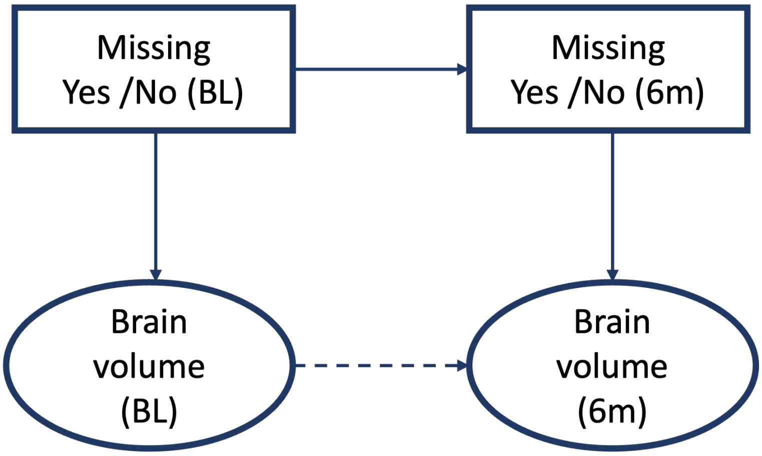


Figure 4.2: Temporal dependency of auxiliary variables (rectangles). The solid line is prescribed, and the dashed line may be inferred from the data. BL and 6m depict baseline and month 6 respectively.

Auxiliary variable	**Target variables**
brainvol.bl.aux	brain.bl
brainvol.m06.aux	brain.m06
brainvol.m12.aux	brain.m12
brainvol.m24.aux	brain.m24
CogScore.m06.aux	Cog.m06
CogScore.m12.aux	Cog.m12
CogScore.m24.aux	Cog.m24
brain68.aux	brain68.bl
snp.aux	SNP

Table 4.7: Auxiliary variables defined for ADNI dataset

Auxiliary variable	**Target variables**
CSF_aux_V00	Abeta.42_V00, Alpha.synuclein_V00, p.Tau181P_V00, Total.tau_V00, tTau.Abeta_V00, pTau.Abeta_V00, pTau.tTau_V00
Biological_aux_V00	Biological_V00
Biological_aux_V08	Biological_V08
UPDRS_aux at V01 to V11	UPDRS1, UPDRS2,UPDRS3, UPDRS (V01 to V11)
MedicalHistory_aux (V01 to V11)	Medical History features at V01 to V11
NonMotor_aux (V02, V04, V06,V08,V10)	NonMotor features (V02, V04, V06,V08,V10)

Table 4.8: Auxiliary variables defined for PPMI dataset

4.3.4 Imposing Constraints on Network Structure

As described earlier it is important to restrict CPDAG equivalence class as much as possible by prior knowledge to allow the correct orientation of as many edges as possible. In our case we specifically imposed the following constraints for BN structures:

- Demographic and other clinical baseline features (age, gender, ethnicity) can only influence other features, but they are not influenced by themselves.
- Medical history in PD can depend on motor, non-motor, and other clinical features.
- Imaging features can be related to each other, but they don't influence other features.
- Clinical diagnosis in AD is dependent on cognitive assessment scores, but not vice versa.

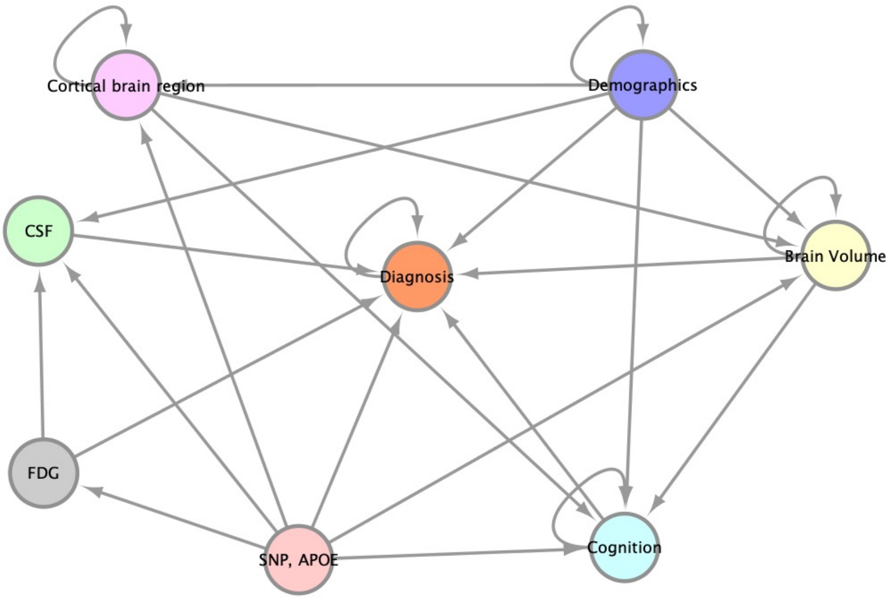


Figure 4.3: Potentially allowed edges between different variable groups in ADNI.

- Clinical outcome measures (e.g. UPDRS for PD) can influence imaging and in PD they can be mutually correlated with the assessment of non-motor symptoms.
- Biomarkers, including genomic features, can influence all features, except for clinical baseline features.
- Longitudinal features must follow the right temporal order, i.e. there are no edges pointing backward in time.
- Auxiliary variable for a particular feature/group can only influence its corresponding feature/group and the auxiliary variable for the same feature/group at the next time point (see last section and Figure 4.2).

Figures 4.3, and 4.4 schematically depict the set of potentially allowed edges, which we defined between variable groups for the AD and PD datasets used in this work.

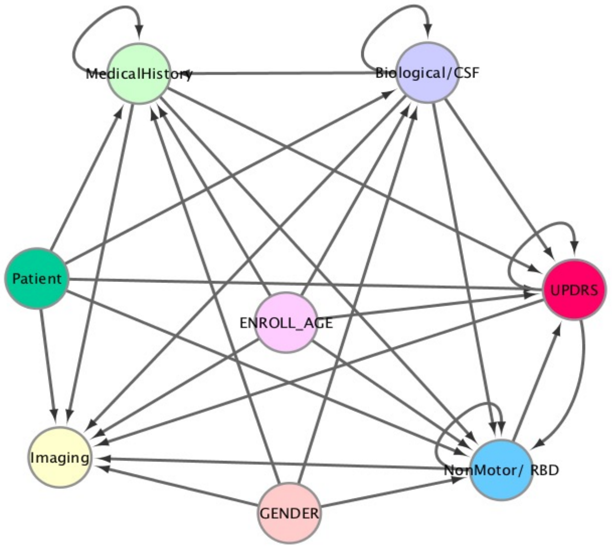


Figure 4.4: Potentially allowed edges between different variable groups in PPMI.

4.3.5 Data Discretization

Structure learning with BNs is only computationally efficient, if all variables follow a Gaussian or multinomial distribution, because then the marginal log-likelihood, integrating out model parameters, can be computed analytically [67]. Since in our case we had highly heterogeneous data, where many features were clearly non-Gaussian, we decided to perform data discretization. In the case of ADNI study, this was done via a supervised, decision tree-based approach [162], where baseline diagnosis of patients (CN, MCI, or AD) was taken as a label. In the case of PPMI study, all patients had a de-novo PD diagnosis. Accordingly, we here employed an unsupervised univariate clustering via gaussian mixture models (GMM) for discretization purposes. Both methods result in a variable number of discrete values for each feature (Tables 4.1, and 4.3). For comparison reasons, we also conducted BN structure learning without any discretization while assuming a Gaussian distribution for each continuous variable, see details in the next section.

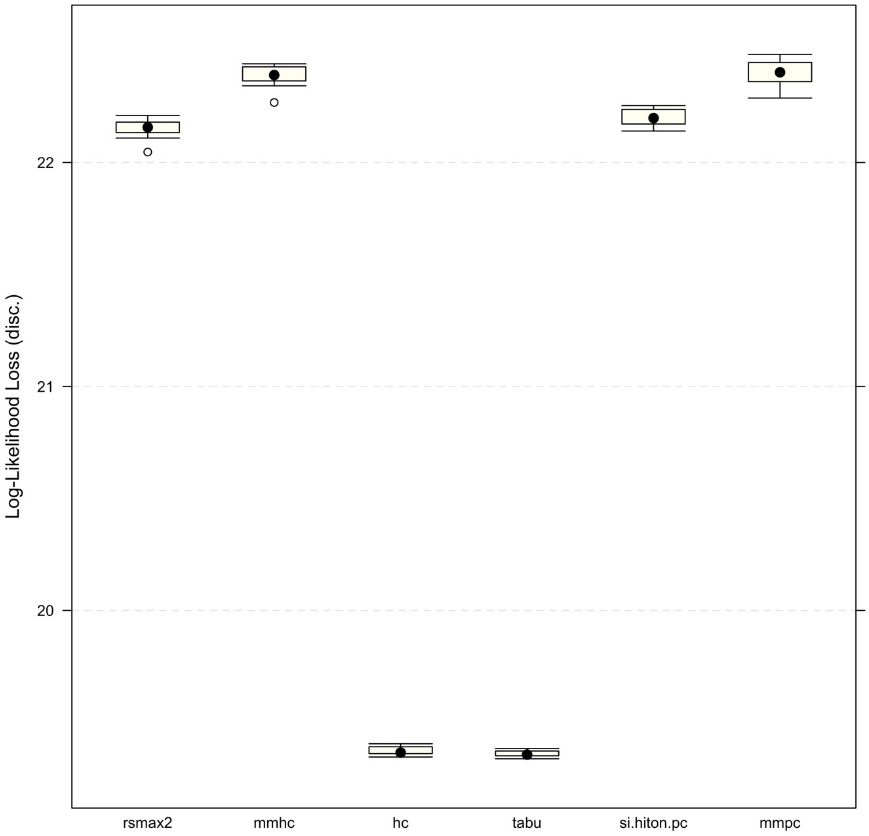


Figure 4.5: Comparison of different BN structure learning algorithms via 10-fold cross-validation for ADNI dataset. The y-axis depicts the negative log-likelihood of the test data.

4.3.6 BN Structure and Parameter Learning

We learned the network on six different algorithms from R-package bnlearn [113] hc, MMHC, tabu search, MMPC, RSMAX2 and SI-HITON-PC(described in detail in Chapter 2). Tabu search was identified as the best performing BN structure learning algorithm for ADNI and hc for PPMI for discretized data (Figures 4.5, 4.6). This is in agreement with recent findings that in most situations score based search methods are superior to constrained-based ones [163]. Given a learned BN topology, parameters can then be inferred using a Dirichlet prior to account for parent-child node configurations that are not observed [109]. BN structure and parameter learning was executed via the R-package bnlearn [164].

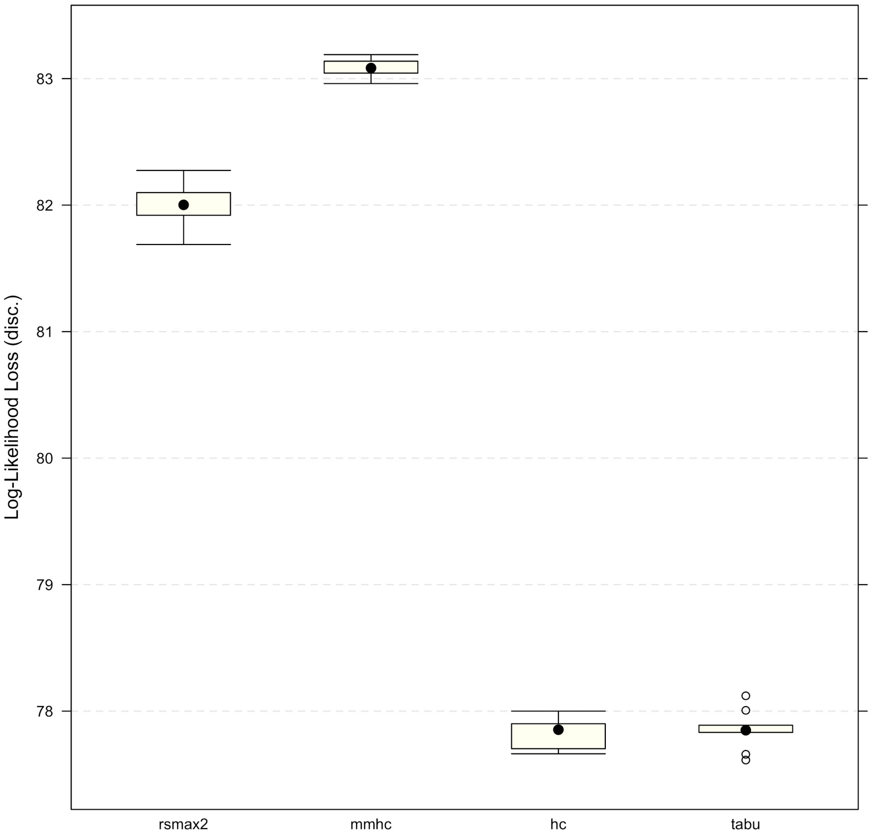


Figure 4.6: Comparison of different BN structure learning algorithms via 10-fold cross-validation for PPMI dataset. The y-axis depicts the negative log-likelihood of the test data. The two Markov Blanket learning algorithms (SI-HITON-PC, MMPC) are not shown, because their implementation in the bnlearn package resulted in an error message.

When omitting data discretization, we end up in a BN with a mixture of Gaussian and discrete nodes (hybrid BN). To allow for a direct comparison with BN structure learning after discretization, we used the same structure learning algorithms. In addition, score-based search algorithms have empirically been found to show a more robust behavior in terms of network reconstruction accuracy than constraint-based methods for mixed discrete/continuous data, specifically for smaller sample sizes [165].

4.3.7 Simulation of Synthetic Patients

As described in Chapter 3, given a BN with learned parameters, a synthetic patient can be simulated by first drawing random values from parent node distributions and subsequently from their child node distributions while conditioning on the values of the parents. Each synthetic patient thus corresponds to a vector of features, which follow the conditional statistical dependencies learned by the BN. If the BN is learned from discretized data, then also each virtual patient's feature vector is discrete.

We generated synthetic data in two ways and compared the results:

- Non-conservative method: Here we directly drew the synthetic data from the BN by the approach described above.
- Conservative method: A general concern at this point is that synthetic subjects could show differences from real subjects either due to insufficient model fit or due to the existence of confounding factors that are not part of the observed data, resulting in biases in BN parameter estimates. To account for this aspect we developed a scoring scheme, which could help to exclude unrealistic synthetic subjects directly after simulation. This was done by training an RF classifier [166], which puts 100 times more weight on correctly classifying original patients than simulated ones. The weighted RF classifier assigns to each synthetic subject a probability/confidence score to fall into the real patient distribution. In this way, we here excluded synthetic subjects that showed a lower than 50% probability to fall into the real subject distribution and could thus be regarded as outliers. The whole procedure of simulating

subjects and excluding seemingly unrealistic ones can be run iteratively until a desired number of synthetic subjects have been generated.

4.3.8 Classifier trained on AD and PD subjects to evaluate synthetic data

To validate our synthetic data generation scheme we generated the same number of synthetic as well as real subjects for ADNI and PPMI and then asked whether a conventional RF classifier was able to separate between synthetic and real subjects within a 10 times repeated 10-fold cross-validation scheme. That means we sequentially left out 1/10 of subjects and trained an RF on the remaining subjects to learn the discrimination between real and synthetic subjects. We used the left-out portion of the data to assess the prediction performance of the RF. We used the pAUC at a pre-specified true positive rate of 90% for real patients as a measure of the prediction performance. That means we looked at AUC-ROC at which the detection rate for a real patient was between 90% and 100%. This was done to account for the fact that misclassification of a synthetic patient as real would be far less relevant than the other way around. Following the implementation in R-package pROC [167] the pAUC is a measure in the interval [0, 1], where 0.5 represents the chance level.

4.3.9 Simulation of counterfactual interventions in BN

Judea Pearl developed a well-established theory for modeling and simulating interventions into BNs [95]. Assume we want to predict the intervention effect of $X_k = x$ on the remaining random variables in the BN, i.e. $P(X_1, ..., X_{k1}, X_{k+1}, ..., X_n | do(X_k = x))$. Pearl demonstrated in his work that this intervention effect can be computed by estimating the conditional probability distribution $P(X_1, ..., X_{k1}, X_{k+1}, ..., X_n | X_k = x)$ within a *mutilated* BN, in which all incoming edges into X_k have been deleted.

In practice, we used logic sampling [168] for estimating the conditional probability distribution $P(X_1, ..., X_{k1}, X_{k+1}, ..., X_n | X_k = x)$ in the *mutilated* BN. Logic sampling instantiates all the nodes in a BN by sampling from the prior distribution and removes all samples that are not compatible with the evidence [169].

4.4 Results

4.4.1 BN structure reflect expected causal associations

To gain a better understanding of the variable dependencies learned by our BN models we performed a non-parametric bootstrap [154] similar to the one explained in Chapter 3. Figure 4.7 and Figure 4.8 depicts the network structure learned by ADNI and PPMI data respectively. As expected, edges connecting variables, which represent the same group of features (e.g. UPDRS, CSF biomarkers, brain volume measurements) at different visits were inferred more stable than edges between different variable groups. That means BN structure learning was able to learn stable longitudinal dependencies in the data. This is for e.g. marked by the connections between variables DX.bl (baseline diagnosis), DX.6 (diagnosis at 6 months), DX.12 (diagnosis at 12 months), and DX.24 (diagnosis at 24 months) in ADNI. Clinical diagnoses at each time point are dependent on cognitive impairment scores at the same time point because the clinical diagnosis of dementia in practice is done on the basis of such tests. In addition, in ADNI stable connections between genotype (SNPs) and baseline diagnosis, cognitive impairment scores (Cog.bl), and amyloid PET scan diagnostics fluorodeoxyglucose (FDG) were found. We investigated the relative influence of individual SNPs in the sparse autoencoder network output to understand these connections better. This was done via the method described in [161], see “Methods” section for more details. Altogether, there was a non-zero influence of all 110 SNPs plus APOE4 status in the SNP group. The most relevant SNP (rs9384488) has been associated with quantitative global cortical Abeta load [170]. Abeta plaques are one of the hallmarks of AD, and Abeta measurements are part of the CSF variable group, hence providing an interpretation of the SNP $\rightarrow$ CSF edge in our BN as well as SNP $\rightarrow$ FDG.

In PPMI the edge of UPDRS1 to non-motor symptoms reflects (found in about 500/1000 BN reconstructions) the fact that the UPDRS scoring system comprises three parts, and the first part captures non-motor symptoms (cognitive function, behavior, and mood) [171]. Similarly, the stable edge between non-motor symptoms and RBD can be explained by the fact that sleeping disorder assessment is part of non-motor symptom-related variables in PPMI. In summary, BN structures learned by our models reflected expected variable dependencies in both datasets.

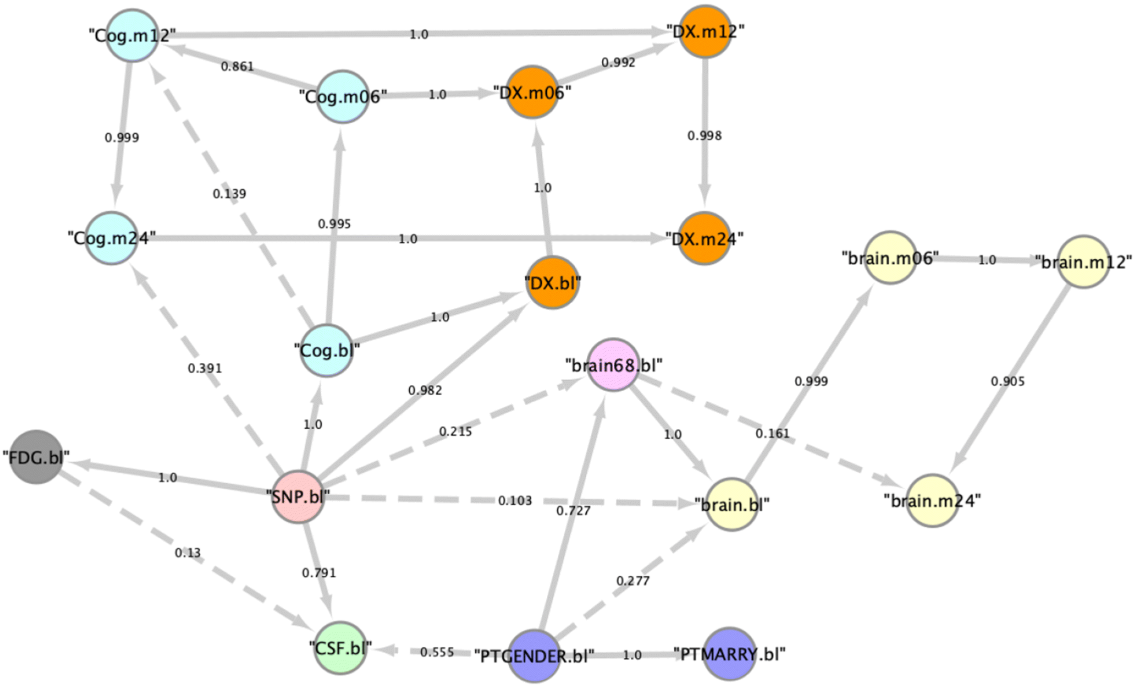


Figure 4.7: Variable dependencies identified in ADNI dataset in more than 100/1,000 bootstrapped BN reconstructions (dashed lines; relative frequency = edge label). Solid edges indicate variable dependencies that are found commonly in bootstrapped BN reconstruction and the final BN topology.

4.4.2 Synthetic AD and PD patients looks realistic

Figure 4.9 and 4.10 demonstrates that for both, ADNI and PPMI, the cross-validated classification performance is used to detect synthetic subjects not clearly better than the chance level. It reflects the performance of both non-conservative and conservative methods. We observe from the figure that the conservative method performs better than the non-conservative method. We also generated more synthetic data using the conservative method. This was done five times and we generated 1368, 2278, 5817, and 10878 synthetic subjects. The RF classifier was trained separately for each case including the actual number of real subjects (689). We observe from Figures 4.11 and 4.12 that the classifier was not able to distinguish between real and synthetic data for all five cases.

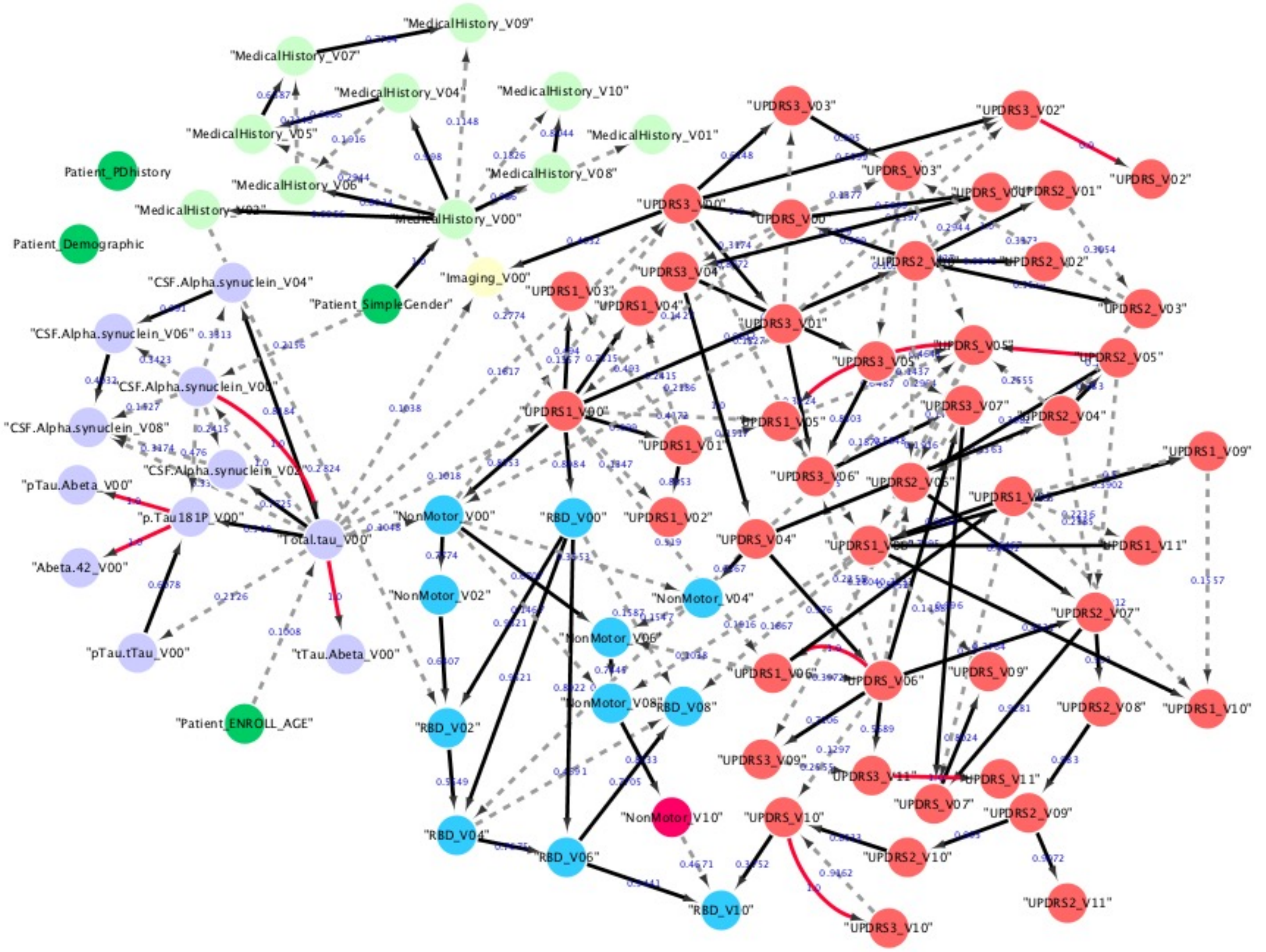


Figure 4.8: Variable dependencies identified in PPMI dataset in more than 100/1000 bootstrapped BN reconstructions (dashed lines) and the final BN (learned on the entire dataset, red lines), respectively. Solid edges indicate variable dependencies that are found commonly in bootstrapped BN reconstruction and the final BN topology. Auxiliary variables are not shown to simplify the representation.

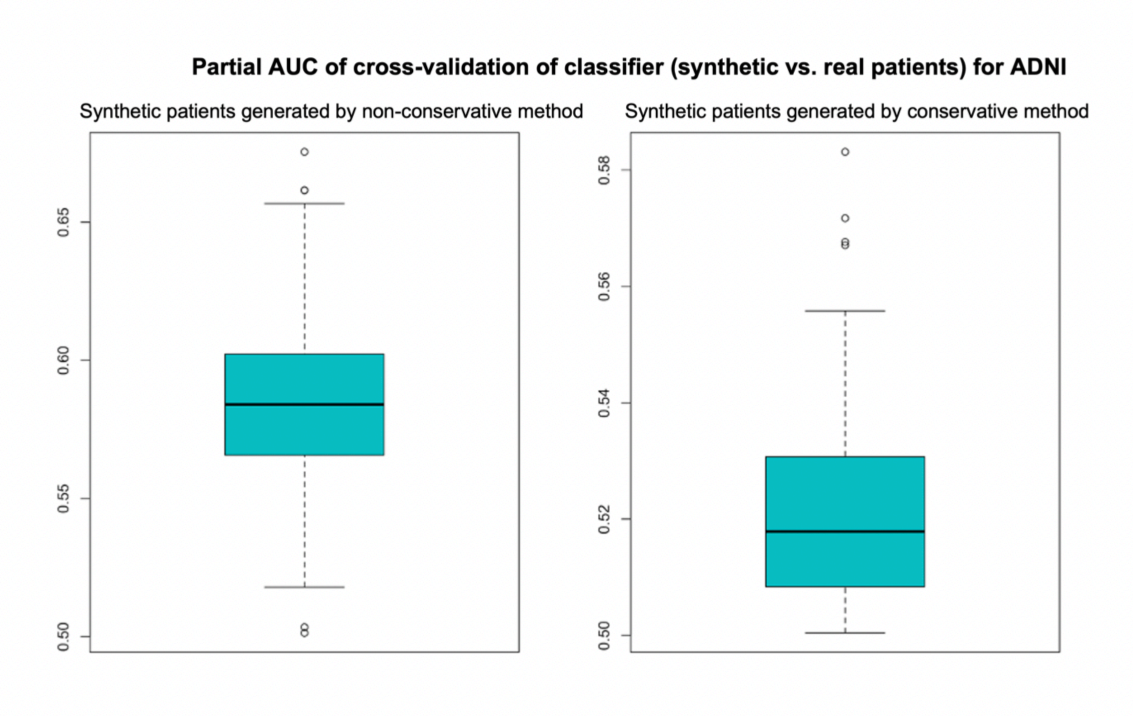


Figure 4.9: Performance of a RF to correctly identify synthetic subjects, measured via the partial area under ROC curve (pAUC) at a pre-specified detection rate of $\geq 90\%$ for real patients. The pAUC was assessed on test sets within 10 repeats of a 10-fold cross-validation procedure. Accordingly, boxplots show the distribution of the 10-fold cross-validated pAUC that was obtained from 10 repeats of the cross-validation procedure. The left plot shows the performance when the SC is obtained by directly drawing from the BN. The right plot shows the performance when using our suggested conservative approach.

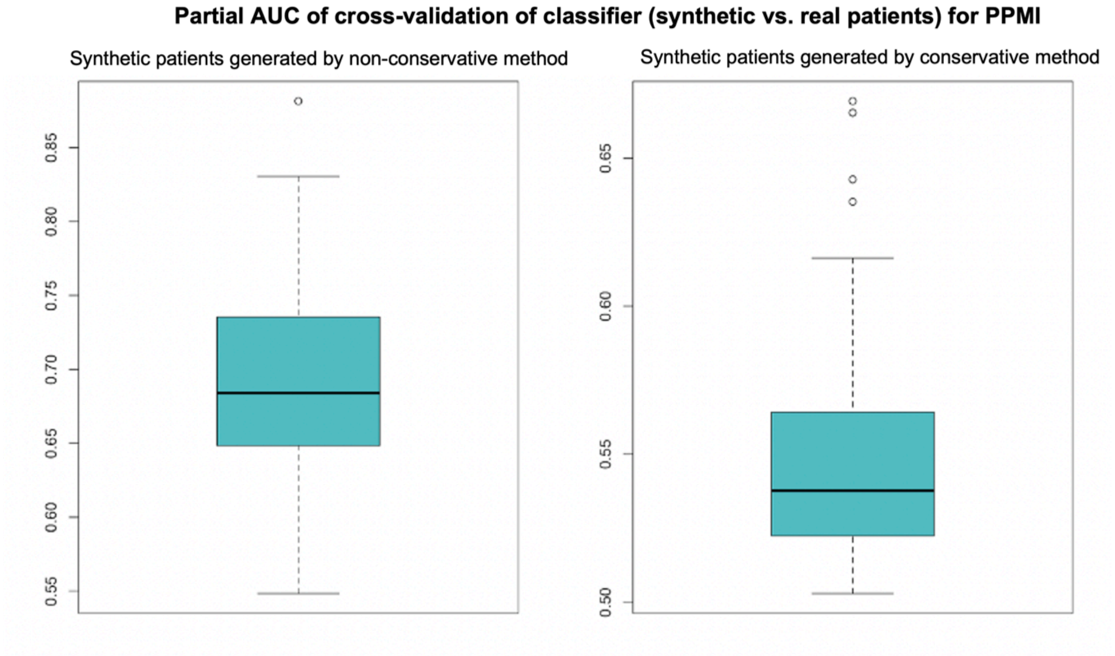


Figure 4.10: Performance of a RF to correctly identify synthetic subjects, measured via the partial area under ROC curve (pAUC) at a pre-specified detection rate of $\geq 90\%$ for real patients. The pAUC was assessed on test sets within 10 repeats of a 10-fold cross-validation procedure. Accordingly, boxplots show the distribution of the 10-fold cross-validated pAUC that was obtained from 10 repeats of the cross-validation procedure. The left plot shows the performance when the SC is obtained by directly drawing from the BN. The right plot shows the performance when using our suggested conservative approach.

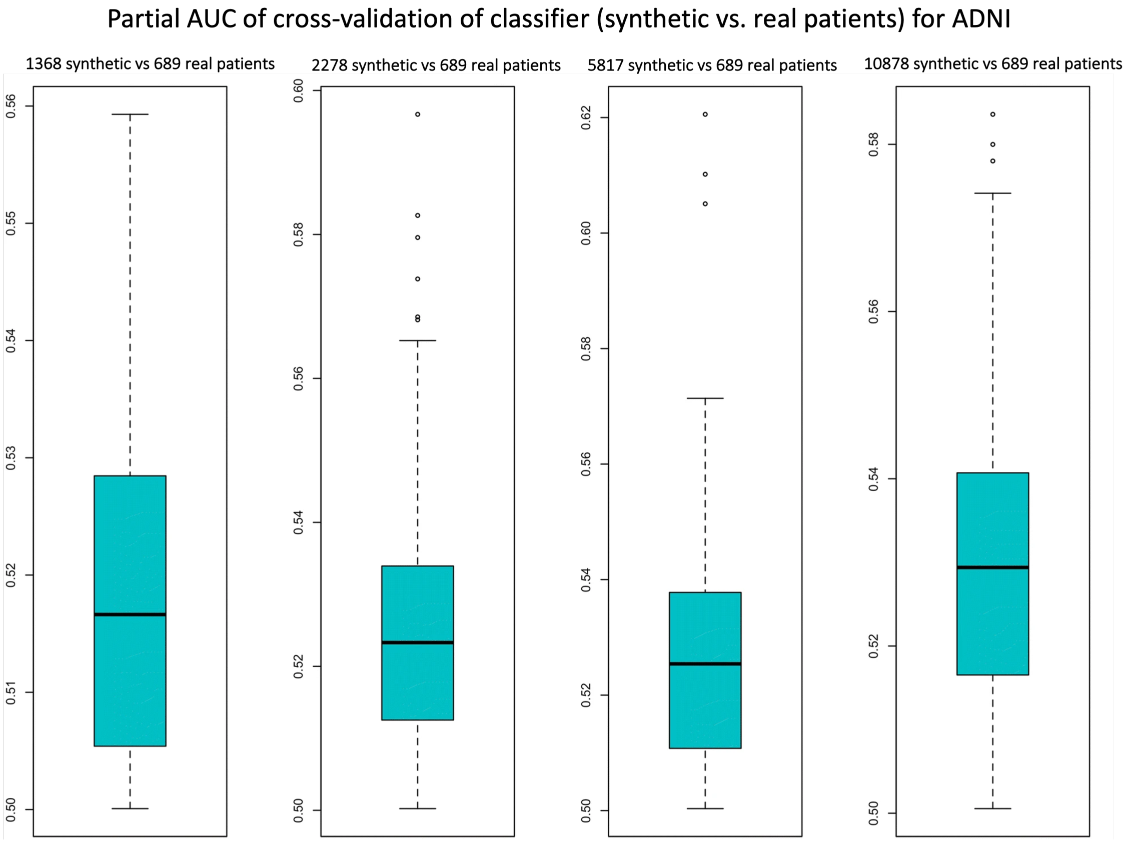


Figure 4.11: Performance of RF classifier to correctly identify a given number of real ADNI patients among synthetic subjects. The performance was measured via the pAUC at a pre-specified detection rate of $\geq$ 90% for real patients. The pAUC was assessed on test sets within 10 repeats of a tenfold cross-validation procedure. Accordingly, boxplots show the distribution of the tenfold cross-validated pAUC that was obtained from 10 repeats of the cross-validation procedure.

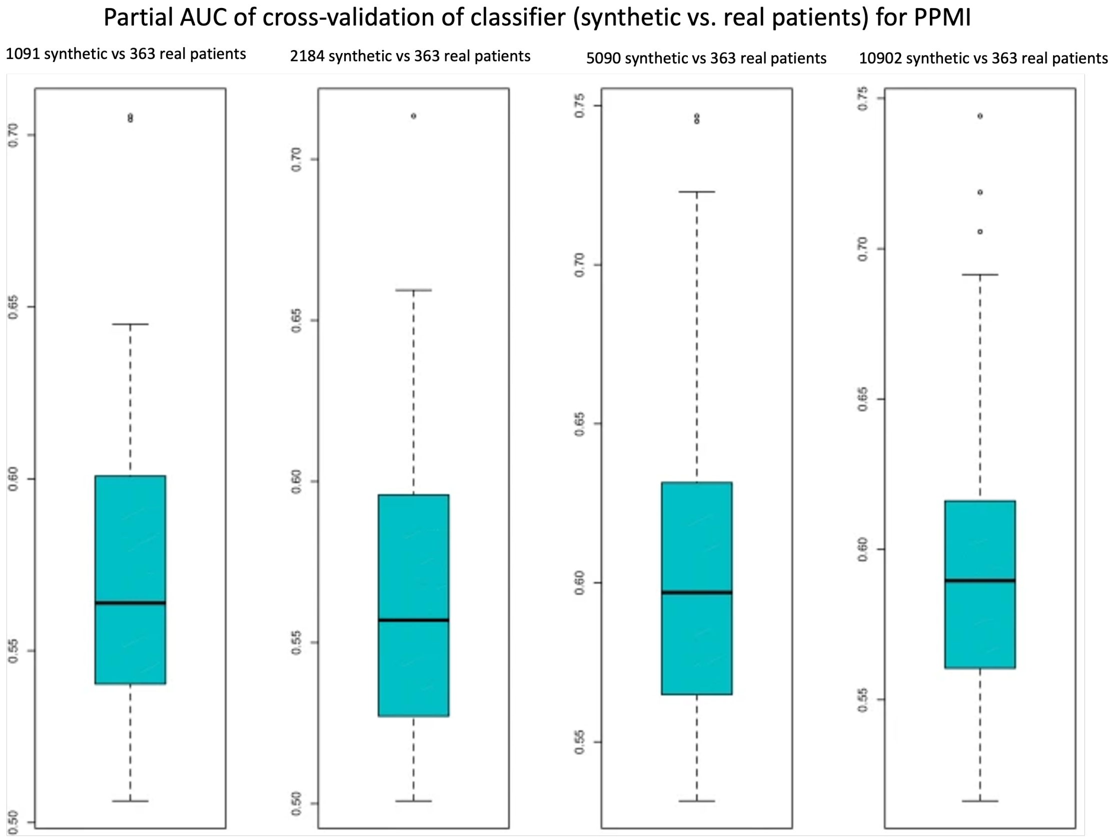


Figure 4.12: Performance of RF classifier to correctly identify a given number of real PPMI patients among synthetic subjects. The performance was via the pAUC at a pre-specified detection rate of $\geq 90\%$ for real patients. The pAUC was assessed on test sets within 10 repeats of a tenfold cross-validation procedure. Accordingly, boxplots show the distribution of the tenfold cross-validated pAUC that was obtained from 10 repeats of the cross-validation procedure.

4.4.3 Simulating an intervention in a SC

As pointed out earlier, the final BN structure learned from ADNI represents expected dependencies between variable groups and indeed all of these dependencies can be regarded as causal (Figure 4.7, solid edges): In particular, note that the effect of genotype (SNPs) on cognitive assessment scores (Cog.bl), amyloid PET scan diagnostics (FDG.bl) and baseline diagnosis can only be interpreted causally. Likewise, the influence of gender on subcortical brain volumes can only be interpreted causally, although there potentially exist mediators such as longevity (women on average live longer than men). To further exemplify the use of our causal BN, we simulated an intervention for dementia and MCI patients at baseline that shifted their cognition scores (Alzheimer's disease assessment scale-Cognitive subscale (ADAS)11, ADAS13, MMSE, clinical dementia rating sum of boxes (CDRSB), FAQ, rey auditory verbal learning test (RAVLT)) towards the median score of the CN patients (n = 423), e.g. the effect via a drug. We encoded perturbed cognition scores via the autoencoder model that we had trained earlier on cognition scores of real subjects. We then simulated the same number of synthetic subject trajectories as real subjects while conditioning on the shift in study baseline cognitive assessment scores. That means we used the perturbed cognition scores of real subjects and then sequentially drew data for each dependent variable using the conditional probability tables (i.e. BN parameters) learned by our model. This implies that the intervened node becomes statistically independent from its parents, i.e. all incoming edges of variable Cog.bl are deleted in the intervened network [95]. Figure 4.13 demonstrates, how the effect of our counter-factual improvement of cognition scores at baseline resulted in an expected significant shift of diagnoses toward CN or MCI throughout the study. Hence, our simulation of a "perturbed" ADNI cohort underlines the validity of our BN model. Altogether this example underlines the validity of our BN models and demonstrates the possibility of qualitatively studying intervention effects in-silico.

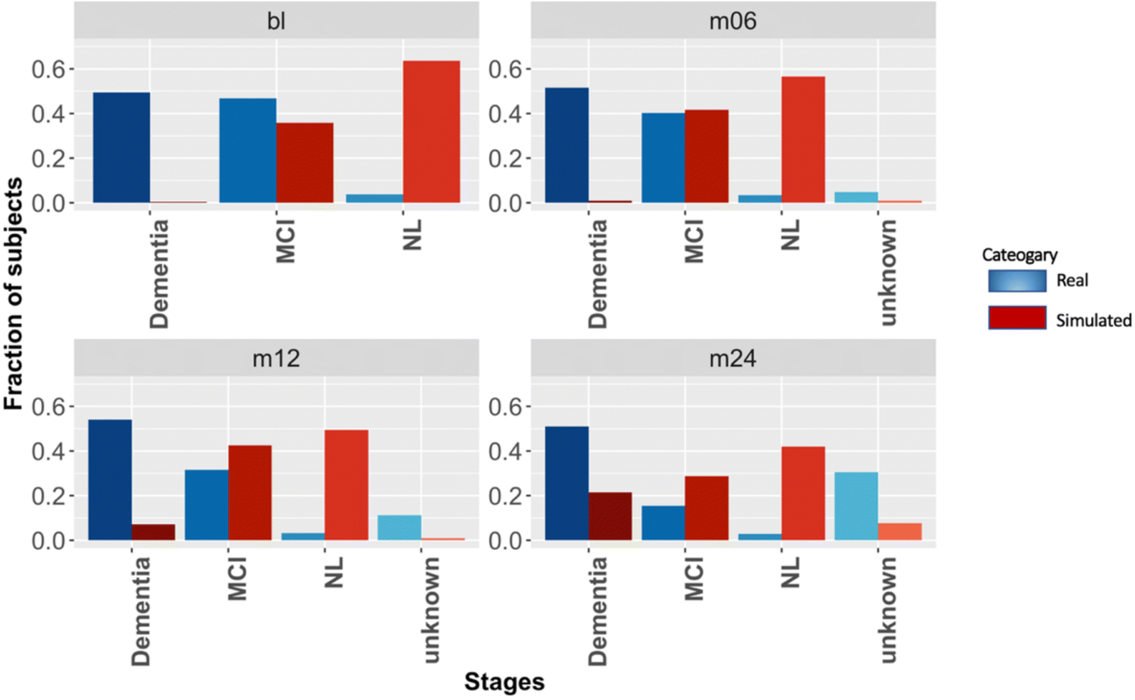


Figure 4.13: Simulation of a SC with an intervention: The figure shows diagnostic labels of 689 real ADNI subjects (red) and of a simulated cohort of the same size (blue) at different visits. The cognitive assessment scores of the simulated cohort have been shifted at baseline. MCI, NL (CN); unknown, unknown diagnosis/diagnosis not reported.

4.5 Conclusion

This work demonstrates an application of a realistic simulation of synthetic clinical subject trajectories across multiple biological scales and data modalities outside the area of mechanistically well-understood biological processes. This was achieved via a combination of deep learning techniques (sparse autoencoders) to significantly reduce the input dimensionality of our data and BN learning. We also showed that our SC approach allows for simulating interventions and studying their downstream effects in a qualitative manner in-silico. Such an approach could help the design of future clinical studies because it allows for assessing, which variables or variable groups are more likely to show differences after a planned intervention (e.g. with a drug).

Our proposed approach is not without limitations: MBN structure learning requires defining variable groups and constraints on the network structure, which implies a detailed understanding of the data. Moreover, we typically need to discretize input data to account for non-linearities between input variables while making BN structure and parameter learning at the same time computationally efficient. BN structure and parameter learning require sufficiently large datasets that are representative of the disease population. In addition, our method uses a sparse autoencoder-based aggregation of input features into variable groups, which naturally implies computational costs and a certain loss of information. Drawings of synthetic patients from the MBN model thus make the re-identification of real patients from the training data relatively unlikely. However, in its current implementation, our approach does not provide strict theoretical guarantees for this situation. But, we would like to point out that privacy-preserving training of neural network models is possible in principle [172] and is discussed in chapter 5. Training all sparse autoencoders via the modified stochastic gradient descent algorithm proposed by Abadi et al. provides theoretical guarantees for the data privacy of our entire MBN model. In future work, we will explore this aspect further and make an according implementation available. Altogether, the work presented here can only be seen as an extension to the proof of concept for the idea of simulating realistic multi-scale, multi-modal SCs, and further methodological advancements are necessary.

Failure will never overtake me if my determination to succeed is strong enough.

Dr. APJ Abdul Kalam

5

Variational Autoencoder Modular Bayesian Network

This chapter is adapted from our work in "Gootjes-Dreesbach, L., Sood, M., Sahay, A., Hofmann-Apitius, M., Fröhlich, H. (2020). Variational autoencoder modular Bayesian networks for simulation of heterogeneous clinical study data. Frontiers in Big Data, 3, 16."

5.1 Motivation behind the developed approach

This method is an extension of the previously developed method with MBNs and sparse autoencoders. We have developed this method as an improvement on the previous method and to serve the purpose of sharing patient data and dealing with the challenges of data privacy. Our focus here is on data-driven, model-based simulations of synthetic patients across biological scales and modalities (e.g., clinical, imaging) where no or little mechanistic understanding is available and required. Our novel proposed method [VAMBN] is a combination of a BN [173] with modular architecture and a VAE [82] encoding defined groups of features in the data. This approach also allows for generating synthetic patients under certain theoretical guarantees for data privacy [174]. VAEs have recently been extended to deal with heterogeneous multimodal and missing data [175], which is a common situation in clinical studies. VAEs are generative because drawings from the latent distribution can be decoded again. Our suggested approach aims to combine the advantages of BNs and VAEs while mitigating their limitations (Figure 5.1). The key differences to the previous approach are that here we used regularized version of classical autoencoders where i) No discretization is required and ii) it is possible to model arbitrary statistical distributions of heterogeneous data types within each module. In consequence, synthetically generated subjects now lie on the same numerical scale as original patient data.

Following the idea of module networks [176, 121], we first define modules of variables that group together according to the design of the study. For example, demographic features, clinical assessment scores, medical history, and treatment might each form such a module. This means that we assume the grouping of variables into modules to be known and defined upfront. Our aim is then to learn an MBN between variables in these modules. Because of its generative property, we use HI-VAEs, rather than regression trees to represent conditional joint distributions of variables within each module [175]. Each HI-VAE is thus only trained on a small subset of variables, hence significantly reducing the number of network weights compared to a full HI-VAE model for the entire dataset and allowing for applying the well-established "do" calculus for simulating interventions [95]. VAMBN also allows for simulating synthetic subjects by first drawing a sample from the BN and second by decoding it through the VAE representing the corresponding module.

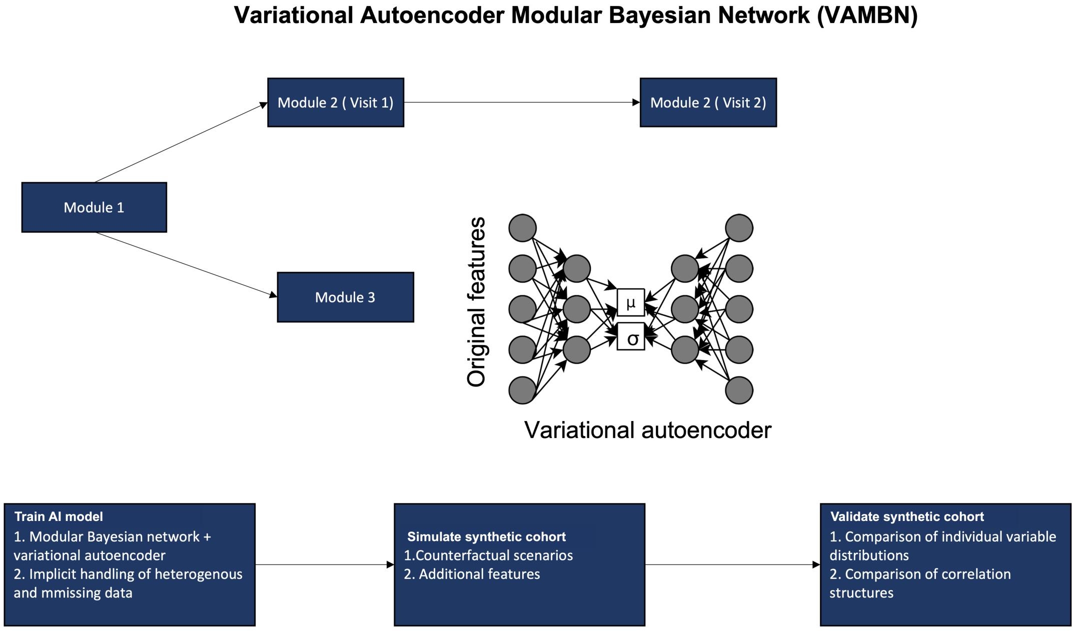

Variational Autoencoder Modular Bayesian Network (VAMBN)
Module 1
Module 2 (Visit 1)
Module 2 (Visit 2)
Module 3
Original features
μ
σ
Variational autoencoder
Train AI model
1. Modular Bayesian network + variational autoencoder
2. Implicit handling of heterogenous and mmissing data
Simulate synthetic cohort
1.Counterfactual scenarios
2. Additional features
Validate synthetic cohort
1. Comparison of individual variable distributions
2. Comparison of correlation structures

Figure 5.1: Conceptual overview about VAMBN approach: In the first step, a low-dimensional representation of known modules of variables is learned via HI-VAEs. The same group of variables (e.g., module 2) may have been assessed at different visits of the study, for example, visits 1 and 2. Accordingly, we get a low-dimensional representation of module 2 at visit 1 and module 2 at visit 2. In a second step, a BN is learned between low dimensional representations of modules, such that the temporal ordering of visits is considered and further constraints explained later are fulfilled. We call the resulting structure an MBN. The MBN explicitly models missing data at specific visits. Synthetic patients can be generated by sampling from the MBN and the HI-VAEs for each module. The "do" calculus allows for the simulation of counterfactual interventions into synthetic cohorts, such as adding features from another dataset. We carefully validate synthetic cohorts by comparison against real patients.

5.2 Methodology

5.2.1 Overview of the datasets used

PPMI

Here, we used data from 362 de novo PD patients and 198 healthy controls. All PD patients were initially untreated and diagnosed with the disease for 2 years or less. They showed signs of resting tremors, bradykinesia, and rigidity. We used 266 clinical variables measured at 11 visits during 96 months comprising demographics, patient PD history, dopamine transporter scan (DaTscan) imaging, non-motor symptoms, CSF biomarkers (Abeta, α-synuclein, dopamine, phospho tau (P-Tau), total tau (T tau)), and UPDRS scores.

ADNI

We used 689 subjects, 26 healthy, 321 MCI, and 336 Dementia, with 1 subject converting from MCI to Dementia subjects at baseline. We used 250 variables, the same set that was also considered in chapter 4.

5.2.2 Heterogeneous incomplete variational autoencoders

As described in Chapter 2, VAEs were originally developed for homogeneous data without missing values. However, clinical data within one and the same module (e.g., demographics) could contain continuous as well as discrete features of various distributions and numerical ranges, i.e., the data are highly heterogeneous. Moreover, there could be missing values. Nazabal et al. [175] extended VAEs to address this limitation. Their HI-VAE approach starts from a factorization of the VAE decoder according to:

$$p(x,z) = p(z) \prod_{j} p\ (X_j|z) \tag{5.1}$$

where $x \in \mathbb{R}_D$ denotes a D-dimensional data vector, and $z \in \mathbb{R}_K$ is its K-dimensional latent representation. Furthermore, x_j indicates the jth feature in x. In the factorization, it is further possible to separate observed (O) from missing features (M):

$$p(x|z) = \prod_{j\in O} p\ (X_j|z) \prod_{j\in M} p\ (X_j|z) \tag{5.2}$$

A similar separation is possible in the decoder step. Accordingly, VAE network weights can be optimized by solely considering observed data (input dropout model). The input dropout model is essentially identical to the approach we described earlier for MBNs.

To account for heterogeneous data types, Nazabal et al. suggest to set:

$$p(x_j|z) = p\ (x_j|\gamma_j\ = h_j\ (z)) \tag{5.3}$$

where $h_j()$ is a function learned by the neural network, and γ_j accordingly models data modality specific parameters (e.g., for real-valued data $\gamma_j =\ (\mu_j(z),\ \sigma^2(z))$. Moreover, the authors use batch normalization to account for differences in numerical ranges between different data modalities. Finally, Nazabal et al. do not use a single Gaussian distribution as a prior for z, but a mixture of Gaussians, i.e.:

$$s \sim Categorical(\pi) \tag{5.4}$$

$$z|s \sim N(\mu(s), I_K) \tag{5.5}$$

where s is K-dimensional.

Nazabal et al. [175] extended VAEs to address the situation of homogeneity by developing the HI-VAE approach and we refer to it for more details about their VAE extension.

Importantly, categorical variables s are added to the MBN graph G as parents of variables encoding modules. In practice, we kept K at 1 for all modules, resulting in a single normal distribution for z, with the exception of the demographic data in both studies. For these modules, K was set to 2. This choice was made after visual inspection of the embeddings for each of the individual variable groups, indicating that for modules containing demographic data and neurological examination, $K =$ 2 was the minimal value for which a sufficient fit to the data was possible. This was likely due to the existence of many categorical features among these variables.

5.2.3 Low dimensional representation of a module

Here, we assume that each low-dimensional representation of a module is the result of a HI-VAE encoding. We identify low dimensional representations with random variables $X = (X_v)\ , v \in V$ indexed by nodes in a DAG, $G = (V, E)$. This means that there is a DAG between low-dimensional representations of modules (MBN). In our case, random variables represented by nodes, either follow a Gaussian distribution (we explain the reasons later), or they could be of categorical nature, i.e., follow a multinomial distribution and not be autoencoded. We impose a restriction at this point, that a discrete node cannot be the child of a Gaussian one. Under this assumption, the conditional log-likelihood of the training data $D = \{x_{vi} | i = 1, ..., N, v \in V\}$ given G can be calculated analytically [177]:

$$log\, p\ (D|G) = \sum_{v \in V} log\, p\ (X_v | X_{pa(v)}) \tag{5.6}$$

$$log\, p\ (X_v | X_{pa(v)}) = \sum_{c \in C} l_c\ (Y_c) \tag{5.7}$$

$$l_c\ (Y_c) = \frac{n_c}{2}\ (log\, |\sum_c | + k\ log\, 2\pi + 1) + n_c\ log\, \frac{n_c}{N} \tag{5.8}$$

where C is the set of possible partitionings of Gaussian variable X_v according to the configuration of its discrete parents, and n_c is the number of patients in partition c. Note that modeling a Gaussian distribution conditional on discrete parents corresponds to a local analysis of variance (ANOVA) model. The associated design matrix is denoted as Y_c, and k is the number of columns of that matrix. $\sum_c$ is the covariance matrix. In a similar way, the local log-likelihood for a discrete node X_v with only discrete parents can be computed. We refer to [177] for more details. By considering, in addition, the number of parameters of the MBN, we can use the BIC to score G with respect to data D.

5.2.4 Modeling missing data in MBNs

As described before, the pattern occurring mostly in longitudinal observational cohorts is MNAR and we make use of auxiliary variables to handle such kind of data. The details are described in Chapter 2. In our MBN framework, auxiliary variables are fixed parents of all nodes, which contain missing values following an MNAR pattern. We also define higher-level missing data nodes that show whether a participant does not have any data for the entire visit. If the auxiliary variable of a node representing an autoencoded variable group is identical to the missing visit node, the auxiliary variable itself is removed from the network and the node is directly connected to the missing visit node instead. These higher-level nodes account for the high correlation between the different auxiliary nodes at a visit. Note that to facilitate modeling in the MBN, auxiliary and missing visit nodes were only introduced for nodes and visits with more than 5 missing data points in total.

5.2.5 MBN Structure learning

Most edges in the MBN structure are not known and hence need to be deduced from data. As discussed earlier that MBN structure learning is NP-hard, the search space of possible network structures should a-priori be restricted as much as possible. We follow two essential strategies for this purpose:

1. We group variables in the raw data into autoencoded modules, as explained above.
2. We impose causal constraints on possible edges between modules.

More specifically, we imposed the following type of constraints that are similar to those imposed in Chapter 4:

- Modules of demographic and other clinical baseline features (e.g., age, gender, ethnicity) can only have outgoing edges.
- Modules representing medical history can only depend on the modules mentioned in 1 and biomarkers.
- Modules of imaging features can be related to each other, but they do not influence other modules.
- Modules of clinical outcome measures (e.g., UPDRS) can influence imaging, and they can be mutually correlated with an assessment of non-motor symptoms.
- Biomarker modules can influence all modules, except for modules of clinical baseline features.
- Longitudinal measures must follow the right temporal order, i.e., there are no edges pointing backward in time.

- Auxiliary and missing visit nodes were connected to their respective counterparts at the next time point, accounting for a correlation between these measures over time, e.g., through study dropout.

Accordingly, we blacklisted possible edges that could violate any of these constraints. We learned the network on six different algorithms from R-package bnlearn [113]; 'hc', MMHC, tabu search, MMPC, RSAMAX2, and SI-HITON-PC(described in detail in Chapter 2). Tabu search was found to be the best learning algorithm [101]. In addition, it should be noted that due to the typically small number, variables in the MBN runtime were not a major concern here.

5.2.6 MBN Parameter learning

Given a graph structure G of a MBN parameters (i.e., conditional probability tables and conditional densities) were estimated via maximum likelihood.

5.2.7 VAMBN: Bringing MBNs and HI-VAEs together

Let $v \in V$ be a node in our MBN and X_v the corresponding random variable. Note that X_v is a low dimensional embedding/encoding of certain variables in the original input space, A_v. The total likelihood $p(X, A|G, \Theta)$ given graph G and model parameters Θ can be written as:

$$p\ (X, A|G, \Theta) = \prod_{v \in V} p(X_v|pa(X_v), \Theta_v) p(A_v|X_v, \Theta_v) \tag{5.9}$$

where $p(A_v|X_v, \Theta_v)$ is the generative model of the data represented by HI-VAE (it is the decoder distribution). Moreover, $pa(X_v)$ denotes all module nodes plus (in our case, one-dimensional) categorical δ variables. Hence, $p(X_v|pa(X_v), \Theta_v)$ is a normal distribution with mean

$$m_v = \Theta_v^{(0)} + \sum_{p \in pa(X_v)} \Theta_v^{(p)} p \tag{5.10}$$

[i.e., modeled via a linear regression with intercept $\Theta_v(0)$ and slope coefficients

$\Theta_v^{(\rho)}]$, and residual variance $v_v = Var(X_v\text{-}m_v)$. Our aim is to find parameters Θ maximizing $logp(X, A|G, \Theta)$. Using the factorization of this quantity and the typical assumption of node-wise statistical independence of parameters [67], we can optimize the total log-likelihood by the following two steps:

1. For all $v \in V : \widehat{\Theta}_v^* = argmax\ logp\ (A_v|X_v, \widehat{\Theta}_v)$. This is achieved via training a HI-VAE model for each module X_v, i.e., optimizing associated network weights $\widehat{\Theta}_v$.
2. For all $v \in V : \widetilde{\Theta}_v^* = argmax\ logp\ (X_v|pa(X_v), \widetilde{\Theta}_v)$. This is achieved by learning the MBN structure G and associated parameters $\widetilde{\Theta}_v$ based on HI-VAE-encoded modules.

Overall, the training of the proposed VAMBN approach thus consists of the following steps:

1. Definition of modules of a variable.
2. Training of HI-VAEs for each module. In practice, the training procedure included a hyperparameter optimization over.

a) Learning rate $\in 0.01, 0.001$

b) Minibatch size $\in 16, 32$

c) Each candidate parameter set was evaluated via a 3-fold cross-validation using the reconstruction loss as an objective function.

3. Definition of constraints for possible edges in the MBN.
4. Structure and parameter learning of the MBN using encoded values for each module: Note that by construction; our model for each variable, X_v follows a mixture of Gaussian distributions. Let $s \sim Categorical(\pi)$ indicate the mixture component. Hence, $X_v|s$ is Gaussian. Introducing s into the MBN thus yields a network with only Gaussian and discrete nodes, and parameter and structure learning can accordingly be performed computationally efficiently, as explained before.

We also considered using $N(m_v, \nu_v)$ as a prior for X_v instead of the original

Gaussian mixture prior to training of HI-VAE models in a second iteration of the entire VAMBN training procedure. In reality, we could not observe a significant increase in the total model likelihood $p(X, A|G, \Theta)$ due to this computationally more costly procedure. The plots illustrated in Figures 5.2 and 5.3 contrast the log-likelihoods of real patients after the initial/base training of VAMBN (red) and one further iteration of the entire VAMBN training (consisting of continued training of all HI-VAE models with a modified prior and re-estimation of the MBN, blue). Log-likelihoods shown for HI-VAE models are averaged over all modules ADNI and PPMI.

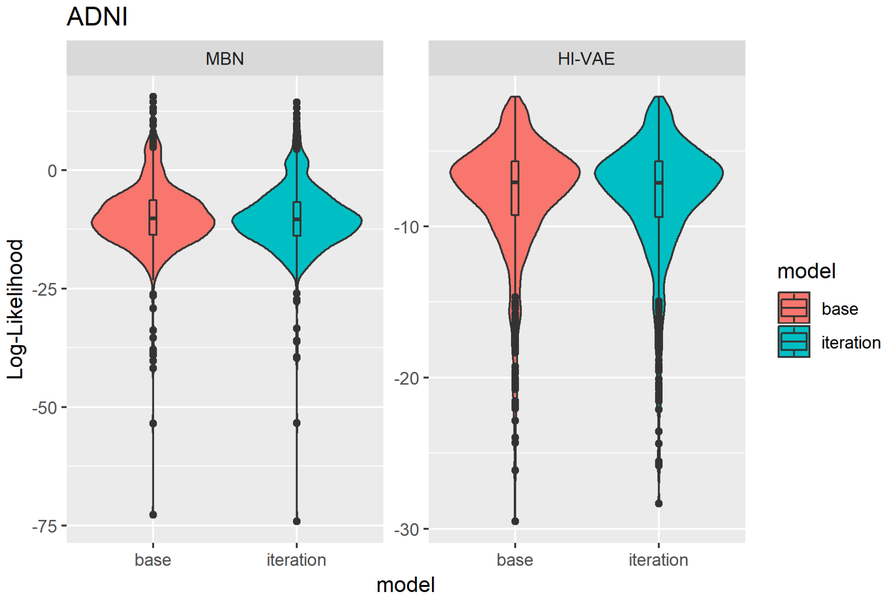


Figure 5.2: Log-likelihoods shown for HI-VAE models are averaged over all modules of PPMI

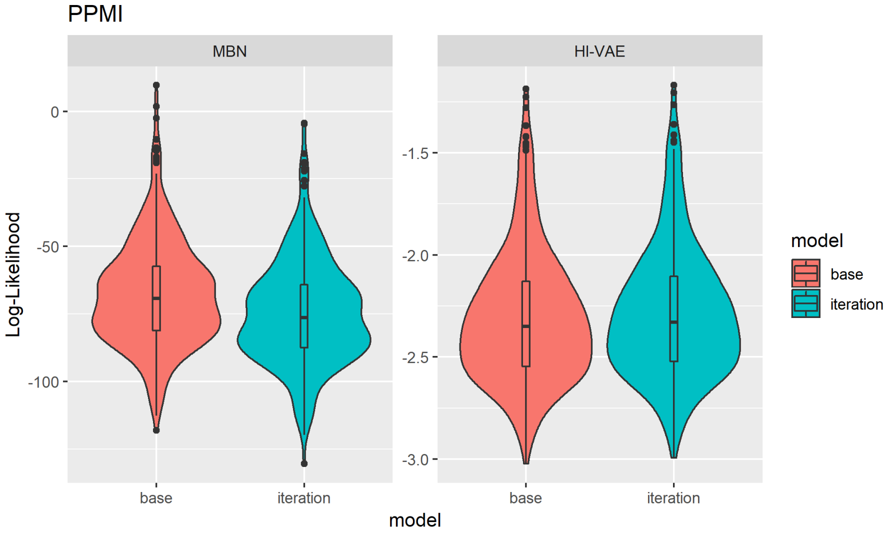


Figure 5.3: Log-likelihoods shown for HI-VAE models are averaged over all modules of ADNI

5.2.8 Simulating SC

Samples were simulated from MBN by following steps:

1. Draw samples from the MBN.
2. Decode MBN samples through HI-VAE. Note that a sample drawn from the MBN represents a vector of latent codes. Decoding maps these codes back into the original input space.

5.2.9 DP respecting model training

One of our motivations for developing VAMBN was to strengthen the mechanism for sharing data across organizations that addresses data privacy concerns. Practically, this could be achieved by sharing either simulated datasets or ready-trained VAMBN models. However, specifically in the latter case, there is the concern that by systematically feeding inputs and observing corresponding model outputs, it might be possible to re-identify patients that were used to train VAMBN models. This is particularly true for HI-VAEs, which encode groups of raw features. To deal with this challenge, we applied the concept of DP. DP is a concept developed in cryptography that poses guarantees on the probability to compromise a person's privacy by a release of aggregate statistics from a dataset [178, 174]: Let A be a randomized algorithm and $0 < \varepsilon$, $0 < \delta < 1$. According to Dwork et al. [174] A: D→R is said to respect (ε, δ) differential privacy, if for any two datasets D1, $D2 \in D$ that differ only in one single patient and for any output of the randomized algorithm S, we have

$$Pr(A(D_1) \in S) \leq e^{\varepsilon} \, Pr(A(D_2) \in S) + \delta \tag{5.11}$$

Abadi et al. [172] showed that it is possible to directly incorporate (ε, δ) differential privacy guarantees into the training of a neural network by clipping the norm of the gradient and adding a defined amount of noise to it.

It is straightforward to incorporate this approach into the training of each VAE. Hence, it helps us to provide guarantees on (ε, δ) differential privacy for the entire VAE because (ε, δ) differential privacy is composable. This means that the property for a system of several components is fulfilled if all of its components fulfill (ε, δ) differential privacy [174].

5.3 Results

5.3.1 VAMBN reflects expected causal relationships in PPMI and ADNI data

As outlined previously, our proposed VAMBN approach results in an MBN that describes conditional statistical dependencies between groups of variables that are encoded via HI-VAEs. The next question is to find how statistically stable the expected causal relationships are. To address this point, we performed a similar non-parametric bootstrap of the MBN structure learning as described in chapter 3 and chapter 4 [179]. We overlayed this bootstrapped network with the MBN learned from the complete data to get an overall impression of the learned VAMBN model as well as the stability of inferred conditional statistical dependencies. Figure 5.4 and Figure 5.5 highlight that, in both ADNI and PPMI, inferred edges agree well with expected causal dependencies. In ADNI (Figure 5.4), the cognitive tests are connected at different time points and they are also connected to the brain volumes. Brain volumes are also further connected at different time points. Age and gender are factors that also affect brain volumes [180, 181]. In PPMI (Figure 5.5), the RBD sleepiness score and non-motor symptoms mutually influence each other, and the same holds true for UPDRS. UPDRS is dependent on age, medical history, and α-synuclein levels in CSF.

Altogether, these examples underline that VAMBN models permit a certain level of interpretation.

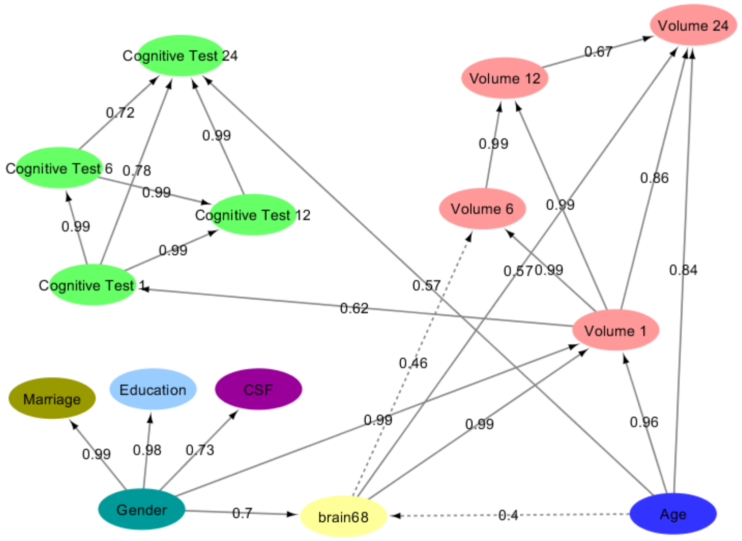


Figure 5.4: Final MBNs learned by VAMBN based on ADNI data. The edges are labeled with the bootstrap frequencies of each connection. For readability, auxiliary variables and missing visit nodes were removed for the visualization.

5.3.2 Evaluation of SC

We validate synthetic patient cohorts by comparing them against original patients:

- Marginal distributions of individual variables.
- Correlation structures.
- Expected differences between patient subgroups, e.g., treated vs. placebo patients.

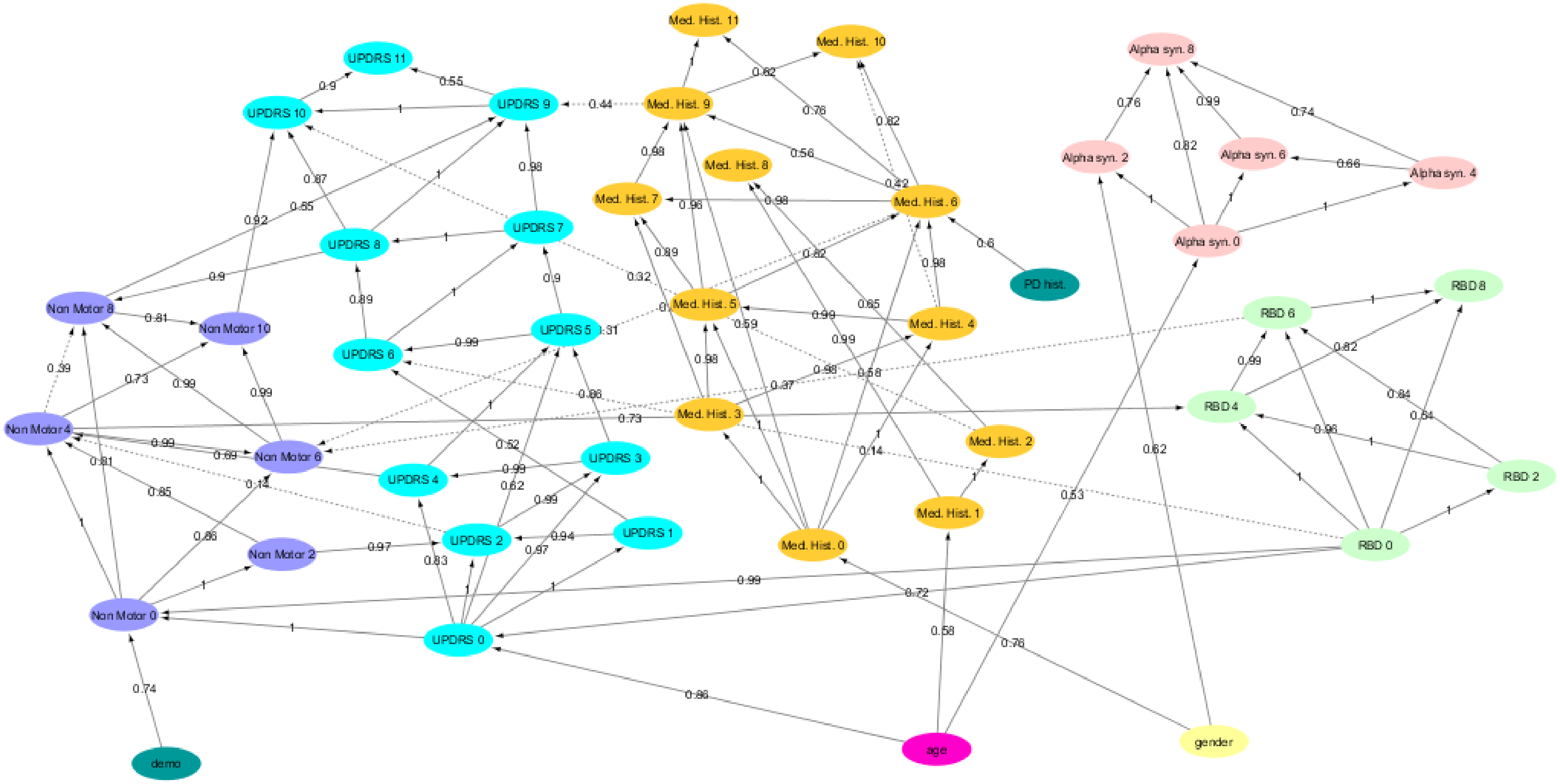


Figure 5.5: Final MBNs learned by VAMBN based on PPMI data. The edges are labeled with the bootstrap frequencies of each connection. For readability, auxiliary variables and missing visit nodes were removed for the visualization.

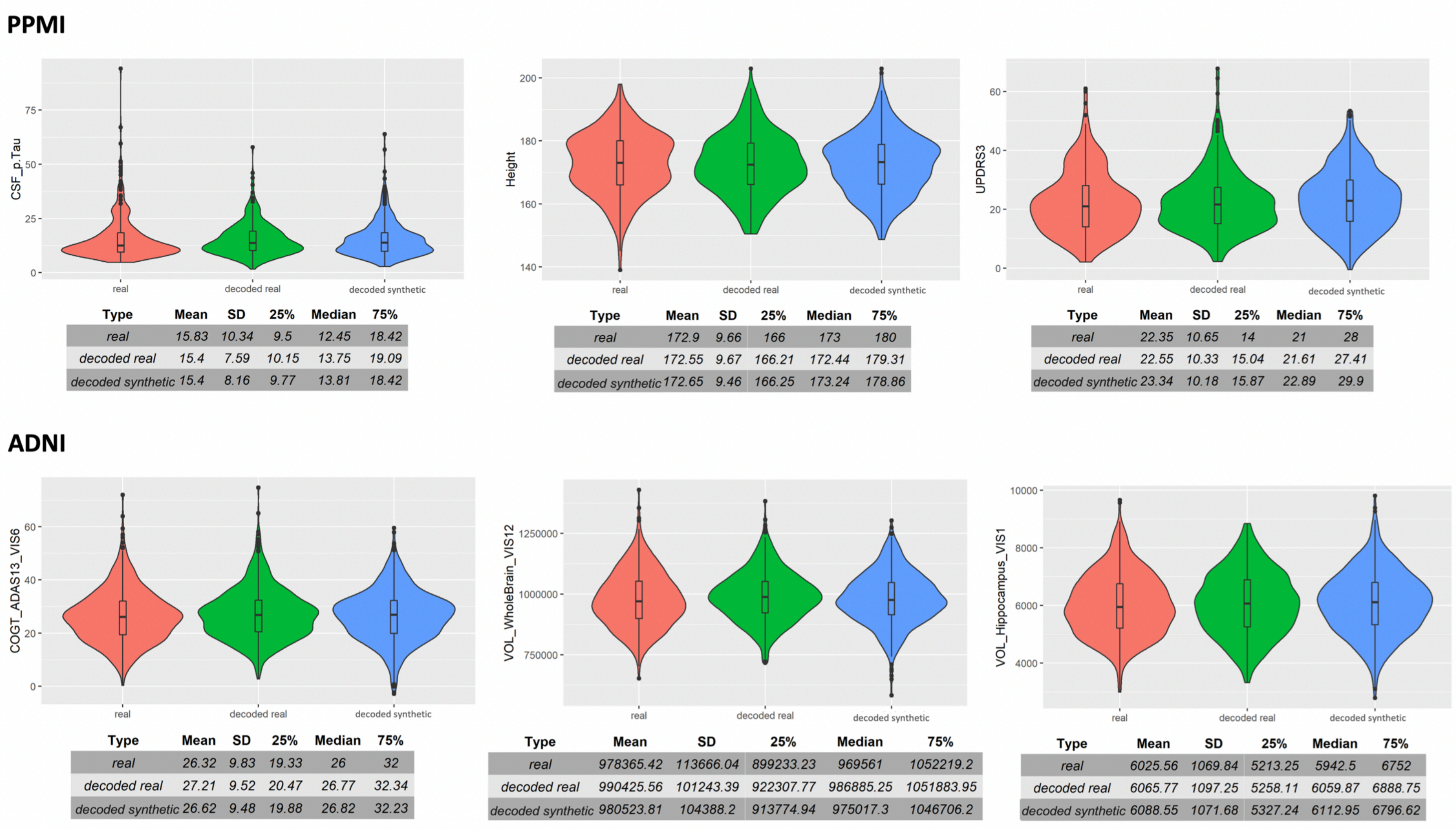
Simulated patient trajectories generated by VAMBN are only useful if they are sufficiently similar to real ones. On the other hand, we clearly do not want VAMBN to simply regenerate the data it was trained on (which would trivially maximize similarity to real patients). It is therefore not straightforward to come up with a criterion or interpretable index to measure the quality of a synthetic patient simulation.

From our point of view, simulated patients should mainly fulfill the following criteria:

- Summary statistics (e.g., mean, variance, median, lower quartile, upper quartile) over individual variables should look similar to real ones.
- Correlations between variables in simulated patients should be close to the ones observed in real ones.
- MBN structures learned from simulated patients should be close to the ones learned from real ones.
- Treatment effects or other expected outcomes should be similar in simulations, also in terms of effect size.

To assess VAMBN with respect to these criteria, similar to the validations described in the previous chapters, we simulated the same number of synthetic subjects as real ones in each study. Figure 5.6 demonstrates that marginal distributions for individual variables were sufficiently similar (but not identical) to the empirical distributions of real data in both studies.

In addition, the empirical distributions of Pearson correlations in simulated and real data were close to each other (Figure 5.7). Interestingly, in both cases (marginal distributions and correlations), the largest differences were observed between HI-VAE-decoded features of real patients and the original features of the same patients. Hence, the majority of the “simulation error” can be attributed to an imperfect fit of HI-VAE models.



Type	Mean	SD	25%	Median	75%
real	15.83	10.34	9.5	12.45	18.42
decoded real	15.4	7.59	10.15	13.75	19.09
decoded synthetic	15.4	8.16	9.77	13.81	18.42

Type	Mean	SD	25%	Median	75%
real	172.9	9.66	166	173	180
decoded real	172.55	9.67	166.21	172.44	179.31
decoded synthetic	172.65	9.46	166.25	173.24	178.86

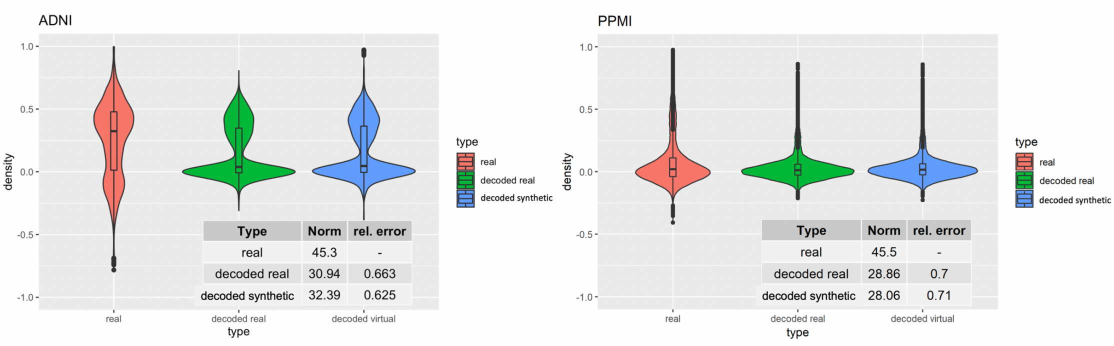
Type	Mean	SD	25%	Median	75%
real	22.35	10.65	14	21	28
decoded real	22.55	10.33	15.04	21.61	27.41
decoded synthetic	23.34	10.18	15.87	22.89	29.9

Type	Mean	SD	25%	Median	75%
real	26.32	9.83	19.33	26	32
decoded real	27.21	9.52	20.47	26.77	32.34
decoded synthetic	26.62	9.48	19.88	26.82	32.23

Type	Mean	SD	25%	Median	75%
real	978365.42	113666.04	899233.23	969561	1052219.2
decoded real	990425.56	101243.39	922307.77	986885.25	1051883.95
decoded synthetic	980523.81	104388.2	913774.94	975017.3	1046706.2

Type	Mean	SD	25%	Median	75%
real	6025.56	1069.84	5213.25	5942.5	6752
decoded real	6065.77	1097.25	5258.11	6059.87	6888.75
decoded synthetic	6088.55	1071.68	5327.24	6112.95	6796.62

Figure 5.6: Examples of real and simulated/synthetic patients for (A) PPMI and (B) ADNI datasets. The figure compares the marginal distributions of selected variables for real patients (red), synthetic patients (blue), and real patients decoded via the HI-VAE model (green). The tables show summary statistics of the distributions.

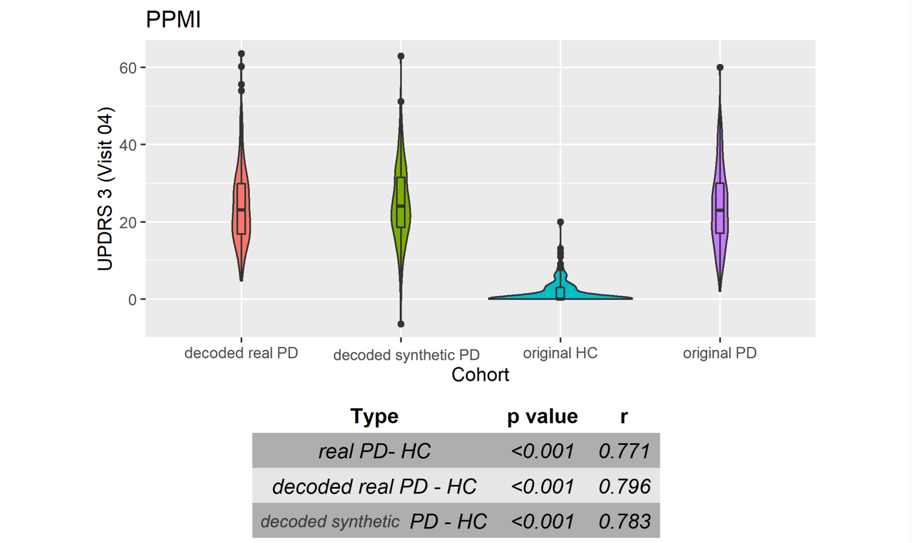


Type	Norm	rel. error
real	45.3	-
decoded real	30.94	0.663
decoded synthetic	32.39	0.625

Type	Norm	rel. error
real	45.5	-
decoded real	28.86	0.7
decoded synthetic	28.06	0.71

Figure 5.7: Distribution of Pearson correlation coefficients between variables in real patients (red), synthetic patients (blue), and decoded real patients (green). Tables show the Frobenius norm of the correlation matrices as well as the relative error, which consists of the norm of the matrix that is the difference between the decoded real or synthetic correlation matrix divided by the norm of the original correlation matrix.

As a final assessment of the quality of synthetic patients, we compared known patient subgroups in simulated and real data. Figure 5.8 demonstrates that, in PPMI, UPDRS-III scores of simulated PD patients showed similar differences to healthy controls than in real PD patients.



Type	p value	r
real PD- HC	*<0.001*	*0.771*
decoded real PD - HC	*<0.001*	*0.796*
decoded synthetic PD - HC	*<0.001*	*0.783*

Figure 5.8: Distribution of PPMI. Distribution of original (purple) and decoded (red) UPDRS-III scores of real PPMI de-novo PD patients at visit 4 in comparison to PPMI healthy controls (blue). MDS-UPDRS-III scores of synthetic PD patients are shown in yellow. The table at the bottom shows differences in MDS-UPDRS-III scores between original PD, decoded real PD, and synthetic PD patients compared to PPMI healthy controls, showing p-value and effect size from three Mann-Whitney U-tests.

Altogether, we thus concluded that VAMBN allows for a sufficiently realistic simulation of synthetic subjects with respect to our three defined criteria. At the same time, we could confirm that indeed none of the simulated patients were a simple regeneration of one of the patients in the training data.

5.3.3 Generalizability of VAMBN models

A relevant question is how generalizable VAMBN models are, i.e., whether they are purely overfitted or whether they can sufficiently describe data in an independent test set. To address this point, we randomly split data in PPMI and ADNI into 80% training and 20% test. VAMBN models were only fitted to the training set. We then recorded the log-likelihood of patients in the training and test sets, indicating a sufficiently good agreement (Figure 5.9). We thus concluded that VAMBN models are generally not overfitted. This means that the previously reported agreement of synthetic and real patients cannot just be the result of overfitting the data with an overly complex model.

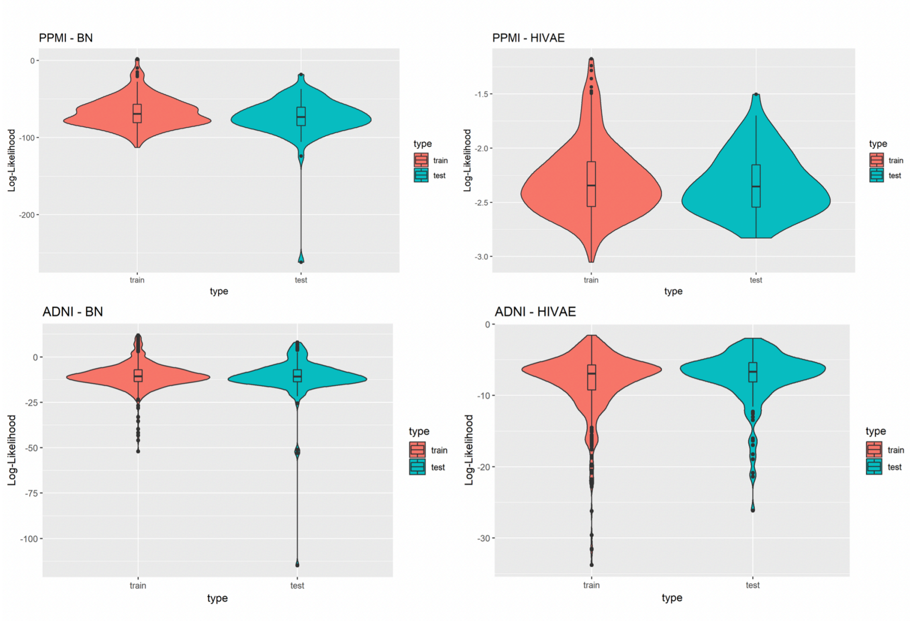


Figure 5.9: This figure compares the log-likelihoods of real patients in a training set (red) and a test set (blue) of PPMI (top row) and ADNI datasets (bottom row) for the MBN and the HI-VAE models. The HI-VAE log-likelihoods are based on the participants included in the respective sets after averaging across all separate HI-VAE models.

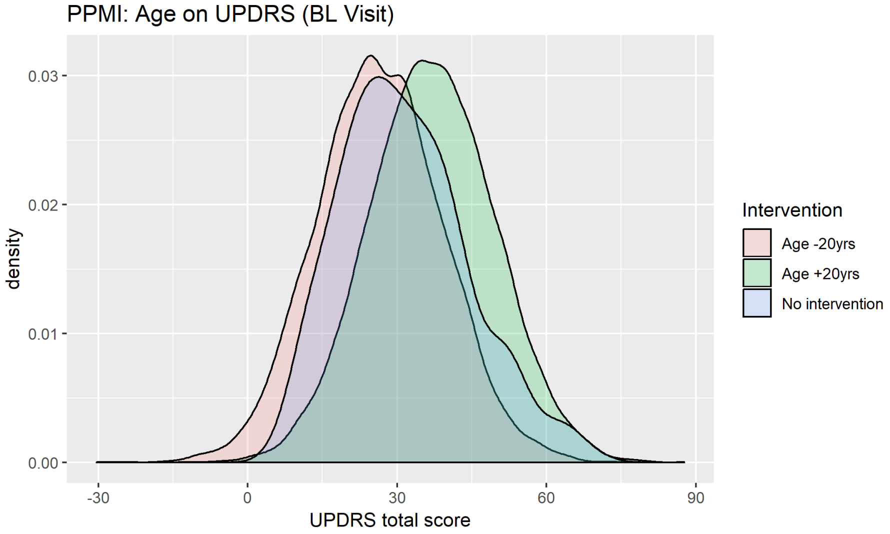


Figure 5.10: Counterfactual simulation of the shift of age 20 years before and after in PPMI.

5.3.4 Simulation of counterfactual scenarios match expectations

Due to its nature as a hybrid of a BN and a generative neural network, VAMBN allows for the simulation of counterfactual scenarios via the "do" calculus, as explained in Chapter 4.

In PPMI, making all patients 20 years younger shifts the distribution of UPDRS-III scores to the left (fewer motor symptoms), whereas making them 20 years older has the opposite effect (Figure 5.10). Again, this effect matches expectations.

These counterfactual simulations exemplify the possibilities of VAMBN and at the same time reconfirm that the model has learned the expected variable dependencies from data because the simulation effects match expectations.

5.3.5 Differential privacy respecting modeling training

As the last point, we investigated differential privacy respecting model training of VAMBN. As indicated in the Methods section, this can be realized by defining a certain privacy loss via constants (ε, δ) for each HI-VAE model trained within VAMBN. Smaller values for these constants generally impose stronger privacy guarantees but make model training harder. To investigate this effect more quantitatively, Figure 5.11 shows the reconstruction errors of the HI-VAE models for the ADNI data at the first visit for "cogtest" module as a function of the number of training epochs and in dependence on different values for ε, δ. It can be observed that in dependence on these constants, longer training, and more data are required to achieve the same level of reconstruction error than for conventional model training without differential privacy.

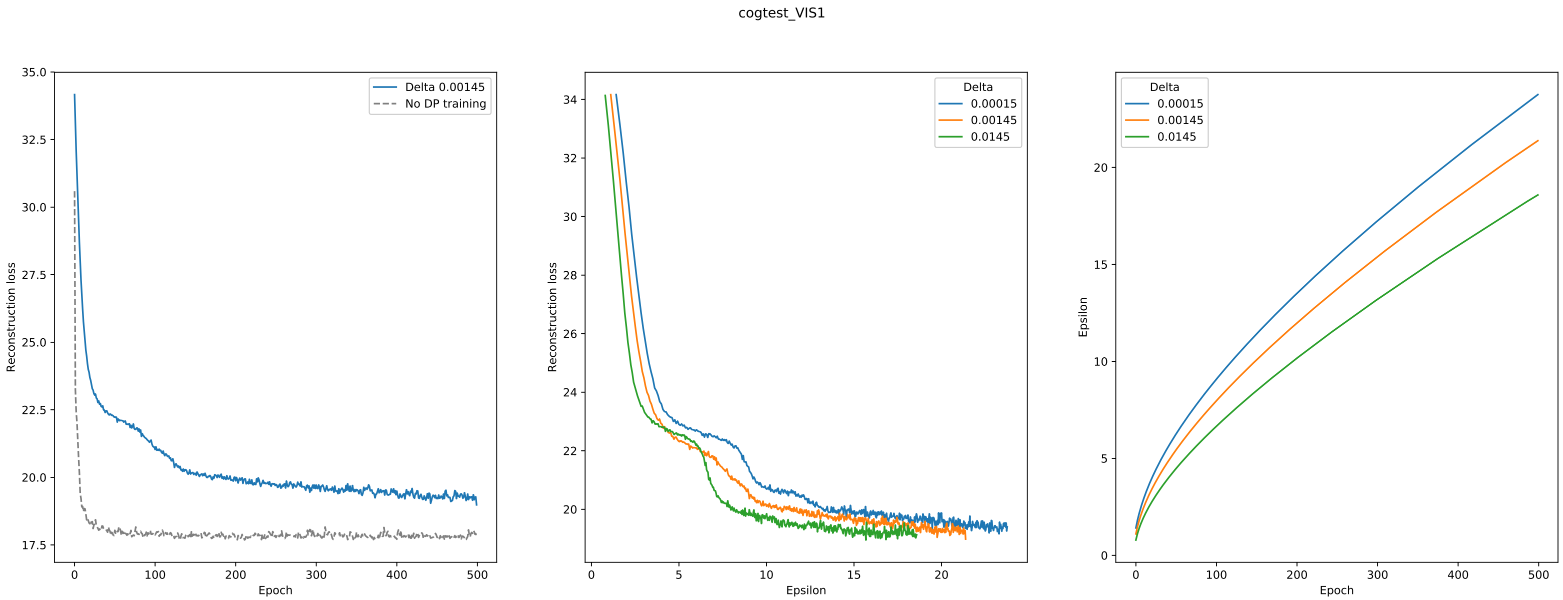


Figure 5.11: This figure shows the effects of DP respecting HI-VAE training on the HI-VAE step of the model. (Left) reconstruction loss change between DP and conventional model training for laboratory data at visit 1 for cogtest module at visit 1 ADNI study; (middle) epsilon plotted against reconstruction loss for different delta values; (right) epsilon over 500 epochs, given different deltas. A noise multiplier of 1.1, norm clipping at 1.6, and a learning rate of 0.01 were used.

5.4 Conclusion

Sensitive patient data requires high standards for protection as like that reinforced by the European Union through the GDPR (`https://eur-lex.europa.eu/eli/reg/2016/679/oj`). However, at the same time, these data are instrumental for biomedical research in the entire healthcare sector. Establishing a mechanism for sharing data across organizations without violating data privacy is therefore of utmost relevance for scientific progress. In this work, we build on the idea of developing generative models to simulate synthetic patients based on data from clinical studies. A recent publication proposed to train GANs based on a few variables recorded from more than 6,000 patients in the Systolic Blood Pressure Trial [75]. In contrast, our work focuses on the realistic situation regarding a much smaller sample size coupled with a significantly higher number of variables, which is common in many other medical fields, such as neurology. Our results demonstrate that VAMBN models generally do not overfit and allow for a sufficiently realistic simulation of synthetic patients. In contrast to GANs, our VAMBN method relies on explicit modeling of time dependencies, as well as missing and heterogeneous data. Moreover, VAMBN models can be interpreted via the MBN structure. In addition, we demonstrated that data privacy respecting model training is in principle possible with VAMBN.

From a user perspective, we see two important aspects for the successful application of our approach:

- A careful understanding of the data and its structure, including the ability to define variable groups.
- A careful check of the quality of synthetic data, using the approaches suggested in this paper.

Taken together, VAMBN is a new method for the simulation of synthetic cohorts for which we see a number of interesting future use cases in healthcare:

- Simulation of counterfactual scenarios to help the design of clinical trials.
- Privacy-preserving sharing of data across organizations to help data scientists understand the structure of sensitive patient data, judge their utility for

modeling purposes and derive statistical hypotheses that can be verified or falsified with available real data.

- Training of AI models that can subsequently be tested with available real data.
- Merging of different synthetic cohorts from the same indication area into a global synthetic meta-cohort based on overlapping variables. This global synthetic meta-cohort could be used to,
    - identify for a specific real patient within the overall distribution a best matching synthetic avatar.
    - efficiently generate control arms for clinical trials.

Of course, our work is not without limitations: Building VAMBN models require (in contrast to GANs) a relatively detailed understanding of data and careful handling of missing values in particular. Our examples have shown that VAMBN models can in practice already be learned from datasets with comparably small sample sizes and many variables. Nonetheless, our method, as with any AI-based approach, is principally dependent on sample size and signal-to-noise ratio in data. In the extreme case of more variables than samples (high dimensional setting), we expect VAMBN to become statistically unstable and overfit. From a technical side, VAMBN implies training multiple neural networks, which usually requires a modern parallel computing architecture. It thus remains a subject of future research to investigate how VAMBN models could be made better accessible to practitioners in order to facilitate their use in a widespread manner.

Overall, we see our work as a useful complement to federated machine learning techniques, which, together with synthetic patient simulation tools, could help to break data silos and thus enhance progress in biomedical research.

Be the change that you wish to see in the world.

Mahatma Gandhi

6 Generation of global meta-cohort for AD with data derived from digital devices

This chapter is an adaptation of our work in “Evaluating Digital Device Technology in Alzheimer’s Disease via Artificial Intelligence, Preprint medRxiv, 2021.”

6.1 Data derived from digital devices and sensor technologies

The medical data that are collected via digital devices like smartphones, wearable devices, and embedded environmental sensors can [182] provide an alternative path to disease assessment as they allow objective, ecologically valid and long term follow up with continuous estimation [183]. The evaluation of this aspect is the goal of innovative medicines initiative (IMI) project, remote assessment of disease and relapse-AD (RADAR-AD), project (`www.radar-ad.org`), and the work described in this chapter is a step towards this direction.

With the lack of a cure for AD, it has become essential to target the disease at an early stage [184]. A prerequisite for the differential diagnosis of disease stages of AD is a distinctive pattern of cognitive and functional impairment [185, 186]. Cognitive impairment is examined through questionnaire-based tests that assess multiple cognitive domains, including attention, memory, language, concentration, etc., e.g., the MMSE [187]. Functional impairment is measured by tests like the FAQ, which is marked by deterioration in activities of daily living (ADLs), such as the use of technology, managing finances, and shopping [188]. However, questionnaire-based assessments only provide a subjective snapshot of a patient's cognitive and functional abilities, which can vary over time. Hence, digital device technologies, including smartphone apps, currently receive increasing interest in the assessment of dementia symptomatology [189, 190]. These technologies can measure features of disease symptoms remotely and deliver real-time data to healthcare providers [191]. Digital Measues (DMs), for example, scores reached in an AR game, could allow for an accurate, quantitative, and objective monitoring of disease symptoms [192].

As an example of a panel of digital technology, in this work, we focused on a digital medical device built by Altoida, Inc. [193] that deploys a battery of immersive AR and motor activities over iOS, and Android smartphones and tablets, including activities which require the user to place and find objects in a virtual environment. The activities are designed to put the user under a cognitive load representative of what they experience when performing complex activity of daily living (ADL). The device generates several scores, including performance in individual tasks and derived overall neurocognitive domain scores (backed by the diagnostic and statistical manual of mental disorders, fifth edition (DSM-5) criteria) that can be used to identify cognitive impairment and neurocognitive disease.

DMs can be collected in two ways [182]:

- Active data collection: It requires the user to be actively involved in inputting the parameters being measured. e.g. digital e-assessment memory cognitive test that examines memory on a tablet to detect AD [194]. The measures derived via this approach are generally targeted at addressing specific metrics that are known to be associated with the disease.

- Passive data collection: In this approach, the measures are derived without active user engagement and it results in continuous real-time data acquisition. e.g. a smartwatch that is used as a step counter that continuously assesses the symmetry and length of the steps or a smart ring-based continuous monitor measuring heart rate variability (HRV). Therefore, active interaction with smart devices can lead to the generation of a high-frequency longitudinal data set that can be mined for signatures of the disease, while users have their own control.

However, the generation of DMs is an expensive process. There have been very few studies regarding this and before any clinical use, they have to be evaluated by assessing their relationship to established clinical scores and understanding their diagnostic benefit. There are very limited studies that focus on the measurement of clinical data and data generated from digital devices together for the same patients. This limits our understanding of data derived from digital devices as compared to clinical data. Considering this challenge, there arises a need to generate a synthetic meta-cohort where synthetic DMs can be added to the established observational cohorts for AD such as ADNI.

6.2 Need of a synthetic meta cohort with Digital Measures (DMs)

Here, we are introducing the concept of the meta cohort. Meta cohort consists of a number of cohorts that are considered as a single entity. As discussed above, we need to evaluate DMs by assessing their relationship with respect to the established questionnaire-based scores such as ADLs because both, DMs and questionnaire-based scores measure the cognitive impairment in the patient. This is a very important step to evaluate the sensitivity of DMs with respect to clinical outcomes. There are several studies that are focusing on measuring data for AD via digital devices but it often becomes difficult to measure the clinical outcomes for the same set of patients due to limited resources, inclusion and exclusion criteria, or due to budget problems. We can generate synthetic features in an already existing cohort that comprises clinical outcomes. This kind of cohort, known as meta cohort can bring all the features together, including DMs and clinical which in turn can assist in establishing and examining connections between these features. Moreover, this process can help us to establish significant DMs, as we can have their direct comparison with the clinical outcomes.

In this respect, RADAR-AD follows the ambition to evaluate a broad panel of digital technologies with respect to their potential for early AD diagnosis while focusing on ADLs.

The overall ambition of the work presented in this chapter was two-fold:

1. Generation of a meta-cohort by addition of DMs as synthetic features for ADNI patients.
2. Understanding the relationship between the DMs produced by the Altoida application and established tests and questionnaires (e.g., MMSE, FAQ/ADLs).

6.2.1 Overview about data from Altoida

Participants were diagnosed with different stages of the disease according to guidelines of the revised national institute on aging (NIA) and Alzheimer's association (AA) [195]. The smartphone application combines data from hand micro-movements$^{\text{TM}}$ and microerrors$^{\text{TM}}$, gait micro-errors$^{\text{TM}}$, visuospatial navigation micro-errors$^{\text{TM}}$, and recent voice parameters. One complete session consists of various motor function activities, a series of complex AR activities, and speech analysis [193].

The Altoida test: The Altoida test is a purely smartphone-sensor-based, digital biomarker-based prediction model, which only includes age, sex, and years of education to personalize neuromotor index (NMI) on an individual level. Using a tablet or smartphone device, a person is asked to perform a series of tasks ranging from simple motoric tasks to complex AR tasks. During these tasks the handheld device collects telemetry and touch data from the built-in sensors, enabling profiling of hand micromovements, screen touch pressures, walking speed, navigation trajectory, cognitive processing speed, and more. The motor activities consist of drawing activities and tapping activities. In the shape drawing activity, the subject is asked to draw various shapes (on the touch screen) using their index finger. In the tapping activity, the subject is first asked to tap a simple series of buttons (left, right) and then a similar series in which buttons are randomly highlighted. During these motor activities, eye tracking can be enabled to get more sensor data. In the AR activities, the subject is asked to walk around the room holding the device in their hands in front of them. On the screen, the environment is shown, augmented with digital objects. The subject is asked to virtually place three objects by clicking a “place” button on the screen while holding the device near a physical surface such as a table or desk. Afterward, the subject is asked to find these three objects by holding the tablet close to the location where they placed the objects. A speech activity can then be (optionally) performed in which the subject is asked to verbally describe an image. All activities above are performed twice in a specific order: motor, AR, speech, motor, AR, speech. For more details on the complete activity battery, please refer to Bugler et al. 2020 [193].

Data analysis for Altoida test: During an Altoida test, the hand-held device is recording data from various sensors. This raw sensor data is difficult to interpret and difficult to compare between patients. For the purpose of Alzheimer's prediction and cognitive domain scoring, the dimensionality of the data by means of feature extraction is reduced.

Recorded data: During an Altoida test (sensor) data from the following sources are recorded:

- ACC, accelerometer, measuring the acceleration in all three axes, relative to the device with the gravity component filtered out.
- ATT, attitude meter or, gyroscope, measuring the angle of rotation over all three axes.
- TOUCH, screen touches combined with an estimate of the applied pressure.
- PATH, the trajectory of the device through the physical space, as estimated by the tablet based on filtering the data of the accelerometer and gyroscope.

Feature extraction: During the Altoida activity batter, the handheld device records data from all available sensors for later analysis. Depending on the available sensors in the device, this data can include accelerometer data, gyroscope data, screen touches, speech audio recording, screen brightness, ambient brightness, eye tracking, and compass data. Thus, a single session of a single subject can lead to millions of data points over all sensors. To reduce the large dimensionality of the raw sensor data, various feature extraction techniques were applied depending on the performed test. For example, in the drawing tests, the participants were asked to draw a specific shape with their finger on the touch screen. Given the raw screen touches, various features such as drawing precision, drawing speed, number of breaks in touches, etc. were extracted. Likewise, during the AR test accelerometer and gyroscope data were collected which lend itself to a Fourier analysis. Thus, the obtained frequency magnitudes could again be used as a feature. The extracted features were subsequently used to assess the overall cognitive performance of a subject. The performance is presented as a set of scores over several cognitive domains.

Feature extraction is a data reduction technique aimed at describing the data as a set of non-redundant and informative metrics. The sensor data described in the previous section are what we consider “raw” data. Using feature extraction, interpretable features can be achieved from this raw data. For example, the data from the accelerometer is just a long list of measured accelerations in all three axes. By itself, this does not say much about the subject. By means of feature extraction, for example, micro-tremors from this data were extracted. These micro-tremors are far more concise and informative than the original raw data.

Sensor data generated from these tests are considered as “raw” and using feature extraction, we can generate DMs. These measures were subsequently used to assess the overall cognitive performance of a subject via several cognitive domain scores related to perceptual motor coordination, complex attention, cognitive processing speed, inhibition, flexibility, visual perception, planning, prospective memory, and spatial memory.

There are 9 derived scores for different cognitive domains explained in the following section:

Cognitive domains: The cognitive performance of a subject ranks the subject's test performance relative to his/her peers. The performance is presented as a set of scores over several cognitive domains. To compute, the recorded sensor data from an Altoida test is reduced into a set of characteristic elements, or data features. For each of these features, a percentile score is computed by comparing the participant's feature value to those of a group of healthy participants of the same age and sex. Per domain, a specific set of these percentile scores are combined to form each of the cognitive domain scores.

Altoida currently reports on the following cognitive domains:

- Perceptual Motor Coordination, motor coordination in response to perceived input.
- Complex Attention, capacity to choose what to pay attention to and what to ignore.
- Cognitive Processing Speed, speed, and accuracy of information processing.
- Inhibition, ability to overlook stimuli that are irrelevant to the task.
- Flexibility Ability, to transition between thinking about two different concepts.
- Visual Perception, visual search speed, visual perception, and efficiency.
- Planning, the process of thinking about the activities needed to achieve the desired goal.
- Prospective Memory, ability to remember to carry out intended actions in the future.
- Spatial Memory, ability to recognize items that previously appeared in physical space.

Altoida is currently working on new cognitive domains to augment the current set of domains. These prospective cognitive domains are:

- Speech and articulation, fluency of speech, and comprehension of visual information.
- Eye movement, eye focus, and hand-eye coordination.

6.3 Methodology

6.3.1 Overview of Altoida data used for this work

For our analysis, we examined 148 participants, out of which 123 were measured once, 20 were measured twice, and 5 were measured thrice, leading to a total of 178 subject records. In addition to DMs, for all 148 participants, traditional MMSE scores were available, which are often used in clinical routines. Table 6.1 provides an overview of data from Altoida that we use in our in terms of baseline summary statistics grouped by diagnostic stages.

	CN, N =58	**MCI, N = 29**	**MCI at Risk for AD, N = 33**	**Prodromal AD, N = 15**	**Dementia,N = 13**
Age	66 (60,72)	66 (61, 72)	72 (66, 77)	69 (64, 72)	73 (65, 77)
Gender					
Male	38 (66%)	15 (52%)	12 (36%)	5 (33%)	8 (62%)
Female	20 (34%)	14 (48%)	21 (64%)	10 (67%)	5 (38%)
Education (yrs)	13 (11, 16)	10 (8, 15)	12 (8, 15)	10 (8, 19)	9 (6, 12)
Amyloid Pos.	0 (0%)	0 (0%)	33 (100%)	15 (100%)	13 (100%)

Table 6.1: Summary statistics of Altoida demographic and amyloid data grouped by diagnostic stages, data are n (%), or median (IQR in brackets), unless specified otherwise

6.3.2 Overview of ADNI data

Table 6.2 provides the summary statistics for baseline ADNI data. Participants in ADNI are on average more educated and older than in Altoida. In ADNI a high fraction of patients within the MCI and demented groups are carriers of at least one APOE4 risk allele (MCI: 52%, dementia: 69%). We examined 1445 subjects, having longitudinal measurements at baseline, months 6, 18, 24, and 36.

	CN, N =440	**MCI, N = 735**	**Dementia, N = 270**
Age	74 (71, 78)	74 (68, 79)	75 (71, 80)
Gender			
Male	223 (51%)	442 (60%)	151 (56%)
Female	217 (49%)	293 (40%)	119 (44%)
Years of Education	16 (15, 18)	16 (14, 18)	16 (12.25, 18)
apoe4			
0	317 (72%)	353 (48%)	85 (31%)
1	113 (26%)	295 (40%)	132 (49%)
2	10 (2.3%)	87 (12%)	53 (20%)
Amyloid			
0	358 (81%)	507 (69%)	167 (62%)
1	82 (19%)	228 (31%)	103 (38%)

Table 6.2: Summary statistics of ADNI demographic data, amyloid and APOE4 status grouped by diagnostic stages, data are n (%), or median (IQR in brackets) unless specified otherwise

The data consisted of three different diagnostic stages:

1. CN. Cognitively normal population.
2. MCI. MCI has been defined in ADNI as follows [196]:1) subjective memory complaints reported by the study participant, study partner, or clinician; 2) memory loss according to an education-adjusted WMS-R Logical Memory Test; 3) global Clinical Dementia Rating (CDR) score of 0.5, and 4) general cognitive and functional performance sufficiently preserved such that a diagnosis of dementia could not be made.
3. AD. For an AD diagnosis, additional criteria according to the MMSE test and the National Institute of national institute of neurological and communicative disorders and stroke - alzheimer's disease and related disorders association (NINCDS/ADRDA) had to be fulfilled.

We considered FAQ and MMSE subitem scores as cognitive tests because i) they are also assessed in the Altoida data (MMSE), ii) reflect ADL (FAQ), and iii) generally have been suggested to reflect disease progression [197, 198, 199].

We also calculated the genetic burden scores of mechanisms involved in AD by encoding the SNPs at the molecular mechanism level. We followed the following steps to achieve this task. The information related to SNP data was downloaded from the ADNI server for different subsets of data:

- ADNI 1: 581,500 SNPs from 757 subjects, measured via Illumina Human610-Quad Bead Chip platform
- ADNI 2/GO: 708,870 SNPs from 432 subjects, measured via Illumina HumanOmniExpress
- ADNI 3: 16,743,712 SNPs from 327 subjects measured via Illumina Omni 2.5M

SNPs were imputed via the Michigan Imputation Server [200] using the haplotype reference consortium (HRC) reference panel, which consists of 64,976 haplotypes [201]. SNPs were considered as reliable imputed if the r^2 was above 0.3 (default setting).

Two major databases, phenome-wide association studies (PheWAS) Catalog [202] and DisGeNet [203] were used to gather AD-associated SNPs. SNPs collected from both these databases were further extended by those SNPs, which were strong in Linkage Disequilibrium ($r^2 > 0.8$) via HaploReg (Version 4.1) [204].

In order to map these SNPs to genes, two steps were performed:

- Mapping to genes in closest chromosomal location via HaploReg and using default settings.
- Via phenome-Wide association studies (eQTL) mapping using gene expression data from brain tissues obtained from the genotype-Tissue expression (GTEx) Portal (GTEx Consortium, 2013). Only cis-eQTL was taken into account.

Subsequently, the multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig) knowledge base [205] was used to find those genes that can be linked to AD-related biological mechanisms. As we required a minimal number of mechanisms, we selected 20 relevant/well-known mechanisms that also had a large number of genes in their corresponding network. Then, the corresponding SNPs were mapped to each of the mechanisms.

In addition to MMSE and the genetic burden scores [205], we used neuroimaging-derived features CSF protein measurements (amyloid-beta, tau, and phospho-tau). We determined the amyloid status of the patients based on the neuroimaging features, Florbetapir (AV45) amyloid positron emission tomography (PET) value of the patients. If patients had an AV45 value >1.11, they were amyloid positive and negative otherwise [206].

6.3.3 Feature description

The description of MMSE subitem scores and FAQ subitem scores are presented in Tables 6.3 and 6.4 respectively.

MMSE Subitem scores	**Description**
MMSE Attention Concentration	Clinical test used to assess mental function
MMSE Language	Tests related to naming a pencil and a watch, repeating words, and carrying out complex commands like drawing a figure measured
MMSE Memory Recall	Registration recall
MMSE Orientation	Testing orientation to time and place
MMSE Working Memory Registration	Testing related to repeating names prompts

Table 6.3: Description of item scores of MMSE in Altoida and ADNI

FAQ Subitem Scores	**Description**
FAQFORM	Assembling tax records, business affairs, or other papers, Partial score of FAQ
FAQBEVG	Heating water, making a cup of coffee, turning off the stove. Partial Score, FAQ
FAQGAME	Playing a game of skill such as bridge or chess, working on a hobby. Partial Score, FAQ
FAQFINAN	Writing checks, paying bills or balancing a checkbook. Partial Score, FAQ
FAQMEAL	Preparing a balanced meal, Partial score of FAQ
FAQTV	Paying attention to and understanding a TV program, book, or magazine, Partial score of FAQ
FAQREM	Remembering appointments, family occasions, holidays, medications, Partial score of FAQ
FAQSHOP	Shopping alone for clothes, household necessities, or groceries, Partial Score of FAQ
FAQTRAVL	Traveling out of the neighborhood, driving, or arranging to take public transportation, Partial score of FAQ
FAQEVENT	Keeping track of current events, Partial Score of FAQ

Table 6.4: Description of item scores of FAQ in ADNI

6.3.4 Motivation for generation of the meta-cohort and synthetic feature generation

As observed in Altoida data, it comprises DMs and clinical outcomes such as item scores of MMSE that were used to assess the overall cognitive performance of a subject. Despite both types of features being available for all the subjects, it lacks the features that reflect ADL and the number of subjects is also limited. However, in ADNI data, we have MMSE features as well as features ADL e.g. FAQ but it lacks DMs. This motivates us to generate a synthetic meta-cohort for ADNI data consisting of DMs as synthetic features.

6.3.5 Overview about analysis strategy

The overall strategy of our approach is outlined in Figure 6.1.

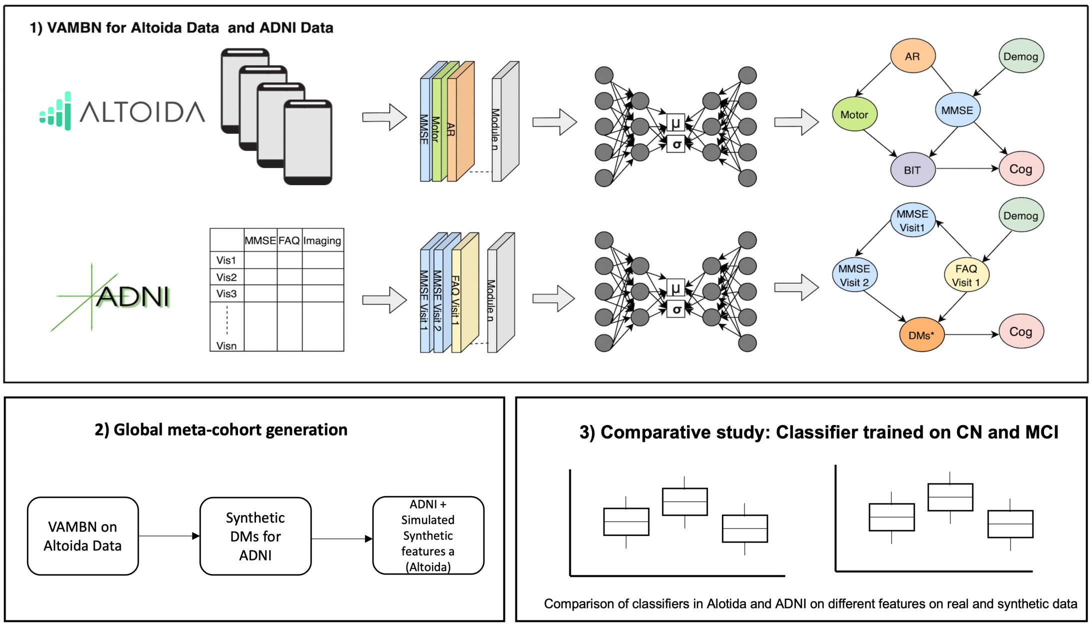

1) VAMBN for Altoida Data and ADNI Data
ALTOIDA
MMSE
Motor
AR
Module n
μ
σ
AR
Demog
Motor
MMSE
BIT
Cog
ADNI
MMSE
FAQ
Imaging
Vis1
Vis2
Vis3
Visn
MMSE Visit 1
MMSE Visit 2
FAQ Visit 1
Module n
MMSE Visit1
Demog
MMSE Visit 2
FAQ Visit 1
DMs*
Cog
2) Global meta-cohort generation
VAMBN on Altoida Data
Synthetic DMs for ADNI
ADNI + Simulated Synthetic features a (Altoida)
3) Comparative study: Classifier trained on CN and MCI
Comparison of classifiers in Alotida and ADNI on different features on real and synthetic data

Figure 6.1: Brief workflow of our methodology is described here in 3 steps. 1) Fitting the VAMBN model for Altoida and ADNI. 2) Generation of global meta-cohort 3) Testing the classifier on Altoida and ADNI on different features on real and synthetic data.

1) Fitting a VAMBN model

The Altoida app. results in 11 individual scores for different digital tasks in the synthetic environment described in Table 6.5. It describes the module names and their description for Altoida data. As the cognitive domains in Altoida measure very similar parameters compared to the MMSE subitem scores, there is a very large number of potential correlations between the Altoida DMs and the 5 different MMSE subitem scores (illustrated in Table 6.3). To address this issue, in this work we employed a VAMBN, which results in a quantitative and visualizable network structure. More specifically, we employed the VAMBN [158] approach which was described in Chapter 5.

Module Name	Description
AR Global Telemetry Variance	The variance in telemetry (accelerometer and gyroscope) over the entire duration of the AR test
AR Intro Read Times	The time the subject required to read the introduction to the AR test
AR Object Finding	Features pertaining to finding an object, such as time required to find the next object, distance traveled while searching, etc.
AR Object Placement	Features pertaining to placing an object, such as time taken to find a suitable surface, distance traveled, holding the device steady, etc.
AR Object Placement FFT	Fast Fourier Transform frequency spectrum analysis of a few seconds prior to placing an object. These are special enough to warrant their own group
AR Place and Find Telemetry Variance	The telemetry variance in the moments before placing and finding objects. In contrast to the global telemetry variance, this disregards the walking periods
AR Screen Button Presses	Touch screen data of button pressed during the AR test, such as pressure, and touch accuracy
Motor Drawing Features	Features pertaining to the finger drawing tests
Motor Tapping Features	Features pertaining to the finger tapping tests
Motor Test Durations	The total duration of each test part
MMSE Attention Concentration	Clinical test used to assess mental function
MMSE Language	Tests related to naming a pencil and a watch, repeating words, and carrying out complex commands like drawing a figure measured
MMSE Memory Recall	Registration recall
MMSE Orientation	Testing orientation to time and place
MMSE Working Memory Registration	Testing related to repeating names prompts

Table 6.5: Description of the individual scores and their module names for different tasks performed in Altoida app. and MMSE subitem scores in the synthetic environment

Fitting a VAMBN model consists of several steps:

- **Defining interpretable variable groups in both Altoida and ADNI:** DMs and motor scores were grouped into different modules. Demographic features (e.g., age, sex, etc.) were not grouped into a module as we wanted to see their individual effects on other features. Modules and their description for Altoida data are presented in Table 6.5 and for ADNI are presented in Table 6.6.

- **Learning of a low-dimensional representation of each module:** This was achieved by training a HI-VAE [207]. The result was a GMM encoding the higher dimensional input data.
- **Learning of an augmented BN connecting the modules:** The BN contained additional auxiliary variables to account for missing values in the data, e.g., due to patient drop-out. We repeated BN learning 1000 times using random subsamples of the data to derive a measure of statistical confidence [175].

The result of the above three steps was a quantitative network model representing patient-level data. In the following section, we will describe the details of the learning procedure.

The VAMBN workflow is elaborated in the following steps:

1. Definition of modules that summarizes the original input features (Tables 6.5 and 6.6)
2. Encoding of modules into lower dimensional latent distributions (multivariate Gaussian) via HI-VAE. The training procedure of HI-VAEs for each module included a hyperparameter optimization over the following parameters:

   a) Learning rate: 0.01, 0.001

   b) Mini batch size: 16, 32

   c) Weight Decay: 0, 0.001, 0.01

   The 3-fold cross-validated reconstruction loss was used as an objective function to evaluate each candidate parameter set.
3. Structure and Parameter Learning: We followed two essential strategies for this purpose:

Module Name	**Description**
csf_VIS1	Abeta, tau and ptau csf biomarkers at baseline
volume_VIS (1,6,12,24)	6 brain volumes, entorhinal, hippocampus, ventricles, whole brain, middle temporal, and fusiform at baseline, month 6, month 12, and month 24 (brain volumes were normalized with respect to intracranial brain volume (ICV))
Imaging_VIS(1)	Imaging measures such as fluorodeoxyglucose-positron emission tomography (FDG PET) and Alzbio3 kits and Florbetapir (AV45) amyloid PET at baseline
ADAM_Metallopeptidase_subgraph	”rs4575098” ”rs2277027” ”rs1422795” ”rs7174386” ”rs12906705” ”rs28455654” ”rs383902”
Amyloidogenic_subgraph	”rs17571” ”rs676134” ”rs2829946” ”rs2830088”
APOE_subgraph	”rs405509” ”rs439401”
Apoptosis_signaling_subgraph	”rs1136410” ”rs1805411” ”rs1469926” ”rs319724” ”rs827423” ”rs6902771” ”rs8006145” ”rs10144225” ”rs10137185” ”rs2667543”
ATP_binding_cassette_transport_subgraph	”rs1045642” ”rs6949448” ”rs2235046” ”rs1128503” ”rs10276036” ”rs1202169” ”rs1202168” ”rs1202167” ”rs1883023” ”rs2777802” ”rs3818689” ”rs2066715” ”rs2066718” ”rs4149308” ”rs2066717” ”rs4149303” ”rs4149301” ”rs2297399” ”rs2297400” ”rs3824479” ”rs2472384” ”rs2253304” ”rs2253182” ”rs2253175” ”rs2253174” ”rs2253172” ”rs2230806” ”rs2243313” ”rs2482420” ”rs2487058” ”rs2487059” ”rs2230805” ”rs3847300” ”rs3847303” ”rs3905000” ”rs2575876” ”rs12826” ”rs7067971” ”rs11190305” ”rs1283816” ”rs1283817” ”rs829079” ”rs1283822” ”rs3752229” ”rs3752232” ”rs3764650” ”rs3752240” ”rs3752242” ”rs2279796” ”rs4147932”
Axonal_guidance_subgraph	”rs1354269” ”rs12364788” ”rs17614100” ”rs7112354”
Caspase_subgraph	”rs2027432” ”rs10159239” ”rs12130711” ”rs1143634” ”rs2276575” ”rs13430599” ”rs10194375” ”rs13426725” ”rs17014923” ”rs6743470” ”rs4663098” ”rs7561528” ”rs744373”
Chemokine_signaling_subgraph	”rs1024611” ”rs991804”
Cholesterol_metabolism_subgraph	”rs5174” ”rs3737983” ”rs2297663” ”rs2297660” ”rs3820198” ”rs7551288” ”rs9371201” ”rs1799986” ”rs12435918”
GSK3_subgraph	”rs2873950” ”rs3108749” ”rs6438552”
Inflammatory_response_subgraph	”rs2243248” ”rs2243290” ”rs7748777” ”rs7759295” ”rs7072793” ”rs6074022” ”rs1569723” ”rs6032678”
Insulin_signal_transduction	“rs1999763”
Interferon_signaling_subgraph	”rs1554606” ”rs8038734”

Module Name	**Description**
Interleukin_signaling_subgraph	”rs4537545” ”rs4129267” ”rs4240872” ”rs7514452” ”rs1800896” ”rs4848300” ”rs17561””rs4848304” ”rs1143634” ”rs2243248” ”rs2243290” ”rs1554606” ”rs10975516” ”rs1330383””rs10815398” ”rs7072793” ”rs4072111” ”rs11857713” ”rs4778636” ”rs7197333”
Lipid_metabolism_subgraph	”rs2228467” ”rs6444175” ”rs11742194” ”rs3846662” ”rs5909” ”rs12435918” ”rs5882”
Matrix_metalloproteinase_subgraph	”rs10836653” ”rs4382897” ”rs17337649” ”rs645419” ”rs2241715”
Nerve_growth_factor_subgraph	“rs3775256”
Tau_protein_subgraph	”rs242557” ”rs3785883”
Wnt_signaling_subgraph	”rs2873950” ”rs3108749” ”rs6438552” ”rs29645” ”rs7901695” ”rs7903146”
MMSE Language_VIS (1,6,12, 24, 36)	Tests related to naming a pencil and a watch, repeating words, and carrying out complex commands like drawing a figure measured at baseline, months 6, 12, 24, and 36. Ranges between 0 and 5.
MMSE Memory Recall_VIS (1,6,12, 24,36)	Registration recall measured at baseline, month 6, 12, 24 and 36. Ranges between 0 and 5. Ranges between 0 and 3.
MMSE Orientation_VIS (1,6,12, 24, 36)	Testing orientation to time and place at baseline, month 6, 12, 24 and 36. Ranges between 0 and 10.
MMSE Working Memory Registration_VIS (1,6,12, 24, 36)	Testing related to repeating names prompts at baseline, month 6, 12, 24 and 36. Ranges between 0 and 3.
MMSE Attention Concentration_VIS (1,6,12, 24, 36)	Clinical test used to assess mental function at baseline, month 6, 12, 24 and 36. Ranges between 0 and 5.

Table 6.6: Modules and their description defined for ADNI data set

a) As explained above, variables in the raw data were grouped into autoencoded modules.

b) Causal constraints as prior knowledge was imposed while learning the augmented BN connected modules in order to restrict the search space and to allow correct causal orientation of as many edges as possible.

After structure learning, the parameters of the MBN were fitted. This was done via maximum likelihood. That means, for each Gaussian node a linear regression was fitted, in which the parents of the node were used as predictors. For each discrete node, parameters were determined via conditional probability tables.

During structure learning, the following causal constraints were imposed on the datasets:

Causal constraints for VAMBN training on Altoida Data

- Demographic features such as age and sex cannot be influenced by any other feature or by each other.
- Amyloid status of a subject cannot be affected by any other feature.
- Cognitive features and digital biomarkers cannot affect clinical diagnosis (it can be only affected by Amyloid).
- No other feature can affect education except age and gender.
- An auxiliary variable representing the missingness of a certain feature can only influence that particular feature at a later visit.

Causal constraints for VAMBN training on ADNI data

- Demographic features such as age and gender cannot be influenced by any other feature or by each other.
- Modules of brain volumes can be related to each other, but they cannot be influenced by modules of MMSE or FAQ features.
- Longitudinal measures must follow the right temporal order, i.e., there are no edges pointing backward in time.

- Modules of genetic burden scores can be related to each other, but they cannot be influenced by other modules except the demographics
- The DMs can only influence each other and no other feature or module.
- Diagnostic status cannot influence any other feature except the acDMs.
- Auxiliary and missing visit nodes were connected to their respective counterparts at the next time point, accounting for a correlation between these measures over time, e.g., through study dropout.

Structure learning was conducted via hc, tabu search, RSAMAX2, and MMHC. The final model was trained via “hc” as it had the lowest negative log-likelihood score.

2) Global meta cohort generation

A ready-trained VAMBN model can be used to infer the value of a specific variable based on the value of other variables within an individual patient using a likelihood weighting algorithm (described in chapter 2). We used the VAMBN model trained on the entire Altoida data to infer DMs in ADNI using the common features that were observed in both datasets (diagnosis, demographics (age, education, gender), MMSE subitems). We used the Bayes likelihood (Bayes-likelihood weighting (lw)) [164] methods from the 'bnlearn' package for this purpose where we average the likelihood weighting simulations using all available nodes as evidence (here we take the common features mentioned above, except node that is being predicted) to compute the predicted values. The value of the predicted variable is the expected value of the conditional distribution. We compute the predicted DMs at each time point (baseline, month 6, 12, 24, and 36) as we also have the values of the common features available at each time point in ADNI. More specifically, we randomly split the entire data into 10 folds and sequentially left out one fold for testing while training the augmented BN on the rest of the data (10-fold cross-validation). The same procedure was repeated 10 times. Splitting of the dataset was done on the subject level rather than on the level of individual data points because the Altoida dataset contains more than one measure for several patients. The aim of the experiment was to predict the latent representation / the low-dimensional code of each DM module. Prediction performances were compared against those of a standard RF regression model [179]. We calculated the NRMSE that represents the prediction performance of the features.

NRMSE is represented by:

$$NRMSE = \frac{1}{Y_{max} - Y_{min}} \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}} \tag{6.1}$$

Here Y_i and $\hat{Y}_i$ denote the real and inferred DM, respectively.

Notably, the approach allows for making predictions in a patient, in which no DM has been observed. This amounts to fixing the value of all nodes in the BN (except the DMs) to their observed values in the patient and subsequently running inference. Accordingly, it is also possible to predict DMs in ADNI based on a VAMBN model learned from Altoida data: For each ADNI patient, we fixed the value of all nodes representing demographic features, diagnostic status, and MMSE scores in the BN and then inferred the most likely value of DMs. This procedure was run for each visit in ADNI data.

10-fold cross-validation errors of VAMBN (more precisely the augmented BN) were compared against a standard RF regression model using 100 regression trees. The following additional hyper-parameters of the RF were considered:

a) mtry = 2, 3, 4. mtry means the number of possible splits at each node or the number of candidate variables considered at each split [208].

b) splitrule = “variance”, “extratrees”, “maxstat”; the three algorithms are described as follows:

   - variance: This split minimizes the weighted variance [208]. This method is favorable towards the variables with many possible splits (e.g. continuous variables or categorical variables with many categories).
   - extratrees: It is known as extremely randomized trees [209]. This algorithm builds an ensemble of unpruned decision trees according to the classical top-down procedure [209] and a number of random splits are

considered for each candidate splitting variable [210].

  - maxstat: It is known as maximally selected rank statistics [211]. Using this splitrule, the optimal split variable is determined via a statistical test for binary splits, which adjusts for the multiple testing of multiple possible split points. Through this approach adjusted p-values can be obtained for continuous covariates.

c) min. node. size = 10, 20. It specifies the minimum number of observations in a terminal node.

These hyper-parameters were tuned via a grid search, in which each hyper-parameter combination was evaluated via an inner 10-fold cross-validation. That means there was a repeated, nested cross-validation procedure.

3) Comparative study: Classifier trained on CN and MCI

To evaluate the quality of synthetic data and synthetic DMs, we performed a comparative study on both Altoida and ADNI data. We trained an RF classifier to classify them into CN and MCI at baseline based on a different set of modalities. All classifiers were adjusted for confounding effects of age and sex. We performed nested cross-validation that nests hyperparameter optimization under the model evaluation procedure, such that the training set from the outer loop is further split into sub-training and validation sets and passed to a parameter optimizing procedure like grid search. The optimizing procedure will then utilize an internal cross-validation loop to test for the optimum parameters on the training set, which will be passed to the outer loop for model evaluation [212].

The internal cross-validation procedure takes an estimator, a parameter grid, and a cross-validation strategy and applies an exhaustive search over the specified grid in search of the best parameters. The external cross-validation procedure takes the optimized estimator from the grid search procedure, the dataset, and a cross-validation strategy and returns the evaluation metrics and predicted probability respectively.

Altoida data: We considered two scenarios. In the first one, classifier was trained and tested on real data and in the second one, it was trained on synthetic data and tested on the same real data as in the first scenario. The same process was followed for all modalities, MMSE, aggregated digtial task scores and digital cognitive domains.

ADNI data: The classifiers were trained and tested on real data for all the modalities separately. There was one more modality, FAQ in addition to the modalities in Altoida. The rationale behind this experiment is that, provided the predicted features were correct, we expect them to discriminate between different diagnostic stages in ADNI, and this should not be significantly worse than Altoida.

6.4 Results

6.4.1 Brief discussion of network structure for Altoida data

Non-trivial dependencies between DMs, MMSE scores, and diagnostic state were obtained in Altoida data. Our analysis of Altoida data resulted in a quantitative network between different groups of variables, which is depicted in Figure 6.2.

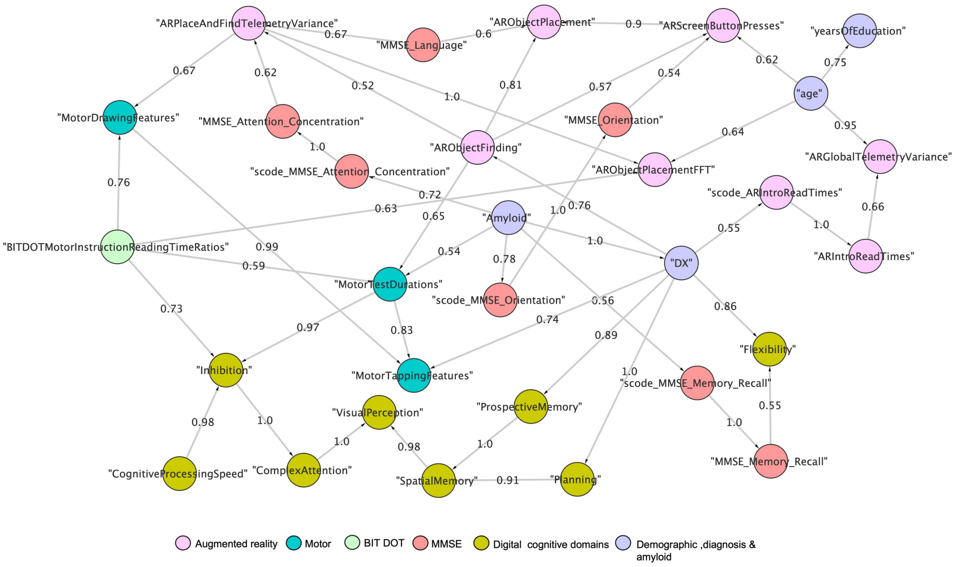


Figure 6.2: Non-trivial dependencies between DMs, MMSE scores and diagnostic state in Altoida data. The depicted network represents dependencies between variables and variable groups learned from Altoida data (a part of the network is represented here). Numbers on edges indicate the level of statistical confidence (bootstrap probability $\geq$ 0.5). A higher value indicates a higher confidence in the existence of a connection. Nodes are color coded by the group they belong to. Nodes isolated from the rest of the network are not shown.

Numbers on edges indicate the level of statistical confidence, i.e., a higher value means a stronger support by the data for the existence of the respective connection. According to our model, the MMSE subitem, "Orientation" is linked to the pressure and accuracy of touch screen button pressing in the AR game (Spearman rank correlation ρ = -0.47, 95% CI [-0.67, -0.21], adj. p < 0.0001). Spearman rank correlation is a measure between -1 and 1, where -1 indicates a perfect anti-correlation and +1 a perfect positive correlation. MMSE language sub-domain score is connected to the telemetry variance observed in object placing and finding (ρ = -0.60, 95% CI [-0.76, -0.38], adj. p < 0.0001) and to the score derived from tasks associated with placing the object in the synthetic environment (ρ = 0.34, 95% CI [0.06, 0.58], adj. p < 0.05). We also observed that the MMSE sub-domain associated with memory recall is connected to the digital cognitive domain that measures the ability to switch between thinking about two different concepts (ρ = 0.31, 95% CI [0.02, 0.55], adj. p < 0.05). The corresponding scatter plots for all these connections are illustrated in Figure 6.3. Overall, our model revealed non-trivial dependencies between DMs and MMSE scores, which are far away from simple one-to-one relationships.

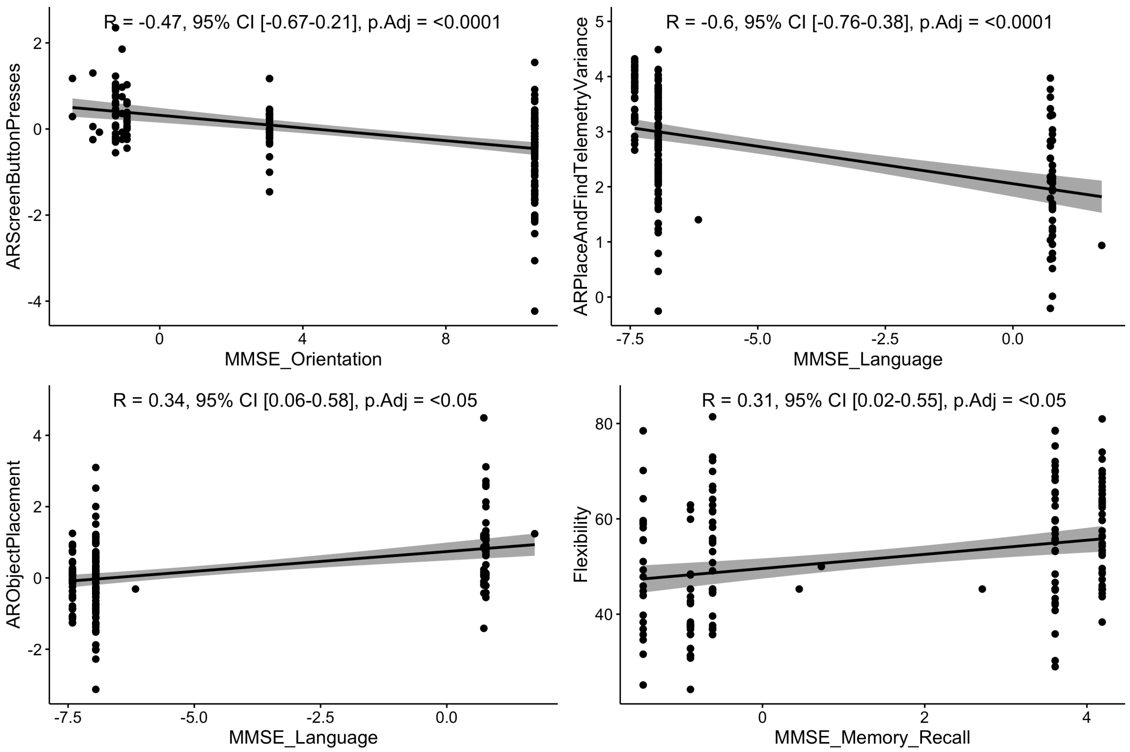


Figure 6.3: Scatter plot between different pairs of variables in Altoida data: MMSE Orientation and AR Screen Button Presses (top left), MMSE Language and AR Place and Find Telemetry Variance (top right), MMSE Language and AR Object Placement (bottom left)) and MMSE Memory Recall Registration and Flexibility (bottom right).

A point of further interest was the dependency of all DMs on the diagnosis, which was reflected via corresponding paths in our VAMBN network. We verified the dependencies of individual DMs between MCI and CN with the help of linear models while correcting for confounding effects of age and sex. This demonstrated highly significant differences between the cognitive domains, “Cognitive Processing Speed” (speed and accuracy of information processing), “Prospective Memory” (ability to remember to carry out intended actions in the future) and “Spatial Memory” (ability to recognize items that previously appeared in physical space) between MCI and CN subjects, see Table 6.7. Likewise, we performed the analysis for MMSE subitem scores, which demonstrated significance in all cases, except for the working memory registration task, which was unconnected to the diagnosis in our VAMBN network (Table 6.8). All DMs showed a significant age dependency which is shown in Table 6.9.

Digital cognitive domains	p.adjusted (CN and MCI)	95% Confidence Interval (CN and MCI)
Perceptual Motor Coordination	0.2067	-0.65, -0.03
Complex Attention	0.2205	-0.65, -0.01
Cognitive Processing Speed	<0.0001	-1.16, -0.58
Inhibition	0.9238	-0.32, 0.37
Flexibility	0.9238	-0.47, 0.21
Visual Perception	0.5791	-0.55, 0.11
Planning	0.2410	-0.65, 0.01
Prospective Memory	<0.0001	-1.27, -0.73
Spatial Memory	<0.0001	-1.29, -0.76

Table 6.7: Significance of the differences in each cognitive domain across CN and MCI (Altoida data). Multiple testing was performed via Holm's correction.

MMSE subitem scores	p. adjusted (MCI and Prodromal AD)	95% Confidence Interval (CN and MCI)
MMSE_Attention_Concentration	<0.0001	-1.24, -0.61
MMSE_Language	0.0001	0.40, 1.05
MMSE_Memory_Recall	<0.0001	-1.40, -0.85
MMSE_Orientation	0.0004	-0.96, -0.30
MMSE_Working_Memory_Registration	0.2428	-0.56, 0.14

Table 6.8: Significance of the differences of each MMSE subitem score across CN and MCI stages (Altoida data). Multiple testing was performed via Holm's correction.

Demographic	Aggregated Digital Tasks	Spearman Rank Correlation Coefficient(ρ)	p.adjusted	95% Confidence interval (lower bound)	95% Confidence interval (upper bound)
AGE	ARScreenButtonPresses_VIS1	0.38	<0.0001	0.10	0.60
AGE	ARGlobalTelemetryVariance_VIS1	-0.40	<0.0001	-0.62	-0.13
AGE	ARObjectPlacementFFT_VIS1	0.32	<0.05	0.03	0.56

Table 6.9: Spearman rank correlation with adjusted p values (Holm's method) for digital tasks and age in Altoida

6.4.2 Assessment of the quality of synthetic DMs

The quality of DMs that were predicted using VAMBN model and RF approach is shown in Table 6.10 (aggregated digital tasks) and Table 6.11 (digital cognitive domains). Estimated NRMSE representing the prediction performance was significantly better for VAMBN trained model as compared to the RF model for most of the features.

Aggregated Digital Tasks	NRMSE: Mean ±SE (VAMBN)	NRMSE: Mean ±SE (RF)
ARObjectFinding	0.1752±0.0018	0.1260 ±0.0009
ARScreenButtonPresses	0.129 ±0.0011	0.1248 ±0.0007
MotorTestDurations	0.1880 ±0.0020	0.2640 ±0.0015
ARObjectPlacementFFT	0.1644 ±0.0022	0.3830 ±0.0014
ARPlaceAndFindTelemetryVariance	0.2144 ±0.0017	0.1570 ±0.0010
ARGlobalTelemetryVariance	0.2227 ±0.0027	0.2375 ±0.0015
ARObjectPlacement	0.1433 ±0.0023	0.2493 ±0.0004
BITDOTMotorInstructionReadingTimeRatios	0.1848 ±0.0030	0.1630 ±0.0014
MotorTappingFeatures	0.2135 ±0.0029	0.3829 ±0.0018

Table 6.10: Prediction error (NRMSE) for aggregated digital tasks using VAMBN and a RF regression model. The results shown were obtained via 10 times repeated 10-fold cross-validation procedure, and the mean and standard error (SE) is of the 10-fold cross-validated error presented.

Digital Cognitive Domains	NRMSE: Mean ±SE (VAMBN)	NRMSE: Mean ±SE (RF)
PerceptualMotorCoordination	0.1896 ±0.0010	1.0212 ±0.0045
ComplexAttention	0.1886 ±0.0028	1.1843 ±0.0057
CognitiveProcessingSpeed	0.1888 ±0.0023	1.1851 ±0.0039
Inhibition	0.2002 ±0.0026	3.1834 ±0.0070
Flexibility	0.2522 ±0.0031	2.3603 ±0.0155
VisualPerception	0.2145 ±0.0023	1.5758 ±0.0060
Planning	0.1638 ±0.0026	1.2976 ±0.0030
ProspectiveMemory	0.2439 ±0.0024	0.7780 ±0.0055
SpatialMemory	0.1976 ±0.0013	1.2946 ±0.0103

Table 6.11: Prediction error (NRMSE) for cognitive domains using VAMBN and a RF regression model. The results shown were obtained via 10 times repeated 10-fold cross-validation procedure, and the mean and SE is of the 10-fold cross-validated error presented.

For the sake of completeness, we also evaluated, in how far MMSE subitem scores could be inferred from DMs, demographic data and diagnostic status. Corresponding results are shown in Table 6.12.

MMSE Subitem Scores	NRMSE: Mean ±SE (VAMBN)	NRMSE: Mean ±SE (RF)
MMSE Attention Concentration	0.4173 ±0.0013	1.7103 ±0.0111
MMSE Language	0.3487 ±0.0011	1.1516 ±0.0053
MMSE Memory Recall	0.4080 ±0.0009	0.8288 ±0.0053
MMSE Orientation	0.3752 ±0.0011	1.9904 ±0.0076
MMSE Working Memory Registration	0.1302 ±0.0002	0.2686 ±0.0090

Table 6.12: Prediction error (NRMSE) for MMSE subitem scores using VAMBN and an RF regression model. The results shown were obtained via a 10-fold cross-validation procedure, and the mean and SE of the 10-fold cross-validated error is presented.

6.4.3 Results of modeling ADNI and synthetic features derived from Altoida

brief discussion of network structure

Non-trivial dependencies of DMs generated were predicted in ADNI data from MMSE subitem scores, FAQ subitem scores, and diagnostic state in Altoida Data. The analysis of ADNI data resulted in a rather large quantitative network model, which can be explored interactively via the web-based tool DigiAD that we developed for this purpose (`https://digi-ad-viewer.scai.fraunhofer.de`). An excerpt of the network is depicted in Figure 6.4.

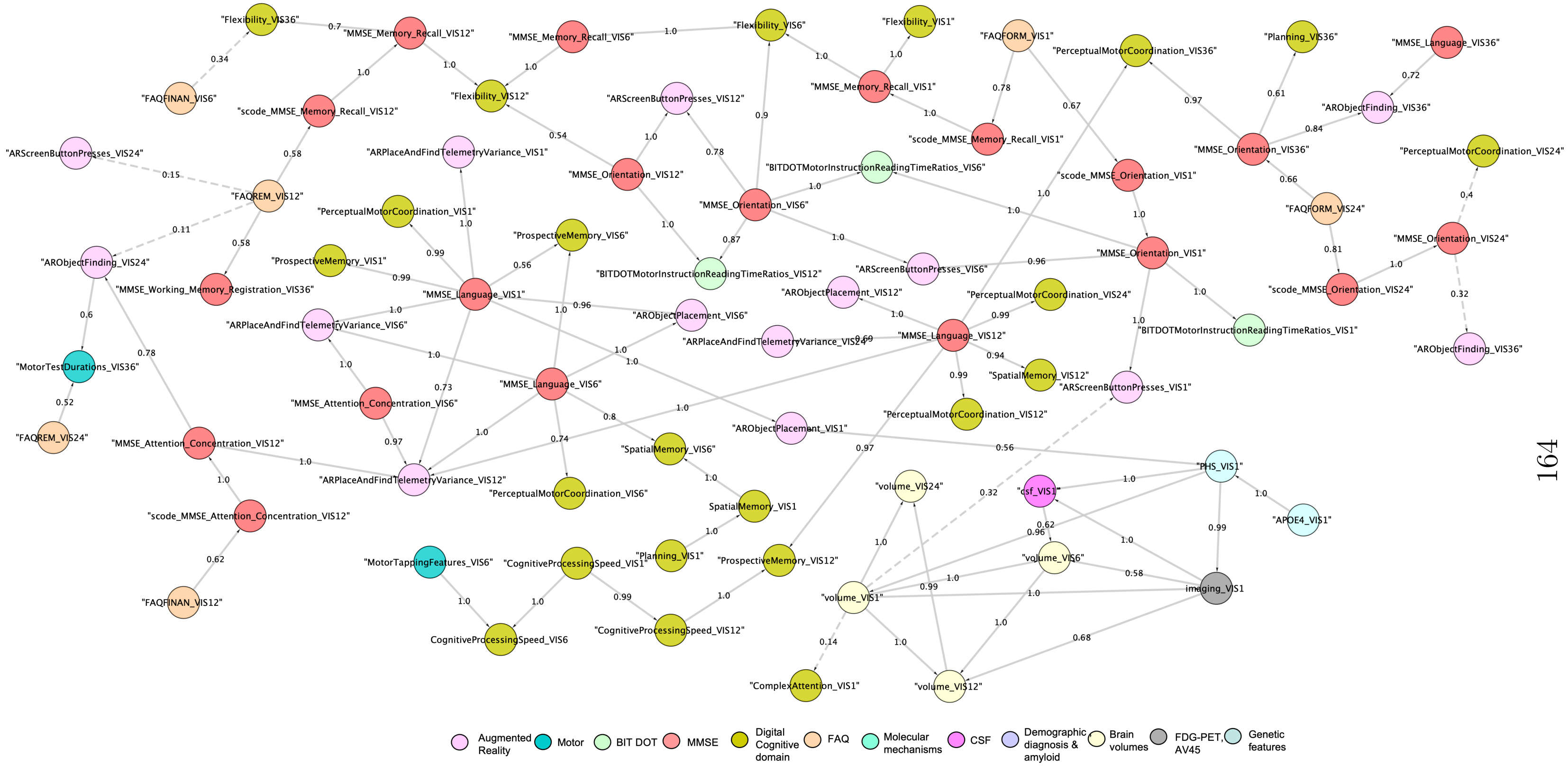

"Flexibility_VIS36"
"MMSE_Memory_Recall_VIS12"
"MMSE_Memory_Recall_VIS6"
"Flexibility_VIS6"
"Flexibility_VIS1"
"FAQFORM_VIS1"
"PerceptualMotorCoordination_VIS36"
"Planning_VIS36"
"MMSE_Language_VIS36"
"FAQFINAN_VIS6"
"scode_MMSE_Memory_Recall_VIS12"
"Flexibility_VIS12"
"ARScreenButtonPresses_VIS12"
"MMSE_Memory_Recall_VIS1"
"ARObjectFinding_VIS36"
"ARScreenButtonPresses_VIS24"
"ARPlaceAndFindTelemetryVariance_VIS1"
"MMSE_Orientation_VIS12"
"scode_MMSE_Memory_Recall_VIS1"
"MMSE_Orientation_VIS36"
"PerceptualMotorCoordination_VIS24"
"BITDOTMotorInstructionReadingTimeRatios_VIS6"
"scode_MMSE_Orientation_VIS1"
"FAQREM_VIS12"
"PerceptualMotorCoordination_VIS1"
"MMSE_Orientation_VIS6"
"FAQFORM_VIS24"
"ProspectiveMemory_VIS6"
"MMSE_Orientation_VIS24"
"ARObjectFinding_VIS24"
"ProspectiveMemory_VIS1"
"BITDOTMotorInstructionReadingTimeRatios_VIS12"
"ARScreenButtonPresses_VIS6"
"MMSE_Orientation_VIS1"
"MMSE_Working_Memory_Registration_VIS36"
"MMSE_Language_VIS1"
"ARObjectPlacement_VIS12"
"PerceptualMotorCoordination_VIS24"
"scode_MMSE_Orientation_VIS24"
"ARPlaceAndFindTelemetryVariance_VIS6"
"ARObjectPlacement_VIS6"
"ARPlaceAndFindTelemetryVariance_VIS24"
"MMSE_Language_VIS12"
"BITDOTMotorInstructionReadingTimeRatios_VIS1"
"ARObjectFinding_VIS36"
"MotorTestDurations_VIS36"
"MMSE_Language_VIS6"
"SpatialMemory_VIS12"
"ARScreenButtonPresses_VIS1"
"MMSE_Attention_Concentration_VIS6"
"PerceptualMotorCoordination_VIS12"
"ARObjectPlacement_VIS1"
"FAQREM_VIS24"
"MMSE_Attention_Concentration_VIS12"
"SpatialMemory_VIS6"
"PHS_VIS1"
"ARPlaceAndFindTelemetryVariance_VIS12"
"PerceptualMotorCoordination_VIS6"
"volume_VIS24"
"csf_VIS1"
SpatialMemory_VIS1
"APOE4_VIS1"
"scode_MMSE_Attention_Concentration_VIS12"
"volume_VIS6"
"MotorTappingFeatures_VIS6"
"CognitiveProcessingSpeed_VIS1"
"Planning_VIS1"
"ProspectiveMemory_VIS12"
imaging_VIS1
"FAQFINAN_VIS12"
"volume_VIS1"
CognitiveProcessingSpeed_VIS6
"CognitiveProcessingSpeed_VIS12"
"ComplexAttention_VIS1"
"volume_VIS12"
Augmented Reality
Motor
BIT DOT
MMSE
Digital Cognitive domain
FAQ
Molecular mechanisms
CSF
Demographic diagnosis & amyloid
Brain volumes
FDG-PET, AV45
Genetic features

Figure 6.4: Non-trivial dependencies of DMs, MMSE scores, FAQ scores, and diagnostic state predicted in ADNI data. The depicted network represents dependencies between variables and variable groups learned from ADNI data. Numbers on edges indicate the level of statistical confidence. A higher value indicates a higher confidence in the existence of a connection. Nodes are color-coded by the group they belong to. Here, shown a part of the model depicts interesting connections between clinical outcomes and DMs. Solid edges represent a bootstrap probability ≥ 0.5, and dashed edges a bootstrap probability < 0.5. Nodes isolated from the rest of the network are not shown.

Importantly, DMs in this network had been predicted for each individual ADNI patient using the AI model trained on Altoida data. The prediction used the diagnostic state, age, education, gender, and MMSE subitem scores of each patient, as explained in the methods section. Accordingly, we confirmed most connections between MMSE subdomains and DMs, which we had previously also observed in the Altoida data, see Table 6.13 for details. The corresponding scatter plots are shown in Figure 6.5.

MMSE subitem scores (from node)	**Digital Measure (to node)**	**Spearman Rank correlation coefficient (ρ)**	**CI [lower ci, upper ci]**	**p.adjusted**
MMSE Orientation (baseline)	AR Screen Button Presses (baseline)	-0.36	[-0.47, -0.25]	<0.0001
MMSE Orientation (month 6)	AR Screen Button Presses (month 6)	-0.47	[-0.56, -0.37]	<0.0001
MMSE Orientation (month 12)	AR Screen Button Presses (month 12)	-0.41	[-0.51, -0.3]	<0.0001
MMSE Language (baseline)	AR Place and Find Telemetry Variance (baseline)	-0.60	[-0.67, -0.51] 6	< 0.0001
MMSE Language (month 6)	AR Place and Find Telemetry Variance (month 6)	-0.53	[-0.62, -0.43]	< 0.0001
MMSE Language (month 12)	AR Place and Find Telemetry Variance (month 12)	-0.62	[-0.70, -0.54]	< 0.0001
MMSE Language (baseline)	AR Object Placement (baseline)	0.35	[0.23,0.46]	<0.0001
MMSE Language (month 6)	AR Object Placement (month 6)	0.18	[0.05,0.29]	<0.0001
MMSE Language (month 12)	AR Object Placement (month 12)	0.19	[0.07,0.31]	<0.0001
MMSE Memory Recall (baseline)	Flexibility (baseline)	0.66	[0.59,0.73]	<0.0001
MMSE Memory Recall (month 6)	Flexibility (month 6)	0.54	[0.45,0.62]	<0.0001
MMSE Memory Recall (month 12)	Flexibility (month 12)	0.49	[0.39,0.58]	<0.0001

Table 6.13: Spearman Rank Correlation between MMSE subdomains and DMs in ADNI with confidence intervals (CI) and adjusted p values (Holm's method)

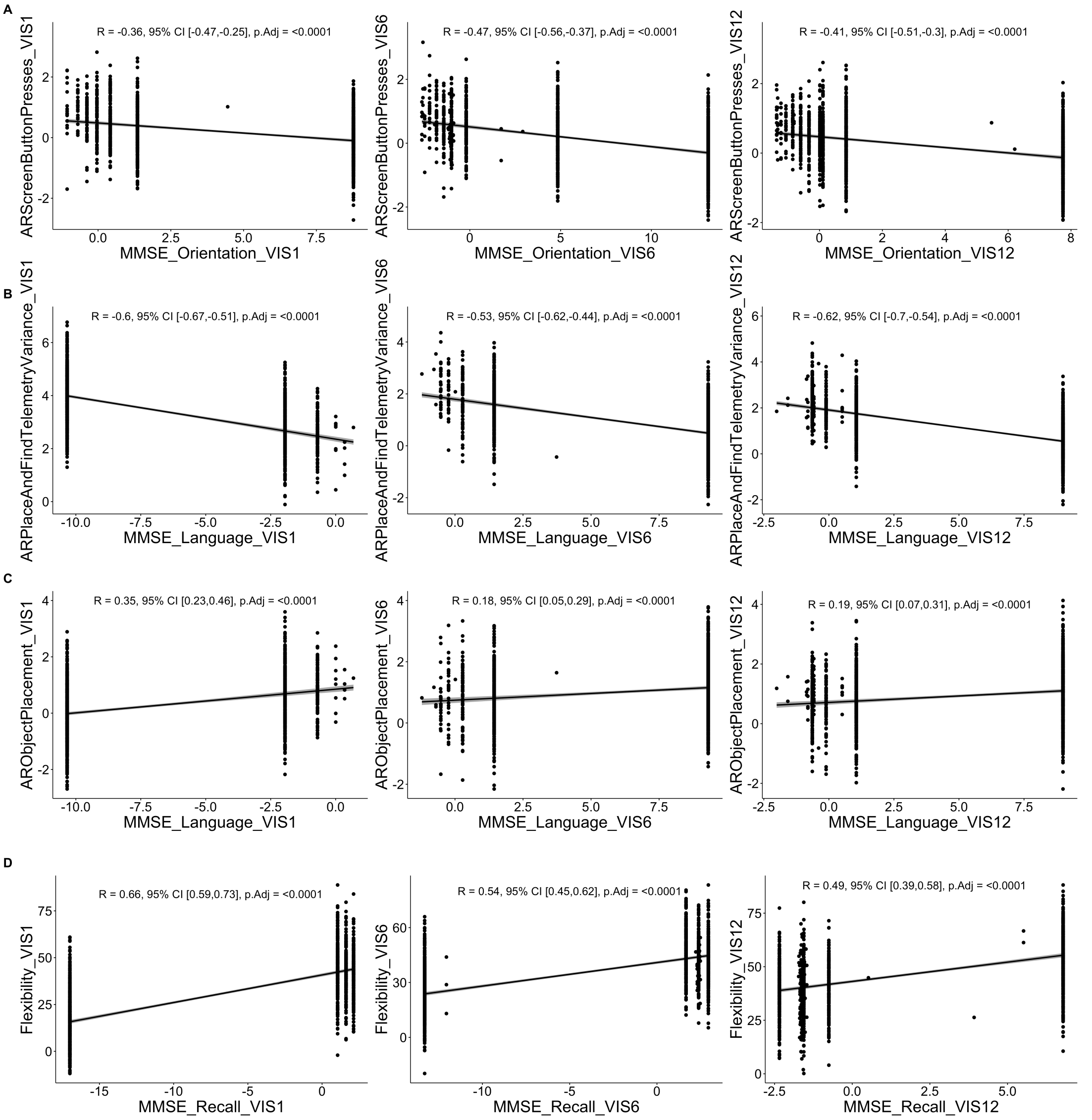


Figure 6.5: Scatter plots between different pairs of variables in ADNI data: A) MMSE Orientation and AR Screen Button Presses (baseline, month 6 and month 12), B) MMSE Language and AR Place and Find Telemetry Variance (baseline, month 6 and month 12), C) MMSE Language and AR Object Placement (baseline, month 6 and month 12), D) MMSE Memory Recall Registration and Flexibility (baseline, month 6 and month 12).

Furthermore in the network, we also observed the expected dependencies of DMs on age. In addition to these expected and confirmed findings, we also predicted new dependencies, for example between the "orientation" domain of MMSE and the difference in time when reading the instructions to the motor tests for the second time (ρ = -0.20, 95% CI [-0.32, -0.08], adj. p < 0.0001) at baseline. Additionally, dependencies were predicted between the "language" domain of MMSE and the motor coordination in response to perceived input (perceptual motor coordination) (ρ = -0.24, 95% CI [-0.36, -0.12], adj. p < 0.0001). We also found novel predicted dependencies between DMs and FAQ subitems. There was a direct link from the FAQ measure related to "remembering appointments, family occasions, holidays, medications" at month 24 to the digital task of the total duration of each motor test at month 36 (ρ = 0.14, 95% CI [0.02, 0.26], adj p = <0.005). Another example was the indirect relationship between the same subdomain of FAQ at month 12 and the cognitive domain "Flexibility" (ρ = -0.24, 95% CI [-0.36, -0.12], adj. p <0.0001). Corresponding scatter plots along with other examples are only unique to ADNI are illustrated in Figures 6.6 and 6.7. Note that FAQ was not assessed in the Altoida data.

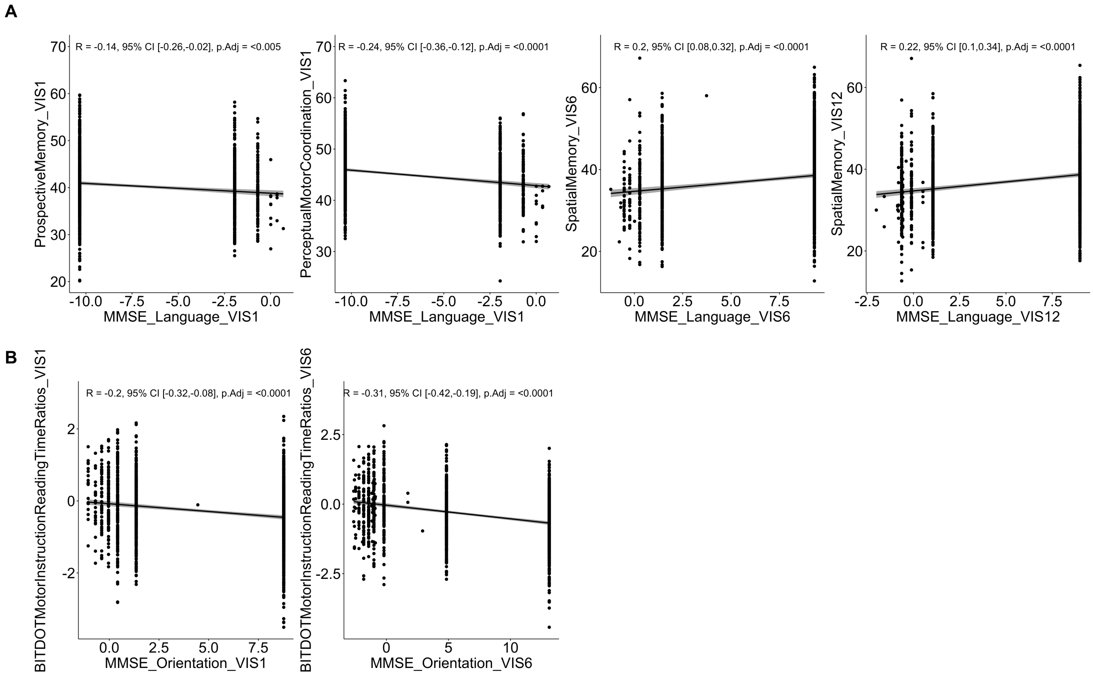


Figure 6.6: Scatter plots between different pairs of variables in ADNI data: A) MMSE Language and Spatial Memory (month 6 and month 12), MMSE Language and Prospective Memory (baseline), MMSE Language and Perceptual Motor Coordination (baseline), B) MMSE Orientation and BITDOT Motor Instructions Reading Time Ratios (baseline, month 6 and month 12).

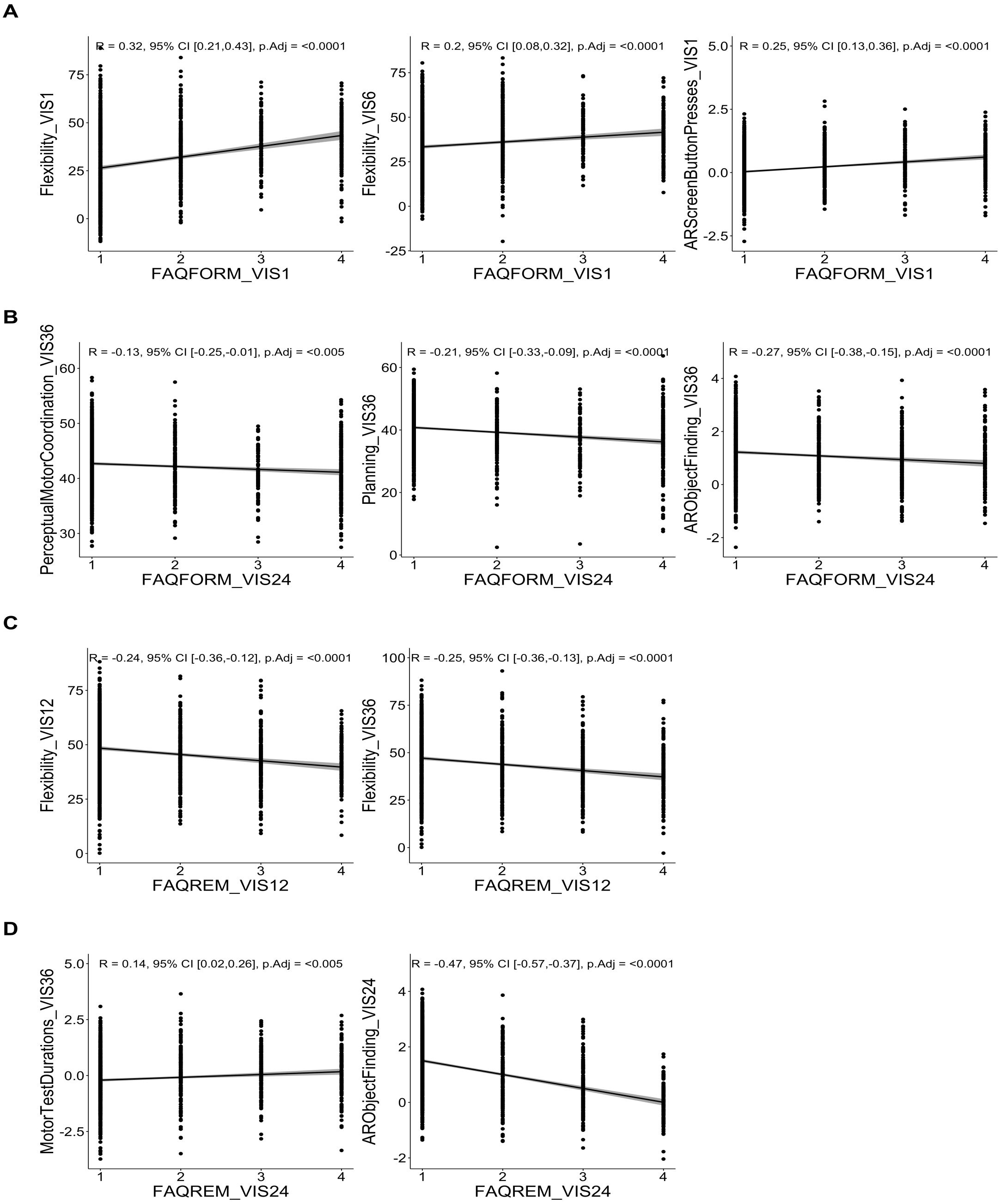


Figure 6.7: Scatter plots between different pairs of variables in ADNI data: A) FAQFORM and Flexibility (baseline), FAQFORM at baseline and Flexibility at month 6), FAQFORM and AR Screen Button Presses (baseline), B) FAQFORM at month 24 and Perceptual Motor Coordination at month 36, FAQFORM at month 24 and Planning at month 36, FAQFORM at month 24 and AR Object Finding at month 36, C) FAQREM and Flexibility (month 12), FAQREM at month 12 and Flexibility at month 36, D)FAQREM at month 24 and Motor Test Durations at month 36, FAQREM and AR Object Finding (month 24)

Interestingly, Our VAMBN model also suggested direct associations of DMs with CSF biomarkers, brain volumes, and molecular mechanisms, but these connections were either not found statistically stable or not significant. We repeated the same analysis using only those feature modalities from ADNI, which were also present in Altoida (i.e., demographic features, MMSE scores, DMs). This generally confirmed the same connections between DMs and MMSE subitem scores discussed before (Figure 6.8).

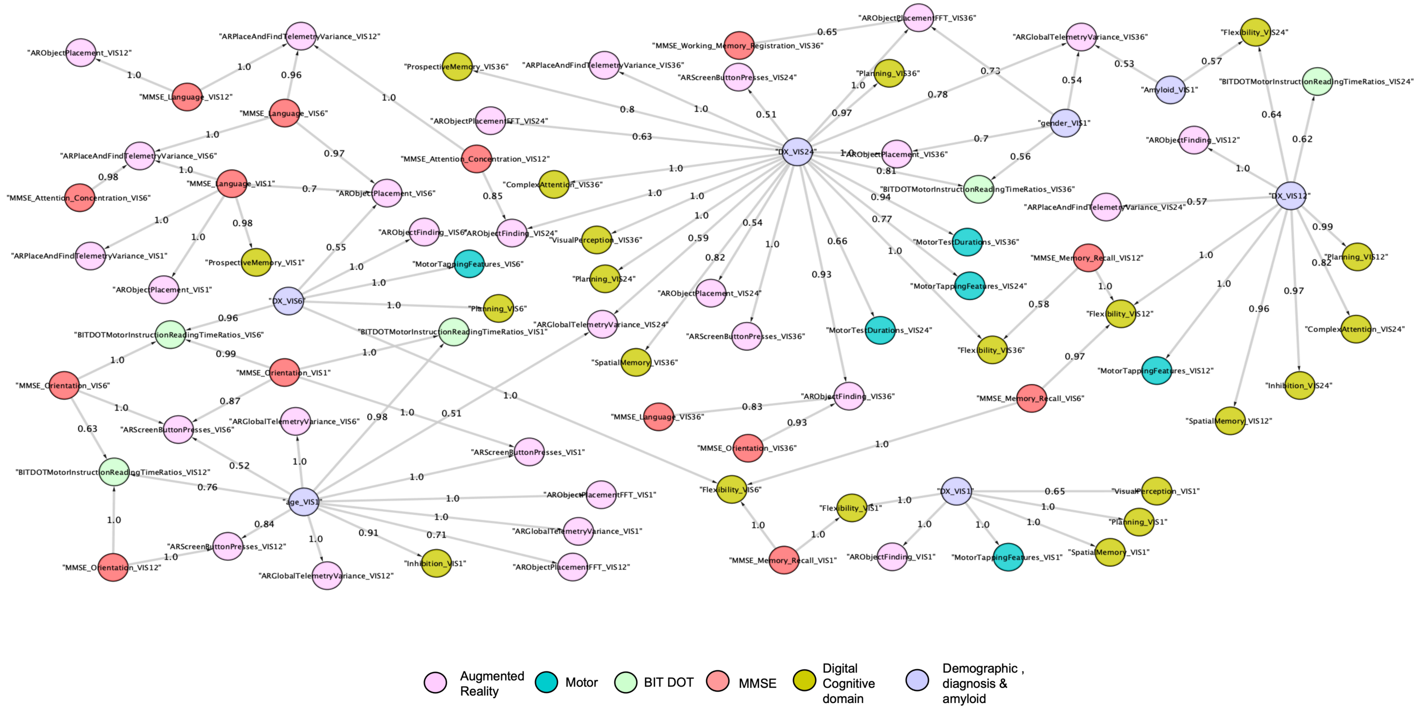

"ARObjectPlacement_VIS12"
"MMSE_Language_VIS12"
"ARPlaceAndFindTelemetryVariance_VIS12"
"MMSE_Language_VIS6"
"ARPlaceAndFindTelemetryVariance_VIS6"
"MMSE_Attention_Concentration_VIS6"
"MMSE_Language_VIS1"
"ARPlaceAndFindTelemetryVariance_VIS1"
"ARObjectPlacement_VIS1"
"ProspectiveMemory_VIS1"
"ARObjectPlacement_VIS6"
"MMSE_Attention_Concentration_VIS12"
"ProspectiveMemory_VIS36"
"ARPlaceAndFindTelemetryVariance_VIS36"
"ARObjectPlacementFFT_VIS24"
"ComplexAttention_VIS36"
"ARObjectFinding_VIS6"
"ARObjectFinding_VIS24"
"MotorTappingFeatures_VIS6"
"DX_VIS6"
"Planning_VIS6"
"BITDOTMotorInstructionReadingTimeRatios_VIS6"
"BITDOTMotorInstructionReadingTimeRatios_VIS1"
"MMSE_Orientation_VIS6"
"MMSE_Orientation_VIS1"
"ARScreenButtonPresses_VIS6"
"ARGlobalTelemetryVariance_VIS6"
"BITDOTMotorInstructionReadingTimeRatios_VIS12"
"age_VIS1"
"MMSE_Orientation_VIS12"
"ARScreenButtonPresses_VIS12"
"ARGlobalTelemetryVariance_VIS12"
"Inhibition_VIS1"
"ARObjectPlacementFFT_VIS12"
"ARGlobalTelemetryVariance_VIS1"
"ARObjectPlacementFFT_VIS1"
"ARScreenButtonPresses_VIS1"
"MMSE_Working_Memory_Registration_VIS36"
"ARScreenButtonPresses_VIS24"
"ARObjectPlacementFFT_VIS36"
"Planning_VIS36"
"DX_VIS24"
"ARObjectPlacement_VIS36"
"BITDOTMotorInstructionReadingTimeRatios_VIS36"
"VisualPerception_VIS36"
"Planning_VIS24"
"ARObjectPlacement_VIS24"
"ARGlobalTelemetryVariance_VIS24"
"SpatialMemory_VIS36"
"ARScreenButtonPresses_VIS36"
"MotorTestDurations_VIS24"
"MotorTestDurations_VIS36"
"MotorTappingFeatures_VIS24"
"Flexibility_VIS36"
"ARObjectFinding_VIS36"
"MMSE_Language_VIS36"
"MMSE_Orientation_VIS36"
"Flexibility_VIS6"
"ARGlobalTelemetryVariance_VIS36"
"gender_VIS1"
"Amyloid_VIS1"
"Flexibility_VIS24"
"BITDOTMotorInstructionReadingTimeRatios_VIS24"
"ARObjectFinding_VIS12"
"DX_VIS12"
"ARPlaceAndFindTelemetryVariance_VIS24"
"MMSE_Memory_Recall_VIS12"
"Flexibility_VIS12"
"Planning_VIS12"
"ComplexAttention_VIS24"
"MotorTappingFeatures_VIS12"
"Inhibition_VIS24"
"SpatialMemory_VIS12"
"MMSE_Memory_Recall_VIS6"
"DX_VIS1"
"Flexibility_VIS1"
"MMSE_Memory_Recall_VIS1"
"ARObjectFinding_VIS1"
"MotorTappingFeatures_VIS1"
"SpatialMemory_VIS1"
"Planning_VIS1"
"VisualPerception_VIS1"
Augmented Reality
Motor
BIT DOT
MMSE
Digital Cognitive domain
Demographic , diagnosis & amyloid

Figure 6.8: Variable dependencies identified via a VAMBN model trained on a subset of features in ADNI, which are also observed in the Altoida dataset. The network depicts those edges, which are found in at least 500 out of 1000 modular Bayesian Network reconstructions, where each modular Bayesian Network has been trained on a a subset of the data randomly drawn with replacement (non-parametric bootstrap). Moreover, we only show edges, which have a preferred direction in more than 50% of the network reconstructions.

6.4.4 Evaluating the fit of the synthetic data

As described in the previous section, synthetic patients were generated for both Altoida and ADNI based on real features and real patients. As VAMBN is a generative model, a trained VAMBN model was used to generate synthetic data. The closer the synthetic data is to the real data, the better the fit of the model. In agreement to our earlier publication [158] we generated as many synthetic patients as real and mathematically assessed the model fit in two ways: a) agreement of the marginal distributions of individual variables; b) agreement of the learned correlation structure. More specifically, for a) we used the KLD as a mathematical measure to quantify the difference. For b) we calculated the Frobenius norm of the difference between the real Pearson correlation matrix and the synthetic one. This quantity was divided by the Frobenius norm of the real correlation matrix, yielding a relative error. The number of these plots is very large. We thus only show selected results here. The distribution plots for selected variables from DMs in the ADNI data are shown in Figure 6.9. The figure shows that the model fits well to the synthetic data and the distribution for the DMs for real and synthetic data are similar. Heatmaps depicting the correlation matrices and the distribution of correlation coefficients are shown in Figure 6.10. This figure illustrates that synthetic data is also able to show very similar correlations in the features as compared to the real data

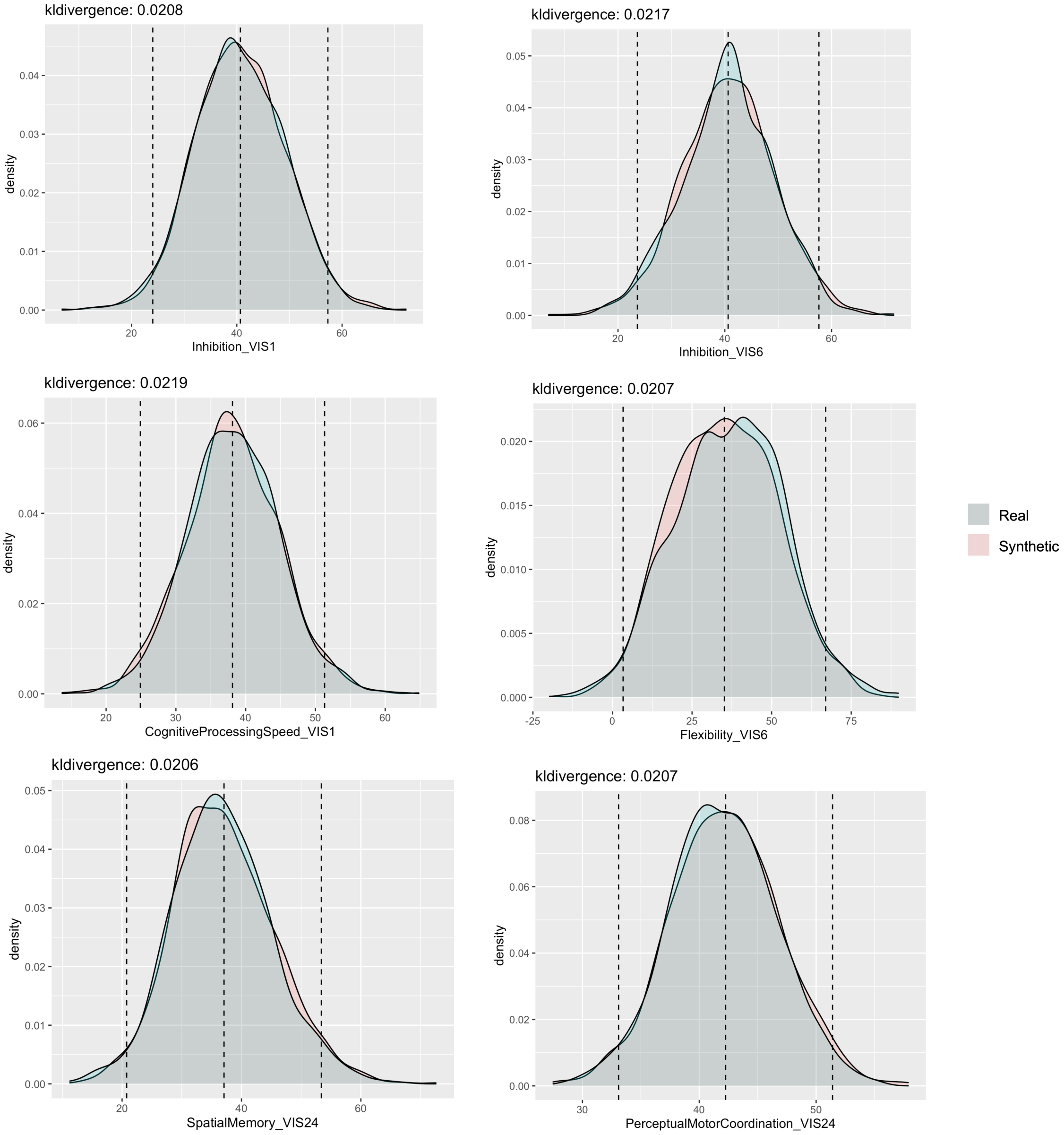
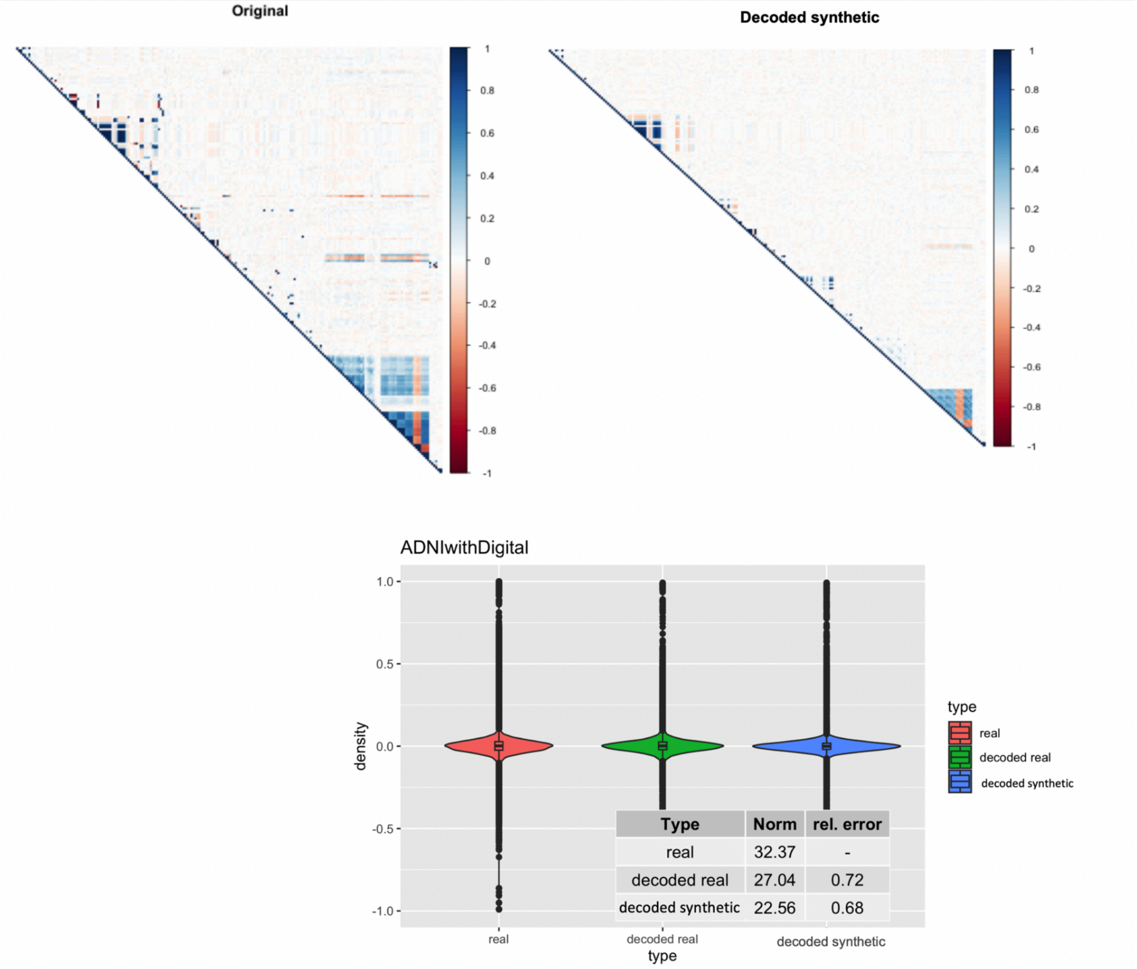


Figure 6.9: Examples of real (pink) and synthetic (blue) patients for ADNI data. The figure compares the distributions of variables related to cognitive domains in DMs for real and synthetic patients. KL-divergence between the real and synthetic patients is mentioned at the top of each plot.



Type	Norm	rel. error
real	32.37	-
decoded real	27.04	0.72
decoded synthetic	22.56	0.68

Figure 6.10: Top: Heatmaps reflecting the Pearson correlation matrix between variables in real data and decoded synthetic data generated via the VAMBN model trained on in ADNI data. Bottom: Distribution of Pearson correlation coefficients between variables in real (red), decoded real (green), and synthetic/simulated patients (blue). Tables show the Frobenius norm of the correlation matrices as well as the relative error, which consists of the norm of the matrix, that is the difference between the real and decoded correlation matrix divided by the norm of the original correlation matrix.

We also trained an RF classifier (as described in Chapter 5) on Altoida to compare real and decoded synthetic data. The classifier was able to separate between synthetic and real participants within 10 times repeated 10-fold cross-validation scheme. It is illustrated in Figure 6.11.

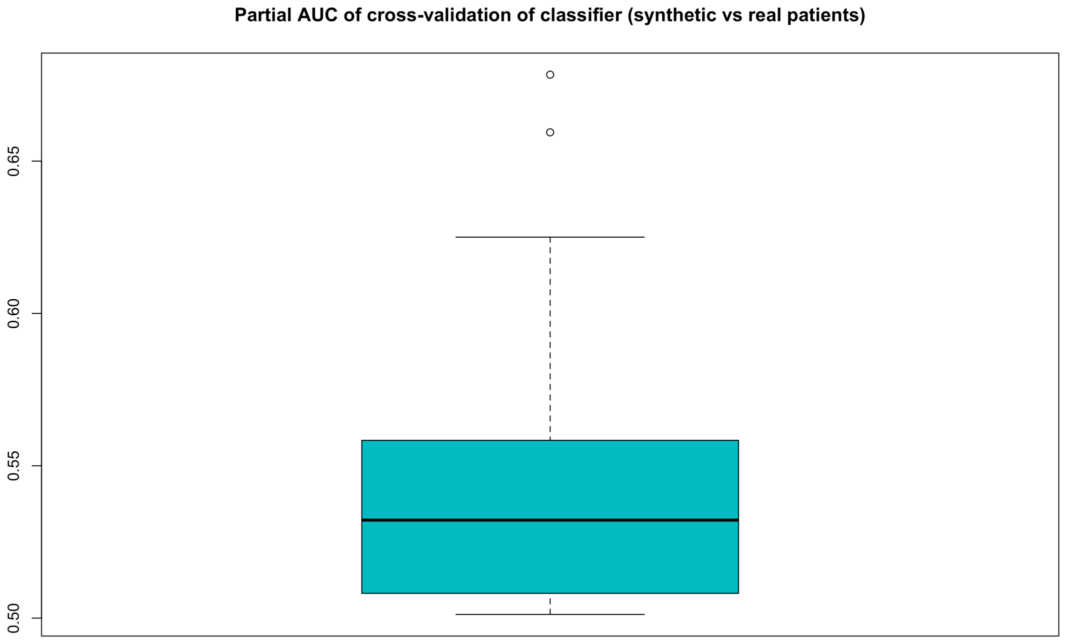


Figure 6.11: Performance of an RF classifier to correctly identify synthetic patients in Altoida, measured via the partial area under ROC curve (pAUC) at a pre-specified detection rate of $\geq$ 90% for real patients. The pAUC was assessed on test sets within 10 repeats of a tenfold cross-validation procedure. Accordingly, boxplots show the distribution of the tenfold cross-validated pAUC that was obtained from 10 repeats of the cross-validation procedure.

6.4.5 Comparative analysis of the utility of synthetic data and synthetically generated DM scores

The different classifiers trained and their results are mentioned in the following section. The results are illustrated in Figures 6.12 and 6.13. Figure 6.12 shows a comparative analysis between the classifiers trained on the same features in Altoida and ADNI data (with synthetic DMs). Figure 6.13 shows the results for the classifier trained on FAQ features in ADNI data.

Altoida data:

- MMSE: We observed that the AUC for classifier trained and tested on real data was 89.7% and that for trained on synthetic and tested on real data was 69%.

- Aggregated Digital task scores: The AUC for the classifier trained and tested on real data was found to be 93.2% and that trained on synthetic and tested on real data was 83%.
- Digital Cognitive Domains: The AUC for classifier trained and tested on real data was found to be 86.7% and that for trained on synthetic and tested on real data was 77.1%.

We observe that the classifier trained on synthetic data and then tested on real data has a lower performance (AUC score) as compared to the classifier trained and tested on real data. The reason could be that there are slight differences between real and synthetic data, specifically with regard to the correlation structure. This leads to the observed behavior.

ADNI DATA:

- MMSE: We observed that the AUC for the classifier trained and tested on real data was 75.7%.
- FAQ: The AUC for the classifier trained and tested on real data was found to be 83.4%.
- Aggregated Digital task scores: The AUC for the classifier was found to be 93.6%.
- Digital Cognitive Domains: The AUC for the classifier was found to be 85.5%.

From the above results, we can see that AUC of the classifier for aggregated digital task scores and digital cognitive domains between Altoida and ADNI are comparable. This shows that the synthetic DMs are also able to distinguish between CN and MCI stages in ADNI. Furthermore, the AUC score of aggregated digital tasks is higher than MMSE score in both Altoida and ADNI data and the DMs including digital cognitive domains have a higher AUC than MMSE and FAQ in ADNI data.

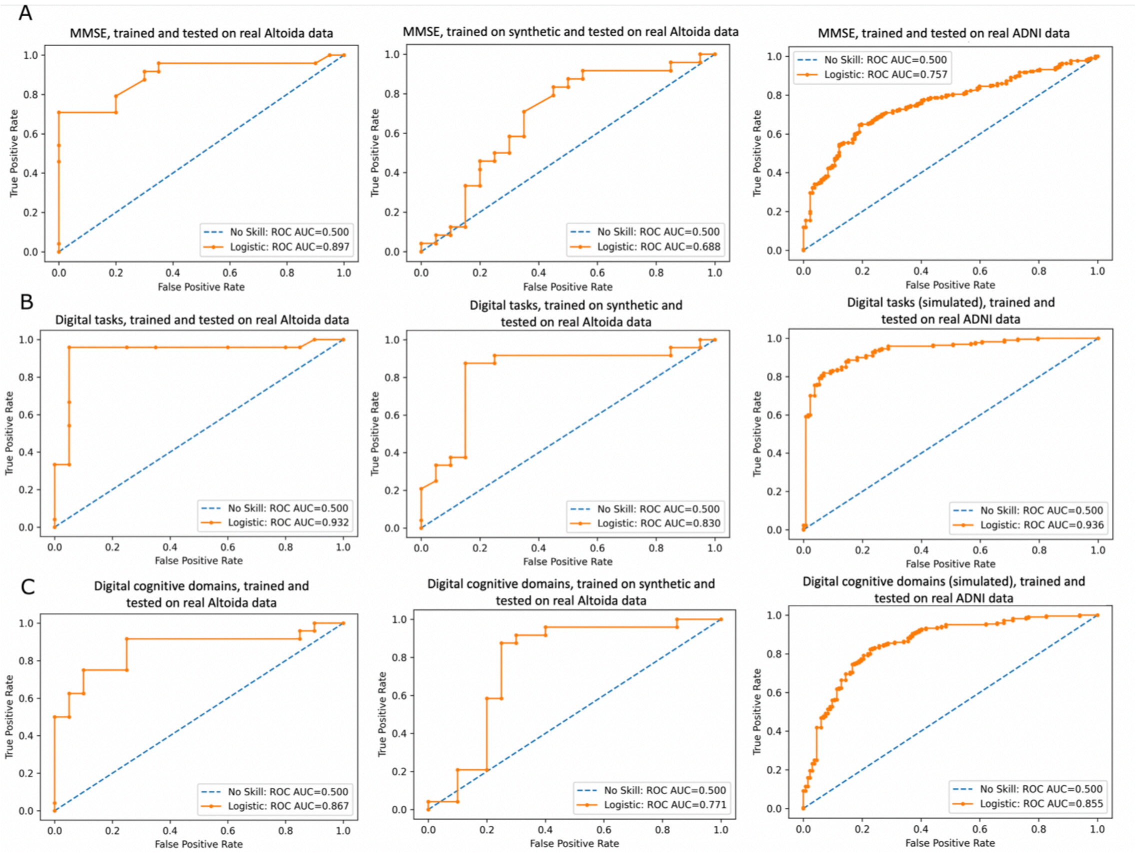


Figure 6.12: Performance of a RF classifier to classify CN and MCI stages A: MMSE features: trained and tested on real data for Altoida data, trained on synthetic data and tested on real data for Altoida data, trained and tested on real data for ADNI data B: Aggregated digital tasks: trained and tested on real data for Altoida data, trained on synthetic data and tested on real data for Altoida data, trained and tested on real data for ADNI data C: Digital cognitive domains: trained and tested on real data for Altoida data, trained on synthetic data and tested on real data for Altoida data, trained and tested on real data for ADNI data

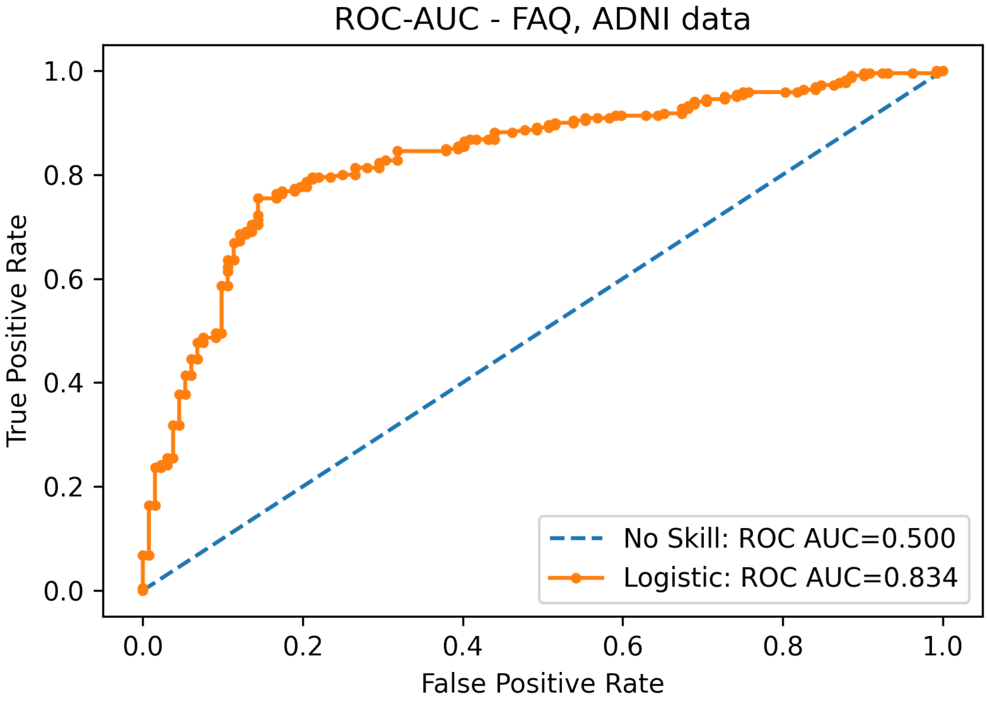


Figure 6.13: Performance of a RF classifier on FAQ data (ADNI to classify CN and MCI stages, trained and tested on real data

6.5 Conclusion

The goal of this work was to simulate a synthetic global meta-cohort, which comprises features that have been observed in another study. We achieved our goal by merging two different cohorts, Altoida and ADNI representing AD based on overlapping variables. The overlapping features in Altoida and ADNI were MMSE, demographic features and diagnostic status. The use case in this work was data collected from digital devices. DMs derived from digital devices currently receive a lot of attention, because they have the potential for a more objective, robust, and sensitive measurement of disease symptoms compared to traditional questionnaire-based assessments. In addition, data from digital devices can be used in outpatient situations and thus allow for continuous monitoring of patients in their natural environment. In the AD field, the potential use of data derived from digital devices for objective assessment of cognitive impairment has been discussed by several authors [193], including the possibility for early disease diagnosis [213]. Specifically, the measures reflecting activities of daily living have been suggested to enable an early disease diagnosis [192]. Despite all their benefits, these measures are currently only available in a few studies, which are, however, not as well characterized and as large as observational cohort studies like ADNI. Therefore, to understand the usability of DMs and their links with the standard questionnaire-based measures in AD, we simulated the DMs in ADNI based on 148 subjects in Altoida.

We were able to disentangle the apparently non-trivial and complex relationships that exist between different types of DMs reflecting performance across distinct neurocognitive domains, as well as individual tasks obtained from a virtual reality game in Altoida with questionnaire-based assessments of cognition (MMSE, FAQ). Our identified associations were statistically stable, significant, and predictable. Additionally, the analysis of classifier on the simulated features in the global meta-cohort showed these simulated features can be used for classification and has significant importance. The classifier trained on DMs in ADNI to classify subjects into CN and MCI also showed a similar comparison to the classifier trained on same features in Altoida. They also illustrate a higher AUC score to classify the subjects than the traditional questionnaire-based scores in both ADNI and Altoida. Altogether, the approach described in this paper could serve as a blueprint to generate a global meta-cohort when the features are of significant importance but are not available in many studies. This approach also helps us to better understand DMs in AD and their links with questionnaire-based scores. Generation of a global meta cohort has several advantages, a few of which are that it can help us to identify the best matching synthetic avatar for a specific real patient within the overall distribution. It could also help to efficiently generate control arms for clinical trials.

The greatest glory in living lies not in never falling,
but in rising every time we fall.

Nelson Mandela

7

Conclusion

7.1 Overview

The present work is an attempt to realistically simulate synthetic data and synthetic subject trajectories across multiple biological scales and data modalities. This aim is derived from the challenge we face regarding data privacy, data sharing, and data silos.

It was important to leverage the multi-scale and multi-level data available from large patient observational cohorts to achieve this goal. In the context of NDDs, this type of data has enormous potential and has previously helped in understanding the patterns underlying the diseases and has also enabled the possibility to discover disease subtypes.

Despite the availability of large amount of data, patient data is highly sensitive and often comes along with a lot of legal and ethical constraints. Due to these constraints, it becomes challenging to share the data outside the organization that is responsible for the data. It is also possible that sometimes data cannot be shared within the same organization. Consequently, this leads to the creation of "data silos" which further slows the pace of the research. As discussed earlier, there are some established methods that try to address this challenge but they come with certain limitations, one of them being the classical anonymization techniques. In order to share the data, data owners try to anonymize the data in a number of ways like removing identifiable features (e.g. names and addresses), adding perturbation to them (e.g., adding noise to the date of births) or grouping the variables into broader categories to ensure more than one individual in each category [44]. Once this kind of information is linked to other datasets (e.g., social media platforms) it might become easier to identify specific individuals [214]. A further concern with anonymization techniques is that fully anonymous data in essence is useless from a data science perspective. We need a certain level of individual-level information in the data if we want to build patient-level models. The other limitations of the methods that try to solve the issue of data silos are that, some of them cannot account for small sample sizes, missing values, or accommodate several variables with different numerical scales and properties. We have discussed these limitations in more detail at the beginning of the thesis.

In our work, we have tried to address these limitations to simulate a synthetic data cohort that could help to solve the problem of data silos. We have addressed an important question in the field of translational research. In the following section, we summarize our methodology and discuss the applications of our approach. We discuss the potential it holds in terms of the advancement of the current state of translational research. We also talk about the advantages and limitations of our approach over the already existing approaches.

7.2 Achievements relative to existing methods for synthetic data generation

As discussed in our work, one of the main roadblocks in preventing the acceleration of scientific research is the challenge that comes with data sharing. There have been advances in science to deal with this challenge but one of the barriers is to share individual patient-level data while preserving patient privacy [75]. The concept of differential privacy poses guarantees on the probability to compromise a person's privacy by a release of aggregate statistics from a dataset [215]. It has been shown that pairs of deep neural networks can be trained with differential privacy. As described in our work earlier, methods have been developed which aim to simulate the data that is similar to the real data. Some of the examples of the existing methods fall into the following groups (1) sampling methods that impose the risk of data leakage and quantifiable privacy e.g. Gibbs Samplers [216], sampling from BN (2) multiple imputation methods e.g. MICE (3) GANs [62](4)VAEs.

The sampling method, IM sampling for synthetic data generation simultaneously estimates the marginal distributions for different variables but it does not capture statistical dependencies across variables and therefore the synthetic data generated by this method may fail to capture the underlying structure of the data. As our method is based on the concept of BN, it overcomes this limitation. It is able to retain the data structure and the statistical dependencies across the variables in the synthetic data.

When BN is used for synthetic data generation, the graph structure and the CPDs are typically inferred from the real data. However, one of the limitations is that graph structure learning is an NP-hard problem that might either be too costly to perform or impossible when the subjects are in small numbers but they have a large number of features. The modular structure in our approach accounts for this problem and it helps to reduce the computational complexity by grouping the variables that share parameters into modules.

Another approach, MICE masks sensitive content in datasets with privacy constraints by treating sensitive data as missing data. The data here is imputed with randomly sampled values generated from models trained on the non-sensitive variables. It is probabilistic in nature, however it does not guarantee the resulting generative model to be a good estimate of the underlying joint distribution of the data. On the other hand, the DAG structure of BN in our method represents a factorization of the joint probability distribution of the variables. Therefore, our method also results in a generative model that is a good estimate of the underlying joint distribution of the data.

Recently, GANs are being widely used to generate synthetic data in the field of biomedicine [65, 217] and can generate realistic data from complex distributions. GANs combined with differential privacy have been shown to provide a technical solution to sharing of biomedical data to facilitate exploratory analysis. A recent publication proposed to train GANs based on a few variables recorded from more than 6,000 patients in the Systolic Blood Pressure Trial [75]. However, the major limitation of this method is that it tends to collapse to a statistical mode of distribution, which could raise concerns regarding the coverage of the distribution of real data by synthetic data. In addition, these methods are also not suited to deal with complexities underlying clinical data collected in observational longitudinal cohort studies [218]. They cannot explicitly model time dependencies while accounting for missing and heterogeneous data. Overlooking missing data can lead to loss of statistical power, bias in the estimation of parameters, can reduce the representativeness of the samples, incorrect estimation of variability in the data, and may complicate the analyses of the study [123]. Therefore, accounting for and overcoming the problem of incomplete observations is essential for longitudinal observational data. One of the approaches that address these challenges is the generation of realistic synthetic data using multimodal neural ordinary differential equations [218]. However, this approach is also not without limitations, one being the computational complexity and their sensitivity to several hyperparameters that should be optimized for optimal performance.

In general, one of the main limitations of deep learning methods is that they require substantial sample sizes and many training parameters. This property could cause hindrances in its application for observational cohort studies and clinical trials with small sample sizes. Our method addresses these limitations, as they can handle longitudinal data from observational clinical studies that have MNAR patterns. The auxiliary variables in our method accounts for missing patterns in the data. It also accounts for complexities underlying the longitudinal data and are able to preserve these complexities in the synthetic data.

The other type of neural networks that are generative in nature are known as

VAEs. They employ variational inference to regularize the encoding distribution and ensure that the generation of new data is less prone to overfitting, however the interpretation of the neural network models is far more challenging than for BNs. A standard VAE does not account for heterogeneous and incomplete data. In our approach, we have used special type of VAEs known as HI-VAE that can account for both heterogeneous and missing data.

Another approach, for synthetic patient generation is Synthea. It is a rule-based approach and an open-source software package that simulates realistic patients but not real, and their associated health data in a variety of formats. It is a tool to generate EHRs. The data generated from this package is free from cost, privacy, and security restrictions. However, it has been established that the synthetic data generated from Synthea is not appropriate for research into diseases that are not covered by the project or research focused on clinical discovery [50]. Usually, Synthea modules are built using clinical care guidelines and standards of care, therefore the data generated via Synthea does not include variations in care that would occur in the real world. The data included in Synthea only focuses on the care provided in the hospitals and settings by the provider and it does not include behavioral therapies and treatments that are administered outside the hospital. Moreover, the data is highly different from clinical study data, which is the subject of this thesis.

Empirically, it has also been found to be challenging to replicate population-level summary statistics with Synthea [219]. As model parameters in Synthea are derived from aggregated population-level statistics of disease progression and medical knowledge, there is a huge dependency on the prior knowledge of the system [220, 221]. This type of modeling aims at understanding the disease and offers interpretability, however, when complex systems need to be modeled, it becomes difficult to avoid simplifications and assumptions and this could cause inaccuracies or reduced utility [222, 223]. As the longitudinal data is highly heterogeneous, relying on population-level statistics does not produce models that are capable of heterogeneous health outcomes [224]. While our approach has the ability to identify aberrations of the real data, they lack the constraints that could avoid nonsensical outputs. In this regard, Synthea has an advantage over our method as care maps could provide a standardized metric to validate synthetic data conforming to medical processes.

Our approach is a data-driven approach and has tried to address the abovementioned limitations. It combines the advantages of existing approaches while mitigating their limitations. Our work focuses on the realistic situation regarding a much smaller sample size (as opposed to GANs) coupled with a significantly higher number of variables, which is common in many other medical fields, such as neurology. We have also shown in our work that data privacy respecting model training is possible. In contrast to GANs, our method relies on explicit modeling of time dependencies, as well as missing and heterogeneous data. Nonetheless, our method, as any AI-based approach, is principally dependent on sample size and signal-to-noise ratio in data. Concretely, the Altoida dataset had only ~ 150 subjects together with a very limited set of features. On the other hand, ADNI (dependent on the respective study) had more than 600 subjects with far more features, including several hundreds of SNPs.

7.3 Achievements summary

This work demonstrates a method based on the concept of BN that helps to unravel the complexities underlying a disease, by establishing connections between different clinical parameters of the disease. The model brings together heterogeneous multi-scale and multi-modal data together accounting for MNAR patterns. One of the vital aspects of our approach is data simulation by modeling the data with the help of BN as it has a generative property. Based on this property, our approach helps to solve the problem of data sharing and data silos by a realistic simulation of synthetic clinical subject trajectories across multiple biological scales and data modalities outside the area of mechanistically well-understood biological processes. BN structure and parameter learning require sufficiently large datasets that are representative of the disease population. BN model thus makes re-identification of real patients from the training data relatively unlikely. To further strengthen this point, we also included the concept of privacy-preserving training of neural network models. Thereafter, it opens the possibility to build synthetic patients and at the same time building realistic versions of clinical studies across multiple disease areas in the future. These synthetic studies could then be shared with the larger research community, even if the raw data cannot be because of legal or ethical constraints. Hence, our method could help to unlock one of the key bottlenecks in biomedical research in data-scarce disease areas. A rigorous empirical evaluation of the re-identification risk is something that has not been covered in this thesis and is subject to future research.

As demonstrated in this work, BNs also opens the door to simulating counterfactual scenarios (where we change the age of the patient and simulate its effect on the clinical biomarker) within a well-established theoretical framework, which could help, for e.g., in the design of clinical trials. Moreover, we have shown that simulated data could be used to learn complex AI models, such as a BN structure, which can subsequently be compared to real data. One of the other main applications that we tested using our method was the counterfactual simulation of features that are learned from other studies. An example of this is simulation of data from digital devices in ADNI from a model that was learned on Altoida data. This simulation gives rise to a synthetic meta-cohort. Another way via which a meta-cohort can be generated is generation of different models from cohorts having the same feature set and trajectories. The data sets with the same features can be combined based on their inclusion and exclusion criteria.

Data derived from digital devices currently receive a lot of attention, because they have the potential for a more objective, robust, and sensitive measurement of disease symptoms compared to traditional questionnaire-based assessments. In the AD field, the potential use of this type of data is for objective assessment of cognitive impairment and has been discussed by several authors [193], including the possibility for early disease diagnosis [213]. Using our approach, DMs were simulated in a study that did not originally have these features. This further lets us find connections between DMs and clinical outcomes.

7.4 Limitations of our approach

Our proposed approach is also not without limitations. As the VAMBN method is a special instance of BN, synthetic data preserves patterns of the real data, but it is not identical. Therefore, it is not necessarily true that follow-up analyses on synthetic data would give the same results as those conducted on real data. To build the model we require a relatively detailed understanding of data (in contrast to GANs) and careful handling of missing values in particular. In the extreme case of more variables than samples (high dimensional setting), we expect our method to become statistically unstable and overfit. From a technical point of view, our method usually requires a modern parallel computing architecture, hence it is also computationally demanding.

7.5 Challenges underlying evaluation of synthetic data

The evaluation of synthetic data comes with several challenges as it is a complicated task to efficiently evaluate the synthetic data [225]. There is no clearly accepted metric and therefore throughout the work of this thesis different alternatives have been tried out, and the resulting inconsistency is a limitation of this thesis. Theoretically, we expect the synthetic data to adhere to the following points (Figure 7):

1. They should have high coverage/support of the real data distribution.
2. They should have low density outside the real data distribution / minimum outliers.
3. Synthetic data points should be “sufficiently distinct” (which could mean different things) from real ones. That means we want to be sure that we don’t re-generate a real patient record

The development of an accepted metric covering all the above criteria including a systematic evaluation of a broader set of synthetic datasets must be subject to future research.

7.6 Future outlook

Our work helps in potentially facilitating data sharing and the development of counterfactual interventions. It can help in facilitating trial design and can give a better idea about inclusion and exclusion criteria. Due to ethical reasons, as real

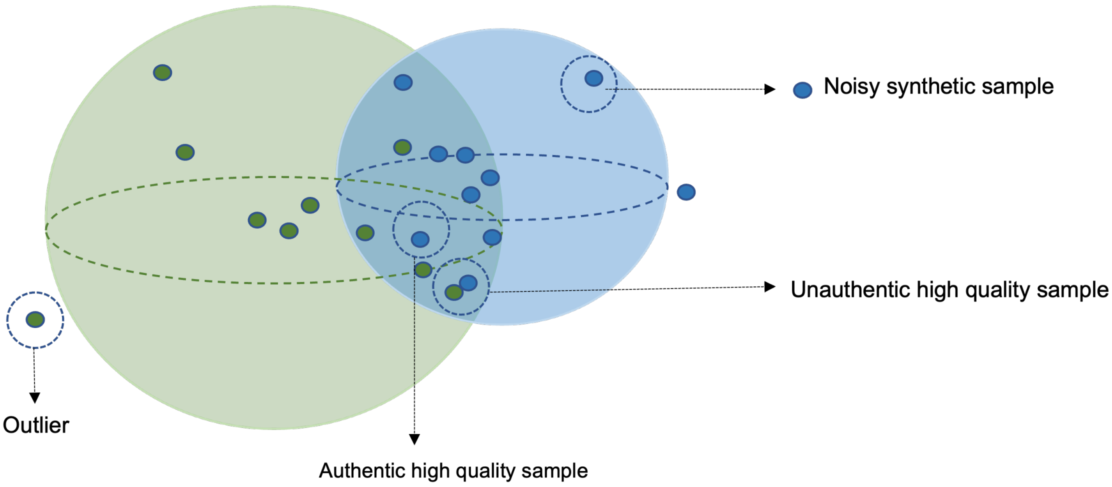

Noisy synthetic sample
Unauthentic high quality sample
Outlier
Authentic high quality sample

Figure 7.1: Pictorial depiction of the criteria that should be fulfilled by real and synthetic data. The green and blue spheres correspond to the real and synthetic distributions, respectively. The Green and blue points represent real and synthetic data respectively. (a) Synthetic data that lies outside the green sphere will look unrealistic or noisy. (b) "Unauthentic" samples or samples that seem to be of high-quality data are generated by overfitted models because they are copied from the training data. (c) High-quality data samples should lie inside the green sphere.

data cannot be used for teaching purposes in schools and colleges, synthetic data can help to solve this problem as it maintains the properties of real data. There is also an opportunity to simulate a scenario of "patients like me". To give an example; we can simulate trajectories for a patient of a certain age and sex who has diabetes. As this data will have a given distribution, this could help us establish a specific subset of subjects having a particular criterion. The benefit is here more for the physician and the individual patient. The physician could potentially use such simulations to get an idea, of how a given patient might progress in his / her disease. Importantly, such a simulation would provide information about a whole distribution of possible developments. Accordingly, the physician could then communicate with the patient to explain the situation in adequate words.

The biggest advantage of synthetic data is that it allows researchers to get access to data in a much easier way (from a legal/contract point of view) than real data. Synthetic data can help to understand the content of real data and its utility for a given analysis task. Although, synthetic data does not replace real data. One of the successful examples of synthetic data initiatives is from the netherlands cancer institute (NKI). The synthetic data generated here is a part of the netherlands cancer registry (NCR) and mimics the structure and some of the statistical patterns of the data. This data not having any real data information can give the researchers insight into the data they want to apply for (`https://iknl.nl/en/ncr/`) synthetic-dataset accessed on 22nd July 2021.

Overall, we see our work as a useful complement to federated machine learning techniques, which could help us understand the complexities underlying the diseases, generate synthetic data, break data silos, and thus enhance progress in biomedical research. However, federated learning comes with its own set of pros and cons. Practically the biggest challenges for federated learning are:

- We need an organizational and legal framework, in which all participating organizations agree on mutual data usage. Typically this requires special contracts, and legal processes are complex and slow.
- Federated learning only works, if data are semantically standardized and harmonized across organizations. This is a huge effort on its own.
- Federated learning is technically challenging and requires the writing of special code. There are libraries for federated learning publicly available, but they are far from being mature.

List of Acronyms

AA	Alzheimer’s Association
Abeta	Amyloid Beta
AD	Alzheimer’s Disease
ADAS	Alzheimer’s Disease Assessment Scale-Cognitive Subscale
ADLs	Activities of Daily Living
ADL	Activity of Daily Living
ADNI	Alzheimer’s Disease Neuroimaging Initiative
AI	Artificial Intelligence
ANOVA	Analysis of Variance
APOE4	Apolipoprotein E4
AR	Augmented Reality
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
AUC	Area Under Curve
AV-45	Florbetapir
BD	Bayesian Dirichlet
BDe	Likelihood Equivalence Bayesian Drichlet
BDeu	Bayes Dirichlet (likelihood) Equivalent Uniform

BIC	Bayesian Information Criterion
BN	Bayesian Network
CC	Complete Case
CDR	Clinical Dementia Rating
CDRSB	Clinical Dementia Rating Sum of Boxes
CN	Cognitively Normal
CPD	Conditional Probability Distributions
CPDAG	Completed Partially Directed Acyclic Graph
CPT	Conditional Probability Tables
CS	Compressed Sensing
CSF	Cerebrospinal Fluid
DAG	Directed Acyclic Graph
DAGAN	Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction
DAG	Directed Acyclic Graphs
DaTscan	Dopamine Transporter Scan
DM	Digital Measure
DNN	Deep Neural Network
DP	Differential Privacy
DPLL	Decomposable Penalized Log Likelihood
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
EHR	Electronic Health Record
ELBO	Evidence Lower Bound

EM	Expectation Maximization
EMCI	Early Mild Cognitive Impairment
eQTL	Phenome-Wide Association Studies
FAQ	Functional Activity Questionnaire
FDG	Fluorodeoxyglucose
GAN	Generative Adversarial Network
GBA	Glucocerebrosidase
GBN	Gaussian Bayesian Network
GDPR	General Data Protection Regulation
GMM	Gaussian Mixture Models
GTEx	Genotype-Tissue Expression
hc	Greedy Hill Climbing
HGC95	Heckermann, Geiger and Chickering
HI-VAE	Heterogeneous Incomplete Variational Autoencoder
HRC	Haplotype Reference Consortium
HRV	Heart Rate Variability
i.d.	identically Distributed
i.i.d.	Independently and Identically Distributed
ICD-10	International Classification of Diseases 10th Revision
IM	Independent Marginals
IMI	Innovative Medicines Initiative
IQR	Inter Quartile Range
KLD	Kullback-Leibler Divergence

LL	Log Likelihood
LMCI	Late Mild Cognitive Impairment
LOCF	Last Observation Carried Forward
lw	Bayes-likelihood Weighting
MAR	Missing at Random
MB	Markov Blanket
MBN	Modular Bayesian Network
MC-medGAN	Multi-categorical medGAN
MCA	Multiple Correspondence Analysis
MCAR	Missing Completely at Random
MCI	Mild Cognitive Impairment
MDS-UPDRS	MDS-Unified Parkinson’s Disease Rating Scale
MDS	Movement Disorder Society
medGAN	Medical Generative Adversarial Network
MICE	Multivariate Imputation by Chained Equations
MLE	Maximum Likelihood Estimation
MMHC	Max-Min Hill-Climbing
MMPC	Max-Min Parents and Children
MMSE	Mini-Mental State Examination
MNAR	Missing Not at Random
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NCR	Netherlands Cancer Registry

NDD	Neurodegenerative Disease
NDDs	Neurodegenerative Diseases
NeuroMMSig	Multimodal Mechanistic Signatures for Neurodegenerative Diseases
NFT	Neurofibrillary Tangles
NIA	National Institute on Aging
NINCDS/ADRDA	National Institute of Neurological and Communicative Disorders and Stroke - Alzheimer's Disease and Related Disorders Association
NKI	Netherlands Cancer Institute
NL	Normal
NMI	Neuromotor Index
NRMSE	Normalized Root Mean Square Error
OOB	Out-of-Bag
P-Tau	Phospho Tau
pAUC	Partial area under ROC curve
PD	Parkinson's Disease
PET	Positron Emission Tomography
PET/SPECT	Positron Emission Tomography/Single Photon Emission Computed Tomography
PheWAS	Phenome-Wide Association Studies
PKPD	Pharmacokinetic-Pharmacodynamic
PPMI	Parkinson's Progression Markers Initiative
pRBD	possible REM-sleep Behavior Disorder
PRS	Polygenic Risk Scores

RADAR-AD	Remote Assessment of Disease and Relapse-AD
RAVLT	Rey Auditory Verbal Learning Test
RBD	Rapid Eye Movement Sleep Behavior Disorder
RBD	REM-sleep Behavior Disorder
RF	Random Forest
ROC	Receiver Operator Characteristic
ROI	Region of Interest
RSMAX2	Restricted Maximization
SC	Synthetic Cohort
SDC	Statistical Disclosure Control
SDL	Statistical Disclosure Limitation
SE	Standard Error
SI-HITON-PC	Semi-InterleavedHiton Parents and Children
SMC	Significant Memory Concern
SN	Subastantia Nigra
SNP	Single Nucleotide Polymorphism
SNpc	Substantia Nigra Pars Compacta
T tau	Total Tau
TREND	Tuebinger Evaluation of Risk Factors for Early Detection of NeuroDegeneration
UPDRS	Unified Parkinson's Disease Rating Scale Score
VAE	Varitational Autoencoder
VAMBN	Variational Autoencoder Modular Bayesian Network

Publication List

Here is a list of peer-reviewed and preprint versions of my publications that formed the basis of this thesis:

1. **Sood, M.**, Sahay, A., Karki, R., Emon, M. A., Vrooman, H., Hofmann-Apitius, M., Fröhlich, H. (2020). Realistic simulation of virtual multi-scale, multi-modal patient trajectories using Bayesian networks and sparse auto-encoders. Scientific reports, 10(1), 10971. https://doi.org/10.1038/s41598-020-67398-

2. Gootjes-Dreesbach, L., **Sood, M.**, Sahay, A., Hofmann-Apitius, M., Fröhlich, H. (2020). Variational Autoencoder Modular Bayesian Networks for Simulation of Heterogeneous Clinical Study Data. Frontiers in big data, 3, 16. https://doi.org/10.3389/fdata.2020.0001

3. **Meemansa Sood**, Mohamed Aborageh, Daniel Domingo.Fernández, Robbert Harms, Thomas Lordick, Colin Birkenbihl, Andrew P Owens, Neva Coello, Vaibhav A. Narayan, Dag Arsland, Maxmilian Bügler, Holger Fröhlich, fort he Alzheimer's Dosease Neuroimaging Initiative, RADAR-AD Consortium. Evaluating Digital Device Technology in Alzheimer's Disease via Artificial Intelligence, medRxiv, 2021

4. **Meemansa Sood**, Ulrike Suenkel, Anna-Katharina von Thaler, Helena U. Zacharias, Kathrin Borckmann, Gehrad W. Eschweiler, Walter Maetzler, Daniela Berg, Holger Fröhlich, Sebastian Heinzel. Bayesian network modeling of risk and prodromal markers of Parkinson's disease, medRxiv, 2022

This publication is currently under review in PLOS ONE journal.

The following is the list of other published peer-reviewed and in review works. These works were not directly related to this thesis but were published during the course of my PhD:

1. Muurling, M., de Boer, C., Kozak, R., Religa, D., Koychev, I., Verheij, H., Nies, V., Duyndam, A., **Sood, M.**, Fröhlich, H., Hannesdottir, K., Erdemli, G., Lucivero, F., Lancaster, C., Hinds, C., Stravopoulos, T. G., Nikolopoulos, S., Kompatsiaris, I., Manyakov, N. V., Owens, A. P., ... RADAR-AD Consortium (2021). Remote monitoring technologies in Alzheimer's disease: design of the RADAR-AD study. Alzheimer's research therapy, 13(1), 89. https://doi.org/10.1186/s13195-021-00825-

2. Emon, M. A., Heinson, A., Wu, P., Domingo-Fernández, D., **Sood, M.**, Vrooman, H., Corvol, J. C., Scordis, P., Hofmann-Apitius, M., Fröhlich, H. (2020). Clustering of Alzheimer's and Parkinson's disease based on genetic burden of shared molecular mechanisms. Scientific reports, 10(1), 19097. https://doi.org/10.1038/s41598-020-76200-

3. de Jong, J., Emon, M. A., Wu, P., Karki, R., **Sood, M.**, Godard, P., Ahmad, A., Vrooman, H., Hofmann-Apitius, M., Fröhlich, H. (2019). Deep learning for clustering of multivariate clinical patient trajectories with missing values. GigaScience, 8(11), giz134. https://doi.org/10.1093/gigascience/giz13

4. Wendland, P., Birkenbihl, C., Gomez-Freixa, M., **Sood, M.**, Kschischo, M., Fröhlich, H. (2022). Generation of realistic synthetic data using Multimodal Neural Ordinary Differential Equations. NPJ digital medicine, 5(1), 122

5. Tamara Raschka, **Meemansa Sood**, Bruce Schultz, Aybuge Altay, Christian Ebeling, Holger Fröhlich. AI reveals insights into link between CD33 and cognitive impairment in Alzheimer's Disease, bioRxiv 2022

References

[1] World Health Organization. *The world health report 2006: working together for health.* World Health Organization, 2006.

[2] Valery L Feigin, Amanuel Alemu Abajobir, Kalkidan Hassen Abate, Foad Abd-Allah, Abdishakur M Abdulle, Semaw Ferede Abera, Gebre Yitayih Abyu, Muktar Beshir Ahmed, Amani Nidhal Aichour, Ibtihel Aichour, et al. Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the global burden of disease study 2015. *The Lancet Neurology*, 16(11):877–897, 2017.

[3] Brittany N Dugger and Dennis W Dickson. Pathology of neurodegenerative diseases. *Cold Spring Harbor perspectives in biology*, 9(7):a028035, 2017.

[4] Gregory D Cuny. Neurodegenerative diseases: challenges and opportunities. *Future Medicinal Chemistry*, 4(13):1647–1649, 2012.

[5] Martin James Prince, Anders Wimo, Maelenn Mari Guerchet, Gemma Claire Ali, Yu-Tzu Wu, and Matthew Prina. World alzheimer report 2015-the global impact of dementia: An analysis of prevalence, incidence, cost and trends. 2015.

[6] Alexandra-Maria Tăuţan, Bogdan Ionescu, and Emiliano Santarnecchi. Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. *Artificial Intelligence in Medicine*, 117:102081, 2021.

[7] Serge Gauthier, Philip Scheltens, and Jeffery Cummings. *Alzheimer's Disease and Related Disorders*. CRC Press, 2005.

[8] Anil Kumar, Jaskirat Sidhu, Amandeep Goyal, Jack W Tsao, and Jacquelyn Svercauski. Alzheimer disease (nursing). 2021.

[9] Vanessa J De-Paula, Marcia Radanovic, and Breno S Diniz. and orestes v. forlenza. *Protein Aggregation and Fibrillogenesis in Cerebral and Systemic Amyloid Disease*, 65:329, 2012.

[10] Alberto Serrano-Pozo, Matthew P Frosch, Eliezer Masliah, and Bradley T Hyman. Neuropathological alterations in alzheimer disease. *Cold Spring Harbor perspectives in medicine*, 1(1):a006189, 2011.

[11] J Gaugler, TJ Bryan James, J Reimer, and J Weuve. Alzheimer's association. 2021 alzheimer's disease facts and figures. *Alzheimer's Dementia: Chicago, IL, USA*, 17, 2021.

[12] Michael T Hayes. Parkinson's disease and parkinsonism. *The American journal of medicine*, 132(7):802–807, 2019.

[13] E Ray Dorsey, Alexis Elbaz, Emma Nichols, Nooshin Abbasi, Foad Abd-Allah, Ahmed Abdelalim, Jose C Adsuar, Mustafa Geleto Ansha, Carol Brayne, Jee-Young J Choi, et al. Global, regional, and national burden of parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 17(11):939–953, 2018.

[14] MC de De Rijk, LJ Launer, K Berger, MM Breteler, JF Dartigues, M Baldereschi, L Fratiglioni, A Lobo, J Martinez-Lage, C Trenkwalder, et al. Prevalence of parkinson's disease in europe: A collaborative study of population-based cohorts. neurologic diseases in the elderly research group. *Neurology*, 54(11 Suppl 5):S21–3, 2000.

[15] E Dorsey, Todd Sherer, Michael S Okun, and Bastiaan R Bloem. The emerging evidence of the parkinson pandemic. *Journal of Parkinson's disease*, 8(s1):S3–S8, 2018.

[16] Konstantina G Yiannopoulou, Aikaterini I Anastasiou, Venetia Zachariou, and Sygkliti-Henrietta Pelidou. Reasons for failed trials of disease-modifying treatments for alzheimer disease and their contribution in recent research. *Biomedicines*, 7(4):97, 2019.

[17] Paolo Stanzione and Domenicantonio Tropepi. Drugs and clinical trials in neurodegenerative diseases. *Annali dell'Istituto superiore di sanità*, 47:49–54, 2011.

[18] Sepehr Golriz Khatami, Sarah Mubeen, and Martin Hofmann-Apitius. Data science in neurodegenerative disease: Its capabilities, limitations, and perspectives. *Current opinion in neurology*, 33(2):249, 2020.

[19] Lori A Whitten. Translational neuroscience and potential contributions of functional magnetic resonance imaging (fmri) to the prevention of substance misuse and antisocial behavior. *Prevention Science*, 14(3):238–246, 2013.

[20] Caitlin Davies, Olivia KL Hamilton, Monique Hooley, Tuula E Ritakari, Anna J Stevenson, and Emily NW Wheater. Translational neuroscience: the state of the nation (a phd student perspective). *Brain Communications*, 2(1):fcaa038, 2020.

[21] Karsten Specht. Current challenges in translational and clinical fmri and future directions. *Frontiers in psychiatry*, page 924, 2020.

[22] Jamie M Zoellner and Kathleen J Porter. Translational research: Concepts and methods in dissemination and implementation research. In *Nutrition in the Prevention and Treatment of Disease*, pages 125–143. Elsevier, 2017.

[23] Nina Fudge, Euan Sadler, Helen R Fisher, John Maher, Charles DA Wolfe, and Christopher McKevitt. Optimising translational research opportunities: a systematic review and narrative synthesis of basic and clinician scientists' perspectives of factors which enable or hinder translational research. *PLoS One*, 11(8):e0160475, 2016.

[24] Elias A Zerhouni et al. Translational and clinical science-time for a new vision. *New England Journal of Medicine*, 353(15):1621, 2005.

[25] K Snape, RC Trembath, and GM Lord. Translational medicine and the nihr biomedical research centre concept. *QJM: An International Journal of Medicine*, 101(11):901–906, 2008.

[26] Steven H Woolf. The meaning of translational research and why it matters. *Jama*, 299(2):211–213, 2008.

[27] Heidi Hörig, Elizabeth Marincola, and Francesco M Marincola. Obstacles and opportunities in translational research. *Nature medicine*, 11(7):705–708, 2005.

[28] Linda L Restifo and Gerald R Phelan. The cultural divide: exploring communication barriers between scientists and clinicians, 2011.

[29] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[30] Yuri YM Aung, David CS Wong, and Daniel SW Ting. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *British medical bulletin*, 139(1):4–15, 2021.

[31] Vasant Dhar. Data science and prediction. *Communications of the ACM*, 56(12):64–73, 2013.

[32] J Leek. The key word in "data science" is not data, it is science, simply statistics, 2013.

[33] Jiaying Liu, Xiangjie Kong, Feng Xia, Xiaomei Bai, Lei Wang, Qing Qing, and Ivan Lee. Artificial intelligence in the 21st century. *IEEE Access*, 6:34403–34421, 2018.

[34] Trevor JM Bench-Capon and Paul E Dunne. Argumentation in artificial intelligence. *Artificial intelligence*, 171(10-15):619–641, 2007.

[35] Alexander L Fogel and Joseph C Kvedar. Artificial intelligence powers digital medicine. *NPJ digital medicine*, 1(1):1–4, 2018.

[36] Abraham Verghese, Nigam H Shah, and Robert A Harrington. What this computer needs is a physician: humanism and artificial intelligence. *Jama*, 319(1):19–20, 2018.

[37] Skyler Norgaard, Ramyar Saeedi, Keyvan Sasani, and Assefaw H Gebremedhin. Synthetic sensor data generation for health applications: a supervised deep learning approach. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1164–1167. IEEE, 2018.

[38] Steven E Dilsizian and Eliot L Siegel. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports*, 16(1):1–8, 2014.

[39] Sarah Houlton. How artificial intelligence is transforming healthcare. *Prescriber*, 29(10):13–17, 2018.

[40] Samira Saifi, Allen J Taylor, Joseph Allen, and Robert Hendel. The use of a learning community and online evaluation of utilization for spect myocardial perfusion imaging. *JACC: Cardiovascular Imaging*, 6(7):823–829, 2013.

[41] Charles E Kahn Jr. From images to actions: opportunities for artificial intelligence in radiology, 2017.

[42] LD Jones, D Golan, SA Hanna, and M Ramachandran. Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern? *Bone & joint research*, 7(3):223–225, 2018.

[43] Carlo Fabrizio, Andrea Termine, Carlo Caltagirone, and Giulia Sancesario. Artificial intelligence for alzheimer's disease: Promise or challenge? *Diagnostics*, 11(8):1473, 2021.

[44] Giske Ursin, Sagar Sen, Jean-Marie Mottu, and Mari Nygård. Protecting privacy in large datasets—first we assess the risk; then we fuzzy the dataprotecting privacy in large datasets. *Cancer Epidemiology, Biomarkers & Prevention*, 26(8):1219–1224, 2017.

[45] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071, 2011.

[46] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[47] Jörg Drechsler. *Synthetic datasets for statistical disclosure control: theory and implementation*, volume 201. Springer Science & Business Media, 2011.

[48] Bill Howe, Julia Stoyanovich, Haoyue Ping, Bernease Herman, and Matt Gee. Synthetic data for social good. *arXiv preprint arXiv:1710.08874*, 2017.

[49] Joshua Kim, Carri Glide-Hurst, Anthony Doemer, Ning Wen, Benjamin Movsas, and Indrin J Chetty. Implementation of a novel algorithm for generating synthetic ct images from magnetic resonance imaging data sets for prostate cancer radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 91(1):39–47, 2015.

[50] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 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*, 25(3):230–238, 2018.

[51] Kudakwashe Dube and Thomas Gallagher. Approach and method for generating realistic synthetic electronic healthcare records for secondary use. In *International Symposium on Foundations of Health Informatics Engineering and Systems*, pages 69–86. Springer, 2013.

[52] Donald B Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468, 1993.

[53] Roderick JA Little. Statistical analysis of masked data. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, 9:407–407, 1993.

[54] Gregory J Matthews and Ofer Harel. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5:1–29, 2011.

[55] DB Rubin. Multiple imputation for nonresponse in surveys, new york: John-wiley & sons, 1987. *Google Scholar| Crossref*.

[56] Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1):1, 2003.

[57] CKCN Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[58] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.

[59] Yarin Gal, Yutian Chen, and Zoubin Ghahramani. Latent gaussian processes for distribution estimation of multivariate categorical data. In *International Conference on Machine Learning*, pages 645–654. PMLR, 2015.

[60] David B Dunson and Chuanhua Xing. Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009.

[61] Ashar Ahmad and Holger Fröhlich. Towards clinically more relevant dissection of patient heterogeneity via survival-based bayesian clustering. *Bioinformatics*, 33(22):3558–3566, 2017.

[62] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[63] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan:

Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

[64] Ramiro Camino, Christian Hammerschmidt, and Radu State. Generating multi-categorical samples with generative adversarial networks. *arXiv preprint arXiv:1807.01202*, 2018.

[65] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017.

[66] Ashar Ahmad and Holger Fröhlich. Integrating heterogeneous omics data via statistical inference and learning techniques. *Genomics and Computational Biology*, 2(1):e32–e32, 2016.

[67] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[68] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[69] Thom Benjamin Volker and Gerko Vink. Anony mice d shareable data: Using mice to create and analyze multiply imputed synthetic datasets. *Psych*, 3(4):703–716, 2021.

[70] Yang Lei, Joseph Harms, Tonghe Wang, Yingzi Liu, Hui-Kuo Shu, Ashesh B Jani, Walter J Curran, Hui Mao, Tian Liu, and Xiaofeng Yang. Mri-only based synthetic ct generation using dense cycle consistent generative adversarial networks. *Medical physics*, 46(8):3565–3581, 2019.

[71] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, et al. Dagan: deep de-aliasing generative adversarial networks for fast com-

pressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2017.

[72] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference*, pages 464–483, 2020.

[73] Ho Bae, Dahuin Jung, Hyun-Soo Choi, and Sungroh Yoon. Anomigan: Generative adversarial networks for anonymizing private medical data. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pages 563–574. World Scientific, 2019.

[74] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[75] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

[76] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[77] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[78] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. In *International Conference on Machine Learning*, pages 4006–4015. PMLR, 2017.

[79] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*,

pages 214–223. PMLR, 2017.

[80] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[81] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv preprint arXiv:2003.05991*, 2020.

[82] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[83] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[84] A Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. 2018. *URL https://arxiv.org/pdf/1807.03653. pdf.*

[85] Beata Nowok, Gillian M Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in r. *Journal of statistical software*, 74:1–26, 2016.

[86] Matthias Templ, Bernhard Meindl, Alexander Kowarik, and Olivier Dupriez. Simulation of synthetic complex data: The r package simpop. *Journal of Statistical Software*, 79:1–38, 2017.

[87] David Heckerman. A tutorial on learning with bayesian networks. *Innovations in Bayesian networks*, pages 33–82, 2008.

[88] Zhifa Liu, Brandon Malone, and Changhe Yuan. Empirical evaluation of scoring functions for bayesian network model selection. In *BMC bioinformatics*, volume 13, pages 1–16. Springer, 2012.

[89] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan kaufmann, 1988.

[90] Kevin B Korb, Lucas R Hope, Ann E Nicholson, and Karl Axnick. Varieties of causal intervention. In *Pacific Rim international conference on artificial intelligence*, pages 322–331. Springer, 2004.

[91] Chengwei Su, Angeline Andrew, Margaret R Karagas, and Mark E Borsuk. Using bayesian networks to discover relations between genes, environment, and disease. *BioData mining*, 6(1):1–21, 2013.

[92] Ildikó Flesch and Peter JF Lucas. Markov equivalence in bayesian networks. *Advances in probabilistic graphical models*, pages 3–38, 2007.

[93] Federico Castelletti, Guido Consonni, Marco L Della Vedova, and Stefano Peluso. Learning markov equivalence classes of directed acyclic graphs: an objective bayes approach. *Bayesian Analysis*, 13(4):1235–1260, 2018.

[94] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[95] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19:2, 2000.

[96] Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

[97] L Antonio Pereira Silva, J Batista Nunes Bezerra, M Barbosa Perkusich, K Costa Gorgônio, H Oliveira de Almeida, and Angelo Perkusich. *Continuous learning of the structure of Bayesian networks: A Mapping Study*. IntechOpen, 2019.

[98] Xiao-guang Gao, Zhi-gao Guo, Hao Ren, Yu Yang, Da-qing Chen, and Chuchao He. Learning bayesian network parameters via minimax algorithm. *International Journal of Approximate Reasoning*, 108:62–75, 2019.

[99] Zhiwei Ji, Qibiao Xia, and Guanmin Meng. A review of parameter learning methods in bayesian network. In *International Conference on Intelligent Computing*, pages 3–12. Springer, 2015.

[100] Segev Wasserkrug, Radu Marinescu, Sergey Zeltyn, Evgeny Shindin, and Yishai A Feldman. Learning the parameters of bayesian networks from uncertain data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12190–12197, 2021.

[101] Yu Hong, Xiaoling Xia, Jiajin Le, and Xiangdong Zhou. Learning bayesian network structure from large-scale datasets. In *2016 International Conference on Advanced Cloud and Big Data (CBD)*, pages 258–264. IEEE, 2016.

[102] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[103] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010.

[104] Doug Fisher, Hans-J Lenz, Hans-J Lenz, et al. *Learning from data: artificial intelligence and statistics V*, volume 5. Springer Science & Business Media, 1996.

[105] Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[106] Fred Glover. Tabu search: A tutorial. *Interfaces*, 20(4):74–94, 1990.

[107] Nir Friedman, Iftach Nachman, and Dana Pe'er. Learning bayesian network structure from massive datasets: The" sparse candidate" algorithm. *arXiv preprint arXiv:1301.6696*, 2013.

[108] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

[109] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[110] Wray Buntine. Theory refinement on bayesian networks. In *Uncertainty proceedings 1991*, pages 52–60. Elsevier, 1991.

[111] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[112] José A Gámez, Juan L Mateo, and José M Puerta. Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*, 22(1):106–148, 2011.

[113] Marco Scutari and Robert Ness. bnlearn: Bayesian network structure learning, parameter learning and inference. *R package version*, 3:805, 2012.

[114] Robert W Robinson. Counting unlabeled acyclic digraphs. In *Combinatorial mathematics V*, pages 28–43. Springer, 1977.

[115] Zhuang Zhang, Jie Zhang, Zhen Wei, Hao Ren, Weimei Song, Jinhua Pan, Jinchun Liu, Yanbo Zhang, and Lixia Qiu. Application of tabu search-based bayesian networks in exploring related factors of liver cirrhosis complicated with hepatic encephalopathy and disease identification. *Scientific reports*, 9(1):1–8, 2019.

[116] Stefano Beretta, Mauro Castelli, Ivo Gonçalves, Roberto Henriques, and Daniele Ramazzotti. Learning the structure of bayesian networks: A quantitative assessment of the effect of different algorithmic schemes. *Complexity*, 2018, 2018.

[117] T Verma and Judea Pearl. Equivalence and synthesis of causal models proceedings of the sixth annual conference on uncertainty in artificial intelligence, 1991.

[118] Marco Scutari. Bayesian network constraint-based structure learning algorithms: Parallel and optimised implementations in the bnlearn r package. *arXiv preprint arXiv:1406.7648*, 2014.

[119] Dimitris Margaritis. Learning bayesian network model structure from data. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2003.

[120] Christopher Meek. Causal inference and causal explanation with background knowledge. *arXiv preprint arXiv:1302.4972*, 2013.

[121] Eran Segal, Dana Pe'er, Aviv Regev, Daphne Koller, Nir Friedman, and Tommi Jaakkola. Learning module networks. *Journal of Machine Learning Research*, 6(4), 2005.

[122] Meemansa Sood, Akrishta Sahay, Reagon Karki, Mohammad Asif Emon, Henri Vrooman, Martin Hofmann-Apitius, and Holger Fröhlich. Realistic simulation of virtual multi-scale, multi-modal patient trajectories using bayesian networks and sparse auto-encoders. *Scientific reports*, 10(1):1–14, 2020.

[123] Derrick A Bennett. How can i deal with missing data in my study? *Australian and New Zealand journal of public health*, 25(5):464–469, 2001.

[124] Joseph G Ibrahim and Geert Molenberghs. Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–43, 2009.

[125] Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402, 2013.

[126] Daniel O Scharfstein, Joseph Hogan, and Amir Herman. Randomized trials in orthopaedic surgery: advances and future directions: on the prevention and analysis of missing data in randomized clinical trials: the state of the art. *The Journal of Bone and Joint Surgery. American volume*, 94(Suppl 1):80, 2012.

[127] Peter C Austin, Ian R White, Douglas S Lee, and Stef van Buuren. Missing data in clinical research: a tutorial on multiple imputation. *Canadian Journal of Cardiology*, 37(9):1322–1331, 2021.

[128] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[129] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, 19(1):1–14, 2018.

[130] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[131] Roberta Balestrino and AHV Schapira. Parkinson disease. *European journal of neurology*, 27(1):27–42, 2020.

[132] Eldbjørg Hustad and Jan O Aasly. Clinical and imaging markers of prodromal parkinson's disease. *Frontiers in Neurology*, 11:395, 2020.

[133] Ronald B Postuma and Daniela Berg. Advances in markers of prodromal parkinson disease. *Nature Reviews Neurology*, 12(11):622–634, 2016.

[134] Ronald B Postuma, Dag Aarsland, Paolo Barone, David J Burn, Christopher H Hawkes, Wolfgang Oertel, and Tjalf Ziemssen. Identifying prodromal parkinson's disease: pre-motor disorders in parkinson's disease. *Movement Disorders*, 27(5):617–626, 2012.

[135] Ronald B Postuma and Daniela Berg. Prodromal parkinson's disease: the decade past, the decade to come. *Movement disorders*, 34(5):665–675, 2019.

[136] Daniela Berg, Ronald B Postuma, Charles H Adler, Bastiaan R Bloem, Piu Chan, Bruno Dubois, Thomas Gasser, Christopher G Goetz, Glenda Halliday, Lawrence Joseph, et al. Mds research criteria for prodromal parkinson's disease. *Movement Disorders*, 30(12):1600–1611, 2015.

[137] Daniela Berg, Per Borghammer, Seyed-Mohammad Fereshtehnejad, Sebastian Heinzel, Jacob Horsager, Eva Schaeffer, and Ronald B Postuma. Prodro-

mal parkinson disease subtypes—key to understanding heterogeneity. *Nature Reviews Neurology*, 17(6):349–361, 2021.

[138] Alessandro Tessitore, Mario Cirillo, and Rosa De Micco. Functional connectivity signatures of parkinson's disease. *Journal of Parkinson's disease*, 9(4):637–652, 2019.

[139] Richard Nathaniel Rees, Alastair John Noyce, and Anette Schrag. The prodromes of parkinson's disease. *European Journal of Neuroscience*, 49(3):320–327, 2019.

[140] Moran Artzi, Einat Even-Sapir, Hedva Lerman Shacham, Avner Thaler, Avi Orr Urterger, Susan Bressman, Karen Marder, Talma Hendler, Nir Giladi, Dafna Ben Bashat, et al. Dat-spect assessment depicts dopamine depletion among asymptomatic g2019s lrrk2 mutation carriers. *PloS one*, 12(4):e0175424, 2017.

[141] Kan Li and Sheng Luo. Functional joint model for longitudinal and time-to-event data: an application to alzheimer's disease. *Statistics in medicine*, 36(22):3560–3572, 2017.

[142] Shashank Khanna, Daniel Domingo-Fernández, Anandhi Iyappan, Mohammad Asif Emon, Martin Hofmann-Apitius, and Holger Fröhlich. Using multi-scale genetic, neuroimaging and clinical data for predicting alzheimer's disease and reconstruction of relevant biological mechanisms. *Scientific reports*, 8(1):1–13, 2018.

[143] Boris Hayete, Diane Wuest, Jason Laramie, Paul McDonagh, Bruce Church, Shirley Eberly, Anthony Lang, Kenneth Marek, Karl Runge, Ira Shoulson, et al. A bayesian mathematical model of motor and cognitive outcomes in parkinson's disease. *PLoS One*, 12(6):e0178982, 2017.

[144] Yue Qiu, Liang Li, Tian-yan Zhou, and Wei Lu. Alzheimer's disease progression model based on integrated biomarkers and clinical measures. *Acta Pharmacologica Sinica*, 35(9):1111–1120, 2014.

[145] Jorge L Bernal-Rusiel, Douglas N Greve, Martin Reuter, Bruce Fischl, Mert R Sabuncu, Alzheimer's Disease Neuroimaging Initiative, et al. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *Neuroimage*, 66:249–260, 2013.

[146] Daniela J Conrado, Timothy Nicholas, Kuenhi Tsai, Sreeraj Macha, Vikram Sinha, Julie Stone, Brian Corrigan, Massimo Bani, Pierandrea Muglia, Ian A Watson, et al. Dopamine transporter neuroimaging as an enrichment biomarker in early parkinson's disease clinical trials: a disease progression modeling analysis. *Clinical and translational science*, 11(1):63–70, 2018.

[147] Christine A Bevc, Jessica H. Retrum, and Danielle M. Varda. New perspectives on the "silo effect": initial comparisons of network structures across public health collaboratives. *American journal of public health*, 105(S2):S230–S235, 2015.

[148] Ilse Van Roessel, Matthias Reumann, and Angela Brand. Potentials and challenges of the health data cooperative model. *Public health genomics*, 20(6):321–331, 2017.

[149] Francesco Pappalardo, Giulia Russo, Flora Musuamba Tshinanu, and Marco Viceconti. In silico clinical trials: concepts and early adoptions. *Briefings in bioinformatics*, 20(5):1699–1708, 2019.

[150] Rezzak Yilmaz, Ulrike Suenkel, TREND Study Team, Ronald B Postuma, Sebastian Heinzel, and Daniela Berg. Comparing the two prodromal parkinson's disease research criteria—lessons for future studies. *Movement Disorders*, 36(7):1731–1732, 2021.

[151] Sebastian Heinzel, Daniela Berg, Thomas Gasser, Honglei Chen, Chun Yao, Ronald B Postuma, and MDS Task Force on the Definition of Parkinson's Disease. Update of the mds research criteria for prodromal parkinson's disease. *Movement Disorders*, 34(10):1464–1470, 2019.

[152] Zoubin Ghahramani. Learning dynamic bayesian networks. In *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pages 168–197. Springer, 1997.

[153] Sebastian Heinzel, Velma TE Aho, Ulrike Suenkel, Anna-Katharina von Thaler, Claudia Schulte, Christian Deuschle, Lars Paulin, Sari Hantunen, Kathrin Brockmann, Gerhard W Eschweiler, et al. Gut microbiome signatures of risk and prodromal markers of parkinson disease. *Annals of neurology*, 90(3):E1–E12, 2021.

[154] Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with bayesian networks: A bootstrap approach. *arXiv preprint arXiv:1301.6695*, 2013.

[155] Shuyi Ji, Zizhao Zhang, Shihui Ying, Liejun Wang, Xibin Zhao, and Yue Gao. Kullback-leibler divergence metric learning. *IEEE Transactions on Cybernetics*, 2020.

[156] Michael Greenacre and Jorg Blasius. *Multiple correspondence analysis and related methods*. Chapman and Hall/CRC, 2006.

[157] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[158] Luise Gootjes-Dreesbach, Meemansa Sood, Akrishta Sahay, Martin Hofmann-Apitius, and Holger Fröhlich. Variational autoencoder modular bayesian networks (vambn) for simulation of heterogeneous clinical study data. *BioRxiv*, page 760744, 2019.

[159] Arthur W Toga and Karen L Crawford. The alzheimer's disease neuroimaging initiative informatics core: a decade in review. *Alzheimer's & Dementia*, 11(7):832–839, 2015.

[160] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[161] Tamás D Gedeon. Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems*, 8(02):209–218,

1997.

[162] Usama Fayyad and Keki Irani. Multi-interval discretization of continuous-valued attributes for classification learning. 1993.

[163] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253, 2019.

[164] Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

[165] Vineet K Raghu, Allen Poon, and Panayiotis V Benos. Evaluation of causal structure learning methods on mixed data types. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, pages 48–65. PMLR, 2018.

[166] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[167] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1–8, 2011.

[168] Max Henrion. Propagating uncertainty in bayesian networks by probabilistic logic sampling. In *Machine intelligence and pattern recognition*, volume 5, pages 149–163. Elsevier, 1988.

[169] Changhe Yuan and Marek J Druzdzel. Importance sampling algorithms for bayesian networks: Principles and performance. *Mathematical and Computer Modelling*, 43(9-10):1189–1207, 2006.

[170] Vijay K Ramanan, Shannon L Risacher, Kwangsik Nho, Sungeun Kim, Shanker Swaminathan, Li Shen, Tatiana M Foroud, Hakon Hakonarson, Matthew J Huentelman, Paul S Aisen, et al. Apoe and bche as modulators of cerebral amyloid deposition: a florbetapir pet genome-wide association study. *Molecular psychiatry*, 19(3):351–357, 2014.

[171] Claudia Ramaker, Johan Marinus, Anne Margarethe Stiggelbout, and Bob Johannes Van Hilten. Systematic evaluation of rating scales for impairment and disability in parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 17(5):867–876, 2002.

[172] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[173] David Heckerman. Bayesian networks for data mining. *Data mining and knowledge discovery*, 1(1):79–119, 1997.

[174] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer, 2006.

[175] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020.

[176] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166–176, 2003.

[177] Bryan Andrews, Joseph Ramsey, and Gregory F Cooper. Scoring bayesian networks of mixed variables. *International journal of data science and analytics*, 6(1):3–18, 2018.

[178] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[179] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application.* Number 1. Cambridge university press, 1997.

[180] Ruth Peters. Ageing and the brain. *Postgraduate medical journal*, 82(964):84–88, 2006.

[181] Lisa L Barnes, Robert S Wilson, Julia L Bienias, Julie A Schneider, Denis A Evans, and David A Bennett. Sex differences in the clinical manifestations of alzheimer disease pathology. *Archives of general psychiatry*, 62(6):685–691, 2005.

[182] Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. Digital biomarkers for alzheimer's disease: the mobile/wearable devices opportunity. *NPJ digital medicine*, 2(1):1–9, 2019.

[183] Antoine Piau, Katherine Wild, Nora Mattek, Jeffrey Kaye, et al. Current state of digital biomarker technologies for real-life, home-based monitoring of cognitive function for mild cognitive impairment to mild alzheimer disease and implications for clinical care: systematic review. *Journal of medical Internet research*, 21(8):e12785, 2019.

[184] Sergey O Bachurin, Svetlana I Gavrilova, Anna Samsonova, George E Barreto, and Gjumrakch Aliev. Mild cognitive impairment due to alzheimer disease: Contemporary approaches to diagnostics and pharmacological intervention. *Pharmacological research*, 129:216–226, 2018.

[185] Anna-Mariya Kirova, Rebecca B Bays, and Sarita Lagalwar. Working memory and executive function decline across normal aging, mild cognitive impairment, and alzheimer's disease. *BioMed research international*, 2015, 2015.

[186] Ronald C Petersen. Mild cognitive impairment. *CONTINUUM: Lifelong Learning in Neurology*, 22(2 Dementia):404, 2016.

[187] Ingrid Arevalo-Rodriguez, Nadja Smailagic, Marta Roqué i Figuls, Agustín Ciapponi, Erick Sanchez-Perez, Antri Giannakou, Olga L Pedraza, Xavier Bonfill Cosp, and Sarah Cullum. Mini-mental state examination (mmse) for the detection of alzheimer's disease and other dementias in people with mild cognitive impairment (mci). *Cochrane Database of Systematic Reviews*, (3), 2015.

[188] Gad A Marshall, Dorene M Rentz, Meghan T Frey, Joseph J Locascio, Keith A Johnson, Reisa A Sperling, Alzheimer's Disease Neuroimaging Initiative, et al. Executive function and instrumental activities of daily living in mild cognitive impairment and alzheimer's disease. *Alzheimer's & Dementia*, 7(3):300–308, 2011.

[189] Marie Mc Carthy, Darragh Walsh, and Jamie Tallon. Can wearables and sensor data be used to add context to activities of daily living questionnaires?(poc). 2018.

[190] Marie Mc Carthy and P Schueler. can digital technology advance the development of treatments for alzheimer's disease?, 2019.

[191] Melinda M Davis, Michele Freeman, Jeffrey Kaye, Nancy Vuckovic, and David I Buckley. A systematic review of clinician and staff views on the acceptability of incorporating remote monitoring technology into primary care. *Telemedicine and e-Health*, 20(5):428–438, 2014.

[192] Marijn Muurling, Casper de Boer, Rouba Kozak, Dorota Religa, Ivan Koychev, Herman Verheij, Vera JM Nies, Alexander Duyndam, Meemansa Sood, Holger Fröhlich, et al. Remote monitoring technologies in alzheimer's disease: design of the radar-ad study. *Alzheimer's research & therapy*, 13(1):1–13, 2021.

[193] Maximilian Buegler, Robbert L Harms, Mircea Balasa, Irene B Meier, Themis Exarchos, Laura Rai, Rory Boyle, Adria Tort, Maha Kozori, Eutuxia Lazarou, et al. Digital biomarker-based individualized prognosis for people at risk of dementia. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 12(1):e12073, 2020.

[194] P Murali Doraiswamy, Vaibhav A Narayan, and Husseini K Manji. Mobile and pervasive computing technologies and the future of alzheimer's clinical trials. *Npj Digital Medicine*, 1(1):1–4, 2018.

[195] Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: to-

ward a biological definition of alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

[196] Mark W Bondi, Emily C Edmonds, Amy J Jak, Lindsay R Clark, Lisa Delano-Wood, Carrie R McDonald, Daniel A Nation, David J Libon, Rhoda Au, Douglas Galasko, et al. Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *Journal of Alzheimer's Disease*, 42(1):275–289, 2014.

[197] Douglas J Gelb. Measurement of progression in alzheimer's disease: a clinician's perspective. *Statistics in medicine*, 19(11-12):1393–1400, 2000.

[198] Gwanghee Han, Michio Maruta, Yuriko Ikeda, Tomohisa Ishikawa, Hibiki Tanaka, Asuka Koyama, Ryuji Fukuhara, Shuken Boku, Minoru Takebayashi, and Takayuki Tabira. Relationship between performance on the mini-mental state examination sub-items and activities of daily living in patients with alzheimer's disease. *Journal of Clinical Medicine*, 9(5):1537, 2020.

[199] Young Min Choe, Boung Chul Lee, Ihn-Geun Choi, Guk-Hee Suh, Dong Young Lee, Jee Wook Kim, Alzheimer's Disease Neuroimaging Initiative, et al. Mmse subscale scores as useful predictors of ad conversion in mild cognitive impairment. *Neuropsychiatric Disease and Treatment*, 16:1767, 2020.

[200] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284–1287, 2016.

[201] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279, 2016.

[202] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H

Ramirez, Erica Bowton, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*, 31(12):1102–1111, 2013.

[203] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855, 2020.

[204] Lucas D Ward and Manolis Kellis. Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, 40(D1):D930–D934, 2012.

[205] Daniel Domingo-Fernández, Alpha Tom Kodamullil, Anandhi Iyappan, Mufassra Naz, Mohammad Asif Emon, Tamara Raschka, Reagon Karki, Stephan Springstubbe, Christian Ebeling, and Martin Hofmann-Apitius. Multimodal mechanistic signatures for neurodegenerative diseases (neurommsig): a web server for mechanism enrichment. *Bioinformatics*, 33(22):3679–3681, 2017.

[206] Sebastian Palmqvist, Philip S Insel, Henrik Zetterberg, Kaj Blennow, Britta Brix, Erik Stomrud, Niklas Mattsson, Oskar Hansson, Alzheimer's Disease Neuroimaging Initiative, et al. Accurate risk estimation of β-amyloid positivity to identify prodromal alzheimer's disease: cross-validation study of practical algorithms. *Alzheimer's & Dementia*, 15(2):194–204, 2019.

[207] Sayeh Bayat, Ganesh M Babulal, Suzanne E Schindler, Anne M Fagan, John C Morris, Alex Mihailidis, and Catherine M Roe. Gps driving: a digital biomarker for preclinical alzheimer disease. *Alzheimer's Research & Therapy*, 13(1):1–9, 2021.

[208] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 9(3):e1301, 2019.

[209] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[210] Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

[211] Marvin N Wright, Theresa Dankowski, and Andreas Ziegler. Random forests for survival analysis using maximally selected rank statistics. *arXiv preprint arXiv:1605.03391*, 2016.

[212] Gavin C Cawley and Nicola L C Talbot. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Technical report, 2010.

[213] Fredrik Öhman, Jason Hassenstab, David Berron, Michael Schöll, and Kathryn V Papp. Current advances in digital cognitive assessment for preclinical alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 13(1):e12217, 2021.

[214] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20(1):1–40, 2020.

[215] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[216] Yubin Park, Joydeep Ghosh, and Mallikarjun Shankar. Perturbed gibbs samplers for generating large-scale privacy-safe synthetic health data. In *2013 IEEE International Conference on Healthcare Informatics*, pages 493–498. IEEE, 2013.

[217] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[218] Philipp Wendland, Colin Birkenbihl, Marc Gomez-Freixa, Meemansa Sood, Maik Kschischo, and Holger Fröhlich. Generation of realistic synthetic data using multimodal neural ordinary differential equations. *NPJ digital medicine*, 5(1):1–10, 2022.

[219] Jeremy Georges-Filteau and Elisa Cirillo. Synthetic observational health data with gans: from slow adoption to a boom in medical research and ultimately digital twins? *arXiv preprint arXiv:2005.13510*, 2020.

[220] Byeong Soo Kim, Bong Gu Kang, Seon Han Choi, and Tag Gon Kim. Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system. *Simulation*, 93(7):579–594, 2017.

[221] Daniel Bonnéry, Yi Feng, Angela K Henneberger, Tessa L Johnson, Mark Lachowicz, Bess A Rose, Terry Shaw, Laura M Stapleton, Michael E Woolley, and Yating Zheng. The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal of Research on Educational Effectiveness*, 12(4):616–647, 2019.

[222] David Hand. What is the purpose of statistical modeling? 2019.

[223] Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, Gorka Epelde, et al. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Medical Informatics*, 8(7):e18910, 2020.

[224] Junqiao Chen, David Chun, Milesh Patel, Epson Chiang, and Jesse James. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (synthea) using clinical quality measures. *BMC medical informatics and decision making*, 19(1):1–9, 2019.

[225] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.