Clustering and prediction of electronic health record data from mental health patients in a Finnish healthcare environment

Oskar Niemenoja

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.
Helsinki 3.6.2019

Supervisor

Prof. Jari Saramäki

Advisor

Dr. Anna-Maria Lahesmaa-Korpinen

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Oskar Niemenoja

Title Clustering and prediction of electronic health record data from mental health patients in a Finnish healthcare environment

Degree programme Life Science Technologies

Major Complex Systems **Code of major** SCI3060

Supervisor Prof. Jari Saramäki

Advisor Dr. Anna-Maria Lahesmaa-Korpinen

Date 3.6.2019 **Number of pages** 47 **Language** English

Abstract

Electronic health records contain a wealth of information of interest to both the patient and the service provider but are historically not designed for easy computational analysis. In this study, we introduce the concept of *treatment pathways* as treatment sessions related to one single initial diagnosis. We explore three mixed-data clustering methods on the mental health patients from the Finnish occupational health population from one health service provider, and identify these treatment paths from electronic health record data. Based on these clusters we create two predictive models to predict treatment pathway length and duration of any possible sick leave of a patient.

We demonstrate how these clustering and predictive models work on health record data and validate the results statistically and with expert evaluation. We show that different clustering methods produce very different outcomes in terms of the size and number of different diagnoses contained in a treatment pathway. The expert-evaluated error rates for these models range from 0.66% to 39.83% for the number of appointments that are incorrectly clustered. The predictive models are shown to be an adequate tool to predict the lengths of the treatment pathway and sick leave. Additionally, these methods perform well at identifying unusually large values for these measures, making them useful in identifying patients at high risk early on in the treatment pathway.

Overall, the study demonstrates the feasibility of the selected methods on large-scale electronic health record data, provides results for clustering and assessing the quality of these clusters and serves as a base for predictive models based on these clusters. The results overall are promising and function as an initial study into further structuring and predicting medical data on a large scale.

Keywords Electronic Health Record, mental health, machine learning, clustering, regression, treatment pathway

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Oskar Niemenoja

Työn nimi Potilastietojärjestelmän aineiston klusterointi ja ennustaminen suomalaisessa terveydenhoitoympäristössä mielenterveyspotilaille

Koulutusohjelma Life Science Technologies

Pääaine Complex Systems **Pääaineen koodi** SCI3060

Työn valvoja Prof. Jari Saramäki

Työn ohjaaja LkT. Anna-Maria Lahesmaa-Korpinen

Päivämäärä 3.6.2019 **Sivumäärä** 47 **Kieli** Englanti

Tiivistelmä

Potilastietojärjestelmät sisältävät paljon erilaista tietoa, jonka hyödyntäminen on sekä potilaan että terveyspalveluiden tuottajan etu. Näitä järjestelmiä on kuitenkin harvoin tuotettu data-analyysin mahdollistamiseksi. Tässä tutkimuksessa esittelemme *hoitopolun* käsitteen, joka sisältää kaikki hoitotoimenpiteet yhteen ensidiagnoosiin liittyen. Tutkimme kolmea erilaista klusterointimenetelmää Suomen työterveyshuollon piirissä oleville mielenterveyspotilaille yhden palveluntuottajan piirissä sekä menetelmien soveltuvuutta potilastietojärjestelmien aineistoille. Näiden pohjalta luomme kahdella menetelmällä ennustemallit, joissa mallinnetaan hoitoketjun pituutta sekä mahdollista potilaan sairaspoissaolojakson pituutta.

Osoitamme, miten nämä klusterointi- ja ennustemenetelmät toimivat potilastietoaineistolla ja arvioimme tulokset tilastollisesti sekä asiantuntija-arvioin. Tutkimuksessa näytämme, että erilaiset klusterointimenetelmät tuottavat hyvin erilaisia tuloksia sekä klustereiden koon että niiden sisältämien diagnoosien toimesta. Virheluvut näille ovat asiantintija-arvion mukaan pienimmillään 0.66% ja suurimmillaan 39.83%, kun arvioitiin virheellisesti klusteroitujen käyntien osuutta koko aineistosta. Ennustemallit todettiin toimivaksi työkaluksi ennustettaessa hoitopolkujen ja potilaan sairaspoissaolojakson pituutta. Erityisen hyvin nämä mallit tunnistavat poikkeuksellisen pitkiä jaksoja tästä aineistosta, jolloin ne soveltultuvat hyvin erityisen suuressa riskissä olevien potilaiden seulontaan hoitoketjun aikaisessa vaiheessa.

Kokonaisuudessaan tutkimus esittelee valittujen menetelmien soveltuvuutta suuren mittaluokan potilastietojärjestelmäaineistolle, tuloksia tiedon klusteroinnille ja menetelmät niiden arviointiin sekä tutkii ennustemallien käytettävyyttä hoitoketjujen ominaisuuksien arvioinnissa. Nämä tulokset ovat kaikkiaan lupaavia ja toimivat pohjana jatkotutkimuksille tutkittaessa terveystiedon strukturointia sekä ennustettavuutta laajemmassa mittakaavassa.

Avainsanat potilastietojärjestelmä, mielenterveys, klusterointi, koneoppiminen, klusterointi, regressio, hoitopolku

Acknowledgements

My heartfelt thanks to my professor and supervisor Jari Saramäki for his words of advice and always having time and interest in my work and studies. You are without a doubt one of the best teachers I had and made writing this Thesis all the more enjoyable.

Anna-Maria, your guidance as my advisor was incredible. Thank you for all your help without which this Thesis would not have been possible. Your enthusiasm and support are an example to others. A special thank you to Elina for her contribution whenever she was present.

Pia, Eeva, Tuija, and Hanna, your support, interest, words of guidance, and encouragement during this work have been invaluable. The support from the company has been overwhelming, and I very much appreciate each of you donating your valuable time for this project. Tuija and Anita, a special thank you for your contribution via interviews and reviews of the data. Such help has been extraordinary, and your expertise is incredible.

Amanda, thank you for all your love and support along the way. My family and friends, I feel truly blessed with such a wonderful group of people with whom to share my journey. To Aalto university and all the teachers and assistants along the way, thank you for giving me the knowledge and guidance to be where I am now. All the beautiful people at Aalto University student union and the bioinformatics guild Inkubio, rock on. You have made these years the best I could have hoped for.

Helsinki, 3rd June 2019

Oskar Mikael Niemenoja

Contents

Abbreviations

Healthcare

ALS	Amyotrophic Lateral Sclerosis, motor neurone disease
ATC	Anatomical Therapeutic Chemical Classification System
BDI	*Beck Depression Inventory* questionnaire
EHR	Electronic Health Record
ICD-9 / ICD-10	International Classification of Diagnoses, 9th / 10th version
ICPC-2	International Classification of Primary Care, 2nd version

Machine learning

CART	Classification and Regression Tree
ILP	Inductive Logic Programming
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
NLP	Natural Language Processing
RF	Random Forest
SMC	Simple Matching Coefficient
SVM	Support Vector Machine
SVR	Support Vector Regression

1 Introduction

Electronic health records contain a wealth of data that could be used for insight into patient relationships for healthcare providers. Most electronic health records are, however, designed to be used mainly in patient care and for storing relevant data and metadata for each patient. The data are often multi-relational and lack a suitable structure for analysis around any one issue. As such, the records can be ill-suited for traditional data analysis. An interesting application for healthcare providers is the analysis of the relevant personal treatment sessions related to a single initial diagnosis. Producing such results needs a model that can both identify treatment sessions as associated with a single diagnosis and then use these identified clusters to predict variables of interest. In this study, we explore different options for producing such a model.

Another feature of electronic health data is the mixed nature of the data. In other words, the data are often represented both numerically and categorically, both carrying relevant information for each problem in question. In the case of Finnish electronic health records, the data can also be quite large in size. Both of these features make the data unique to work with and need special attention for machine learning applications.

Machine learning is a subfield of artificial intelligence used to predict specific patterns and features in the data. Machine learning algorithms can be broadly fit into three categories: *supervised learning* for data that has prior outcome variables in the training data, *unsupervised learning* for data that has no prior labels in the training data and *active learning* in which a human will aid the model in predicting the labels. These can be further broken down to more specific fields, for example, *semi-supervised learning* for combining data with some but not all outcome variables present or *reinforcement learning* in which the model is given feedback on its performance in prior classifications. In this study, we will be using unsupervised and supervised machine learning methods to cluster and predict variables related to different treatment pathways. [1]

The goal of this Thesis was to study the feasibility of various clustering and predictive methods for analyzing mental health treatment pathways when applied to a large-scale electronic health record database. Based on these, we proposed a framework for clustering various health records to clusters called *treatment pathways* for mental well-being. We showed that multiple clustering methods provided a viable tool for analyzing electronic health records, but the performance and quality of such clustering methods left further room for improvement. We also used these clusters in predictive analysis on the data to predict sick leave and treatment path length and outliers. The results were promising, with good results especially in outlier detection, but at the same time, the desired outcome variables related to the clusters were very unequally distributed. Thus most of the predictions were similar due to a small number of large outliers. Still, given the large scale of the data and limited previous studies on the subject, this study serves as a successful initial survey into the topic of large-scale EHR data analysis.

2 Background

Mental health-related problems were the second most significant reason for occupational absence in Finland in 2017 [2]. Around 5% of the general adult population has clinical depression, with the prevalence rising to about 10% when only considering people in primary care in Finland (perusterveydenhuolto) [3]. The patients often require highly differentiated forms of care, as the individual responses to different medication and psychotherapy vary greatly. The national recommendations for good care (Käypä hoito -suositukset) recommend a mix of medication and psychotherapy for most cases of depression and mental issues [3]. However, these guidelines are not always followed through, and the electronic health records (EHR) collected from such treatment are often poorly suited for analyzing and monitoring the effectiveness of care. Moreover analyzing health records is, in general, a challenging area for conventional machine learning methods as the databases are often large, multidimensional, relational, and contain multiple types of data.

Medical records in an electronic database are often not grouped on a per-diagnosis -basis but are instead arranged in a multidimensional relative model with limited links between them. These data may not have any markers denoting which of the diagnoses one particular session is related to, and one session may be used to check up on multiple diagnoses. To effectively monitor treatment related to each specific diagnosis, we must be able to first categorize different treatment sessions as related to a particular initial diagnosis. We define this grouping as the "treatment pathway" of the patient for each specific diagnosis. Figure 1 shows which data are considered when deciding if the items fall under the same treatment pathway. Generally, all data related to the treatment of a single condition are seen as belonging to this pathway, including the number and occupation of treatment personnel seen, length of treatment in days, and locations of individual treatment sessions with possible self-help and virtual sessions.

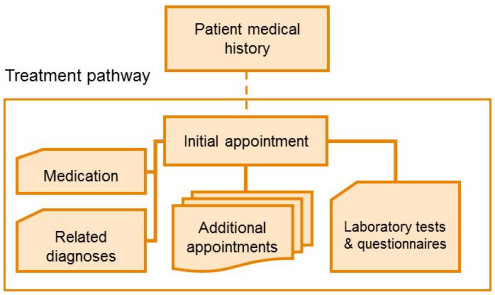

Figure 1: The data that are seen as belonging to the same treatment pathway.

Additionally, any prescribed medication, laboratory tests and questionnaires such as the Beck Depression Inventory (BDI) response patterns can be stored. The treatment path can also contain additional diagnoses along with the initial diagnosis if these are related to the initial disease. Such diagnoses can either be somatic

diseases, meaning a physical symptom arising from a mental health condition, or related mental health issues that are treated as one. The initial diagnosis may also be clarified to better suit the underlying disease by assigning a more specific diagnosis to replace a more general one.

Any items with a close relation to each other concerning the described metrics are labelled as belonging to the same treatment pathway. Medically speaking, given an initial diagnosis for any mental health-related problem, we aim to find and predict the response and future visits to medical personnel, future diagnoses, possible medication, and linked treatment. As a matter of example, we may imagine a patient having multiple diagnoses for different conditions with dozens of visits to a clinic between them. Given a sufficiently experienced doctor, the task of grouping these into meaningful categories in which each visit is linked to an initial diagnosis is a relatively easy task. However, given the enormous scale of the typical EHR system, this is not feasible beyond a small number of patients. Machine learning methods give us a cost-efficient alternative to analyzing, clustering and predicting treatment pathways.

Given a sufficiently efficient and accurate computational method for identifying patient treatment pathways in EHR data, medical personnel could better measure and predict the effectiveness of care for each patient. This also enables us to better analyse large groups of patients retrospectively to study the quality of care and adherence to general guidelines for patients on a more general level. The method can also be used to study other metrics from the data such as medicine performance, abnormal behaviour of medical personnel, or possible undiagnosed conditions.

Predicting future treatment pathways for patients currently receiving care also enables effective planning of personnel workload and improves patient care. Kela, the national social benefits provider in Finland, paid out sick leave benefits for mental health issues for a total of 195 million euros in 2017, a quarter of all the sick leave benefits paid [2]. According to an interview with the chief psychologist and psychotherapist at a large Finnish healthcare provider, PhD Tuija Turunen, in many cases, the initial visit of the patient for mental health issues is for reasons unrelated to those issues. Instead, the medical condition is noticed during other operations. Mental health problems can also manifest themselves as different conditions, such as pain or headache. Improving early detection for these types of cases improves care by eliminating unnecessary visits and steering the patient to the correct specialist sooner. Early detection also reduces the psychological distress of the patient, reduces the need for long sick leaves and is also preventative on a larger scale. An effective treatment on adults also supports the mental well-being of children and relatives as well. [2][3][4]

2.1 Electronic health records in Finland

The International Organisation for Standards defines electronic health records in its standard ISO/TR 20514:2005 as a “repository of information regarding the health status of a subject of care, in computer processable form” [5]. Häyrinen, Saranto & Nykänen [6] have conducted a systematic approach to analysing EHR structure, definition, data types and access as well as related research. They concluded that

EHRs comprised a broad range of different information systems in primary, secondary and tertiary care and typically covered all medical reports related to one person for any one medical service provider as well as billing information and metadata. These may include personal data such as an address, gender, height, age, and treatment information such as past and ongoing medication, diagnoses, and possible specialized data such as laboratory results. The data are collected over long periods and are inherently sparse with emphasis on documenting existing medical conditions. As such, the data can be biased as these are sampled from predominantly non-healthy individuals. Data sets also usually cover only one medical service provider. In this case, an individual's treatment path may appear broken, if they have changed the medical service provider. EHRs can also contain different agreements between the patient and the health care provider based on their compliance with their data being used in studies as well as personal care. This Thesis only considers patients who have given their consent to use their medical data in academic studies.

The Finnish healthcare system is divided between public and private service providers. Still, even within the same sector, the transfer of patient records between any two units not in the same organization is limited. A national initiative to unify and collect medical records in Finland has been put forward by the government (Potilastiedon arkisto, Kanta-palvelu) to remedy this in which all national healthcare providers are directed to submit patient data to a national database [7]. The service combines electronic recipes and electronic health record data nationally, and users have the option to choose to share this data across all health service providers in Finland. However, the more common instances where patient data is traded are company acquisitions or mergers. Historical data acquired this way may differ significantly from the practices of the acquiring organization. The same data, such as blood pressure measurements, may be written in multiple ways even in the same database due to multiple data migrations from multiple sources throughout the database and different practices of the original organizations. This issue brings forth the need to unify and mine the data to structuralise it further, for example by implementing natural language processing (NLP) to various data fields to unify large data sources across time [8].

2.2 Mental health treatment procedures in occupational healthcare

The national outline for treatment procedures of mental health issues is outlined by the national recommendations for good care (Käypä hoito). In acute care for depression, the guidelines recommend both medicinal care and effective psychotherapy. These are equally effective in treating mild or moderate depression, but concurrent use is the most effective form of treatment. With more severe cases of depression, medicinal care is always needed. Persons with three or more diagnosed cases of depression are more susceptible to relapse, in which case long-term medicinal care is in order. In occupational healthcare, the central role is in preventive care and rehabilitation, in which the aim is to have the person returning to work as a form of treatment. [3]

The treatment of depression is divided into three clinical phases. The most effective care characterizes the first acute care phase, either medical, therapeutical or both and aims to remove symptoms of depression on the patient. The duration of this care is individual, but in 40-50% of cases, the symptoms are mitigated in 6 to 8 weeks. The next phase, follow-up care, continues for half a year after the acute care phase. During this phase, drug dosage is not reduced, and regular checkups and psychotherapy are continued. The purpose of this phase is to mitigate the possibility of a relapse. If the patient does not show any symptoms after half a year, the treatment can be discontinued. The third possible phase, maintenance treatment, is reserved to people with the reoccurring disease, usually with three or more cases of depression over their lifetime. The aim is to reduce the possibility of any future relapse. The treatment is done using the same dosage and treatment as with acute and follow-up care and can last for years. Beginning or ending this care is always done with the consultation of a doctor. [3]

The usual, most extensive treatment path usually begins with an unrelated routine health check in professional healthcare. Mental health issues are screened with annual or per-demand questionnaires along with other diseases or via discussion with the occupational nurse. The decrease in mental health is picked up by an occupational nurse or a doctor who forwards the patient to a psychologist. If the diagnosis is confirmed after 3 to 5 meetings, the patients are referred to the national support program for mental health via Kela that partakes in the cost of the treatment. The care continues for 1 to 3 years with guidelines for 80 meetings in the first and second year and 40 in the last alongside with medication. [4]

According to Turunen, the steps above are the most typical, but also an intensive way to treat mental health problems. However, early detection and timely access to treatment are the most critical parts of the treatment plan. Even a short intensive-care psychotherapy treatment in early phases of the diseases is proven to be rehabilitating and improve the condition of the patient early on. Thus, early detection of the disease and the guidance of patients to the right professionals are important in the treatment. The main goal with occupational healthcare with mild to moderate depression is enabling the person to return to work as soon as possible, as stable lifecycle and work-life are rehabilitating to the patient. [4]

Another interview with the medical director of occupational health of southern Finland at the same Finnish healthcare provider, Dr Anita Riipinen, sets the usual number of treatment sessions around 10-20 for the total length of the care. She stated that common problems with the treatment plans are misdiagnoses or understatement of the symptoms, in which mental health symptoms are diagnosed for example as general fatigue, sleep deprivation, or other general issues in fear of the stigma associated with a depression diagnosis. Mental health diagnoses can also be unwanted by the patients in mild cases, as any mental health diagnosis in patient history can affect external factors such as insurance cost and availability later in life. Thus some cases can purposefully be left undiagnosed, further complicating care. [9]

2.3 Previous studies on the subject

Machine learning applications have often been tested on medical data with small data sets. Few tests have been done concerning identifying or working with treatment pathways, however. Many tests have differing goals, and these are seldom shared between the studies.

Brett, Beaulieu-Jones *et al.* [10] and Miotto *et al.* [11] simultaneously utilized deep learning with electronic health records in stratifying personal phenotypes. They applied denoising autoencoders models to medical history data and genotypes to arrive at possible phenotypes for each patient. The tool used by Brett & Beaulieu-Jones was also used to predict the prevalence of ALS (amyotrophic lateral sclerosis, motor neurone disease) in the population. Both found that a denoising autoencoder applied to a sparse EHR data set yielded good approximations of the phenotype and was considered a cost-effective tool in analysing large masses of patients.

Prediction of the probability of a person having a particular disease from EHR data is a popular problem in data science. Banfield and Raftery [12] demonstrated the use of model-based clustering by analysing the effect of glucose treatment on diabetics. Some public data sets have been used by multiple studies. Hunt & Jorgensen [13] and van de Velden *et al.* [14] used a publicly available data set of 303 American heart disease patients to predict disease prevalence from EHR records with different clustering methods. Both studies explored the feasibility of distance-based clustering methods on mixed biological data. Zhao & Weng [15] used an unsupervised Bayesian network on an EHR enriched with outside PubMed data to approximate the prevalence of pancreatic cancer on a data set of patients, where the weighted joint model outperformed a purely Bayesian classifier as well as two other conventional classifiers, k-nearest neighbours and support vector machine.

Michiels *et al.* [16] used EHR data to predict flu outbreaks before the epidemic breakout, with their autoregressive Poisson likelihood model being able to approximate flu outbreak a week in advance and with reasonable probability in predicting the duration of the outbreak. Peissig *et al.* [17] used relational machine learning methods with inductive logic programming (ILP) to analyze if patients had any undiagnosed conditions present in their data. They concluded that relational learning is a viable approach to EHR-based phenotyping of diseases.

Multiple studies have been conducted concerning applying natural language processing approaches to EHR data. These studies mainly try to curate and process open data fields for the use of external systems or try to assess the severity of diagnosis based on open text fields written by doctors. Wang *et al.* [8] developed a novel approach to NLP to detect relevant free text fields linked to a diagnosis and gained a 64% precision with 87% recall in identifying relevant articles. Xia *et al.* [18] successfully applied similar methods to a larger sample with better accuracy and identified disease severity based on free text fields. Gøeg *et al.* [19] developed a clustering method which they successfully used in estimating the similarity between different treatment frameworks and clinical models between service providers, enabling better co-operation between different EHRs.

The common element between these tests is that EHR data is a popular tool in

studies for demonstrating the feasibility of methods in small scale medical environment. However, the applications are often data-specific and related to a single problem. Analysis of treatment paths has been limited, even more so in the context of large-scale EHR systems with thousands of patients with millions of visits between them.

2.4 Clustering

Clustering methods usually work by dividing a given space into partitions based on some predefined metric. These partitions can be either *hard* (mutually exclusive) or *soft* (partially overlapping). Different partition methods perform differently on different data sets, and it is usually up to interpretation as to which algorithm performs the best. In the case of EHR, a large portion of the data is coded into categorical data as opposed to purely continuous values. For example, data such as diagnoses, medication, and laboratory type have distinct values with no specific order. Such a data set where the data contains both continuous and categorical entries is called *mixed data* and needs a specific approach for analysis. Most clustering algorithms can work on only continuous or categorical data, but some methods are suited for use for mixed data. In this study, we will be focusing on three such methods: similarity-based metrics, dimension reduction, and model-based clustering. [14]

2.4.1 Similarity-based metrics

Similarity-based metrics are a logical counterpart to traditional distance metrics in clustering. In these methods, we aim to create a matrix of similarity or dissimilarity between individual entries on a data set which can then be used to cluster the data set much in the same way as with a distance matrix. A widely used method was first proposed by Gower in 1971 [20]. Given two vectors x and y of k variables each, the Gower's similarity coefficient is given by the weighted average of the distances of each pair of variables. This can be formulated as

$$g(x, y) = \frac{\sum_{i=1}^{k} s_i w_i}{\sum_{i=1}^{k} \delta_i w_i} \tag{1}$$

where the subscript i denotes each pair of variables for (x_i, y_i), $w_i \geq 0$ denotes the weight of the variable i in the sum, s_i is the similarity between (x_i, y_i) and δ_i is a measure of comparability between the i-th variables on the vectors. The weight w_i can be an arbitrary value of weighting, with higher priority given to characters that are known to contain more information on the similarity of the vectors (x, y). The value of δ_i denotes if the values can be compared, and gets a value of 0 when variables (x_i, y_i) have no comparable similarity metric and one otherwise. Thus the sum $\sum_{i=1}^{k} \delta_i w_i$ is the weighted number of comparable dimensions on the two vectors, giving a scalar value with which to average the similarities. When $\delta_i = 0$ we set $s_i = 0$ as well. If $\sum_{i=1}^{k} \delta_i w_i = 0$ the value of similarity is not defined, but Gower's solution sets the similarity as 0 for conventionality. [20]

The similarity s_i is calculated differently for different groups of variables. For continuous variables it is given by range-normalised Manhattan distance, $s_i = 1 - |x_i - y_i|/R_i$ where R_i is the range of variable i. For binary and categorical data, the distance is the simple matching coefficient for two cells, (x_i, y_i). Thus $s_i = 1$ when the dimensions are matching and 0 when the value differs for the dimensions. The similarity measures are then combined using the Equation 1, which results in a

Gower similarity coefficient GS. As most distance algorithms expect the lower value to mean higher similarity, however, we can deduct this value from 1 to arrive at the Gower *dissimilarity* coefficient, $GD = 1 - GS$. This dissimilarity value can then be used as a substitute for distance metric in conventional distance-based clustering algorithms such as k-means, partitioning around medoids (PAM) or hierarchical clustering. [20]

For clustering applications, similarity-based metrics are easily understandable. Moreover, the implementation is usually quite straightforward. This makes similarity-based metrics a good benchmark for comparing with other analyses on the data.

2.4.2 Dimension reduction

Dimension reduction methods aim to present the data in a subdimensional space such that the number of random variables is reduced. The resulting subspace aims to capture the maximal amount of information related to the variance in the features present while making the data set easier to work with. The two most widely used methods are *Principal Component Analysis* (PCA) for numerical data and *Multiple Correspondence Analysis* (MCA) for categorical data. The two share a link in that MCA can be represented as PCA of a matrix of dummy variables, where each set of column-variable pairs is coded into a contingency table, and PCA is performed on the resulting table. A method for dimension reduction for mixed data is then to concatenate the continuous variables with the dummy variables and perform PCA on the resulting matrix. This can be shown to be equal to PCA on purely numerical values and MCA on purely categorical values [21]. The idea has been independently proposed by multiple authors as a dimension reduction method for mixed data, most notably Hill & Smith [14], de Leeuw & van Rijckevorsel [22], Kiers [23] and Pagès [24], while the process itself is usually referred to as either PCAMIX (Kiers) or Factor Analysis of Mixed Data (FAMD, Pagès) on technical terms. In this study, we will refer to the method as FAMD. [14]

FAMD is analogous to PCA on a weighted matrix $\mathbf{X}\mathbf{D}_\Sigma^{1/2}$, where $\mathbf{X}$ is the initial data matrix. $\mathbf{D}_\Sigma$ is a matrix with diagonal elements of $s_1^2, \cdots, s_K^2; \pi_{K+1}^2, \cdots, \pi_{K+Q}^2$, where s_i is the standard deviation of the i-th continuous variable, π_i being the number of objects in the associated category of the i-th dummy variable, K is the number of continuous variables and Q is the number of categorical variables on the data. In other words, the data are ordered so that numerical variables are followed by categorical variables, and from this matrix, the diagonal matrix of squares of standard deviations and number of objects in each category is calculated. Then, PCA is performed on the weighted matrix $\mathbf{X}\mathbf{D}_\Sigma^{1/2}$. PCA converts a set of correlated values to a matrix of linearly uncorrelated values, which are conventionally called principal components. The resulting primary components are orthogonal and ordered so that the first dimension f_1 holds the maximum number of information on the variance of the data, the second dimension f_2 the second most information and so on. Therefore we can reduce the complexity of the data by choosing only to use n first dimensions that capture the largest amount of information on the variance of the data and discard the rest, effectively reducing dimensionality on the original data

set. [14]

Assume a data set consisting of K quantitative variables, $k = 1, \cdots, K$ and Q qualitative variables, $q = 1, \cdots, Q$. The first primary component of the PCA method is denoted as f_1. This primary component is maximised with respect to

$$\sum_{k \in K} r^2(v_k, f_1) + \sum_{q \in Q} \eta^2(v_q, f_1), \tag{2}$$

where r is the simple correlation coefficient or fraction of matching values between the variable v_k and the primary component f_1, where $v_k, k \in K$ is the k-th continuous variable of the set of all continuous variables, η is the correlation ratio between the variable v_q and the principal component f_1 and where $v_q, q \in Q$ is the q-th categorical variable of a set of all of the categorical variables [24]. The first part of the sum maximizes the correlation between continuous variables and the second part between categorical variables. The contribution of each variable is the same. The second primary component f_2 is orthogonal to the first, and again maximises information related to the Equation 2, and so on continuing for all of the other primary components. Thus the FAMD maximises information related to the correlation of both the continuous and categorical variables. [14][24]

The resulting matrix has the attractive property of containing only numerical coordinates for each data point in the data set. Thus we can apply any continuous distance-based clustering method such as k-means to the data set. Moreover, we can analyze and visualize the most weighted variables on the data with conventional PCA analysis tools to check the behaviour of the data. The approach has some shortcomings, however. Vichi *et al.* (2001) brought up the so-called masking problem for cases where distance-based clustering is applied after dimension reduction [21]. Dimension reduction aims to maximize the variance in the original set of variables while the following clustering step aims to minimize within-group variance and maximize the between-group variance. Without carefully considering the underlying variables it is possible to end up with a situation where we are removing the underlying cluster structure with the dimension reduction before applying the clustering itself. Such a case could occur, for example, if a set of strongly correlated variables are not connected to the cluster structure but are well expressed in the smaller dimensional projection. [14]

2.4.3 Model-based clustering

Model-based clustering is based on the assumption that sample observations arise from a distribution that consists of a mixture of multiple individual components [12][25]. The first models making use of finite mixture models for classification contain the works of Everitt [26], Banfield & Raftery [12] and Hunt & Jorgenssen [27]. Model-based clustering takes a probabilistic approach to clustering and, rather than assigning each observation a hard class label, the method assigns a list of probabilities for each observation of that observation belonging to each of k latent categories on the data. Rather than fitting centroids such as k-means or hierarchical distance-based structure on the data, the models fit a group of multivariate distributions on the

data. Assume that we have n observations $x = x_1, x_2, ...x_n$, of which we construct a joint distribution that is a mixture of G components, each of which is a multivariate distribution with density function $f_k(x_i|\mu_k, \Sigma_k), k = 1, \cdots, G$. The weighted mixture model is then

$$f(x|\pi, \mu, \Sigma) = \prod_{i=1}^{n} \sum_{k=1}^{G} \pi_k f_k(x_i|\mu_k, \Sigma_k) \tag{3}$$

where π_k are the weight coefficients for the components such that $\sum^k \pi_k = 1$ and f_k is any probability distribution characterized by mean μ_k and covariance matrix Σ_k [26]. For f_k, normal, gamma and Poisson distributions are widely used. Maximum likelihood estimation for the parameters is difficult, but the usual solution is to apply the expectation maximization algorithm developed by Dempster, Laird and Rubin [28] onto the data. The EM algorithm expects that there is some missing data, namely the class labels that can be estimated to form the complete data set. The EM algorithm provides an iterative approach for estimating the missing data and tuning the model with repeating estimation (E) and maximization (M) steps until convergence is reached. The mathematical formulation of the EM algorithm is outside the scope of this Thesis, but detailed accounts of EM algorithm with model-based clustering can be found in Dempster *et al.* [28], Banfield *et al.* [12] and Melnykov *et al.* [25].

After s iterations for an unobserved class label $\gamma_{ik} = E(\xi_{ik}|x_i)$, where $\xi_{ik} = 1$ if x_i belongs to a given cluster and zero otherwise, we arrive at estimated values for the mean and covariance matrix for each x_i as

$$\gamma_k^{(s)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ik}^{(s)}, \tag{4}$$

$$\mu_k^{(s)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(s)} x_i}{\sum_{i=1}^{n} \gamma_{ik}^{(s)}}, \text{and} \tag{5}$$

$$\Sigma_k^{(s)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(s)} (x_i - \mu_k^{(s)})(x_i - \mu_k^{(s)})'}{\sum_{i=1}^{n} \gamma_{ik}^{(s)}}. \tag{6}$$

The components of $\gamma_k^{(s)}$ are ellipsoids centered at $\mu = \mu_1, \mu_2, ...\mu_G$ [26]. However, the ellipsoids can have varying forms characterized by the heterogeneous covariance matrix Σ_k. We can decompose the covariance matrix as $\Sigma_k = \lambda_k D_k A_k D_k^T$ where D_k is the matrix of eigenvectors of Σ_k, A_k is a diagonal matrix with elements proportional to the eigenvalues of Σ_k on the diagonal and λ_k is a scalar [12]. Thus the geometric attributes of the component are encapsulated with these parameters. D_k controls the orientation of the principal components, A_k defines the shape of the density contours, and the volume of the cluster is given by λ_k. With some applications these are often represented with three letters in order, denoting the volume, the shape and the orientation of the ellipsoid. These letters usually are denoted to have values of "E" for "equal", "V" for "variable" and "I" for "coordinate axis" that define the order of freedom of the three parameters [29].

Imagine the records in an EHR database as p-variate occurrences centred on each person and diagnosis of a mixed data type. The cluster can be seen as a distribution of visits centred around a mean with a common covariance matrix. Thus a model-based approach can help us partition the data probabilistically. More often than not, we can have multiple diseases being treated simultaneously or old diseases affecting new diagnoses and treatments. Thus a shift from a hard partition into a probabilistic one can bring a more realistic picture of the clustering of the treatment paths. This is dependent, of course, on if these probabilistic distributions capture an apparent effect on the data and those individual treatments are independently distributed around a mean, which is not necessarily given.

2.5 Prediction

Predicting outcome variables from the data can be done via machine learning models trained on a labelled data set. For our case, the methods are supervised learning methods. *Classification* models aim to classify a point in the data to a prior set of distinct classes while *regression* models try to predict a numerical value to the output variable. The training data for supervised learning problems contains both the input variables as well as the outcome variable or variables. This data is usually represented as a matrix of individual training vectors divided into different features. Supervised learning aims to fit a model or a function to the data so that it can be used to correctly predict the output variable of an entirely new set of input data. [1]

Given the varied nature of medical data, it would be beneficial to be able to use any selected methods on both classification and regression tasks on the data set. We may be inclined to, for example, find out both the number and type of any upcoming medications for a patient as both numerical and categorical variables. Some models are able to address both classification and regression problems, either with an extension of the algorithm or by some small variation in the model. However, these can seldom be used simultaneously, and different outcome variables need a new model to be trained on the data, so in theory, there is little benefit to using the same model for classification and regression both. In practice, however, initializing and maintaining multiple libraries and methods can be unwanted. For this reason, in the scope of this study, we are focusing on exploring methods that can be used for both classification and regression tasks. Moreover, any selected models need to be able to work on mixed data types and have adequate performance for the tasks at hand.

2.5.1 Random Forests

Random Forests are an extension of decision trees into the domain of machine learning. Conventional decision trees are used to create a tree-like structure in which decision nodes or branches are used to divide points on the feature space into final classes called leaf nodes. Decision trees have many attractive features for machine learning. They are efficient to compute, intuitive to understand and can be grown arbitrarily large until each point in the training data is correctly classified, resulting in a perfect fit for the training data. This, however, introduces a significant bias into the data,

and the trees are prone to suffer from overfitting. This means that the trees perform very well in the context of the training data but given new data that is even slightly out of the scope of the training data, the model underperforms heavily. This is a real problem in machine learning as generally we want good performance on out-of-sample data points that are typical of real-world data applications. This feature of the model is often called robustness.[30]

Early solutions to the problem were pre-pruning or post-pruning the trees, in which the growth of the tree was stopped or cut at some point to allow for a more general representation of the data. This had some downsides, namely that the pruning point was hard to determine and any gain in generality came at the expense of accuracy of the model. In 1995 Ho [30] introduced the idea that using multiple decision trees increases both the generality to the model while simultaneously increasing the accuracy of the model. The idea was later built upon by Amit and Geman [31], and Breiman [32], who formulated the model that is most commonly used today. The idea that additional trees increase generality while improving accuracy was formally proven by Breiman by studying the error function of the random forests as more trees are added, but the same phenomena had been noticed and used before in the context of multiple classifiers for a data set [30][32].

The idea of random forests algorithm is to create or grow multiple decision trees with the same data set and combine these individual trees to create the final model. This combining is often done on voting on the final values or choosing the mode of the trees as the final tree. The requirement for such models is that the trees are uncorrelated. The usual approach is to use a method called *bootstrap aggregating* or *bagging* to the data in which multiple samples with replacement are drawn from the data set to train the model, thus increasing invariability to the input data. However, for most of the data, some features are more critical in identifying the final class than others. Over multiple rounds of voting, this would make such features over-represented in the data, thus making the individual trees correlated. This is mitigated by employing an additional step to the bagging phase called *feature bagging* in which the data is sampled not only concerning the data set but also the individual features. Such a method was first suggested by Ho and later generalized by Breiman, and most contemporary methods for random forests are built on the principle of feature bagging. [32][30]

Random forests can be used for both classification and regression tasks. The tree can get both categorical and numerical values in the leaf nodes, and we can use either as the outcome variable for the algorithm. The classification labels are given by a tree whose leaf nodes are categorical, or if many leaf nodes share the same path, that which is the most common among the values. For a regression tree, the outcome value of a given problem is the numerical value of the leaf node or, in case of multiple numerical values sharing the same decision tree, the average of the leaves for that case. With both methods the random forest algorithm then grows multiple trees randomly and combines these individual numerical or categorical values, either by voting or averaging across the values, to arrive at a classification or regression estimate for any one given task. [33]

Another feature of random forests built on Breiman's theory, which utilise the

traditional CART (Classification and Regression Tree) algorithm is that the decision space can be thought of like an n-dimensional space that is split by hyperplanes into partitions, representing the splitting rules on the features. For any feature f consisting of d dimensions the tree can, therefore, have 2^d different splits done on it, each corresponding to a combination of categorical values. As the value is exponential, some tree-building algorithms limit the number of dimensions for a categorical feature to some reasonably computable number. This is, for example, the case with the `randomForest` package on `R` that is a port of the original `Fortran` implementation of Breiman's algorithm. Some of the features in the data set of this study are very high-dimensional, for example, when considering the place of the appointment out of a network of over 200 possible hospitals. A common workaround is to convert such d-dimensional variables into d substitute or dummy variables that have a value of 1 if the categorical value of the vector is corresponding to the feature dimension and 0 otherwise. This, however, means that any d-dimensional variable is represented as d individual features on which the feature space is sampled, thus increasing the weight of the variable. We can, for example, imagine a data set of 2 numerical variables and one 98-dimensional categorical variable. If the categorical variable is coded into dummy variables and the features are sampled from these 100 new features the numerical variables are going to account for only 2% of the final results as opposed to the original 66.6%. This can be accounted for by lowering the weight or probability of sampling of such features by $\frac{1}{d}$ to not make these values over-represented. [32]

2.5.2 Support Vector Machine

Support-vector machines (SVM) or support-vector networks are a popular tool for supervised classification and regression problems. As the name suggests, the method involves using vectors of the data points on the training set to partition and classify the data set. The idea was first introduced by Vapnik in 1963 [34] and later generalized in 1995 by Vapnik & Cortes [35] for use in machine learning.

The main idea behind SVM's is to fit hyperplanes onto the k-dimensional data set in such a way that the hyperplanes split different partitions with the widest possible margin. Figure 2 shows a simple example with two groups in a 2-dimensional space. Formally the support-vector machine is defined by taking a data set of n points represented as p-dimensional real vectors $(\overline{x_1} \dots \overline{x_n})$ each with a group label $y_i, i \in (1 \dots n)$. For simplicity, we set this label to be either -1 or 1 for each vector. The objective is to separate the points with a hyperplane. The plane can be written as the set of points $\overline{w} \cdot \overline{x} - b = 0$, where $\overline{w}$ is a normal vector to the hyperplane and $\frac{b}{||w||}$ is the offset of the plane from origo along the vector $\overline{w}$. To separate the two classes we introduce two constraints for both of the classes which ensure that each point lies on the correct side of the hyperplane. For labels $y_i = 1$ we must have $\overline{w} \cdot \overline{x} - b \geq 1$ and for $y_i = -1$ we have $\overline{w} \cdot \overline{x} - b \leq -1$. Vapnik uses an expression

$$y_i(\overline{w} \cdot \overline{x_i} - b) \geq 1 \tag{7}$$

where y_i is the class label of the vector, namely 1 or -1. As the width of the margin is

$\frac{2}{||w||}$ and as we try to find the largest possible margin the problem becomes minimizing $||w||$ subject to the constraint $y_i(\overline{w} \cdot \overline{x_i} - b) \geq 1$. An important feature to note is that the hyperplane is completely characterized by the points bordering the margin $\overline{x_i}$. Vapnik calls these points the *support vectors* of the hyperplane, thus giving a name to the algorithm. [35]

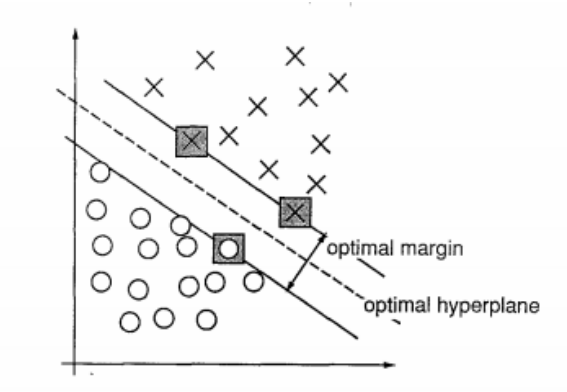

Figure 2: A simplified example of a 2-dimensional plane with a separable hyperplane partitioning the space. The plane is chosen so that the margin enveloping the hyperplane is as large as possible. The hyperplane is defined by support vectors, here marked with grey squares. Image from Vapnik & Cortes (1995) [35]

Following the formulation by Vapnik, the minimization problem can be solved by a Lagrange multiplier making it a quadratic programming problem, which when solved yields a simplified form $\overline{x_i} = \Sigma_{i=1}^{n} \alpha_i y_i \overline{x_i}$, where $\alpha \geq 0$ and $\alpha > 0$ only when the vectors $\overline{x_i}$ lie on the margin boundary. The constant b can then be calculated as $b = \overline{w} \cdot \overline{x_i} - y_i$ for all $\overline{x_i}$ that lie on the boundary, fully characterizing the equation. However, this is only feasible for clean cuts on the data. Vapnik and Cortes suggested a way to account for a so-called *soft margin* that allows for a number of points on the wrong side of the margin by characterizing the hypervector as $max(0, 1 - y_i(\overline{w} \cdot \overline{x_i} - b))$ and minimizing

$$[\frac{1}{n} \Sigma_{i=1}^{n} max(0, 1 - y_i(\overline{w} \cdot \overline{x_i} - b))] + \lambda ||\overline{w}||^2 \tag{8}$$

in which the parameter λ measures the bias of the algorithm to maximise the margin size while sacrificing possible points appearing on the wrong side of the margin, with larger values meaning larger margin but possible misplaced vectors on the margin. The solved function behaves the same, but α is limited by an upper bound of $\frac{1}{2n\lambda}$. [35]

The described algorithm works for any linear combination of support vectors, but fails for any nonlinear solutions. To get around this Vapnik suggested a kernel transformation of the vector space into higher-dimensional space $\phi(\overline{x_i})$ that satisfies $k(\overline{x_i}, \overline{x_j}) = \phi(\overline{x_i}) \cdot \phi(\overline{x_j})$. The problem is then calculated as usual, but using the transformed space $\phi(\overline{x_i})$ instead of $\overline{x_i}$. As an example switching into polynomial

kernel $k(\overline{x_i}, \overline{x_j}) = (\overline{x_i} \cdot \overline{x_j})^d$ can solve a polynomial problem where a linear solution is not possible. [35]

The method was also later generalized for regression problems making it a powerful and well-established tool for our analysis with the EHR data set. Such a machine was formulated by Drucker *et al.* [36], and is subsequently called support vector regression (SVR). The regression uses the same general methods applied by the SVM method but uses a different starting point for the algorithm. We want to find the function $f(x)$ that fits our input data with some pre-defined error value ϵ. We can define such a function in the vector space as $f(x) = \overline{w} \cdot \overline{x} + b$, where $\overline{x}$ is the input vector, $\overline{w}$ is a normal vector to the hyperplane and b is a bias term of the data. Then finding the general margin on the data for points that deviate at most ϵ can be thought of an SVM optimization problem by, again, minimising $\frac{1}{2}||w^2||$ under the constraint $|y - f(x)| \leq \epsilon$. We can now apply the SVM algorithm to the same problem, reaching a general solution for $f(x)$ as a result. For nonlinear functions, we can again employ a kernel trick on the feature space. Training the model then corresponds to solving $f(x)$ via the SVM method, which can then be used to calculate new values for any new set of points. [36]

2.6 Validation of clusters

While validation of prediction output is a straightforward statistical exercise, comparing different clustering outputs is often a subjective task. To introduce a measure of comparability to this, we need a standardized test setting for comparing different clusters resulting from different methods. A simple measure is the similarity of the final clusters. If we have access to a baseline or want to compare different results against each other, we can use a standardised metric to compare the values. In this study, we can use such measures to compare different partitions obtained using different methods with each other. This information, alongside the expert review, can help differentiate between methods and approach a possible best solution for the application.

2.6.1 Adjusted Rand Index

An interesting problem related to clustering is measuring the correspondence of different partitions. Suitability of partitions is a largely subjective task, but comparing partitions can bring insight into how similar or dissimilar clusterings are compared to a baseline or each other. Multiple methods exist for comparing partitions, but a widespread one is some variation of *pair counting based measure*, possibly the most used being the one often credited to William Rand based on his work of evaluating criteria for clustering methods [37][38]. The method often called the *Rand index* or *RI* was later improved upon by Hubert [39] into the version commonly used now. Suppose that we have an n object set $S = O_1 \dots O_n$ that is partitioned in two ways, one into R groups $U = u_1 \dots u_R$ and the other into C groups $V = v_1 \dots v_C$ so that U and V are subsets of S, and $\{R, C\} \leq n$. Now let us define the following rules for any pairs of objects [39]:

A Number of pairs of objects that are placed in the same class in U and the same class in V

B Number of pairs of objects that are placed in different classes in U and in different classes in V

C Number of pairs of objects that are placed in different classes in U and the same class in V

D Number of pairs of objects that are placed in the same class in U and different classes in V

The *Rand index* is then

$$R = \frac{A+B}{A+B+C+D} = \frac{A+B}{\binom{n}{2}}, \tag{9}$$

as the total number of pairs of objects is $\binom{n}{2}$. In other words, the index denotes the number of pairs that are in the same cluster or both in different clusters in the data set, divided by the total number of pairs in the data. This value is bound between $[0, 1]$. The agreements can also be expressed in a contingency table in which each cell denotes the number of objects that are shared between partitions u_i and v_j. An example of this can be seen in Figure 3.

Partition V

Partition U	Class	v_1	v_2	...	v_C	Sums
	u_1	n_{11}	n_{12}	...	n_{1C}	$n_{1\cdot}$
	u_2	n_{21}	n_{22}		n_{2C}	$n_{2\cdot}$
	.	.	.		.	.
	.	.	.		.	.
	.	.	.		.	.
	u_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R\cdot}$
	Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot C}$	$n_{\cdot\cdot} = n$

Figure 3: Notation for comparing two partitions. Each cell n_{ij} denotes the number of objects that are common to classes u_i and v_j. Image from Hubert (1985) [39]

However, the original *Rand index* is troublesome as a raw measure. Given any two partitions where the number of partitions is large when compared to the number of data entries, there is a large probability that some matching pairs are found at random [38]. Thus the lower bound for the index is not truly 0, but some statistical value given the partition. Hubert approached the problem by assuming a generalized model for randomness in which the partitions U and V are picked at random subject to both having the original number of classes and object. This gives the expected value of object pairs of type **A** as $E(\Sigma_{i,j}(\binom{n_{ij}}{2})) = \Sigma_i(\binom{n_{i\cdot}}{2})\Sigma_j(\binom{n_{\cdot j}}{2})/\binom{n}{2}$. The formula is equal to the number of pairs that can be formed from the data $\Sigma_i(\binom{n_{i\cdot}}{2})$ multiplied by the number of object pairs that can be constructed from columns $\Sigma_j(\binom{n_{\cdot j}}{2})$ divided by the total number of pairs $\binom{n}{2}$. Combining this with the Rand

index Hubert defined the *adjusted Rand index, ARI* as $\frac{\text{Index - Expected Index}}{\text{Maximum Index - Expected Index}}$ or

$$\frac{\Sigma_{i,j}(\binom{n_{ij}}{2}) - \Sigma_i(\binom{n_{i.}}{2})\Sigma_j(\binom{n_{.j}}{2})/\binom{n}{2}}{\frac{1}{2}[\Sigma_i(\binom{n_{i.}}{2}) + \Sigma_{.j}(\binom{n_{.j}}{2})] - \Sigma_i(\binom{n_{i.}}{2})\Sigma_j(\binom{n_{.j}}{2})/\binom{n}{2}} \tag{10}$$

where $n_{ij}, n_{.j}$ and $n_{i.}$ are values from the contingency table. The ARI is bound between [-1,1] with values near -1 denoting unexpectedly dissimilar partitions, values around 0 near the expected value and values near 1 denote very similar results. [39]

3 Research material and methods

The data set was processed on a 16-core Windows machine, using the R statistical language for analysis. The basic analysis pipeline for the study is shown in Figure 4. The data were first preprocessed, followed by clustering with three methods. The results were validated and used to create a new data set in which each treatment path was labelled with a cluster number. This was followed by a new preprocessing step during which the data was prepared to be used for the prediction of various outcome variables related to the treatment paths. After this, these models were analysed, and the results validated.

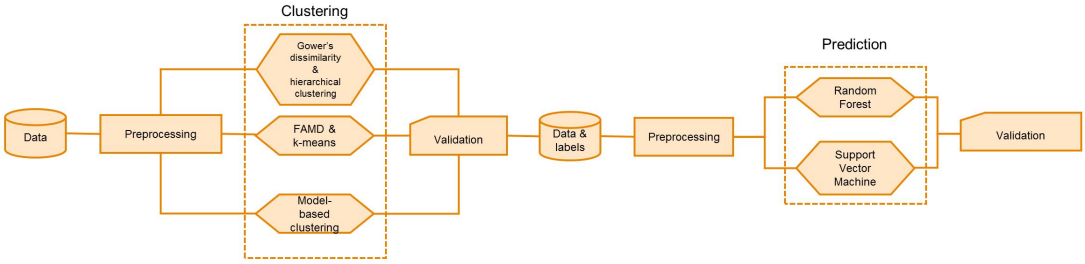

Figure 4: A graph showing the basic analysis pipeline for the process of clustering the EHR data set and predicting outcome variables based on the labels produced by the clustering method.

3.1 Data preprocessing

The database consisted of information on over 1.2 million patients who have used the medical services for a large Finnish healthcare provider since the beginning of the year 2012. The data set had been de-identified for the use of analytics and did not contain names or SSIDs from which an individual could be identified. Otherwise, the data set contained mostly complete medical records of the patients, including such data as appointment history, laboratory and imaging results, diagnoses, age, sex, and past as well as current medications. Occupational information such as the general area of industry and sick leave data was present for some of the population. The data set also contained data on the performing doctors and nurses, such as fields of expertise and location of the appointment. Not all features were used in the analysis but rather were selected on a per-application basis.

The data set for the study was limited to the individuals having used mental health services at some point in their life over the age of 18 and within the occupational healthcare services. The patients were identified by having a diagnosed ICD code of specific mental disorders in their patient history. The codes can be seen in Table 1. The codes were selected in such a way that they represented a majority of mental health cases for the service provider. The data set consisted of 62.9% female and 37.1%

male patients. The patients had, on average, 61.2 individual visits to professionals during their whole lifespan for any disease. For simplicity, medication for the patient was considered as the number of antidepressants the person had been prescribed during an appointment. This included psychoanaleptics with an ATC code N06 and its subtypes N06A, N06B and N06C that are used in the treatment of depression and mental disorders.

ICD-10	Explanation
F10-F19	Mental and behavioral disorders due to psychoactive substance use
F32	Major depressive disorder, single episode
F33	Major depressive disorder, recurrent
F34.1	Dysthymic disorder
F41.9	Anxiety disorder, unspecified
F43	Reaction to severe stress, and adjustment disorders

Table 1: The ICD-10 codes that a person must have to be included in the study. All codes include all subtypes of the hierarchical ICD-10 tree. For example, the code F32 includes the subtypes F32.0, F32.1, F32.2, F32.3, F32.4, F32.5, F32.8, F32.81, F32.89 and F32.9 as well. F10-F19 includes all of the codes and subcodes between the two values. Specific references can be found in the ICD-10 specifications [40].

For the model, we selected the 11 most important features to be used for clustering the data. These features were selected in co-operation with medical professionals in the house. The dimensions consisted of data on four major fields: personal data, such as randomly generated person identifier, age and gender; occupational data such as field of the profession; appointment-linked data such as the general field of the appointment, location of the appointment and unit of care; and medical data associated with the visit. The chosen data dimensions are showcased in Table 2 along with data classes. As with any data set of similar use case, the data was not purely numerical but a mix of numerical and categorical data. This put any applications for the data in the field of *mixed data analytics*.

Dimension	Data class	Explanation
Person identifier	Categorical	Numerical identifier
Industry of the patient	Categorical	Industry category, eg. "Heavy industry"
Gender	Categorical	Sex of the patient
General type of the visit	Categorical	Occupational healthcare, private appointment or public service
Unit of care	Categorical	Unit code of the unit of care
Primary diagnosis	Categorical	ICD-10 code
Secondary diagnosis	Categorical	ICD-10 code
Product code	Categorical	Product code for the treatment
Age of the patient	Numeric	Numerical value
Visit timestamp	Numeric	Days since the beginning of the data set on 1.1.2012
Number of prescribed medication	Numeric	A numerical value if the patient was prescribed medication during the appointment

Table 2: Classes of the data dimensions used in clustering of the data. The data values were abridged and masked such that for example "Heavy industry" would be marked with a letter "C" and a certain unit of care would be marked with an abbreviated label, for example, "1HEL" for the primary hospital in Helsinki.

To deal with the sparsity of the data, the missing values in the data were re-coded as categorical variables of the class "missing" or in case of numerical values as a number 0. In other words, the absence of data was considered meaningful in the scope of this study. Consider for example a case that a missing value for a second diagnosis rarely means that the status of the second diagnosis is unknown but instead that the person was not diagnosed with anything other than the primary diagnosis. The logic is embedded in the use of EHR systems where only the relevant fields of the visit are usually filled during each visit. While some missing data points undoubtedly were genuinely missing, the overwhelming majority of data was considered to mean the absence of that particular issue associated with the disease and not an absence of the data.

3.2 Clustering

The clustering was done with three distinct methods to analyse the performance and viability of the methods for clustering of the data: Gower's dissimilarity matrix and subsequent hierarchical clustering, dimension reduction and the subsequent k-means clustering, and the model-based clustering methods. The Gower dissimilarity matrix was calculated using the `daisy` [41] package. The dissimilarity matrix was suitable

for use with hierarchical models, so the `hclust` [42] method from the `R base` package was used. For FAMD the dimension reduction was calculated with the `FactoMineR` [43] package which resulted in a numerical transformed matrix of coordinates. We used a simple `k-means` [42] clustering from `R base` to cluster these. The package `MixAll` [44] was used for the *model based* method. The parameters used are displayed in Table 3. For any value not mentioned, the default value was used.

Library	Version	Method	Parameters
R base	3.4.3	hclust	method: "complete"
R base	3.4.3	kmeans	centers = 14 058
cluster	2.0.6	daisy	metric = "gower" type: list(logratio = 3) stand = TRUE
FactoMineR	1.39	FAMD	npc = 3
MixAll	1.2.0	clusterMixedData	models = c("categorical_pk_pjk", "gaussian_pk_sjk") nbCluster = 14 058 strategy = clusterFastStrategy()

Table 3: R packages and the parameters used for the runs. Default parameters were used if not specified. [41][42][43][44]

The data set was sampled to contain fewer than 46 340 visits. This was due to limitations in the selected libraries, namely in the `daisy` package used to calculate the Gower dissimilarity matrix. The method stores the length of the dissimilarity matrix as an unsigned 32-byte integer. This means that for any $n \times n$ matrix the maximum number of individual appointments is then $\sqrt{2^{31}} \approx 46\,340.95$ with any matrix size larger than this resulting in an integer overflow. To keep the results comparable and due to long processing times, the sampled data set was the same for all of the methods. The samples were chosen according to the following algorithm:

1. Create an empty data set of length 0 as the sampled data set
2. Create a list of all the appointments grouped by patients
3. Shuffle the list of patients
4. Calculate the length of the sampled data set.
   (a) If the length is less than 46 340 continue the algorithm
   (b) If the length is more than 46 340 terminate the algorithm
5. Add the next patient from the list of the patients with all of the related visits to the sampled data set

6. Return to step 4.

Usually defining the number of clusters is a central problem in data science. In this case, however, deducting the number of classes was quite straightforward. To find k treatment paths among the patients, we should arrive at the number of unique diagnoses across all the patients. Thus the number of clusters was set to be the set of all unique combinations of persons and diagnoses present in the data, $\binom{\text{[number of patients]}}{\text{[number of diagnoses]}}$. Thus k or *centres* was set as 14 058. The running time for the scripts was quite long, with Gower and FAMD processing the data for around half an hour with a 16-core computer and the mode-based method processing the same data set in around three hours.

The methods were chosen so that the whole data set was processed at a time, rather than one person at a time or a similar split. This was done for several reasons. First, it was assumed that the individual differences in the data were the largest separating factor so that most methods would naturally split the data at different patients. Second, the performance of the clustering algorithm was considered. A single clustering pass that finds $k \cdot n$ partitions in the data is more effective than k individual passes each finding n passes, as even with the most optimized algorithm there is always some computational overhead involved in setting up a new model and initializing the clustering for a new patient. For a large data set of millions of patients where performance is important, this issue could be magnified. Lastly, if calculated for different persons, the clustering algorithm works on slightly different criteria for each individual. When hierarchical clustering is run separately on two patients, the resulting splits can be quite different, whereas when clustering one large matrix of multiple people, the splitting rules will be similar for each patient. This increases the comparability of the results of each patient.

However, this choice in running a single pass on a large data set of multiple people also introduced some complications. Some of these assumptions turned out to be false, and for example, the natural separation between persons was not as strong as believed. As per the definition of a treatment path, a cluster that spans multiple persons was impossible. To mitigate the issue, any such partitions were cut into separate clusters according to the different person identifiers, and a new boolean column was set to the result set to differentiate between these values. A value of 1 meant that the values had been cut and a value of 0 meant that the values were the original clusters. While this increased the number of clusters and conversely meant that the resulting clusters would be smaller, the process upheld the definition of the treatment path and enabled the use of the clustering for predictive measures. A possible future improvement would be to re-run the clustering on the results and combining possibly split values back together. All clusters that had no mental health diagnoses after the clustering were removed. Strictly this was done by removing any cluster that had no diagnoses under the ICD-10 code F00-99. This ensured that we were only looking at mental health related treatment paths.

To improve readability, a final step was added where the data were un-abridged where such abbreviations had been done to the values. The short category labels were replaced with the full names, and some readability changes were made, such

as changing the NA letter to a simple dash, '-'. This was done to make estimating the accuracy of the results by professionals more accessible and intuitive. Some additional columns such as explanatory text fields were also added to the data that had not been used in the clustering step but were available for the appointments to improve the readability of the data by humans further.

3.3 Validation of clustering results

The validation of the clustering results was done in two steps. First, we calculated the *adjusted Rand index* (ARI) for each pair of methods and assessed the similarity of the methods. This was done to get a general idea of the similarity of the methods. The `R` package `mclust`[45] was used for this step. This gives us a measure of similarity, but the quality of clustering is largely a subjective measure as well. Mathematical and statistical tools offer us little insight into whether the results obtained are sensible in the medical sense. For example, a computer might cluster two visits closely together by resemblance. Further analysis is then needed to asses whether the visits truly are related or simply impossible from a medical point of view. A person might, for example, have back pain and diagnosis on mild depression done on the same day in the same hospital, and for that reason, the visits have high similarity. A medical expert is needed to asses if such back pain can be the source or result of the depression or if the two are unrelated.

Two experts were used for validation, the first being the medical director of occupational health and the second the leading psychologist and psychotherapist for the healthcare provider. The experts were asked to assess the number of incorrectly clustered appointments from the data, which gives us the number of false positives in the data as a measure of the quality of the data. The amount of data chosen for validation was chosen such that the combined number of appointments and clusters was around 750. This could mean, for example, 650 appointments across 100 clusters or 500 appointments across 250 clusters. This was chosen to limit the workload on the experts for the data but is large enough to be a good representation for the whole data. It is important also to note that in reality, the clusters are not mutually exclusive even though the clustering results are represented as such. Many diseases can be treated with one visit, and one underlying visit can be either a source or an explaining factor to multiple other diseases. The model-based clustering method is capable of producing probabilistic clustering estimates, but this was not used for this estimate. Thus a person might cluster one appointment as belonging to multiple groups, but with our model, it was only set in one.

Some additional general notes related to the different clustering results were collected from the experts. These included a general review on the quality of data and their feasibility in the further analysis as well as any possible other notes, such as those on missing or incorrect data. These were included in the results as well. Finally, a brief visual analysis of the clusters was done on the results as well, as the *FAMD* method provided a numerical space on which to plot the data according to different clusters. This enabled us to compare the methods visually.

3.4 Prediction

For generating successful insight into future and ongoing treatment paths, we need to be able to predict both numerical and categorical outcome variables from the initial visit of any one treatment path. Such outcome variables are for example the length of sick leave and the total number of visits during the treatment pathways as well as categorical variables such as possible future diagnoses or the most visited healthcare unit for the upcoming treatment pathway. The methods chosen need to be able to predict any such outcome variables from the input data.

Two models were chosen for the prediction on the data; Random Forest (RF) prediction and Support Vector Machine (SVM). Two R libraries were chosen for the task, namely the `randomForestSCR` package for RF analysis and the `e1071` package for the SVM analysis. The parameters for both libraries are shown in Table 4. Both of the methods are supervised machine learning methods, and thus need labelled training data for fitting the models. Because of this, we used the class labels obtained from the clustering step to create a data set for training and validation for the prediction tools.

Library	Version	Method	Parameters
randomForestSRC	2.5.1	rfsrc	ntree = 1 000
e1071	1.6-8	svm	kernel = “linear”
			C = 0.5

Table 4: R packages and the parameters used for the prediction methods. Default parameters were used if not specified. [46][47]

For this study we inspected five outcome variables, namely *the length of possible sick leave in days* and *the number of appointments for any treatment pathway* as numerical values and *if the person will require sick leave at all during the treatment pathway*, *if the sick leave will be 10 days or longer* and *if the treatment path will be 10 appointments long or longer* as categorical variables. The data set was grouped and divided by a cluster label, and the outcome variables were calculated for each cluster and added as individual variables onto the data set. The data set was also enriched with some past data, namely by adding columns for the number of past visits related to mental health issues, past frequency of visits and the cumulative number of sick leave days so far. We also discarded all visits save for one corresponding to the first visit for each treatment pathway. This was done by ordering clusters by date and saving only the first appointment. Thus we were left with a data set of n rows, where each row corresponds to an initial visit related to a mental health related treatment pathway and n is the number of identified clusters. The models were then trained on this data set.

3.5 Validation of prediction results

Training a machine learning algorithm is typically done by dividing the data into two parts called *training* and *testing* data sets. Training data is complete with the outcome variables and is used to train or fit the model to predict the outcome values. The remaining test data set is one where we know the labels beforehand, but use the trained methods to predict values for the data. The difference between the predicted values and the real values can then be used to estimate the performance of the model. To eliminate any bias resulting from chance in choosing the test sample, we can split the data set into k equal parts and run the test k times, each time using $k-1$ parts as the training data and the last one as testing data set. The final error values were averaged from these results. In our case, we used 10-fold training and validation of the data. For multiple outcome variables, multiple models had to be trained. This was done by replacing the outcome variable with any one value we want to predict and keeping the rest of the data set as is.

For the prediction results the validation was done by estimating the *Mean Absolute Error* or *MAE* and *Median Absolute Deviation* of *MAD* values on regression data and precision, recall, accuracy and balanced accuracy values for classification data. Mathematically these can be expressed as $MAE = \frac{\Sigma_{i=1}^{n}|y_i - x_i|}{n}$ and $MAD = median_i(|y_i - x_i|)$, where x and y are vectors an $i \in \{1, \cdots, n\}$. For precision, recall, accuracy and balanced accuracy values we calculate true positive (tp), false positive (fp), true negative (tn), and false negative (fn) rates on the data. We then get precision as $\frac{tp}{tp+fp}$, recall as $\frac{tp}{tp+fn}$, accuracy as $\frac{tp+tn}{tp+tn+fp+fn}$, and balanced accuracy as $\frac{\frac{tp}{tp+fn}+\frac{tn}{tn+fp}}{2}$. Balanced accuracy is a more robust measure of accuracy when the data set is not balanced so that each value is as likely, thus giving as an unbiased measure of accuracy on the whole data set. Additionally we can compare the results to a trivial classifier, that outputs all variables to a single class. Comparing the accuracy of such a classifier with the methods used can highlight the bias in the original data as well as be used to asses the overall usability of the models.

4 Results

The overall results for the data were promising. All of the chosen clustering methods were able to generate meaningful output, and prediction accuracy was adequate. Given the challenging nature of the data, this was a positive result. However, there were also issues with the methods which will be outlined along with the results.

4.1 Clustering treatment paths

The clustering results were quite different for the three methods. The adjusted Rand index or the ARI index for the methods can be seen in Table 5. The values are very low for any two methods. FAMD is very dissimilar with both the Gower and model-based method, with a similarity of 7% and 15%. Values of 0% denote a random spread of classes, so the clusters are very different. Gower and model-based methods produced somewhat more similar results, but the dissimilarity between the methods was still quite high. This is interesting, as the ARI can also be expressed as the probability of two models sharing the same clustering. In essence, FAMD shares only 7% and 15% of the clusterings with either of the other methods, while Gower and model-based methods themselves also only share around 43% of the labels. Thus, we are left with three methods where most of the pairs of objects do not share the same clusters.

	Model-based	FAMD
Gower	0.431	0.152
FAMD	0.070	

Table 5: The pairwise adjusted Rand index values of the different clustering methods. The value is analogous to the probability of pairs of methods sharing the same partitioning. The low values indicate that the partitions are rather dissimilar.

Because the topic of the study was limited to mental health issues, and because after the clusters were obtained, any clusters not containing mental-health related diagnoses were discarded and not used in the results. This meant that the three clustering methods resulted in different results and data sets, both in size and content. All in all, FAMD method found 4 551 visits across 3 216 clusters with mental health-related issues, Gower 10 273 visits across 2 221 clusters and the model-based method 18 226 visits across 1 883 treatment pathways. FAMD had on average clusters of 1.42 visits per cluster, Gower 4.63, and model-based method 9.68 visits per cluster. FAMD had, on average, 1.65 different diagnoses per cluster, Gower 4.24 diagnoses, and model-based method had, on average, 6.81 different diagnoses per cluster. Of these, mental health-related diagnoses for each clustering method were 1.48 for FAMD, 2.14 for Gower and 2.53 for the model-based method. The values are displayed in Table 6. The original number of appointments in the data was $n = 46\,317$, and the total number of unique diagnoses was 3 662.

	FAMD	Gower	model
Unique persons	771	771	771
Appointments	4 551	10 273	18 226
Clusters	3 216	2 221	1 883
Avg. size of clusters	1.42	4.63	9.68
Unique diagnoses	504	1 163	1 580
Diagnoses per cluster	1.65	4.24	6.81
Of which mental health related	1.48	2.14	2.53

Table 6: Values of various indicators on the partitions of the three methods.

Based on these numbers, we are able to generalise the features of each of the methods. On average, the FAMD method created more but relatively small groups of one or two appointments, both diagnosed with mental health-related issues. The Gower and model-based methods were more prone to adding more appointments to the clusters with the model-based method creating fewer, larger clusters. The number of mental health related diagnoses is quite similar, however, so the difference in size is mainly due to the addition of additional visits related to other diagnoses to the treatment paths. This can mean, for example, that model-based method had more related diseases or check-ups clustered into the treatment paths than the other two methods. We can inspect these values also visually. Figure 5 shows the frequency plot of cluster sizes for all three methods. FAMD results in most clusters with a small size. The model-based method resulted in much larger treatment pathway clusters, having a large tail of a small number of very long treatment pathways. Gower's method is between the two with some very large clusters as well, but more small clusters than the model-based method.

Interestingly FAMD has no clusters at all of size 16 or larger with clearly the most mass concentrating at ten appointments or less. Most of the clusters for model-based and Gower methods lie below the size of 25 visits. However, we can see that both the model-based and Gower methods result in a small number of very large clusters which can be seen as outliers in the data. These outliers may also affect the prediction accuracy for models of such clusters.

After removing non-mental health related treatment pathways, only 4 399 appointments out of the original 46 317 were present in all three partitions. 41 918 appointments were missing from at least one of the final results of one of the methods, meaning that they had been pruned for at least one method for belonging in a treatment pathway with no mental health related diagnoses. Model-based and Gower methods shared 8296 appointments, FAMD and Gower 4418 and FAMD and model-based method 4468 appointments between them. These are displayed in Table 7. For FAMD this means that roughly 3.3% of the appointments are not shared by either of the other two methods and for the model-based method, 45.8% of the appointments are unique to that method alone. The results are not as surprising as they seem. As the FAMD method had very few appointments with any other

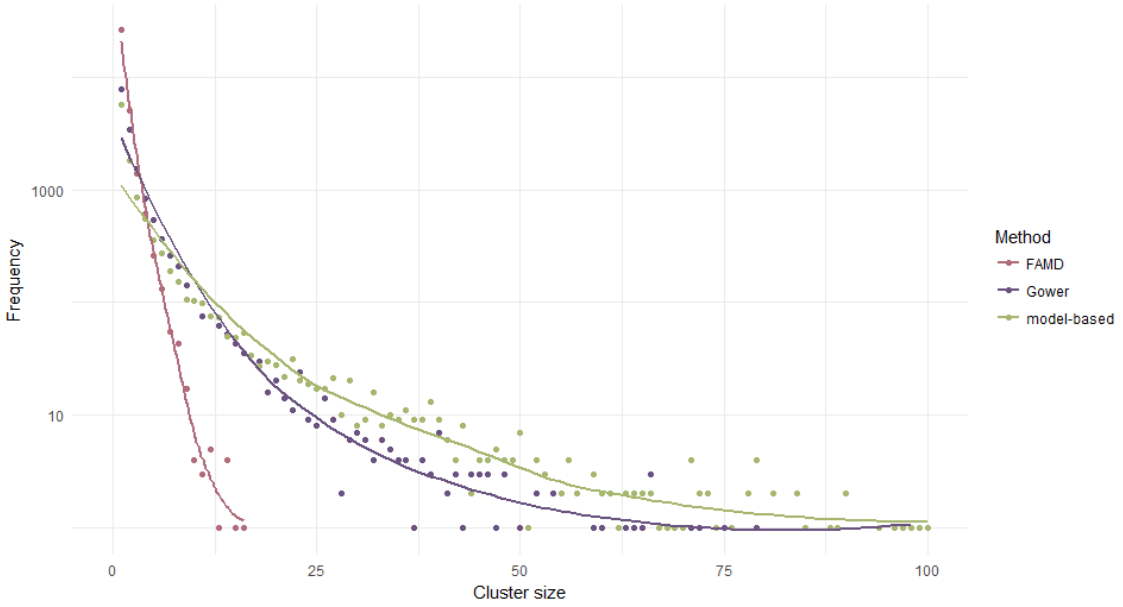

Figure 5: The size of the individual clusters for the three clustering methods. Individual dots represent the frequency of specific cluster sizes with the trend lines denoting the differences on the data. The y-axis is displayed on a logarithmic scale.

diagnosis than those related to mental health issues, most of the appointment data not related to mental health were easily excluded from the results. Almost all appointments in the clusters produced by the FAMD method were included in either of the two other methods. The number of unique appointments increasing with larger cluster size is also natural, with the model-based method naturally having a large number of unique appointments due to the method producing overall the largest clusters with the highest number of appointments. Still, the results can be seen as a measure of dissimilarity between the methods. For example, 19.2% of the appointments found in the clusters produced by the Gower's method were not found with the model-based method, while the model-based method included nearly 10 000 appointments that are not included in either of the two other methods. Thus the resulting clusters are not only different permutations of the same data but contain vastly different appointments as well. In the context of treatment pathways, this is analogous to different methods producing different relationships between diseases and appointments, where some data are seen as belonging to a certain treatment pathway in one method but are not in another method.

	Model.based	FAMD
Gower	8 296	4 418
FAMD	4 468	NA

Table 7: Number of appointments that are shared between the final partitioning of the clustering methods after the non-mental health related clusters are removed from the data set.

As we chose to cluster all of the patients on one pass, in some cases, multiple patients were clustered into the same treatment pathway. These were manually separated from each other as a postprocessing step. The overlap between patients introduced by the methods was more noticeable than estimated, however. These accounted for 12.3% of the appointments with Gower hierarchical clustering, 84.7% with FAMD and 91.6% for model-based clustering. Gower dissimilarity matrix thus seems to follow the split along the different persons better than FAMD and model-based methods, while FAMD and model-based methods provide clusters that are less dependent on the values of one dimension. This is also intuitively plausible, as both FAMD and model-based methods transform the space when clustering the methods. FAMD creates a transformed space in which to cluster the appointments while the model-based method fits multinomial distributions onto the data. Both of these have a more abstract level of distance between different elements of any one dimension than the Gower method, that constructs the dissimilarity matrix from the similarity of the dimensions. Thus the clusters on this plane are more prone to follow the separation of variables on, for example, the dimension denoting personal id's.

As an additional visualisation step, we can use the reduced dimension space produced by the FAMD method to plot each of these clustering approaches on a plane. Such visualisation can be seen in Figure 6. We can see how FAMD creates more mixed results, whereas Gower and model-based methods have a more uniform overall image, denoting the larger clusters. This is further apparent when we compare the zoomed-in pictures, where Gower and model-based clustering methods share similar cluster size and position, but FAMD is more prone to slice larger clusters into smaller ones. An interesting note is that the clustering method of Gower's dissimilarity matrix and model-based method are unrelated to the dimension reduction of the FAMD method, but both can be easily represented in the reduced space to with visible clusters. This suggests that the observed dissimilarity between FAMD and the other two methods has more to do with the cluster size and less with the underlying method being based on entirely different phenomena. In other words, all three methods measure roughly the same similarity but with different parameters and subsequent outcomes.

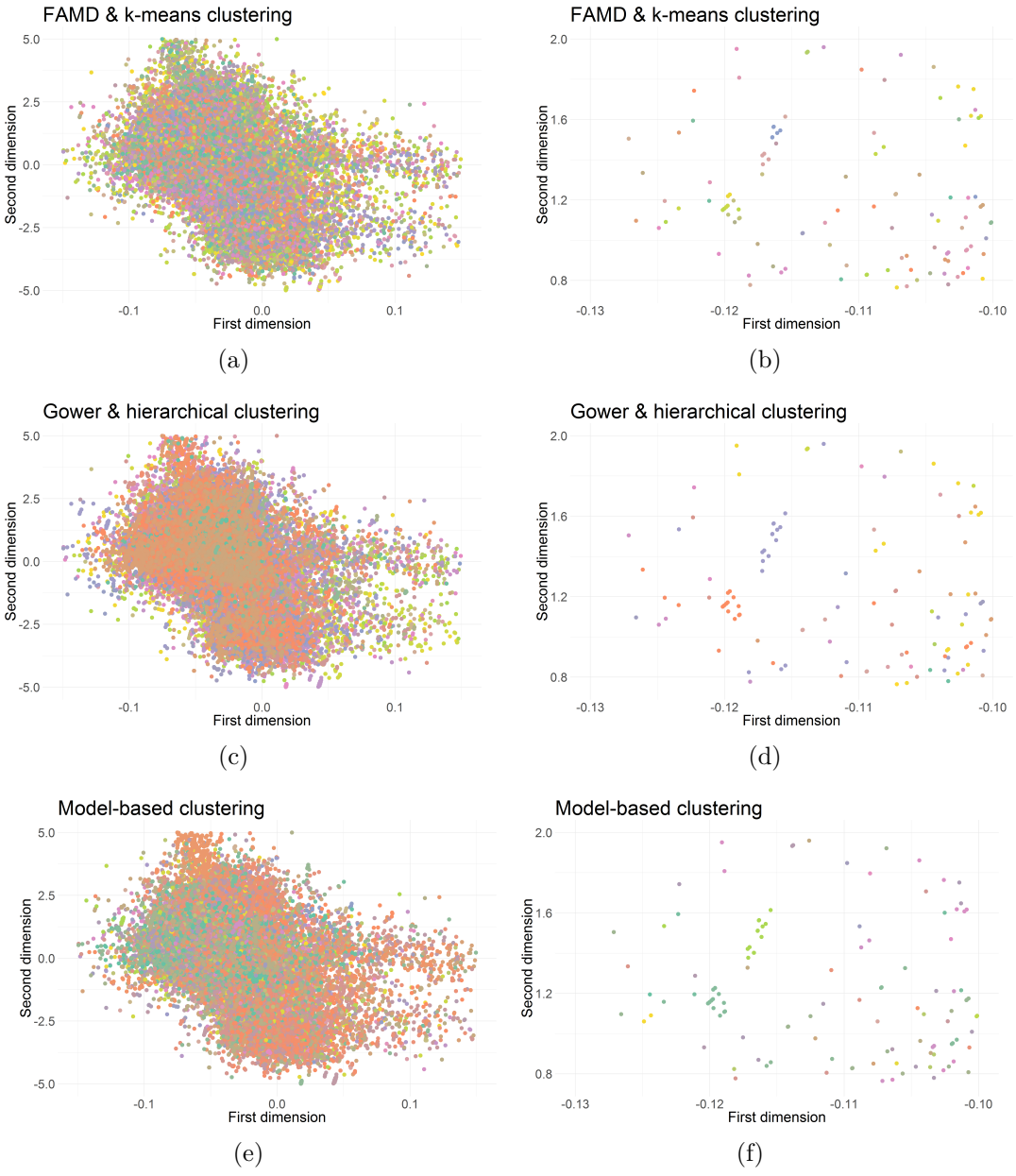

Figure 6: A visualization of the different clustering methods. The graphs are represented in the FAMD reduced space with each color representing a different cluster. The Figures are as follows: (a) describes the FAMD method around the origo; (b) describes the FAMD method zoomed in onto a smaller section of the space; (c) describes the Gower method zoomed out; (d) describes the Gower method zoomed in to details at the same coordinates; (e) describes the model-based method zoomed out and (f) describes the model-based method zoomed in to details. The Figures (b), (d), and (f) show show the many small clusters of the FAMD method and the fewer, larger clusters of the Gower and the model-based methods clearly. The color space is quite crowded as the number of clusters is quite high. Here the clusters are not limited to only those containing mental health visits but contain all appointments across the data.

Overall then we can deduce that the results of the clusters were rather different. On average, the FAMD method created a large number of small groups, while model-based and Gower methods produced larger groups with a few very large treatment pathways as outliers. The pathways shared very few clustering values and even differed much in the appointments that were included in the treatment pathways. Thus we can conclude that the methods differ not only in the clustering results but in content as well, and as such give very different representations of the data.

4.1.1 Validation

The human grading of the data was done by two medical experts, first of which was a chief psychologist and psychotherapist, and the second one a regional medical director of occupational health. The validation results were subjective and so differed quite radically, with near tenfold difference in magnitude of observed error between the two experts. The results can be seen in Table 8. Expert 1 validated the model to be much more accurate than an expert 2. This was later confirmed to be mainly due to different interpretations of the data and treatment pathways. Expert 1 put more emphasis on if the appointments could be connected at all, whereas expert two was stricter on a mental health treatment consisting only of diagnoses related to mental health. This means that the error rate for expert 2 consists mainly of appointments for reasons other than mental health while for expert 1 the treatment paths can have additional, related appointments included. The experts assessed the number of appointments across all of the clusters that were incorrectly clustered in the treatment pathway. These contained any such entries that were not related to the overall treatment pathway or were related to the treatment of some other disease. These values were calculated both across all appointments and again across all clusters. The error ratio of the former was calculated as the number of incorrect appointments in the complete data, which gave a more general measure of error over the whole data set, while for clusters the error rate signified the number of clusters that had any incorrect values in the clusters. This signified a measure of reliability on any single clustering result being correct.

	FAMD		Gower		Model-based	
	Expert 1	Expert 2	Expert 1	Expert 2	Expert 1	Expert 2
Appointments	456		649		693	
Incorrect	3	22	16	170	120	276
Error ratio	0.66 %	4.82 %	2.47 %	26.19 %	17.32 %	39.83 %
Clusters	294		104		73	
Incorrect	1	14	7	37	15	28
Error ratio	0.34 %	4.76 %	6.73 %	35.58 %	20.55 %	38.36 %

Table 8: Expert validation on the clustering results. *Expert 1* is a psychologist and psychotherapist while *Expert 2* is an occupational doctor. "Appointments" denote the total number of appointments in the data, and "incorrect" denotes the number of appointments that are assessed as not belonging to the cluster and are thus errors in the clustering results. "Error ratio" is the ratio of such errors of all of the appointments, giving us a measure of an error on the data. The same measures are calculated for clusters as well in the lower part of the table so that for clusters, the incorrect value is calculated as the number of clusters that have any incorrect values in them. This gives a measure on the reliability of any single clustering result, while the appointment-wise results give a more general error measure over the whole data set.

Some general trends can be seen from the data. FAMD has the lowest error on both per-appointment and per-cluster rates. Model-based methods have the largest error with Gower sitting between the two. The Gower method was assessed to be much more inaccurate by expert 2 than expert 1, who gave the method relatively low error values. For the model-based method, both experts assessed the method to have high error, with expert 1 judging 17.3% of all of the appointments as incorrectly clustered, and expert 2 judging 39.8% of the appointments as such. For the Gower method, both experts assumed the ratio of incorrect clusters as being higher than that of the error ratio calculated across all of the appointments. This suggests that the Gower method tends to have the incorrectly clustered appointments spread out between many clusters rather than a few clusters containing many incorrect appointments. In other words, on average, the number of clusters that have incorrectly clustered appointments is larger than the ratio of incorrect appointments inside them. For FAMD and model-based methods, these values were quite similar, and no such trend can be seen.

As seen previously from Table 5, the clusters resulting from the three clustering methods were quite different. Surprisingly, the error rates between the methods were quite similar. The similar error scores, especially for Gower and model-based method, seem to suggest ambiguity in the interpretation of the results. In other words, deciding if one particular appointment is part of a treatment path or not is somewhat subjective and, especially, ruling out any single disease as not affecting a given

mental health problem is difficult. For example, if one method has clustered sleeping problems with a case of depression while another method has instead clustered stress symptoms to the treatment pathway, both can be explanatory factors for the disease. This was mirrored in the written answers as well, with many cases being noted as difficult to rule in or out to the treatment pathway with the given information. Such cases, however, were not counted towards the error values in the validation.

Additionally, a single appointment can be related to multiple diseases. Because the clusters created by the Gower and FAMD method are hard and exclusive, and by extension, because we chose hard clusters for the model-based method as well for comparability, some information is always lost in the treatment pathways. While we can imagine there being an optimal way of clustering all related appointments, diagnoses, laboratory results and medications, using hard clusters, we are always left with different interpretations of the same treatment pathway. In this sense, treatment pathways are not exclusive, and multiple plausible treatment pathways can be created from the same data set. While this seems counter-intuitive to the notion of defining a problem, in fact, this makes the results more usable in a medical sense. Different treatment pathways constructed from the same data using different methods can highlight different sides of the disease and bring into light different aspects for quality control and tracking of the treatment. This was highlighted in the written results, with notes commenting on some types of treatment pathways being more visible in some clustering method.

Some other general notes by the experts about the data were also made. These were related to individual treatment paths, often describing somatic symptom disorders related to each diagnosis or other issues during the treatment. These included, for example, notes on the musculoskeletal system or possible drug or alcohol abuse. Abnormal reactions to medication or heavy medication usage were also noted. Additionally, some notes commented on possible missing data. The notes also contained some related thoughts on the results, such as how the results could be used to validate the quality of individual treatment paths or their use for further studies into somatic symptom disorders. Any such notes were later discussed internally.

Overall the error rates were quite varying, with FAMD and Gower performing very well according to expert 1 and the model-based method performing the worst according to both of the experts. Surprisingly, however, as a subjective opinion, expert 1 valued the model-based method as the best of the three as the information brought by combining similar but not strictly mental health-related appointments often brought new insight into individual treatment paths. This was because such a model was more informative on the mental condition and identified other possibly related factors in comparison to the clusters restricting treatment pathways to mental health related visits alone. This is in line with the general idea that mental health problems can be intertwined with other medical issues and getting insight into such these can help understand the treatment path as a whole [4]. Thus the interpretation of the error values is not straightforward and is dependent on the context of the goal. For any task requiring small error rates, the FAMD method produces the best results, while for a more context-aware analysis, we might rather use model-based or even Gower method for clustering the data. The larger model-based and Gower method

-based clusters naturally contain more data than FAMD, thus also increasing the risk of any errors in the data. Simultaneously these additional pieces of data can be of interest if they are correctly clustered into plausible treatment pathways.

4.2 Predicting future treatment from initial diagnosis

As per the expert opinion that the model-based method provided the clinically most interesting results, it was chosen as the benchmark for creating the prediction models. The labels acquired from the model-based method were used to create a modified data set based on the first visit of the treatment pathway. As was seen from Figure 5, the size of individual clusters was focused on the smaller side with some more extensive treatment paths. The percentile graph of the sick leave and the number of appointments can be seen in Figure 7. As can be seen, over 75% of all patients were prescribed no sick leave with only the top few percents being prescribed very long sick leaves. The largest values, however, are in the hundreds. The number of appointments is similarly distributed with more than 60% of treatment paths having ten appointments or less. The largest ten per cent of values for both categories, however, are very high with the top one per cent having treatment path lengths of over 100 appointments and sick leave of several hundred days. Thus the data is quite biased, which brings some concern as to the comparability of the data.

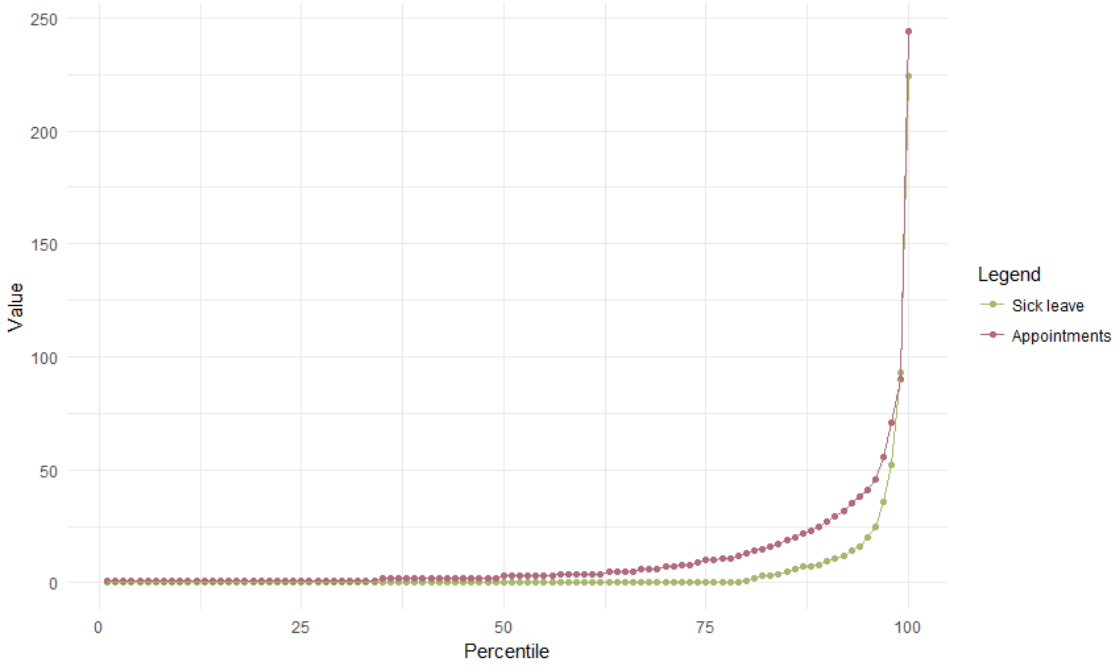

Figure 7: The n-th percentiles of sick leave length and number of appointments related to each treatment path. The x-axis represents the percentile rank and the y-axis the value of the variable. In the case of sick leave data, this is the length of possible sick leave, and for appointments, it is the number of total appointments throughout the treatment pathway.

We calculated the mean absolute error and median absolute deviation values of the regression results on the length of the sick leave and the number of upcoming appointments throughout the treatment pathway with the RF and SVM methods, which are depicted in the Table 9. We can see how RF outperforms SVM for both of the cases, though for the prediction of the number of appointments this difference is quite small. The average length of sick leave was 4.24 days and the number of appointments 9.68, while the median value for the length of the sick leave was 0 days and the median treatment path length was 3 appointments. The MAE values are also quite high related to the data averages. The median error values are, however, quite modest, with the deviation in error rate for the length of the sick leave almost zero with the RF method. For the number of appointments, the median of error values is half of that of the mean, suggesting a highly skewed error distribution for both of the cases. The high MAE values are partly explained by the wide distributions in both the length of sick leave and the number of appointments. Very large outliers create a bias on the data, overshadowing the otherwise quite modest error rates for both the length of the sick leave and the number of appointments. It is important to note, however, that most of the patients are prescribed no sick leave at all. Thus simply guessing a sick leave length of 0 days would give us a median error of 0. The median error rate for the number of appointments is high as well, with error rates larger than the median value of the number of appointments. This suggests that overall, the models do not perform exceptionally well in regression tasks, mainly due to the large outliers in the data.

	RF		SVM	
	MAE	MAD	MAE	MAD
Length of sick leave	3.71	0.05	4.28	1.11
Number of appointments	8.15	4.03	8.28	4.00

Table 9: Mean Absolute Error (MAE) and Median Absolute Deviation (MAD) error values of the prediction of sick leave length and the number of appointments in the treatment path.

Another interesting and often more important task for medical use is detecting outliers. As such, it may be essential to know if the person will be on sick leave for a long time or if the treatment pathway itself will be longer than usual. Table 10 depicts the precision, recall, accuracy, and balanced accuracy values of methods when used to predict outliers or presence of values in the data. The values were calculated as a classification task on the data set. A new feature was created for each treatment pathway depicting if that pathway was over the threshold in length, and the model was set to predict the value of this feature. Alongside the classifiers, a trivial classifier was created that simply voted for the median value for each of the classification tasks, which was a truth value of “False” for all three cases. This is analogous to the program predicting no outliers for any data point. The simple

accuracy is shown for this model as well. To be viable, a method should beat this benchmark considerably in accuracy. For the first case of detecting if the person will be on sick leave at all some progress is seen, with the RF model achieving a 90% accuracy and balanced accuracy and the SVM model a slight improvement over the trivial classifier at 83% accuracy and 82% balanced accuracy. For detecting large outliers of over ten sick leave days, both RF and SVM beat the dummy model slightly in accuracy performance but fare considerably better when we study the balanced accuracy. As the dummy model votes negative for all cases, the true positive rate is 0. Thus the balanced accuracy of the dummy model is half of the accuracy score or 45.5% for long sick leaves. The case is similar for detecting outliers for the number of appointments, with RF and SVM gaining similar accuracy scores but much better balanced accuracy scores than the dummy model with a balanced accuracy score of 38.5%.

	RF				SVM				Dummy
	P.	R.	A.	B.A.	P.	R.	A.	B.A.	A.
Length of sick leave $\geq$ 0 days	0.71	0.89	0.90	0.90	0.32	0.81	0.83	0.82	0.79
Length of sick leave $\geq$ 10 days	0.70	0.65	0.94	0.81	0.68	0.58	0.93	0.78	0.91
Number of appointments $\geq$ 10	0.79	0.45	0.77	0.69	0.57	0.47	0.79	0.68	0.77

Table 10: Classification accuracy on sick leave length and the number of appointments for both the RF and SVM classifiers. Precision (*P.*) and recall (*R.*) values are stated for each as well as the accuracy (*A.*) and balanced accuracy (*B.A.*) values, which weights the accuracy of the model for very imbalanced data sets. For reference the accuracy rates for a dummy classifier are presented, that assumes each value to be negative.

For the applications of the models, the positive and negative accuracies are equally important. Thus the balanced accuracy rate gives a good estimation of the performance of the models. We can achieve high accuracy in only estimating negative values, but this type of model is useless if we are interested in the positive values as well. The predictive ability of the models used perform quite well from this perspective.

The precision and recall scores provide us with further insight into the data. The precision values for the RF models are around 70% to 80% for all of the cases. Thus the ratio of true positives out of all of the predicted positive values is quite high. SVM is much more varied with precision for whether the sick leave is longer than 0

being as low as 32%. Length of sick leave of over 10 days and treatment pathways of over 10 appointments have better values but still fall behind RF. Thus a large portion of the predicted positive values is, in fact, false positives. For the recall values, the values are more similar across the methods. RF and SVM both have a recall value around 85% for detecting non-zero sick leave lengths, meaning that they capture a substantial part of all of the positive values in the data. Examining the outliers of sick leave length, the recall values are between 58% and 65% and for outliers of the number of appointments just below 50% for both models. This general trend can also be seen from the balanced accuracy values with the non-zero sick leave length resulting in the highest recall values and lower values for long sick leave and a large number of appointments.

Thus it can be seen that the quality of models is nuanced. The models perform about the same as a trivial classifier for raw accuracy but are more capable when considering that the models have to be able to classify both positive and negative values. Still, the precision and recall values leave room for improvement, especially for the SVM model, with RF having large problems with recall values for detecting the large outliers as well. Still, both models have balanced accuracy values of over 50% and can be seen as useful in predicting the desired outcome variables. Overall, the SVM performed somewhat worse on the classification tasks than the RF. However, it is important to note that across all the tests, the SVM method was considerably faster, with the RF method finishing each model in around a minute while the SVM model was trained in a few seconds.

Random Forest has the added benefit of providing us with a straightforward measure of variable importance. The Breiman-Cutler permutation importance calculation is performed by taking the out-of-bag error values of random trees, removing individual features one at a time from the feature space and re-calculating the error rate [32][46]. The feature importance is calculated as the difference between the two values. The model is implemented in the library `randomForestSRC`. The importance of values can be seen in Figure 8, where the values are calculated from the model used to predict if the person will be on sick leave during the treatment path. The values represent the importance of each variable on the length of the sick leave.

The initial sick leave prescribed during the first session of the treatment path seems to be the main explanatory factor for the values. For a rather biased data set this is unsurprising. The results also suggest that sick leave is prescribed rather early in the treatment path. The other values, however, are unexpected. High values for the doctor specialisation features are most likely explained because not all doctors prescribe sick leave regularly, but rather sick leave is usually prescribed by an occupational doctor. Interestingly enough, the primary and secondary diagnoses have a negative effect on the outcome of the length of the sick leave. Thus there is no variance in need for sick leave between patients with the same diagnoses. The other variables carry little to no weight as well, with the only non-negative values being the number of prior appointments and the average time between past visits. Thus information on the past number of appointments and the frequency of these improve model performance. Variables such as gender, age or even the prior tendency for sick

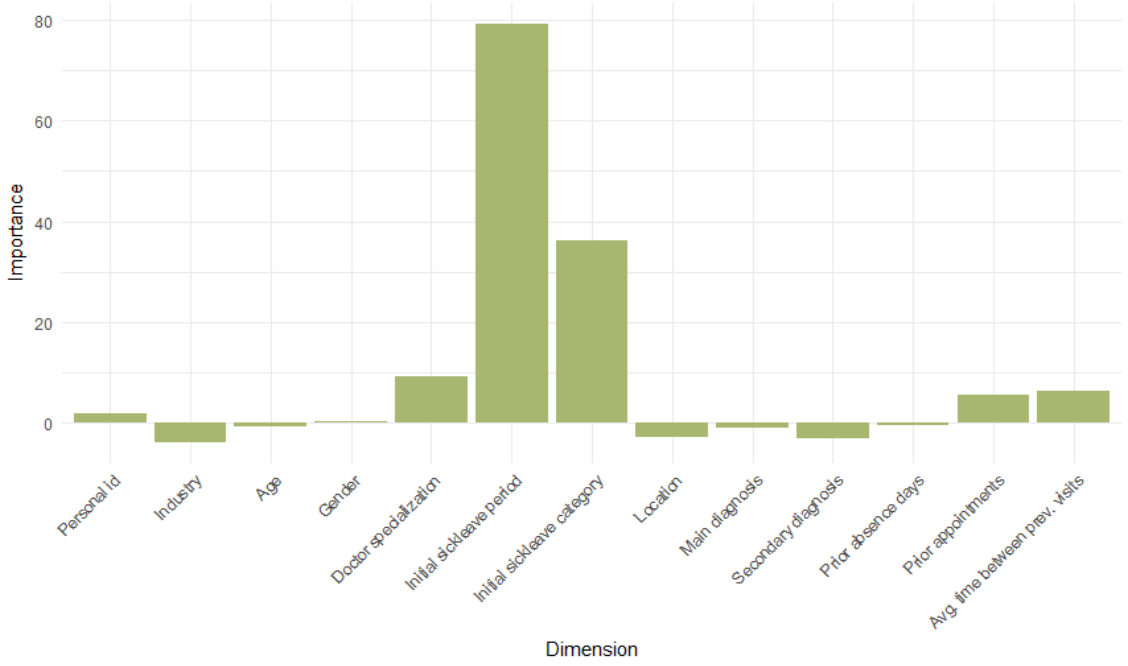

Figure 8: Importance of the various dimensions used for the predictive models. The values represent the importance of each variable on the length of sick leave prescribed. The model employs a Breiman-Cutler permutation importance calculation to asses the importance of the variables. [32][46]

leave seem not to affect the outcome very much. Conversely, it can be interpreted that doctors put little weight on patient history when prescribing sick leave. This is interesting as the data is readily available to the doctor via the EHR system during each appointment.

Overall the RF method performed somewhat better in most of the tasks compared to SVM, but the SVM method was much faster. The overall error rates are quite small for classification tasks but perform worse on regression tasks. The initial sick leave period seems to be an essential factor in deciding the future treatment path of the person with surprisingly little effect on personal medical history or demographic background.

4.3 Errors and confidence

Data quality was surprisingly hard to assess for EHR data throughout the study. This was mainly due to a large number of missing data for each appointment. The doctor usually writes down as much information as is required by the guidelines of each service provider, but rarely more. In other words, any fields that are not enforced are not usually filled. This may occur for example in the case of follow-up checks for the patient after the initial appointment, where the information from the initial visit is not repeated for each different entry. Some cases were also identified

where all appointment data is missing. This can happen for some treatment paths where the treatment sessions are not enforced to be recorded at all or very lightly. In such cases accurately clustering these appointments is hard, and the task of assessing if these treatment paths are correct more complicated. While the results were deemed meaningful, at least a part of the error values are probably due to the inaccurate base data for the analysis.

During the validation of the clustering results, we only assessed the false positives of the data. This was done to keep the amount of work for medical professionals feasible. However, this left out important data on the quality of the results, most importantly if some values were missing from the treatment paths. A more work-intensive method would be to construct pre-made clusters with medical experts as a baseline for validation. For the scope of this project, this was deemed too heavy, as building a clustering by hand with any significant number of appointments is rather tedious. However, given this golden baseline, we could then compare the results with any similarity measure, for example, ARI, and deduce the best clustering as that being the closest one to the baseline. Such a baseline could also be used as model data for any predictive models in the future.

Additionally, the clusters used were hard clusters in the sense that one appointment could only be included in one treatment pathway. In reality, this is not the case, and multiple diseases can be treated with a single appointment. Out of the clustering methods tested the model-based method is capable of soft clustering, which instead of class labels assigns probabilities of each appointment belonging in each treatment pathway. Such soft clusters could be used to construct more truthful representations of treatment pathways. Such a study was ruled out from the scope of this Thesis but might prove an interesting starting point for future studies.

The amount of data used for the analysis was sampled down due to the limitations of both the used hardware and software. This brings an inherent error in the analysis. The total number of individual appointments for our data set is 5 240 100 out of which our sample was 46 340, giving us a confidence level of 99% with a margin of error of 0.6% on the data. The sample was therefore quite sufficient, but for many applications, more extensive coverage is attractive. Further issues to mitigate this effect with the chosen methods could be running the analysis multiple times over different subsets of the data or improving computational capacity.

The choice of a single clustering pass over a data set of multiple people turned out to be a significant problem. As discussed above, the choice to cluster a data set comprising of multiple patients was done to improve performance and comparability, and because the risk of treatment pathways mixing was estimated to be small. This turned out to be a false assumption, with a major part of the FAMD and the model-based method producing clusters spanning multiple patients. This introduced some error into the clustering results, as any splitting of clusters increased the number of clusters and made the resulting clusters smaller as a result. All results from the clustering step underestimate the size of the clusters and overestimate the number of the clusters slightly. The same error is then carried over to the prediction results, as the data sets for prediction models were constructed from the class labels obtained from the clustering step. For future studies, this issue can be mitigated at least in

two ways. The first is a model where each patient is treated individually and the clusters formed from all of the data related to that individual and the repeated for each customer. Another choice would be to add another clustering step after the splitting of clusters that aims to combine any split clusters for a patient back together. In both cases, care would have to be taken so that the issue is correctly addressed. For person-wise clustering, the problem is with heavier performance and ensuring comparability between treatment pathways between patients. For the additional clustering step, the issue is ensuring that the resulting clusters are meaningful and the clustering step increases information of the clusters rather than pairs clearly incorrect clusters together. However, if done correctly, either could increase the quality of the results.

The predictive models have room for optimization as well. The models, as such, had quite a small amount of training data. With a larger pool of classification results, the error rates for the prediction models could be brought down as the training data size for the models grows. Another approach would be tuning the parameters of the models more, or in the case of the random forest, model, increasing the number of trees further. Both of the steps could be used in unison for more accurate results. The data model of the prediction task was quite simple, as well. With a deeper understanding of the patient history and the relevance of the data to future diagnoses, the model could be improved further. This is the most logical first step for any possible future work on the subject.

A large portion of difficulty in working with the data was with the mixed data type. An interesting approach would be to convert the data set wholly to either numerical of categorical data. As the data is already mainly categorical, some performance improvements could maybe be achieved if we converted the numerical values into categorical, which would enable us to work on a broader selection of tools for the data. However, as of now, both numerical and categorical data carry value for the resulting clusters and predictions and keeping both is justified in the scope of this work.

5 Conclusions

We applied three distinct clustering methods to a large EHR data set to identify treatment pathways in mental health efficiently. The methods, FAMD with k-means clustering, Gower with hierarchical clustering, and model-based approach, all resulted in meaningful results with distinct features concerning the cluster size, the number of clusters and the inclusion of additional appointments without a mental health diagnosis. The FAMD method produced the most but smallest clusters while Gower and model-based methods created fewer but larger treatment pathways with additional diagnoses between the mental health diagnoses. Expert validation of the clustering results assigned the FAMD method as the most accurate representation of the treatment pathways. However, it was noted that the external data provided by, for example, the model-based method was beneficial in studying the treatment pathway as a whole. Thus the different methods are not mutually exclusive but might fill different roles. While the FAMD method gives the most accurate results and might be useful for strict clustering of mental health appointments alone, the broader clustering techniques could be used for a more general overview on the quality of treatment and its effects. However, overall, the clusters produced by the three methods were very different. According to the ARI measure, the clusters are fundamentally different. Also, when comparing the features and contents of each clustering result, we can see that these vary across the results as well. Overall then we arrive at three very different ways to construct treatment pathways from the data set, with the FAMD method being the most accurate but the model-based method being the most interesting for treatment path analysis according to the experts.

The predictive power of the models used was mediocre for the data with quite large error rates in regression tasks. Classification tasks, however, yielded quite good balanced accuracy values on predicting outliers and non-zero values on sick leave length and treatment pathway length. The tests suggest that employing simple machine learning methods to electronic health data can yield better-than-random prediction values on various outcome variables related to our chosen treatment pathways. Especially the good results in outlier detection can be useful for future applications of these methods. Such outlier information can be useful, for example, for finding patients needing specialized care along with their treatment, or steering patients with long treatment paths more quickly to the correct experts. In general, the random forest algorithm performed better than the SVM did, but the performance of both was good enough to warrant possible future research into applications on treatment pathway prediction. The performance of SVM, however, was much better with running times of seconds instead of minutes in training the models. Moreover, the clustering results from the treatment pathway analysis can be used for more kinds of analyses than those outlined in this study. For example, the number and type of probable upcoming medication along the treatment path can be calculated and predicted used the same methods outlined in this Thesis, thus broadening the possibilities of data analysis on the previously unstructured electronic health data.

The feature analysis revealed that there is surprisingly little effect on patient background on prescription of sick leave. The initial sick leave and doctor occupation

hold the most weight, but diagnosis or personal features such as age, gender and treatment history hold little to no prediction capability. The results also show that the number of prior sick leave days is a poor indicator for predicting length of the sick leave period. Instead, the number of prior appointments and the frequency of these improve model performance. Possible strategic or operative decisions made purely on historical or demographic data alone are, therefore, likely to be unrelated to the sick leave rates of the treatment paths identified in the study. However, the relatively high mean error rates on the regression predictor create some uncertainty in these results.

Overall, we tested the feasibility of the available R libraries on large-scale medical data of mental health patients. The results were promising with some concern for performance issues related to the mixed data nature of electronic health data and the large scale of the related data sets. However, medium-scale studies clearly are feasible, with much room for optimisation on the performance of the models. The results of this study can also be generalised for larger data sets, such as complete EHR systems. The results in clustering and prediction showed that meaningful relationships and models can be constructed from electronic health record data to be used for the analysis of treatment pathways.

References

[1] S B Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. Technical report, 2007.

[2] Kansaneläkelaitos and Timo Partio. Kelan kuntoutustilasto, Suomen virallinen tilasto: Sosiaaliturva 2017. Technical report, 2017.

[3] Suomalaisen Lääkäriseuran Duodecimin ja Suomen Psykiatriyhdistys ry:n asettama työryhmä - Depressio. Technical report, 2016.

[4] Tuija Turunen. Personal interview, 21.2.2019.

[5] ISO/DTR 20514. Health Informatics: Electronic Health Record; Definition, Scope, and Context. 2004.

[6] Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International Journal of Medical Informatics*, 77(5):291–304, 2008.

[7] Terveyden ja hyvinvoinnin laitos. Potilastiedon arkisto, Potilastietojärjestelmien käyttötapa. Technical report, 2016.

[8] Zhuoran Wang, Anoop D. Shah, A. Rosemary Tate, Spiros Denaxas, John Shawe-Taylor, and Harry Hemingway. Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning. *PLoS ONE*, 7(1):e30412, jan 2012.

[9] Anita Riipinen. Personal interview, 22.2.2019.

[10] Brett K. Beaulieu-Jones and Casey S. Greene. Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of Biomedical Informatics*, 64:168–178, dec 2016.

[11] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6(1):26094, sep 2016.

[12] JD Banfield. Model-based Gaussian and non-Gaussian clustering. *JSTOR*, 1993.

[13] Lynette Hunt and Murray Jorgensen. Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):352–361, 2011.

[14] Michel van de Velden, Alfonso Iodice D'Enza, and Angelos Markos. Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, (September), 2018.

[15] Di Zhao and Chunhua Weng. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of Biomedical Informatics*, 44:859–868, 2011.

[16] Barbara Michiels, Van Kinh Nguyen, Samuel Coenen, Philippe Ryckebosch, Nathalie Bossuyt, and Niel Hens. Influenza epidemic surveillance and prediction based on electronic health record data from an out-of-hours general practitioner cooperative: Model development and validation on 2003-2015 data. *BMC Infectious Diseases*, 17(1), 2017.

[17] Peggy L. Peissig, Vitor Santos Costa, Michael D. Caldwell, Carla Rottscheit, Richard L. Berg, Eneida A. Mendonca, and David Page. Relational machine learning for electronic health record-driven phenotyping. *Journal of Biomedical Informatics*, 52:260–270, dec 2014.

[18] Zongqi Xia, Elizabeth Secor, Lori B. Chibnik, Riley M. Bove, Suchun Cheng, Tanuja Chitnis, Andrew Cagan, Vivian S. Gainer, Pei J. Chen, Katherine P. Liao, Stanley Y. Shaw, Ashwin N. Ananthakrishnan, Peter Szolovits, Howard L. Weiner, Elizabeth W. Karlson, Shawn N. Murphy, Guergana K. Savova, Tianxi Cai, Susanne E. Churchill, Robert M. Plenge, Isaac S. Kohane, and Philip L. De Jager. Modeling Disease Severity in Multiple Sclerosis Using Electronic Health Records. *PLoS ONE*, 8(11):e78927, nov 2013.

[19] Kirstine Rosenbeck Gøeg, Ronald Cornet, and Stig Kjær Andersen. Clustering clinical models from local electronic health records based on semantic similarity. *Journal of Biomedical Informatics*, 54:294–304, apr 2015.

[20] J C Gower. A General Coefficient of Similarity and Some of Its Properties. Technical Report 4, 1971.

[21] Maurizio Vichi and Henk A L Kiers. Factorial k-means analysis for two-way data. Technical report, 2001.

[22] J. de Leeuw and J. van Rijckevorsel. HOMALS and PRINCALS - Some generalizations of principal components analysis. 1980.

[23] Henk A L Kiers. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56(2):197–212, 1991.

[24] Jérôme Pagès. Analyse factorielle de donnees mixtes: principe et exemple d'application. *Montpellier SupAgro*, 52:93–111, 2004.

[25] Volodymyr Melnykov and Ranjan Maitra. Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116, 2010.

[26] B S Everitt. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, 6(April):305–309, 1988.

[27] Lynette Hunt and Murray Jorgensen. Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, 41(2):154–171, 2003.

[28] A P Dempster, ; N M Laird, and ; D B Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. Technical Report 1, 1977.

[29] Damien McParland and Isobel Claire Gormley. Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification*, 10(2):155–169, 2016.

[30] Tin Kam Ho. Random Decision Forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pages 278–282, 1995.

[31] Yali Amit and Donald Geman. Ullman Shape Quantization and Recognition with Randomized Trees. Technical report, 1997.

[32] Leo Breiman. Random Forests. Technical report, 2001.

[33] Andy Liaw and Matthew Wiener. Classification and Regression by RandomForest. Technical report, 2002.

[34] V N Vapnik and A. Ya. Chervonenkis. Necessary and Sufficient Conditions for the Uniform Convergence of Means to their Expectations. *Theory of Probability & Its Applications*, 26(3):532–553, 2005.

[35] Corinna Cortes, Vladimir Vapnik, and Lorenza Saitta. Support-Vector Networks. Technical report, 1995.

[36] Harris Drucker, Chris J C Burges, Linda Kaufman, Alex Smola, and Vladimir Vapoik. Support Vector Regression Machines. Technical report, 1997.

[37] William M Rand. Objective Criteria for the Evaluation of Clustering Methods. Technical Report 336, 1971.

[38] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. Technical report, 2010.

[39] Lawrence Hubert and Phipps Arabie. The trasition from bargaining to a competitive market. *Journal of Classification*, 2:193:218, 1985.

[40] Terveyden ja hyvinvoinnin laitos. *Tautiluokitus ICD-10. Klassifikation av sjukdomar. Luokitukset, termistöt ja tilasto-ohjeet.* 2011.

[41] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2018.

[42] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[43] Sébastien Lê, Julie Josse, and François Husson. {FactoMineR}: A Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[44] Serge Iovleff. *MixAll: Clustering and Classification using Model-Based Mixture Models*, 2018.

[45] Luca Scrucca, Michael Fop, Thomas Brendan Murphy, and Adrian E Raftery. {mclust} 5: clustering, classification and density estimation using {G}aussian finite mixture models. *The {R} Journal*, 8(1):205–233, 2016.

[46] H Ishwaran and U B Kogalur. *Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2017.

[47] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017.