Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Inger-Ane Sætra Schefte

Characterising and Clustering Co-Occurring Morbidities and Medications in Child and Adolescent Mental Health

Master's thesis in Computer Science
Supervisor: Øystein Nytrø
Co-supervisor: Kaban Koochakpour
June 2023

NTNU
Norwegian University of
Science and Technology

Inger-Ane Sætra Schefte

Characterising and Clustering Co-Occurring Morbidities and Medications in Child and Adolescent Mental Health

Master's thesis in Computer Science
Supervisor: Øystein Nytrø
Co-supervisor: Kaban Koochakpour
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

This master's thesis analyses electronic health records of patients within the Child and Adolescent Mental Health Services (CAMHS) in Norway, with the primary focus on characterising co-occurring morbidities and medication usage among this population. Given the frequent co-occurrence of mental disorders, this research is motivated by the rising prevalence of mental health disorders in young populations and the need for effective tools for diagnostics and treatment. By leveraging pattern recognition and clustering methodologies, the study characterises distinct patterns and patient subgroups with similar diagnoses and treatment paths. The feasibility of this approach in distinguishing subgroups within the CAMHS population is explored and validated.

This research contributes to a deeper understanding of the interplay between mental health disorders and medication usage in young patients. The findings offer insights into the pharmaceutical approaches to managing mental conditions and aid in characterising specific patient subgroups based on their diagnoses and medications. Overall, this study highlights the potential of using machine learning techniques to analyse EHRs and contribute to the improvement of mental health care services for children and adolescents.

Sammendrag

Denne masteroppgaven fokuserer på å analysere elektroniske pasientjournaler (EPJ) til pasienter i Barne- og Ungdomspsykiatrisk Poliklinikk (BUP), med spesiell vekt på å karakterisere samtidige diagnoser og medisinbruk blant disse pasientene. Motivasjonen for forskningen er den økende forekomsten av psykiske lidelser blant unge mennesker og behovet for effektive verktøy for diagnostisering og behandling. Studien bruker mønstergjenkjenning og klyngemetoder for å identifisere mønstre og karakterisere pasientundergrupper med lignende diagnoser og medisinsk behandlingsforløp. I tillegg, diskuteres og vurderes den praktiske anvendeligheten av denne tilnærmingen for å karakterisere pasientundergrupper i BUP.

Resultatene bidrar til en dypere forståelse av samspillet mellom psykiske lidelser og medisinbruk hos unge pasienter. Funnene gir innsikt i farmasøytiske tilnærminger for å håndtere samtidige mentale diagnoser og bidrar til å karakterisere spesifikke pasientundergrupper basert på deres diagnoser og medisinbruk. Denne studien understreker potensialet i å bruke maskinlæringsteknikker til å analysere EPJer og kan bidra til innsikt rundt diagnoser og behandling innen psykisk helsevern for barn og ungdom.

Preface

This Master's thesis represents the results of research conducted in the spring of 2023 at the Norwegian University of Science and Technology (NTNU). It is the product of a collaboration with the IDDEAS project at NTNU and was supervised by Øystein Nytrø and Kaban Koochakpour from the Department of Computer Science.

The primary goal of this research was to analyse co-occurring morbidities and medications in electronic health records of patients within the CAMHS. Through this work, an attempt has been made to characterise patient profiles and subgroups by utilising cluster analysis. Furthermore, the results have been interpreted in the context of Child and Adolescent Mental Health Services (CAMHS). As this is an interdisciplinary endeavour, it required input and feedback from professionals in the medical domain.

Acknowledgements are due to several individuals and groups who contributed to this project. First and foremost, I would like to thank Øystein Nytrø and Kaban Koochakpour for their guidance and support throughout the project. Odd-Sverre Westbye and the rest of the IDDEAS team played a key role in the implementation of the project and provided insightful guidance. Hanne Klæboe Greger from the CAMHS clinic at St. Olavs Hospital in Trondheim offered valuable input and discussions regarding the research results. Technical assistance was also generously provided by the HUNT Cloud team.

During this project, several fellow students contributed to making the process more enjoyable. Stine and Sofie provided much-needed distractions at our study room at Gamle Fysikk, while Andrea and Solveig were indispensable during numerous coffee breaks. Technical discussions with Tiril also added a valuable dimension to the work. Lastly, I would like to acknowledge the support of friends and family, who offered encouragement throughout the duration of this project.

Inger-Ane Sætra Schefte

Trondheim, June 9, 2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Co-occurring morbidities are a common phenomenon in the context of mental health disorders. Investigating patterns of co-occurring morbidities, and their corresponding treatment protocols, can provide valuable insight to mental health care providers like Child and Adolescent Mental Health Services (CAMHS), where the presentation of co-occurring disorders and their associated treatments can influence treatment outcomes. CAMHS are the specialist healthcare instances working with children and adolescents with mental disorders. This study analyses the diagnosis and medication data of patients treated at CAMHS. The included patient population comprises episodes of care of patients who have received at least one diagnosis within CAMHS. This research aims to characterise recurring patterns of diagnoses and associated medication use within this patient population and characterise distinct patient subgroups within CAMHS based on their demographic, diagnostic and therapeutic profiles. The findings of this study can contribute to a more comprehensive understanding of co-occurring disorders and medications within CAMHS, ultimately improving patient care.

As the title suggests, this project's focus is to characterise and cluster co-occurring morbidities and medications in CAMHS. *Morbidity* refers to the condition of suffering from a medical condition. *Medication* refers to drug-based interventions in this context. The objective of *characterising* is to describe the distinct features of patient subgroups within CAMHS. Furthermore, *clustering* involves utilising unsupervised machine learning techniques to group similar records, in this project, records of patient morbidities and medications. By employing clustering, insights into the distinct characteristics of patient subgroups in CAMHS can be derived.

This research involved a collaborative effort with IDDEAS, the Individualised Digital DEcision Assist System, a research group working closely with CAMHS in Norway. This multidisciplinary team, composed of developers, clinicians, researchers, entrepreneurs, and specialists in health informatics, aims to develop and evaluate an IDDEAS clinical decision support system (CDSS) designed explicitly for CAMHS [IDDEAS, 2022b]. The IDDEAS CDSS is envisioned to provide knowledge-based support to clinicians, contributing to improving CAMHS. As information technology has yet to be widely integrated into CAMHS in Norway, the IDDEAS system will be the first CDSS developed directly towards mental health care for children and adolescents in Norway. A key focus of this system is children and adolescents diagnosed with attention-deficit hyperactivity disorder (ADHD), representing approximately 29% of mental disorders within this demographic group [IDDEAS, 2022b].

This chapter introduces the project and provides an overview of the research's background motivation, goals, and research questions. Furthermore, it outlines the research methods and presents an overview of the thesis structure, briefly describing the subsequent chapters.

1.1 Background and Motivation

Co-occurring morbidities are frequently observed within the CAMHS population, where multiple mental disorders tend to manifest simultaneously. The ability to differentiate and accurately diagnose these disorders is crucial for clinicians and professionals working in CAMHS. Appropriate diagnoses and tailored treatments depend on effectively identifying and distinguishing symptoms associated with each disorder. However, challenges can arise due to the inherent overlap in symptoms among various conditions. Ensuring precise diagnostics optimises treatment outcomes. Therefore, characterising patient subgroups by specific combinations of diagnoses and medications within CAMHS can prove valuable as it enables the recognition of prevalent combinations of disorders and medications, facilitating more efficient recognition and treatment.

This research aims to characterise co-occurring morbidities and medication within the CAMHS population. It seeks to characterise distinct patient subgroups through the application of clustering techniques. Moreover, it intends to validate the results within the context of CAMHS and discuss the characterised phenomena. The data records of interest are of patients admitted for treatment at the CAMHS clinic at St. Olavs Hospital in Trondheim receiving one or more unique diagnoses. Additionally, associated prescribed medications and demographic information about the patient form part of this analysis. The relevant information

will be extracted from the database the IDDEAS project utilises, containing data from the electronic health record (EHR) system, BUPdata, used at the CAMHS clinic at the St. Olavs Hospital at the time of data collection.

This research's motivation stems from characterising co-occurring morbidities and medications in CAMHS, providing insights into medical practices within CAMHS. Validating results in the context of CAMHS is an integral part of this process to ensure clinical applicability. As such, findings will undergo assessments in consultation with clinicians at the CAMHS clinic at St. Olavs Hospital. These professionals' feedback will be invaluable for evaluating the clinical feasibility of applying clustering techniques to clinical data and assessing the chosen algorithm's effectiveness in characterising patient subgroups. Interactions and validations with clinicians represent key elements of this research, serving as sources of clinical insights and driving the discovery of relevant results.

1.2 Goals and Research Questions

Building upon the motivational context presented in the previous section, this section outlines the specific goals and research questions defined for this study.

The overall goal of the research is:

Goal *To analyse co-occurring morbidities and medication in EHRs of patients in CAMHS, investigate if patient profiles and subgroups can be characterised by cluster analysis, and interpret the results in the context of CAMHS.*

The goal of this master's thesis is to conduct an analysis of co-occurring morbidities and medications in CAMHS by utilising EHRs of patients in CAMHS. In this context, *morbidity* refers to the state of experiencing a particular medical condition or disease, while *medication* pertains to pharmaceutical treatments. *Characterising* assesses the objective of describing distinct features. The research seeks to characterise patient profiles and subgroups through cluster analysis. This clustering analysis may uncover trends or phenomena within the field of youth mental health. This project will also evaluate the appropriateness and efficacy of clustering techniques on the dataset and assess the validity of the findings within the clinical domain. The evaluation will comprise a dual focus, encompassing both a clinical perspective, where input from clinicians will be vital in providing clinical context to the findings. Additionally, a technical perspective, assessing the process and technical aspects of the research, will be evaluated. Overall, this research aims to contribute to characterising co-occurring morbidities and medications in CAMHS and potentially improve patient care.

Further, the research questions framing this thesis are presented. These questions aim to assess the application of clustering to analyse diagnoses and medications and how cluster analysis can characterise patient subgroups according to diagnoses and medications. The research questions are defined as follows:

Research question 1 *How can clustering be used to analyse diagnoses and medications?*

This question centres around the utilisation of clustering techniques. The objective is to explore and identify a suitable method that can be applied to analyse diagnoses and medications. *Clustering* enables the grouping of similar records based on specific criteria or features. In this context, the focus is to understand how this technique can characterise patterns and relationships within diagnosis and medication data.

Research question 2 *How can the results of the cluster analysis characterise patient subgroups according to diagnoses and medications?*

This research question delves into how the findings from the cluster analysis established in the first research question can contribute to the characterisation of patient subgroups according to their diagnoses and medications. As highlighted by the goal, the research will utilise data from EHRs in CAMHS and characterise patient subgroups within this context.

1.3 Research Method

This section summarises the research methodology applied in the experiments and why this methodology has been chosen. The methodology is described in detail in Chapter 6.

To address the project's goals and research questions, a systematic approach was adopted to extract, process, and analyse data from the EHRs of patients within CAMHS. The focus was on understanding the data within its clinical context, identifying a suitable approach for characterising co-occurring morbidities and medications reflected in these records, and applying the identified methodology. Subsequently, the characterised subgroups and the results were validated and contextualised within the scope of CAMHS.

Clustering analysis was used to cluster diagnoses and medications to ascertain the feasibility of characterising patient subgroups of diagnoses and drugs using the selected algorithm. The process was implemented in multiple stages. It started with extracting relevant data, followed by thorough cleaning and preprocessing to prepare the data for analysis. Subsequently, an exploratory data analysis (EDA)

was performed to gain insight into the data. Furthermore, pattern mining and clustering techniques were employed to identify patterns and characterise patient subgroups within the dataset. Initially, a total of 16,202 episodes of care were considered, leading to a refined subset of 8,499 episodes that met the criteria of having more than one diagnosis or medication for the subsequent clustering experiment. This subset facilitated the identification of patterns and characterisation of subgroups within the cohort, potentially offering novel insights into patient diagnostics and treatment practices within CAMHS.

The clustering algorithm employed for the experiments is MASPC (Maximal-frequent All-confident pattern Selection and Pattern-based Clustering). This algorithm operates in two distinct phases: the initial phase entails identifying frequent patterns, while the subsequent phase involves clustering the records based on the identified patterns. The utilisation of MASPC on the dataset facilitates the identification of patterns of diagnoses and medications, allowing for the clustering of a diverse range of combinations of diagnoses and medications within the dataset. The main goal of the clustering experiment is to characterise subgroups within the dataset and validate their clinical relevance rather than establishing definitive truths. This analysis provides insights into patient subgroups based on patterns of diagnoses and medications, enhancing the understanding of the dataset and its implications about the CAMHS practice.

The results of the methodology need to be validated in clinical and technical contexts. The Silhouette Index (SI) and Calinski-Harabasz Index (CI) have been used as validity metrics to find the optimal number of clusters and assess the clustering results' quality. For the clinical application and validation of the results, choices along the way in the research process and the analysis results have been presented to professionals to receive feedback and interpret the result in the clinical context. The input received during the presentation of the results has played an essential role in validating the findings and providing valuable insights into the alignment between the results and the clinical reality within CAMHS.

1.4 Thesis Structure

This section outlines the structure of this thesis.

The introductory chapter is followed by Chapter 2, which presents the clinical background theory needed to comprehend the clinical aspects of the research. This chapter equips the reader with an understanding of the medical domain relevant to this research.

Chapter 3 introduces the technical background theory of the experiments, pre-

senting key concepts employed in the subsequent experiments.

Chapter 4 offers an overview of related work, specifically focusing on the use of machine learning techniques in analysing co-occurring morbidities. Clinical literature on co-occurring morbidities in the mental health domain is also presented.

Chapter 5 presents the dataset and research environments utilised, along with legal agreements and approvals required for data access. Subsequently, Chapter 6 details the research methodology, justifying the chosen research approaches.

Chapter 7 explains the experimental design, detailing the experimental aims, plan and setup that led to the results presented and illustrated in Chapter 8, Experimental Results.

Chapter 9 presents the clinical validation of the results within the context of CAMHS, drawing upon input and feedback from clinicians during the presentation of the results.

Chapter 10 provides a discussion of the research pertaining to project planning and execution. This chapter encompasses various elements, including a discussion of the project's timeline, the clinical validation during experiments, experimental limitations, an evaluation of the experimental aims, a discussion of technical aspects such as data cleaning and preprocessing, and the application of the MASPC algorithm.

Chapter 11 serves as the final chapter, where the research is concluded and summarised. It provides an overview of the contributions made by the study and identifies potential areas of future work. In this chapter, the goals and research questions are evaluated and analysed.

Chapter 2

Clinical Background

This chapter elaborates on essential clinical background theory, providing a foundation for understanding the research methodology. The specialist health service in Norway, including CAMHS, utilised the International Classification of Diseases, Tenth Revision (ICD-10) diagnostic code system and the ICD-10 multiaxial classification system during the time of data collection, making introducing and explaining these concepts necessary. Specifically, the chapter presents clinical diagnostics in CAMHS, the use of ICD-10 and the multiaxial classification system as key diagnostic frameworks. Proceeding further, the phecodes system is introduced. This system establishes a mapping system that correlates ICD-10 diagnoses with clinical phenotypes. Lastly, the chapter presents the Anatomical Therapeutic Chemical (ATC) classification system, a categorisation system for drugs and medications. This chapter lays the groundwork for understanding the clinical part of this research by providing an understanding of these systems.

2.1 Clinical Diagnostics in CAMHS

This section explores the concepts related to clinical diagnostics in CAMHS during the time of data collection. In CAMHS, two important tools were used for the accurate identification and classification of diagnoses: ICD-10 and the multiaxial classification system. This section explores these diagnostic frameworks and their use in CAMHS. The fall specialisation project inspires some of the content in this section.

2.1.1 ICD-10

The ICD-10 is the 10th revision of the international classification of diseases and health-related problems and represents a universally recognised system of cataloguing illnesses and health-related issues. Administered by the World Health Organization (WHO), this classification tool has been the official standard for the classification of mental disorders in Norway since 1997. The oversight and implementation of ICD-10, including tailoring its application to suit the Norwegian health service, is executed by the *Direktoratet for e-helse* (English: Directorate for Electronic Health) [Malt and Braut, 2022].

The coding scheme employed by ICD-10 involves a letter from A to P, followed by two numeric characters. Compared to the ninth revision of the classification system, ICD-9, ICD-10 provide more categories for classification and the possibility to specify classifications further using decimals [Sosial- og helsedirektoratet, 1999]. In CAMHS, the ICD-10 codes starting with the letters F, R, and Z are more frequent. However, diagnoses under other codes are also present.

F-Codes: These codes represent a section within ICD-10, specifically Chapter 5, encompassing diagnoses F00-F99 related to mental and behavioural disorders. Within this spectrum, F90-F99 is a subgroup that covers diagnostic codes for behavioural and emotional disturbances that usually occur in childhood and adolescence, and these diagnoses are prevalent within CAMHS. Each F-code corresponds to a specific diagnosis and should be assigned only when sufficient information is available to confirm a particular diagnosis with certainty [Sosial- og helsedirektoratet, 1999]. For instance, F90 serves as the designated code for hyperkinetic disorders.

R-Codes: These codes primarily describe symptoms rather than denoting any specific disorders. Within the context of CAMHS in Norway, R-codes are used temporarily until the medical team gathers enough information for an accurate diagnosis [Direktoratet for e-helse, 2018].

Z-Codes: These codes outline reasons for patient contact. In Norwegian CAMHS, Z-codes should only be utilised when there is an absence of meaningful diagnoses or R-codes [Direktoratet for e-helse, 2018].

A general rule is that diagnostic codes should be prioritised over R-codes and Z-codes, with R-codes being given precedence over Z-codes. This rule ensures that the most accurate and specific diagnosis is selected wherever possible.

2.1.2 Multiaxial-Classification System

CAMHS in Norway uses the ICD-10-based multiaxial classification of child and adolescent psychiatric disorders. The system is developed by WHO and has been adjusted for use in CAMHS in Norway [Direktoratet for e-helse, 2018]. The multiaxial classification comprises six distinct axis, with each disorder assigned to a specific axis. One of the advantages of the multiaxial system lies in its capacity to describe and diagnose complex cases involving multiple concurrent diagnoses, thereby offering a more comprehensive understanding of a patient. This research has focused on diagnoses coded in axis 1, 2, 3 and 4.

The multiaxial classification system consists of the following six axes:

- **Axis 1**: Clinical psychiatric syndromes
- **Axis 2**: Specific disorders of psychological development
- **Axis 3**: Mental developmental disabilities
- **Axis 4**: Somatic conditions
- **Axis 5**: Abnormal psychosocial situations
- **Axis 6**: Global assessment of psychosocial disability (CGAS)

Axis 1 focuses on clinical psychiatric disorders and covers a broad spectrum of mental health conditions, such as hyperkinetic disorders, mood disorders, anxiety disorders, psychotic disorders, and personality disorders [Direktoratet for e-helse, 2018]. Within CAMHS, Axis 1 diagnoses are more prevalent, with the primary diagnosis of most patients falling under this category. Axis 2, on the other hand, is concerned with specific disorders of psychological development, including learning disorders and speech disorders. When a patient has an Axis 1 diagnosis, concurrent diagnoses on Axis 2 can provide a more comprehensive understanding of their overall mental state. Axis 3 refers to the intellectual level of a patient, addressing diagnoses related to mental retardation. Lastly, Axis 4 is concerned with somatic conditions. Diagnoses from ICD-10 chapters I-IV, VI-XVII, and XIX-XX can be considered under Axis 4 [Direktoratet for e-helse, 2018].

It should be noted that for Axis 1, 2, 3, 4 and 5, clinicians have the option to assign the code *x-000* if no condition is detected or *x-999* if there is insufficient information to give a specific code. In these codes, x represents the axis number. For instance, 1-000 indicates that no condition is detected in Axis 1, while 1-999 indicates insufficient information to assign a code for Axis 1 [Direktoratet for e-helse, 2018].

2.2 PheWAS Phecode Map

This section outlines the Phecode Map 1.2, a resource for mapping ICD-10 diagnoses to corresponding phecodes and phenotypes. This mapping system is a part of the Phenome-Wide Association Studies (PheWAS) catalogue [Denny et al., 2013]. Each entry within the phecode system comprises an ICD-10 diagnosis, along with its linked phecode that denotes the associated phenotype. The term *phenotype* describes an individual's characteristics resulting from the individual's genes (genotype) and the environment. In the context of this research, the phecode mapping system has been employed to convert and categorise diagnoses into broader phenotype groupings, thereby facilitating a more comprehensive analysis. This is further described in Chapter 6.

The phecode system was developed upon ICD-9, the 9th revision of the ICD. However, numerous healthcare institutions, including the Norwegian CAMHS, primarily utilised the successor, ICD-10, between 1997 and 2018. Wu et al. [2019] present their work on mapping phecodes initially developed for ICD-9 to ICD-10 diagnoses. Their efforts resulted in the mapping of 76.2% of the ICD-10 codes to a corresponding phecode. A set of these mappings between ICD-10 and phecodes is illustrated in Table 2.1. It's important to clarify that phecodes do not equate to the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5) diagnostic system, despite apparent similarities in certain mappings.

Phecodes and phenotypes represent biologically relevant characteristics that are biologically relevant for biomedical research. Consequently, the phecode system is a valuable asset in identifying biologically meaningful phenotypes that can be used in research. The phecode system was primarily created to facilitate PheWAS using EHRs [Wu et al., 2019]. According to Wu et al. [2019], are phecodes particularly advantageous for high-throughput genotype-phenotype studies within EHRs.

Table 2.1: ICD-10 codes with associated phecodes and phenotypes.

ICD-10	**ICD-10 String**	**PheCode**	**Phenotype**
F80.2	Receptive language disorder	315.2	Speech and language disorder
F81.0	Specific reading disorder	315.1	Learning disorder
F90.1	Hyperkinetic conduct disorder	313.1	Attention deficit hyperactivity disorder
F90.9	Hyperkinetic disorder, unspecified	313.1	Attention deficit hyperactivity disorder
F91	Conduct disorders	312	Conduct disorders
F95	Tic disorders	313.2	Tics and stuttering

2.3 Anatomical Therapeutic Chemical Classification System

The Anatomical Therapeutic Chemical (ATC) Classification System is developed by WHO for categorising drugs according to their therapeutic properties. The system is widely used in the healthcare sector and has a hierarchical structure that allows for the comparison of drugs with similar therapeutic effects. Each drug is assigned a unique code of seven characters that describe its anatomical, therapeutic, and pharmacological classification [WHO Collaboration Centre for Drug Statistics Methodology, 2022].

The first level of the system consists of fourteen main groups, represented by letters A to V, which are based on the substance's area of effect. For example, substances in the N-group affect the nervous system and are frequently prescribed in CAMHS. The second level of the hierarchy describes the therapeutic or pharmacological effect of the substance, and the third level represents the chemical, pharmacological, or therapeutic subgroup. For instance, group N06 is for psychoanaleptics, with subgroups N06A (antidepressants) and N06B (psychostimulants used for ADHD and nootropics). A substance with multiple therapeutic areas may have more than one ATC code [WHO Collaboration Centre for Drug Statistics Methodology, 2022]. The hierarchical structure of the ATC classification system is illustrated in Figure 2.1.

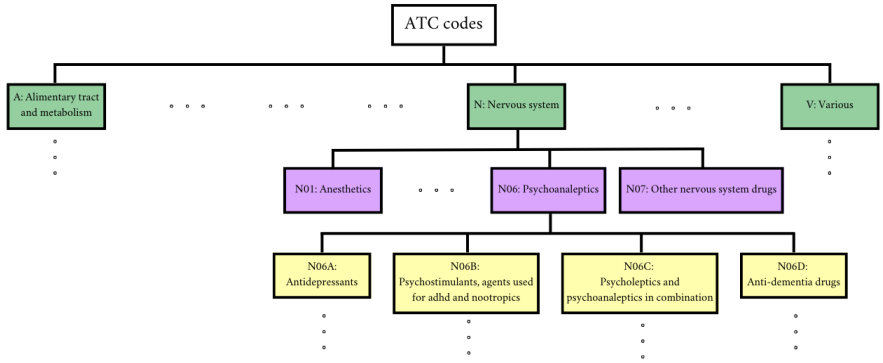


Figure 2.1: The ATC classification system hierarchy.

Chapter 3

Technical Background

This section introduces the key technical concepts and data analysis techniques applied in this research. The initial discussion pertains to machine learning and clustering. This is followed by a description of the hierarchical agglomerative clustering (HAC) method, an algorithm utilised in the research methodology. Further in the chapter, the focus shifts to clustering validity indexes, which are useful in validating the clusters generated in this study. Concepts and algorithms pertinent to mining frequent itemsets are also presented, given their importance in the applied methodology. Familiarity with these concepts is essential for comprehending the research presented. In essence, this chapter establishes the foundation for the methods employed in this research.

3.1 Machine Learning and Clustering

This section introduces machine learning (ML) and ML concepts, as these are crucial for understanding the experiments and choice of methodology presented in Chapter 6. ML is a sub-field of artificial intelligence (AI) that focus on making programs that make it possible for computers to learn from data [Géron, 2019, p. 4]. ML techniques are based on statistics and mathematical computations that allow computers to find patterns and connections in large amounts of data and information.

Supervised ML and unsupervised ML are two subgroups of machine learning. Supervised machine learning aims to train models on labelled data to predict and classify previously unseen objects. For instance, an email spam filter trained

on pre-labelled emails, categorised as either spam or not, can predict whether incoming emails should be directed to the spam folder.

In contrast, unsupervised models do not rely on pre-labelled data but seek to explore and uncover patterns within the data itself. Clustering algorithms are examples of unsupervised machine learning [Géron, 2019]. Clustering algorithms partition data records into clusters based on some similarity measure. A record in a cluster is more similar to the records within the same cluster than those in other clusters. Given the wide array of clustering algorithms, each exhibiting optimal performance on different data types, it is critical to align the selection of a clustering algorithm with the nature of the data [Frades and Matthiesen, 2010].

A clustering analysis consists of multiple steps. Initially, the features of the data intended for analysis must be identified and selected. Here, it is important to ensure that the chosen features represent valuable information because they will influence the resulting clusters. Second, choosing a similarity measure that defines how similar or dissimilar the records are is necessary. The selection of this measure should reflect the type of data in use and adequately capture the relationships within the data. Common similarity measures encompass Euclidean distance, Cosine similarity, and Pearson's correlation. The next step involves choosing an appropriate clustering algorithm tailored to the requirements of the task. Finally, the results generated from the clustering must be analysed [Frades and Matthiesen, 2010].

3.2 Hierarchical Agglomerative Clustering

Hierarchical agglomerative clustering (HAC) is a clustering algorithm that creates a hierarchical cluster. HAC initially computes a distance matrix indicating the pairwise distances between all clusters [Frades and Matthiesen, 2010]. HAC begins by treating each data record as a separate cluster. Then, the two clusters exhibiting the smallest distance are combined into a single cluster, prompting an update of the distance matrix to reflect the new distances relative to the other clusters. This approach iteratively merges smaller clusters into larger ones until an appropriate number of clusters is achieved. The distance is calculated based on a selected similarity measure and linkage rule. This technique results in a dendrogram that shows how the clusters are related in a tree-based structure of merged clusters [Frades and Matthiesen, 2010].

In addition to selecting a similarity measure, a linkage rule must be chosen to determine the appropriate distance to include in the distance matrix. Various approaches exist for deciding the distance between clusters, with three popular approaches being single-linkage, complete-linkage, and average-linkage. Single-

linkage calculates the distance between clusters as the shortest distance between any two points in separate clusters, unlike complete-linkage, where the distance is the greatest distance between any two points belonging to different clusters. Average-linkage computes the distance as the average distance between all pairs of members from two distinct clusters [Frades and Matthiesen, 2010].

3.3 Cluster Validity Indexes

Cluster validity indexes are used to assess the quality of clustering results and measure how well data points are grouped into clusters based on various criteria. Examples of such criteria may be compactness, separation and connectivity. There are several cluster validity indexes, each with different strengths and limitations.

Cluster validity indexes can broadly be divided into external and internal indexes. The external cluster validity indexes compare the clustering result with external criteria, such as the true labelling or desired label, and evaluate the performance of the clustering algorithm based on these external criteria. This differs from internal validity methods that do not need external information to evaluate the clustering result. Instead, internal methods evaluate clustering results based on similarity and dissimilarity measures between the data points [Nidheesh et al., 2020]. This section presents the Silhouette Index and the Calinski-Harabasz Index that both are internal validity indexes.

3.3.1 Silhouette Index

The Silhouette Index (SI) is a metric used to evaluate the quality of a clustering result and measures how well each data point in a cluster is separated from other clusters Rousseeuw [1987]. This index is based on the distance between the data points in the same cluster and between the data points in the neighbouring cluster. The SI score ranges from -1 to 1, where -1 is the worst possible value. Typically, a higher SI indicates better clustering results as it indicates that the data points in the cluster are close and that the clusters are well separated. SI scores above 0 indicate good clustering, and larger values indicate that the clusters are compact and well separated [Zhong et al., 2020]. A SI score close to 0 indicates overlapping clusters [Pedregosa et al., 2011].

3.3.2 Calinski-Harabasz Index

The Calinski-Harabasz Index (CI), also known as the variance ratio criterion, is another cluster validity index measuring the ratio of the sum of between-cluster

dispersion to the sum of within-cluster dispersion for all clusters [Caliñski and Harabasz, 1974]. Here the dispersion is calculated as the sum of distances squared [Pedregosa et al., 2011]. Higher scores indicate better clustering with well-defined and separated clusters [Zhong et al., 2020].

3.4 Mining Frequent Itemsets

Frequent itemset mining involves identifying sets of items that occur together in a dataset. It is a technique used in analysis where finding patterns in large datasets are beneficial. The goal is to discover associations and correlations between items that can help make predictions and enhance decision-making processes. The process involves identifying frequent itemsets that meet a user-defined support threshold, representing the minimum frequency of occurrence for a set of items to be considered frequent. This section will explore the concepts related to mining frequent itemsets and the algorithms FPGrowth, FPMax, and Apriori used to achieve this task.

Frequent itemsets are sets of items frequently occurring in a dataset and reveal patterns in large amounts of data. What is considered frequent is decided by a user-defined threshold value that decides how often an itemset should appear to be frequent. A definition of the term itemset is needed to understand what a frequent itemset is. An itemset is a set of items and is a subset of a large set of all possible items. Consider the list of items $S = \{x_1, x_2, x_3, ...\}$ that are present in the dataset $\mathbf{D}$. Every itemset $I \subseteq S$ is considered an itemset of $\mathbf{D}$ [Zaki and Meira, 2020, p. 219-220]. For example, is $I_1 = \{1, 3\}$ an itemset of $S = \{1, 2, 3, 4\}$.

By this definition, if the list of items S contains all possible values in dataset $\mathbf{D}$, all records contain an itemset of S. As itemsets and patterns that occur frequently are of interest, a threshold value defining how often an itemset should appear to be relevant is defined. This threshold is called the minimum support value and defines which itemsets should be considered frequent. The relative support of an itemset I is the number of records in $\mathbf{D}$ that contains I divided by the total number of records in $\mathbf{D}$. Therefore, the relative support of an itemset indicates how often the itemset occurs in the dataset. An itemset is considered to be frequent if the support value of the itemset is greater than or equal to a user-defined minimum support ($minSup$) value, $sup(I, \mathbf{D}) \geq minSup$ [Zaki and Meira, 2020, p. 219-220]. The support of itemset I in the dataset $\mathbf{D}$ is given by the following formula:

$$sup(I, \mathbf{D}) = \frac{Number\ of\ records\ containing\ I}{Total\ number\ of\ records}.$$

A frequent itemset is considered a maximal frequent itemset (MFI) if it is not a subset of any other frequent itemset [Gouda and Zaki, 2001]. Considering three frequent itemsets $I_1 = \{A, B\}$, $I_2 = \{A, C\}$ and $I_3 = \{A, B, C\}$, only I_3 would be considered to be maximal because it is the only frequent itemset that is not a subset of another frequent itemset. Looking at MFIs rather than frequent itemsets can often be beneficial when dealing with large amounts of data because the number of frequent itemsets may be large. The list of MFIs is usually shorter, and all frequent itemsets can be derived from that list [Gouda and Zaki, 2001].

3.4.1 FP-Growth and FPMax Algorithms

FPGrowth and FPMax are tree-based algorithms that mine frequent and maximal frequent itemsets, respectively. Both algorithms use a recursive approach to construct a frequent pattern tree (FP-tree), where each node represents an individual item and its corresponding support. In FPGrowth, the FP-tree calculates the support of itemsets and returns a list of frequent itemsets with their corresponding support values [Zaki and Meira, 2020, p. 233-236]. FPMax is an extension of the FPGrowth algorithm modified to mine maximal frequent itemsets (MFIs) [Grahne and Zhu, 2003]. As it generates only MFIs, it is more efficient than FPGrowth in terms of memory and computational resources.

3.4.2 Apriori Algorithm

The Apriori algorithm utilises a breadth-first approach to mine frequent itemsets. Considering two itemsets X and Y, the algorithm leverages the principles that if Y is frequent, then all subsets $X \subseteq Y$ are also frequent and that if Y is not frequent, then all supersets $X \supseteq Y$ are not frequent. The input of Apriori is a dataset $\mathbf{D}$, a set of items S and a *minSup* value. It starts by mining and calculating the support of all 1-item itemsets. It continues pruning the 1-item itemsets with $support < minSup$ and mines all 2-item itemsets that are supersets of the 1-item itemset with sufficient support values. Then, it continues pruning the 2-item itemsets with low supports and mines 3-item itemsets that are supersets of the itemsets with sufficient support values. The algorithm continues with this process until no new candidates are added. Apriori returns the set of frequent itemsets in $\mathbf{D}$ with a $support \geq minSup$ [Zaki and Meira, 2020, p. 225-227].

Chapter 4

Related Work

This chapter delves into the literature related to this research and is divided into three sections. The first section adopts a clinical focus, reviewing studies and statistics centred around comorbidities in mental health. *Comorbidity* refers to the simultaneous presence of two or more medical conditions in a patient. This exploration provides insights into the clinical aspect of the study, setting the stage for a deeper understanding of the complexities and challenges associated with co-occurring mental health disorders. The second section has a technical perspective, examining the application of machine learning techniques in the analysis of co-occurring morbidities. The technical review aims to underscore the role of machine learning methods in the field. The third section briefly reviews comorbidity indexes used to evaluate an individual's comorbidity. These sections provide a comprehensive overview of the study's clinical and technical foundations. The objective of the literature review is not a thorough examination of all related literature. Instead, the focus has been to deepen the understanding of the project's clinical components, investigate technical strategies applied to similar issues, and explore diverse methodologies to study comorbidities.

The reviews concentrate on comorbidities, omitting medications for a few reasons. First, relevant literature simultaneously addressing both aspects and offering valuable technical insights proved challenging to locate. Second, it was anticipated that methods designed to analyse comorbidities could also be employed to study both comorbidities and medications. Lastly, medications, serving as a treatment modality for conditions, were viewed as supplemental data, enriching the overall information rather than acting as the primary basis for the analysis.

4.1 Co-Occurring Morbidities in CAMHS

Mental health disorders in children and adolescents often co-occur with multiple morbidities. Co-occurring morbidities are the simultaneous presence of two or more distinct conditions in an individual. These co-occurring conditions can encompass a range of disorders, including mood disorders, anxiety disorders, ADHD, conduct disorders, and substance abuse. Understanding the prevalence, patterns, and implications of co-occurring morbidities in CAMHS are important, as it can impact treatment planning, intervention strategies, and overall outcomes for children and adolescents referred to mental health services. This section reviews some of the current literature on co-occurring morbidities within the CAMHS context.

Hansen et al. [2018] conducted a study that sheds light on the prevalence and patterns of neurodevelopmental and non-neurodevelopmental co-occurring morbidities among children referred to two CAMHS outpatient clinics in Norway. Through parent interviews with 407 children, 66.3% boys, it was found that 55.5% of the participants had a neurodevelopmental disorder, with 21.2% having more than one additional neurodevelopmental disorder (homotypic comorbidity). Additionally, 58% of the children had more than one non-neurodevelopmental disorder (heterotypic comorbidity). Notably, a higher proportion of girls with neurodevelopmental disorders exhibited comorbid anxiety or depression disorders, with 44% of the comorbid girls also having an anxiety disorder. These findings underscore the significance of addressing co-occurring morbidities within the CAMHS population and highlight the need for tailored intervention strategies that consider the complexity of multiple mental health conditions in children and adolescents.

The literature, as highlighted by CADDRA [2020] and Coghill et al. [2021], underscores the fact that ADHD is often accompanied by co-occurring conditions, emphasising the importance of considering comorbidities during the diagnostic process. In children aged 6 to 12, it has been found that 11 to 30% of individuals diagnosed with ADHD also exhibit comorbid anxiety, conduct disorders, autism spectrum disorder, or tic disorders. Furthermore, over 30% are diagnosed with learning disabilities or oppositional defiant disorder. For adolescents aged 13 to 17, 11 to 30% are diagnosed with comorbid anxiety, depression, oppositional defiant disorder, conduct disorder, substance use disorders, autism spectrum disorder, or tic disorders. Similarly, over 30% are diagnosed with learning disabilities [CADDRA, 2020, p.14-15]. These statistics demonstrate that the prevalence of specific comorbid disorders in association with ADHD varies across different age groups, with certain disorders being more prevalent in younger children and others becoming more prominent as they grow older. Notably, CADDRA [2020] and

Coghill et al. [2021] emphasise that clinicians may need to make decisions regarding which condition to prioritise for treatment but underscore the importance of addressing comorbid disorders concurrently.

CADDRA [2020] highlight the challenge in diagnosing ADHD when it co-occurs with other conditions. Comorbidities can complicate the diagnostic process, as ADHD shares overlapping symptoms with other diagnoses, and no pathognomonic symptoms are exclusive to ADHD. This underscores the necessity for clinicians in CAMHS to understand the symptoms, specific characteristics, and developmental patterns associated with ADHD and its comorbidities. Such knowledge can help accurately recognise and differentiate specific cases, especially considering the diverse symptoms and characteristics that can arise across age groups and genders within the CAMHS population. By gaining a good understanding of symptomatology and patterns, clinicians can enhance their diagnostic precision and improve the identification and treatment of ADHD cases that co-occur with other conditions in CAMHS.

In summary, the studies discussed shed light on the prevalence, patterns, and challenges associated with co-occurring morbidities in the CAMHS population. The study by Hansen et al. [2018] reveals that many children referred to CAMHS in Norway have co-occurring neurodevelopmental and non-neurodevelopmental disorders. This underscores the need for tailored intervention strategies to address the complexity of multiple mental health conditions in children and adolescents. Furthermore, CADDRA [2020] and Coghill et al. [2021] emphasise that ADHD often co-occurs with other disorders, making correct diagnostics challenging. CAMHS clinicians must comprehensively understand the symptoms, characteristics, and developmental patterns of ADHD and its comorbidities to identify and treat cases accurately. This knowledge is essential for effectively treating co-occurring disorders, considering the diverse presentations observed across age groups and genders within the CAMHS population.

4.2 Machine Learning on Comorbidities

This section provides an overview of some existing literature on machine learning approaches for analysing co-occurring morbidities. The literature review was used to study related work in the field and to gain insight into what has already been done. As noted in the introduction, the objective of the literature review was not to conduct a comprehensive examination of all related literature but to investigate technical strategies applied to similar problems. The searches for papers were carried out using two search engines, PubMed and ACM, between January 31st and February 1st, 2023. Additionally, three new papers were incorporated

into the review in March 2023. The inclusion criteria for paper selection involved the utilisation of clinical data, a specific focus on co-occurring morbidities, and the application of analytical or machine-learning techniques. Papers published before 2017 and meta-analysis papers were excluded. Furthermore, the technical nature of the paper had to be apparent from the abstract, and the availability of the paper was also a prerequisite. The snowballing method was employed to identify additional relevant papers, ultimately yielding seven papers reviewed in this section. The primary objective is to explore relevant methodologies and algorithms utilised in previous studies rather than delving into the specific medical findings resulting from these approaches.

The selected papers apply different algorithms and methods for exploring co-occurring disorders, each with varying domains and approaches. Slaby et al. [2022] propose a rule-based phenotype algorithm based on criteria of inclusion and exclusion classifying ADHD and comorbid conditions. Using health data from the Children's Hospital of Philadelphia (CHOP), they also mined keywords specific to phenotypes and keywords specific to medication that were additionally used to try to improve the results. They found that including phenotype keywords did not improve the results. However, the inclusion of medication keywords improved the results. The results generally received good PPV (Positive Predictive Value) scores. The kind of text mining used in their experiment was a simple version of extracting keywords, which may have impacted the result of text mining not yielding better results for all keywords.

In their studies on comorbidity, Swain et al. [2022] and Kashihara et al. [2021] employ various statistical methods to examine comorbidity. Swain et al. [2022] applies Latent Class Analysis Clustering (LCA) to analyse the comorbidities of individuals with osteoarthritis. The study identifies five distinct clusters by utilising data from a sample size of 221,807 individuals with osteoarthritis and an equivalent number of patients without the condition. The findings reveal a higher prevalence of multimorbidity among individuals with osteoarthritis. On the other hand, Kashihara et al. [2021] adopts a Gaussian graphical mixture model-based clustering technique to categorise self-reported symptoms of patients with mental disorders into clusters of diagnoses. The study identifies four clusters characterised by transdiagnostic disorders. The model combines statistical methods with clustering techniques, and Kashihara et al. [2021] highlights the advantages of using GMM-based clustering; the possibility of further network-based analysis of the resulting network offering additional insights into the data.

Another study, by Nitin et al. [2022], examine comorbidities of developmental language disorders (DLD) using the Automated Phenotyping Tool for Identifying Developmental Language Disorder (APT-DLD), an algorithm developed by

Walters et al. [2020]. This algorithm is designed to classify comorbidities related to DLD using EHRs, and in this study, it identified 37 phenotypes associated with DLD. To address the issue of overlapping diagnoses, the researchers map the diagnoses to phecodes based on the Phecode Map 1.2, described in Section 2.2. However, for this particular study, the mapping was done specifically for ICD-9 codes.

Wartelle et al. [2021] presents an alternative approach to cluster ICD-10 diagnosis codes into multimorbidity patterns using hierarchical agglomerative clustering. To reduce the number of diagnoses, the study groups the ICD-10 diagnoses based on their letter and first number, resulting in the grouping of diagnoses F90-F98 into one cluster. They propose a novel measure of the relative risk of co-occurrence of diagnoses to determine which clusters should be merged at the hierarchical steps. The study utilises data from an emergency department in France and identifies 16 clusters of ICD-10 diagnoses that frequently co-occur. One identified cluster contains most of the F-diagnoses and is named the cluster of mental disorders and at-risk behaviours.

Zhong et al. [2020] and Zhong et al. [2022] propose novel algorithms for clustering diagnosis data from EHRs, along with demographic information. Zhong et al. [2020] introduces the MASPC algorithm, which identifies patterns of diagnoses in EHRs and clusters the records based on a binary representation of these patterns combined with demographic information. The algorithm is discussed in detail in Section 6.3, and experimental results show its superiority over selected baselines in terms of the SI and CI measures described in Section 3.3. On the other hand, Zhong et al. [2022] presents the Demographics and Diagnosis Sequences Clustering Algorithm (DDSCA), which clusters sequences of diagnoses along with demographic information. Unlike Zhong et al. [2020], this algorithm takes into account the order of the diagnoses and outperforms selected baselines in terms of ASPJ (Average Sum of Pairwise Jaccard distance), ASPWE (Average Sum of Pairwise Weighted Edit distance), and ASPLCS (Average Sum of Longest Common Subsequence).

In summary, the reviewed papers employ various methods to cluster co-occurring morbidities. Among the seven papers, five utilise unsupervised methods. It is noteworthy to observe the diverse approaches taken to address the problem. Papers such as Slaby et al. [2022], Swain et al. [2022], and Nitin et al. [2022] focus on studying comorbidity in relation to specific diagnoses, such as ADHD in the case of Slaby et al. [2022]. Conversely, Kashihara et al. [2021], Wartelle et al. [2021], Zhong et al. [2020], and Zhong et al. [2022] explore co-occurring diagnoses in a more general context, encompassing all diagnoses.

4.3 Comorbidity Indexes

This section briefly overviews alternative methods for assessing an individual's comorbidity, specifically through the Comorbidity Polypharmacy Score (CPS) and the Charlson Comorbidity Index (CCI). Unlike the methods surveyed in Section 4.2, these indices yield numerical values signifying the overall severity of a patient's comorbidities. CPS employs polypharmacy and medications as a tool to evaluate the treatment intensity for a patient's conditions, thereby quantifying the severity of comorbidities. This straightforward score is derived by enumerating all comorbid conditions and medications prior to hospitalisation [Stawicki et al., 2015]. CCI aims to forecast patient mortality associated with comorbidities. It assigns a score between 1 and 6 to each condition based on the mortality risk associated with each disease before summing these scores [Charlson et al., 1987].

While both scores contribute to evaluating an individual's comorbidity, this study aims to identify frequently co-occurring morbidities and medications across the entirety of the CAMHS population and characterise subgroups. Within this context, a simple score may not provide sufficient detail. However, recognising such scores is important as they expand the comprehension of potential solutions in the field.

Chapter 5

Dataset

This chapter describes the environments utilised for data analysis, the agreements and approvals signed to gain data access, and the specific dataset employed in this research. It also provides basic statistics pertaining to the dataset before applying the cleaning processes described in Section 7.2.4. Please be aware that some content within this section is derived from the previous fall specialisation project. The specifics of the data extraction process are elaborated in Section 7.2.3. This chapter, serves as a preparatory stage, providing the reader with an overview of the dataset.

5.1 Environments

This section presents an overview of the various environments and systems utilised in the research, including the HUNT Cloud, DBeaver and the HUNT Workbench.

The HUNT Cloud is the main gateway to accessing the required tools for this study. It is a scientific cloud computing solution owned by the Norwegian University of Science and Technology (NTNU) [HUNT Cloud, 2023b]. HUNT Cloud provides a dedicated laboratory for this project, securely storing all sensitive data and offers a range of tools to analyse and work with the data without exporting sensitive patient information outside of the cloud. This laboratory environment was accessed to explore and extract the required data from the database. DBeaver, a database management tool, was utilised with an SSH connection to HUNT Cloud, providing a visual interface to better understand and navigate the data [DBeaver, 2023]. It is important to note that the extracted data remained

stored within HUNT Cloud, as the raw patient data should never leave the secure HUNT Cloud environment. Access to the laboratory environment requires connecting through a VPN provided by HUNT Cloud.

When the desired data was extracted from the database within the laboratory, HUNT Workbench was utilised. HUNT Workbench provides a secure environment for writing code and performing data analysis and offers an array of analytic tools, including Jupyter Notebook, Python, R, RStudio, and MATLAB [HUNT Cloud, 2023a]. In the specific context of this research, Jupyter Notebooks and the Python programming language were used to implement the desired methods and install the necessary packages for conducting data analysis.

5.2 Data Approval and Agreements

This section provides an overview of the legal approvals and agreements that have been procured and formalised to access the database of the EHR system BUPdata. The system was used within CAMHS at the St. Olavs Hospital when the data analysed in the research was collected. The section includes details on the agreements pertinent to both the overarching IDDEAS project and this specific subproject. Some information in this section is derived from the fall specialisation project.

Considering the sensitive nature of patient data stored within the BUPdata database, a thorough data access approval process was undertaken for the IDDEAS project, resulting in access to data of patients referred to treatment in the CAMHS clinic at St. Olavs Hospital in Trondheim from January 1, 1992, to March 5, 2018. Acquisition of this access involved an approval process by the Regional Committees for Medical and Healthcare Research Ethics (REK). Any health and medical research in Norway necessitates preliminary authorisation from the Regional Committee for Medical and Health Research Statistics (REK). This approval must be secured before initiating any project [De nasjonal forskningsetiske komiteene, 2014].

The IDDEAS project was reviewed by REK on 09.10.2021 (Case 2018/2186). REK made the following decision, which was based on the potential societal benefits if the project proves successful:

The project falls outside the scope of the Health Research Act, cf. §2, and can therefore be carried out without the approval of REC. Exemption from the duty of confidentiality is granted cf. regulation 02.07.2009 nr. 989, Delegation of authority to the regional committee for medical and health research ethics pursuant to the Health Personnel Act §29, first paragraph, and the Public Administration

Act §13d, first paragraph.

A comprehensive risk and vulnerability analysis was conducted to ensure data security and privacy. The project underwent a Data Protection Impact Assessment (DPIA) to assess risk, privacy considerations, and compliance with the General Data Protection Regulation (GDPR) [IDDEAS, 2022a]. Maintaining patient anonymity is a priority, and the data should not allow for the identification of specific individuals. Strict measures are implemented to ensure patient confidentiality and protect sensitive information from being linked to identifiable individuals.

In addition to the general data access approval for the IDDEAS project, a non-disclosure agreement (NDA) and a HUNT Cloud User Agreement were also signed to gain data access for this subproject. The health data analysed in these experiments are anatomised but still potentially re-identifiable. This means that if one has knowledge of a patient and its characteristics, it is possible to identify the patient [Solheim, 2022]. Therefore, IDDEAS team members wanting access to the dataset must sign an NDA. In addition to the NDA, a HUNT Cloud User Agreement had to be signed to gain digital laboratory access to HUNT Cloud.

5.3 Description of the Dataset

This section provides an overview of the dataset and the extracted features, aiming to familiarise readers with the data employed in the research. The dataset includes structured information about patients, episodes of care, diagnoses, and prescriptions within the CAMHS clinic at St. Olavs Hospital in Trondheim. Specifically, it encompasses patients who received diagnoses in Axis 1, 2, 3, or 4 from October 3rd, 1956, to October 4th, 2018. The data was extracted from the EHR system, BUPdata, which was utilised at the CAMHS clinic at St. Olavs during that period. The extraction process focused on capturing diagnoses and prescriptions, thereby gathering data on all diagnoses and medications associated with each episode of care recorded in the database. To gain insights into the nature of the data and the clinical processes involved, consultations were conducted with both clinicians and the referenced work by Solheim [2022] studied.

Firstly, it is necessary to define the term *episode of care*, which is also referred to as an *episode* in this report. An episode of care refers to one or more contacts or hospital stays associated with a patient's referral to CAMHS. For instance, if a patient is referred to CAMHS and accepted for assessment and treatment, all clinical appointments related to the assessment and treatment are considered part of the same episode. The duration of an episode can vary widely, ranging from a few days to several years. Some patients may enter CAMHS, receive a diagnosis,

and conclude the episode, while others may require additional appointments and assessments to receive an accurate diagnosis and appropriate treatment. It is important to acknowledge that a patient may have multiple episodes of care within CAMHS.

The dataset analysed in this research comprises ICD-10 codes associated with diagnoses for each episode. For this research, diagnoses from Axis 1, 2, 3, and 4 have been extracted for analysis, as psychosocial situations (Axis 5) and the CGAS score that assesses the level of functioning (Axis 6) were not considered relevant for this specific research. By input from a clinician, it should also be mentioned that most diagnoses in Axis 4 are typically assigned by other specialist clinics rather than CAMHS. It is not uncommon that children receive somatic diagnoses in other clinics and that specialists at these clinics discover or suspect additional mental disorders and subsequently refer the patient to CAMHS for further evaluation and treatment. As a result, the somatic diagnoses obtained from other clinics will accompany the patient to CAMHS and be recorded within the CAMHS system. These diagnoses will be included in the patient's medical records within CAMHS, providing an overview of their healthcare journey and ensuring that all relevant information is accessible to the clinicians at the CAMHS clinic.

Now, some general statistics about diagnoses, demographics, and prescriptions within the dataset are presented. The objective is to provide insight into the distribution and prevalence of various attributes and features inherent in the dataset. It is important to recognise that these statistics reflect the dataset's initial state prior to the application of the cleaning and grouping processes outlined in Chapter 6. Therefore, these analyses furnish a preliminary overview of the raw data, setting the stage for more refined analyses following the data preprocessing steps.

5.3.1 Diagnoses

The dataset comprises 1,365 unique diagnoses. Figure 5.1 illustrates the distribution of all unique ICD-10 diagnoses occurring more than 200 times in the dataset. Note that the x-axis is scaled logarithmically. It is important to note that if a patient receives the same diagnosis multiple times within an episode, it is only counted once. The figure shows that F900, disturbances of activity and attention, is the most frequently occurring diagnosis in the CAMHS dataset. Additionally, F321, moderate depressive episode, F952, combined vocal and multiple motor tics Tourette's syndrome, F431, Post Traumatic Stress Disorder (PTSD), and F901, hyperkinetic conduct disorder, are also prevalent within the dataset. For the clustering experiment, the diagnoses have been grouped into phenotypes. This is

elaborated in Section 6.4.

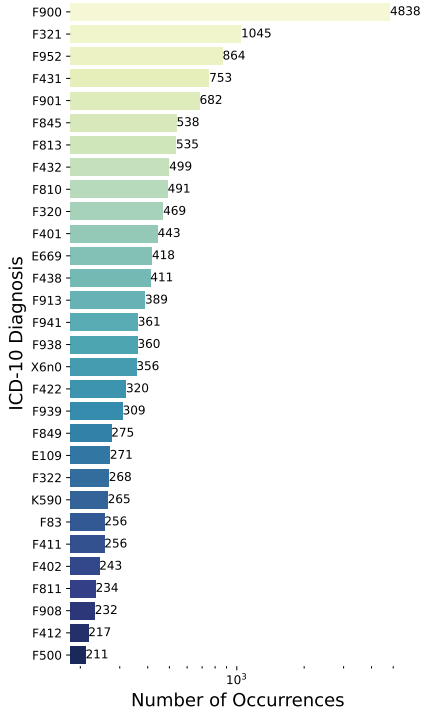


Figure 5.1: ICD-10 diagnoses distribution of episodes in the dataset. If a patient receives the same diagnosis multiple times within an episode, it is only counted once. The x-axis scale is logarithmic and only diagnoses that appear in more than 200 episodes are shown.

5.3.2 Prescriptions

The dataset also includes information about prescriptions issued by clinicians within the CAMHS clinic to identify patterns and characterise subgroups of medical treatment combined with diagnoses. The drug information of the medication typically consists of a trading name, an ATC code, and an ATC name. For the experiment, only the ATC code has been analysed, while the other fields have been employed to address missing values. As described in Section 2.3, the ATC code effectively describes the therapeutic effect of a substance. It has been chosen for the research dataset due to its ability to group drugs based on their intended use while simultaneously limiting the number of unique medications stemming

from different brands or minor variations. Figure 5.2 presents the distribution of ATC codes in the dataset before the truncation process described in Section 7.2.4. It is important to note that if the same ATC code is prescribed multiple times within an episode, it is only counted once. From the figure, it is evident that drugs falling under the ATC code system's N-group (nervous system) are particularly prevalent within the dataset, indicating their significant usage within the CAMHS clinic.

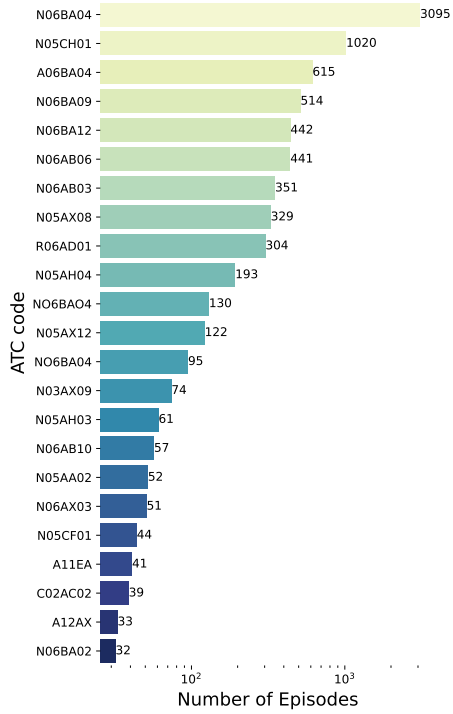


Figure 5.2: ATC code distribution of episodes in the dataset. If a patient is prescribed medications with the same ATC code multiple times within an episode, it is only counted once. The x-axis scale is logarithmic, and only ATC codes that appear in more than 30 episodes are shown.

5.3.3 Demographic Information

For the purpose of characterising patient subgroups, demographic features such as age and gender have been included in the experiments. These features provide basic information about the patients, and their inclusion may reveal interesting

patterns and subgroups related to diagnoses and prescriptions within different patient profiles. Figure 5.3 displays the age distribution of the patients in the dataset. As observed in the figure, most patients are above the age of eight, with a notable increase in the number of patients above thirteen years old. For the clustering experiment, these ages have been grouped into specific age groups, as elaborated in Section 7.2.4.

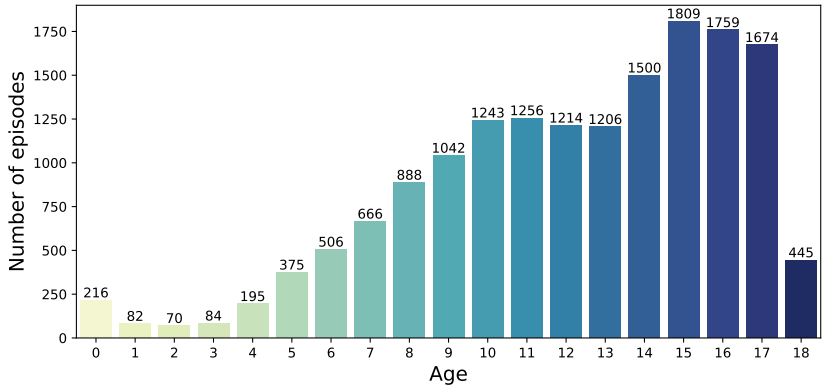


Figure 5.3: Age distribution in the dataset. The age is calculated based on the date the first diagnosis in the episode was given.

Figure 5.4 illustrates the gender distribution within the dataset. In this figure, it should be noted that *M* denotes boys, *F* represents girls, and *0* signifies missing values in the gender field of the database. It is important to clarify that the *0* category does not imply an undefined gender but a missing value, as the EHR system only allowed for selecting either male or female as gender options. From the figure, it is evident that there is a slightly higher number of males compared to females in the dataset.

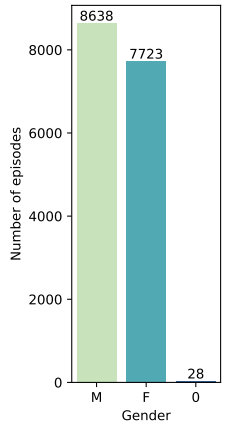


Figure 5.4: Gender distribution in the dataset. F represents girls, M represents boys, and 0 represents missing gender information.

Chapter 6

Methodology

This chapter outlines and justifies the methodology and techniques employed in the research. The first section presents the framework of the project and the process of identifying the scope and area of focus. Then, a discussion of the requirements of the clustering algorithm to be used in the research is provided. Furthermore, the algorithm of choice, MASPC, and the reasoning behind choosing this algorithm are explained. The chapter justifies the choices made in the experimental architecture to facilitate a clear understanding of the research findings. As the research involves applying computer science techniques to medical data, input and insights from clinicians and medical professionals have been incorporated to ensure informed decisions regarding the data and its applicability to the research. It is worth noting that gaining knowledge of the mental health field and the data is a comprehensive process involving analysing the data, studying relevant aspects of medicine related to CAMHS, and consulting with clinicians and specialists in the mental health field.

6.1 Project Framework, Problem Identification, and Adaptation

This section discusses the exploratory processes undertaken to identify a research area that aligns with the frameworks of the project and aims to shed light on the initial stages of the project, involving data exploration and topic selection. As noted in Chapter 1, this thesis is a collaborative effort with the IDDEAS project, which has access to an extensive database of EHRs from CAMHS pa-

tients. The IDDEAS team and the database were provided as the initial foundation for this project, with the responsibility of identifying a precise research focus and problem, evaluating the problem and exploring potential solutions. Additionally, proposing a possible solution, applying the solution and subsequently evaluating this solution was a part of the task. An in-depth analysis of the available data was carried out to facilitate these tasks, supplemented by dialogues with clinicians.

The original scope of the thesis was to investigate natural language processing (NLP) techniques applied to clinical notes in CAMHS. The chosen scope was based on several factors: the promised access to clinical text notes, the interest expressed by the IDDEAS team, and the opportunity to work with NLP techniques. This was also the focus of the fall specialisation project that studied NLP theories, language modelling and its application to clinical text. However, due to the unavailability of the required text data, the research scope had to be revised to focus solely on structured data. Despite the initial intention and efforts to access clinical notes, access was not granted within the anticipated timeframe. The decision to change the scope was made on January 20th, one week into the designated period for the master's thesis. While this change was disappointing, it presented an opportunity to explore alternative areas within the field of clinical data analysis. Even though NLP was not directly relevant to the new scope of the master's thesis, the insights gained from exploring clinical practices in CAMHS proved valuable in the redefined scope.

Following the change of scope, an in-depth exploration of the dataset was undertaken, and dialogues with clinicians were initiated to identify a new direction for the research. Initially, when the focus was on NLP, treatment approaches in CAMHS were analysed, given the expectation that clinical notes would reveal varied treatment methods within CAMHS. Even though direct access to these text notes was not granted, the database was found to house information about prescriptions and medication in CAMHS, an undiscovered aspect of the dataset. Additionally, clinicians indicated co-occurring morbidities as an interesting field for investigation. They suggested that thoroughly examining patient diagnoses and their medical treatments could reveal interesting phenomena within CAMHS, potentially offering valuable insights for the IDDEAS team. As a result, the decision was made to direct the research focus towards the analysis of co-occurring morbidities and medication within CAMHS.

Clustering was selected as the analysis method for this research due to the dataset containing unlabelled records with complex information. Unsupervised machine learning techniques are suitable for exploring such data, as they can uncover hidden patterns and connections without relying on pre-defined labels. By applying

unsupervised techniques, the research aims to reveal interesting patterns and relationships within the dataset and discover potential connections and insights that may not be apparent through traditional analysis methods. Clustering offers a valuable approach to providing a deeper understanding of the data and potentially uncovering novel findings.

6.2 Requirements of Clustering Algorithm

This section presents the requirements that need to be met by the clustering algorithm to be used in this research. As the data related to each episode differs in content, size and uniqueness, this needs to be handled by the clustering algorithm of choice. This section discusses these requirements that again will be evaluated when the chosen clustering algorithm is presented in Section 6.3.

The first requirement of the clustering algorithm is that it needs to be able to consider and handle data containing lists of varying lengths. When a patient enters CAMHS and a new episode of care is started, the patient may receive zero, one or multiple diagnoses. The patient may also be prescribed zero, one or multiple prescriptions. Therefore, the patient data in CAMHS often consists of varying numbers of diagnoses and medications, making it challenging to apply standard clustering algorithms that cluster structured data with one value in each field. One approach is to treat each list as a unique categorical value, but this approach fails to yield good clustering results since only lists with identical elements in the same order would be considered similar categorical values. Another solution is to one-hot encode all diagnoses and medications, but this approach results in high-dimensional data with many unique diagnoses and medications. Therefore, finding a clustering algorithm that can handle lists of varying lengths and still obtain meaningful clustering results is crucial.

Another requirement of the clustering algorithm is that it should be able to handle mixed input datatypes. The demographic data to be clustered comprises categorical and quantitative values, and the clustering algorithm of choice must be capable of handling both. For instance, the gender of the patient is represented by categorical values, such as F (female), M (male), or 0 (unknown). If the algorithm cannot directly process categorical values, they can be converted to one-hot encoded vectors. In addition to gender, the patient's age is another quantitative value in the data. Age can be either a continuous or a discrete quantitative variable, depending on the level of granularity required. To simplify the analysis and identify patterns in the data, the patient's age has been rounded down to the nearest number of years, making it a discrete quantitative variable. Depending on the clustering algorithm, the ages can be grouped into age groups,

resulting in a categorical value. As a result of the demographic data containing both categorical and quantitative values, the algorithm must handle these mixed data types.

The clustering algorithm must also handle large amounts of data. The IDDEAS dataset contains large amounts of data and records, consisting of over 2.5 million entries across various database tables. From these records, a subset of 8,499 records has been carefully selected for analysis. Given the amount of data, employing a clustering algorithm that can effectively scale and handle such volumes of information is essential. The chosen clustering algorithm should scale well and efficiently handle large datasets. It should be able to analyse the selected records efficiently, enabling the identification of meaningful patterns and characterisation of subgroups within the dataset without compromising the quality of results.

In addition to the discussed requirements, the chosen clustering algorithm must exhibit robustness against outliers and noisy data. Given the nature of the research, which involves analysing diagnoses and medications in CAMHS and exploring patient subgroups, the focus is on capturing the broader patterns rather than being overly influenced by noisy records. As the dataset is collected from real-world situations, it is expected to contain outliers and errors. Therefore, the clustering algorithm must not be overly sensitive to outliers and noisy data. It should be capable of adapting and producing meaningful clusters without excessively grouping one outlier into a separate cluster or significantly impacting the overall clustering results. By selecting a clustering algorithm resilient to outliers and noisy data, the algorithm should be capable of detecting the overarching patterns and subgroups within the dataset.

In light of the requirements discussed, the clustering algorithm needs to be able to handle mixed data types and lists of data of varying lengths. It should also scale well for large amounts of data and not be too sensitive to noisy data and outliers.

6.3 MASPC

In light of the requirements described in the former section, the selected clustering algorithm must be able to handle mixed data types and lists of diagnoses and medications of differing lengths. Additionally, the algorithm should scale well for large amounts of data and not be too sensitive to noise. The Maximal-frequent All-confident pattern Selection (MAS) and Pattern-based Clustering (PC) clustering algorithm, developed by Zhong et al. [2020] and briefly discussed in Section 4.2, cluster records containing patient demographics and a varying number of given diagnoses. In their study, they applied the algorithm to cluster patient

demographics and diagnoses across two distinct datasets, suggesting the potential to incorporate medications and laboratory results within the same MASPC algorithm framework. Yang et al. [2023] further validated this proposition by applying the MASPC algorithm to identify specific patterns of ICD-10 diagnoses and medication for patients exposed to COVID-19. Following this, they clustered patients that fell within the COVID-19-resistant patterns, thereby identifying subgroups of individuals who exhibit resistance to the SARS-CoV-2 virus.

The MASPC algorithm can intake a dataset comprising both single-valued and set-valued attributes, returning clusters guided by certain user-defined thresholds. Within the context of this research, the single-valued attributes comprise columns with demographic information, whereas the set-valued attributes represent lists of diagnoses and medication. Moreover, the MASPC algorithm selectively clusters records that contain frequently occurring patterns, thus making it robust to noisy records. Additionally, it is worth noting that MASPC demonstrates linear scalability with the size of the dataset [Zhong et al., 2020], making it scalable for large sets of data. This section aims to describe the theory of the algorithm and its input values, thereby facilitating a better comprehension of the results detailed in Chapter 8.

The MASPC algorithm identifies maximal-frequent all-confident itemsets (MFAs) and subsequently utilises these identified MFAs to cluster records. The algorithm consists of two parts - MAS and PC, which will be discussed in further detail in the subsequent subsections.

6.3.1 Maximal-Frequent All-Confident Pattern Selection

The MAS component of MASPC mines MFAs based on three user-defined threshold parameters. An MFA can be defined as an MFI with an all-confidence exceeding a user-defined threshold. The all-confidence of an itemset I is the ratio of the support of I to the support of the item within I that has the lowest support in the dataset. MAS comprises two phases. In the initial phase, the input dataset is employed to mine MFAs from the set-valued attributes. During the second phase, a particular subset of MFAs from the initial phase is chosen for clustering in the PC phase based on the threshold values. The three user-defined thresholds of MAS are *minSup*, *minAc*, and *minOv*.

The threshold *minSup* dictates the minimum support an itemset must exhibit to qualify as a frequent itemset. This implies that an itemset with a frequency in the dataset greater than or equal to *minSup* is classified as a frequent itemset. MAS exclusively mines MFIs; thus, a frequent itemset must also be an MFI to be considered.

The parameter *minAc* establishes the minimum confidence level for the MFAs. An MFI is regarded as an MFA if its confidence level is greater than or equal to *minAc*. Additionally, the MFA must consist of more than one element.

During the second phase of MAS, the *minOv* parameter directs the MAS algorithm to only select MFAs wherein at least *minOv* records encompass both MFA I and MFA I' if I and I' have common elements. Consequently, *minOv* represents the threshold value for minimum overlap between two MFAs that share items. Throughout this report, the term *pattern* will be used interchangeably with *MFA*, referring to MFAs identified by MAS.

Generally, larger threshold values result in fewer MFAs, while lower thresholds result in more MFAs. The MFAs selected with larger thresholds typically appear in many records, whereas those chosen with lower thresholds tend to yield more MFAs, albeit with lower support in the total dataset. As demonstrated by Zhong et al. [2020], the selection of *minSup*, *minAc*, and *minOv* significantly impacts the quality and efficiency of the clustering. As such, these thresholds should be established based on careful analysis.

6.3.2 Pattern-Based Clustering

The second component of the algorithm, PC, constructs k clusters from the records containing one or more MFAs identified by the MAS phase. PC employs HAC, as detailed in Section 3.2, utilising average-linkage and cosine similarity as the similarity measure. The input of the PC algorithm comprises a dataset containing demographic data, along with the MFAs selected by the MAS phase. Additionally, the user needs to specify a value, denoted as k, which determines the desired number of clusters. The first phase of PC constructs a binary representation of the input dataset indicating the presence or absence of each MFA in each record. Subsequently, the HAC clustering is performed on this binary representation, forming k distinct clusters. It is essential to highlight that PC exclusively clusters records encompassing one or more frequently occurring patterns. Consequently, records lacking such patterns are not included in the clustering process.

6.4 Grouping of Diagnoses Using Phecodes

The goal of this thesis is to investigate co-occurring morbidities and medication using EHRs of CAMHS patients. However, due to the high number of various diagnoses present in the dataset, measures need to be taken to reduce the distribution of unique diagnoses to obtain meaningful results when applying the

MASPC algorithm. This section aims to provide a rationale for mapping ICD-10 diagnoses to phecodes (presented in Section 2.2) and to discuss the strengths and limitations of this approach.

The dataset analysed in this study contains 1,366 distinct diagnoses related to episodes of care. However, given that many of these diagnoses share similarities, the clustering results could be improved by grouping similar diagnoses together to form groups of similar diagnoses. ICD-10 already provides a classification and grouping system for diseases based on the letter and numbers of the diagnoses, which can be utilised to group the diagnoses into larger categories. Wartelle et al. [2021] utilised these groupings to cluster related diagnoses. With the ICD-10 grouping used by Wartelle et al. [2021], diagnoses F90-F98 are grouped in the category *Behavioural and emotional disorders with onset usually occurring in childhood and adolescence.*

However, given that the extracted data is from CAMHS, it is expected that the most frequently occurring diagnoses will be grouped together as the most frequent diagnoses belong to the F-group. For instance, three out of the five most frequently occurring diagnoses are in the F90-F98 group, leading to a few groups with large counts and others with low counts. This will not necessarily yield interesting clustering results because the groups are too broad. However, using phecodes and phenotype mapping provides more subcategories within the F-category of ICD-10, resulting in better dispersion of the frequent mental health-related diagnoses within the dataset. Therefore, in this study, the ICD-10 diagnosis codes will be mapped to phecodes and phenotypes to improve the dispersion of the frequent mental health-related diagnoses within the dataset. Such a mapping was also applied by Nitin et al. [2022] in their research.

Using phecodes to map ICD-10 diagnoses to larger groups has some limitations. One limitation of this approach is that the mapping of ICD-10 diagnoses to phecodes is not perfect, and misclassification can occur. Furthermore, the phecodes system does not cover all diagnoses, leaving some diagnoses out of the analysis. Additionally, some diagnoses may be too rare to be mapped to a phecode, leading to the loss of potentially important information.

The benefit of using phecodes can be observed by looking at the F91- and F92-group of diseases. Diagnoses in these groups are frequent in the dataset and include various conduct disorders. Amongst others, the groups contain F91.0, conduct disorder confined to the family context, F90.1, unsocialized conduct disorder, F92.0, depressive conduct disorder, and F92.9, mixed disorder of conduct and emotions, unspecified. When mapping these ICD-10 diagnoses to phecodes, all four diagnoses are mapped to phecode 312 referring to the phenotype *conduct*

disorder. If left as documented in the medical data, these would all represent different diagnoses and be defined as different categories upon clustering. But when using the phecodes mapping, it provides the benefit of reducing the number of different diagnoses present in the dataset, which may lead to better results when applying clustering algorithms. Additionally, this grouping provides the benefit of more nuances than the ICD-10 groupings utilised by Wartelle et al. [2021], which for example, would see F9 as one category, compared to eight with the phecode categories.

In this study, the benefits of employing mapping of ICD-10 codes to phecodes surpass the drawbacks as it reduces the distribution of categories in the dataset while retaining clinically meaningful information. This is primarily due to the mitigation of disadvantages achieved through an extensive examination of the dataset, resulting in the addition of missing codes and the refinement of existing ones. This is described in Section 7.2.4. Overall, using phecodes can be a valuable tool in analysing the EHRs of patients referred to CAMHS. Still, carefully considering the strengths and limitations is necessary when interpreting the results.

6.5 Evaluation of Clusters

Evaluation of clusters is an essential step in cluster analysis, as it enables assessment of the quality of the obtained clusters. Based on the clustering results, validity metrics allow researchers to make informed decisions about the clustering results, which can be used to adjust variables like thresholds, and the number of clusters to optimise the results. For this research, validity metrics are needed to decide on optimal threshold values and number of clusters, and both SI and CI presented in Section 3.3 have been used to validate clustering results in the experiments. Zhong et al. [2020] also used these measures to validate their results.

For the clustering, internal measures evaluating the results are needed, as no true labelling can be used to validate the results. SI is a widely used cluster validity measure in biomedical data analysis [Nidheesh et al., 2020], and measures how well each data point fits into its cluster. This metric was chosen because it is a widely used metric to evaluate cluster quality, with the advantage that it only takes values between -1 and 1, providing a clear standard for comparison. CI measures the ratio of between-cluster dispersion to within-cluster dispersion across all clusters and is a widely used measure of clustering quality. It offers the advantage of computational efficiency. The combination of these two measures facilitates a more comprehensive assessment of the clustering result, providing insights into both the coherence within clusters and the distinction between clusters.

Chapter 7

Experimental Design

This chapter is dedicated to presenting the experimental plan and setup, with the primary objective of describing the design of the experiments. The purpose is to provide detailed information that enables replication of the experiments and facilitates an evaluation of their quality. Furthermore, the chapter aims to enhance the reader's understanding of the results presented in Chapter 8 by offering insights into the experimental design. Firstly, the experimental plan is outlined, encompassing the definition of the experimental aims, the experimental steps and a presentation of the experimental timeframe and work packages. Subsequently, the chapter describes the experimental setup, providing insights into the execution of the experimental steps and highlighting the tools employed during the process. Finally, the implementation details of the MASPC algorithm are described. This section focuses on the practical aspects of the algorithm rather than the theoretical aspect presented in Section 6.3.

7.1 Experimental Plan

The primary focus of the experiments is to explore patterns and patient subgroups of diagnoses and medication within the dataset, while also assessing the feasibility of the applied methods. This section provides an overview of the experimental plan, beginning with a definition of the experimental aims. Subsequently, the specific steps to be undertaken during the experiments are presented, outlining the procedural details. Furthermore, this chapter also encompasses a timeline for conducting the experiments and the division of work into distinct work packages. It is worth noting that the experimental process will incorporate various evalua-

tion meetings, presentations, and regular communication with clinicians. These measures are implemented to ensure the correctness and meaningfulness of the obtained results.

7.1.1 Experimental Aims

The aims of the experiments are defined in this subsection. The most prominent goal of this research is to investigate if patient profiles and subgroups regarding co-occurring morbidities and medication can be identified by cluster analysis. Item 1 and 2 summarise and describe aims related to the clustering process and the direct result of MASPC. The last two aims delve into the feasibility of the applied methods and the evaluation of the results in the context of CAMHS.

- Identify patterns concerning concurrent diagnoses and medication within patient episodes in CAMHS.
- Characterise subgroups of co-occurring diagnoses and medications of patients with similar demographics.
- Assess the feasibility of using clustering techniques to characterise subgroups of patients with similar demographics who exhibit co-occurring diagnoses and medications.
- Validate if the identified patterns and characterised subgroups pertaining to medications and diagnoses can explain phenomena within CAMHS.

These aims are investigated throughout the experiments and will be evaluated in Section 10.6.

7.1.2 Experimental Steps

There are several steps in a data analysis and clustering experiment. For this research, the experiments can be defined by the presented steps. While presented chronologically, it is important to note that the process is inherently iterative, with continuous review and refinement of the status and outcomes at each step. Completed steps may even require revisiting and modification to optimise the results. The involvement of clinicians throughout the experiment should also be highlighted, providing valuable input to validate and ensure the accuracy of decisions made at each step.

1. **Data selection**. Gaining insight into the data and selecting which data is relevant for the experiments. It includes exploring the data to understand which database tables and columns might be relevant.

2. **Data extraction**. Extracting relevant data for the experiment. Includes extracting the data from the HUNT Cloud laboratory using one or more SQL queries and storing it directly into the HUNT Workbench environment.

3. **Data cleaning and preprocessing**. When utilising a clustering algorithm on a dataset, it is crucial to ensure the presence of valid data. Consequently, it becomes necessary to preprocess and clean the data by eliminating null values and assigning new, appropriate values where needed.

4. **Exploratory Data Analysis (EDA)**. Analyse the dataset that will be used in the experiments to gain valuable insight into the data and characteristics of the dataset. This step facilitates familiarisation with the dataset utilised in the experiments.

5. **Implementation and testing of clustering algorithm**. Implementing the MASPC algorithm and doing test runs to ensure the correctness of the results and implementation.

6. **Clustering the dataset**. Apply the MASPC clustering algorithm to the dataset. Includes finding optimal thresholds and the number of clusters.

7. **Analysing clustering results**. Analysing and visualising the results using graphs. Includes feedback and input from clinicians on the applicability of the results in the context of CAMHS.

8. **Evaluation and discussion**. Evaluating and discussing the clustering results and methodology. Including discussing the choice of evaluation metrics, the technical applicability of the results, and the experimental design and methodology. The experimental results are also discussed and validated in cooperation with clinicians, and key takeaways are defined.

7.1.3 Work Breakdown Structure

This section provides an outline and description of the plan formulated for completing my master's thesis within a designated timeframe of twenty weeks. The purpose of the plan is to obtain a comprehensive understanding of the tasks at hand and to establish a schedule that prioritises the focus areas throughout the semester. It should be acknowledged that the initial plan was developed during the fall specialisation project. However, as elaborated in Section 6.1, there was a change of scope and topic of this master's thesis due to data access issues which required corresponding modifications to the plan.

Figure 7.1 shows a phase-based work breakdown structure (WBS) of the plan for my master's thesis. My master's thesis is divided into three phases; defini-

tion, implementation and documentation. Each phase consists of multiple work packages (WP). Each WP describes a group of tasks to be executed.

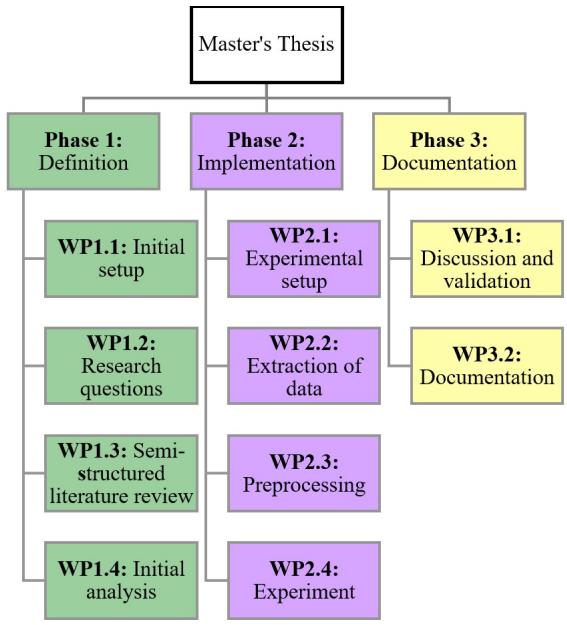


Figure 7.1: WBS of the project.

Phase 1 - Definition:
This phase includes the initial setup of necessary tools, familiarisation with the data, database and tools, definition of scope, research questions, and a literature review.

- **WP1.1 Initial setup:** Set up the required environment in the HUNT Cloud platform to facilitate data exploration and analysis.
- **WP1.2 Research questions:** Define a goal and research questions for the master's thesis.
- **WP1.3 Semi-structured literature review:** Conduct a semi-structured literature review of research and state of the art of machine learning of co-occurring morbidities in the medical field.
- **WP1.4 Initial analysis:** Familiarise and do an initial analysis of the data. Also, analyse the quality of the data and the feasibility of applying

clustering techniques.

It is important to note that certain WPs within this phase will be worked on simultaneously to enhance their outcomes. For instance, WP1.2 and WP1.4 will be conducted concurrently, as initial data analysis is crucial for formulating and defining relevant research questions.

Phase 2 - Implementation:
This phase contains work packages related to the implementation of the experiments. This includes installing necessary tools, querying the data and programming methods to preprocess the data and execute the experiments.

- **WP2.1 Experimental setup:** Installation and setup of the HUNT Cloud Workbench providing necessary tools like Jupyter Notebooks and installed programming languages like Python.

- **WP2.2 Extraction of data:** Extract necessary data from the database.

- **WP2.3 Preprocessing:** Preprocessing of the data. This step involves the cleaning and preprocessing of the extracted data, ensuring its suitability for subsequent analysis.

- **WP2.4 Experiment:** Conduct the experiments. Apply MASPC to the extracted dataset.

Phase 3 - Documentation:
This phase focuses on the documentation and finalisation of the master's thesis, including evaluation and discussion of methodology and results.

- **WP3.1 Discussion and validation:** Analyse and discuss the results and contributions. Includes the participation of clinicians to validate the clinical aspects of the experiments and results.

- **WP3.2 Documentation:** Document the research and discoveries in the master's thesis report.

It is important to note that the documentation process extends throughout the semester, as the report will be continuously worked on. However, it is crucial to allocate dedicated time in the schedule to document the research, refine the writing, and improve the report's overall quality. These additional weeks are necessary to ensure the final thesis report meets the required standards.

7.1.4 Timeline

Figure 7.2 illustrates the timeline allocated for the master's thesis and research. The total timeframe available is 20 weeks, divided among the three defined phases in the WBS. Specifically, six weeks are allocated to the definition phase, ten weeks to the implementation phase, and five weeks to the documentation phase. A more detailed breakdown of the schedule, indicating the latest week by which each WP should be completed, is presented in Table 7.1. The timeline ensures a well-structured and timely completion of the master's thesis.

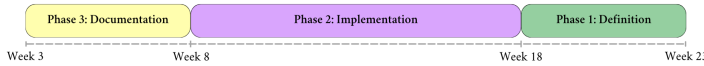


Figure 7.2: Timeline for the project.

Table 7.1: Timelimit for each WP.

Work Package		**Finished by week**
Phase 1	WP 1.1	4
	WP 1.2	8
	WP 1.3	8
	WP 1.4	8
Phase 2	WP 2.1	9
	WP 2.2	10
	WP 2.3	12
	WP 2.4	18
Phase 3	WP 3.1	20
	WP 3.2	23

7.2 Experimental Setup

This section provides an overview of the experimental setup and implementation, including the tools utilised, the data selection process, and a description of the dataset extraction, cleaning, and preprocessing steps. Initial data cleaning was performed within the SQL query used for data extraction, while further cleaning was conducted using Jupyter Notebook and the Python programming language. The process of data extraction, cleaning and preprocessing is described in detail for the purpose of reproducibility.

7.2.1 Tools

This section provides an overview of the tools utilised in the experiments and elaborates on their roles and functionalities. As discussed in Section 5.1, ensuring that sensitive data remains within the HUNT Cloud environment is crucial. The tools mentioned in this section are further elaborated in Section 5.1, but a summary is provided below.

- **Hunt Cloud**: Cloud computing platform providing a laboratory with sensitive patient data.
- **Hunt Cloud Workbench**: This tool provides access to Jupyter Notebooks and the Python programming language and is a secure environment for conducting research within HUNT Cloud. It also offers a visual interface that simplifies navigation through files and folders. Significantly, all the imported Python packages for the research are installed within this environment.
- **DBeaver**: Database management tool utilised for rapid and visual access to the structured database in the HUNT Cloud laboratory.

7.2.2 Data Selection

To select relevant data for analysis, a significant amount of time was dedicated to thorough research of the dataset, including an examination of its features and attributes, as well as an exploration of the interrelationships within the data. The goal was to gain a deep understanding of the underlying meaning of the data. Furthermore, expert input from clinicians was sought to enhance comprehension of the data's inherent characteristics and potential value. Given the large amounts of data, navigating and familiarising oneself with its complexities proved time-consuming.

The data in question possesses diverse levels of quality, necessitating a fine dis-

tinction between erroneous entries and data that might be uninteresting. This discernment requires an understanding of clinical practices within CAMHS as well as a familiarity with the dataset. For example, a thorough knowledge of the ICD-10 diagnostic codes and the multiaxial classification system, discussed in Section 2.1, is critical for accurately selecting diagnoses in Axis 1. In axis 1, both F00 and F99 are considered valid ICD-10 codes. However, within the database records, they are denoted as 00 and 99, respectively. In line with general ICD-10 mapping, the codes 000 and 999 usually denote either the absence of a diagnosis or inadequate information to establish a diagnosis, implying that the codes 00 and 99 should potentially be discarded. However, given that F00 and F99 are valid diagnostic codes explicitly applicable to Axis 1, and clinicians have confirmed that 00 and 99 correspond to F00 and F99, respectively, these codes have been retained in the analysis of diagnoses within Axis 1.

As outlined in Chapter 5, the chosen features for the experiments encompass all diagnoses in Axis 1, 2, 3, and 4 related to a specific episode of care in CAMHS. Furthermore, ATC codes of prescribed medications associated with the episode, along with the patient's gender and age, have also been incorporated as part of the selected features.

7.2.3 Extraction of Data

The data used in the analysis was extracted using two separate SQL queries, one to extract patient demographics and diagnoses related to an episode and one to extract medication information about prescriptions related to an episode. The choice of data resulted from the data selection step, a thorough analysis of which data was available and input from clinicians. The same data could have been extracted using one extended query, but it was more apparent to extract the data utilising two queries. After the two datasets were cleaned separately, they were merged, resulting in a dataset containing the preprocessed and cleaned data ready for the application of MASPC.

Query 1

The following query was executed to extract data about diagnoses for each episode and the patient's demographics. The decisions made during the extraction phase result from a thorough analysis of the available data.

```
1 select
2   pasient.nr as patient,
3   diagnose.sak as episode_id,
4   case
5     when pasient.kjonn = '1' then 'F'
6     when pasient.kjonn = '2' then 'M'
```

```
7       else '0'
8     end as gender,
9     diagnose.akse as axis,
10    case
11      when diagnose.diagnose = '00' and akse = 1 then 'F00'
12      when diagnose.diagnose = '99' and akse = 1 then 'F99'
13      when diagnose.diagnose = '5' and akse = 3 then 'F70'
14      when diagnose.diagnose = '6' and akse = 3 then 'F71'
15      when diagnose.diagnose = '7' and akse = 3 then 'F72'
16      when diagnose.diagnose = '8' and akse = 3 then 'F73'
17      when diagnose.diagnose = '9' and akse = 3 then 'F79'
18      else diagnose.diagnose
19    end as diagnosis,
20    case
21      when diagnose.dato is null then extract(year from age(diagnose.
        endrdato, pasient.fdt))
22      else extract(year from age(diagnose.dato, pasient.fdt))
23    end as age_patient
24  from
25    diagnose left join pasient on diagnose.pasient = pasient.nr
26  where
27    diagnose.diagnose not like '%Z%' and diagnose.diagnose not like '%
        R%' and
28    (
29    (diagnose.akse = 1 and diagnose.diagnose != '999' and diagnose.
        diagnose != '000' and diagnose.diagnose != '1000' and diagnose.
        diagnose != '1999') or
30    (diagnose.akse = 2 and diagnose.diagnose != '999' and diagnose.
        diagnose != '000' and diagnose.diagnose != '2000' and diagnose.
        diagnose != '2999') or
31    (diagnose.akse = 3 and diagnose.diagnose != '30' and diagnose.
        diagnose != '39' and diagnose.diagnose != '99'
32    and diagnose.diagnose != '3999' and diagnose.diagnose != '3000'
        and diagnose.diagnose != '1'
33    and diagnose.diagnose != '2' and diagnose.diagnose != '3' and
        diagnose.diagnose != '4'
34    ) or
35    (diagnose.akse = 4 and diagnose.diagnose not like '%99%' and
        diagnose.diagnose not like '%00%')
36    )
37  order by episode_id, age_patient;
```

Listing 7.1: PostgreSQL query for extraction of diagnoses and demographics.

Given that R-codes describe symptoms and Z-codes reasons for contact, these diagnostic categories have been excluded from extraction as they do not represent a specific diagnosis. Additionally, in the context of Axis 1, 2, and 3, rows containing the diagnoses *000*, *x000*, *999*, and *x999* have been excluded since these codes either denote a lack of detected condition or insufficient data to indicate a diagnosis. For Axis 3, diagnoses coded as *1*, *2*, *3*, and *4* have also been excluded.

These codes correspond to intelligence levels rather than a specific diagnosis, typically representing an IQ above 69. This decision was reached in consultation with clinicians. Regarding Axis 1, rows containing the diagnosis codes *00* or *99* have been replaced with *F00* and *F99*, respectively, as these are valid diagnoses within Axis 1. Additionally, diagnoses on Axis 3 have been adjusted to align with the ICD-10 classifications, as per the mapping of diagnoses provided by a clinician. Table 7.2 shows the mapping of diagnoses in Axis 3 in the dataset to their corresponding ICD-10 diagnoses.

Table 7.2: Mapping of Axis 3 codes used at the CAMHS clinic to ICD-10 diagnoses.

Original diagnosis	**ICD-10 diagnosis**
5	F70
6	F71
7	F72
8	F73
9	F79

Given that episodes of care can vary in length, the patient's age at the beginning of an episode has been calculated by determining the time difference between the patient's birth date and the date of the first diagnosis given in an episode. Some episodes lacked information about the date when the diagnoses were given. In these instances, the date the diagnosis was last updated, indicated by *endrdato*, was employed since most records missing a diagnosis date possessed a date in this field. However, it should be noted that the date fields in the database are somewhat disordered and prone to errors, often leading to *endrdato* coinciding with the date of the diagnosis.

In the initial dataset, the gender of the patient was coded as either 1 or 2. This has been altered to F for female patients and M for male patients. Some records did not provide any gender information and have consequently been mapped to 0. It's important to clarify that the available options for clinicians to record gender were strictly binary: male or female. A missing gender attribute, therefore, does not denote a non-binary gender. Instead, it suggests inaccurately provided information. Table 7.3 presents the mappings conducted for the gender attribute.

Query 2

For the medication and prescription data, the following query was executed. For the clustering task, solely the ATC codes were analysed, whereas the other fields

Table 7.3: Gender mappings in the dataset.

Original Mapping	New Mapping	Meaning
1	F	Girl/female
2	M	Boy/male
Null	0	No gender information available

were utilised during the cleaning process to substitute absent ATC codes.

```
select
  forordning.saknr as episode_id,
  forordning.forordning as regulation,
  resept.resepttype as prescription_type,
  preparat.handelsnavn as trade_name,
  preparat.atckode as atc_code,
  preparat.atcnavn as atc_name
from forordning
left join preparat on forordning.preparatid = preparat.id
left join resept on forordning.nr = resept.forordningnr
order by forordning.saknr;
```

Listing 7.2: PostgreSQL query for extraction of prescriptions.

7.2.4 Data Cleaning and Preprocessing

This section presents the pre-processing and cleaning strategies applied to prepare the extracted data for subsequent analysis. While the extraction process addressed some aspects of data cleaning, further cleaning was necessary. The data extracted encompassed three distinct categories: diagnoses, demographics, and prescriptions. Each category necessitated individual cleaning procedures, implemented using Jupyter Notebook and the Python programming language.

The extracted data was saved in a *.csv* file and imported into a Jupyter Notebook for subsequent cleaning and analysis. The cleaning process adopted an iterative approach, in which each cleaning step was executed, the results evaluated, and new methods or adjustments to existing ones were introduced before reassessing the outcomes. This systematic method facilitated the fine-tuning of cleaning procedures until the data reached a level of reliability and quality suitable for further analysis.

Cleaning of Diagnoses

This subsection describes the cleaning procedures implemented for the extracted diagnostic data. It also addresses the steps taken to clean the diagnoses listed

in the PheWAS catalogue to ensure compatibility with the diagnoses in the extracted dataset.

As detailed in Section 6.4, the diagnoses in the dataset were mapped to phecodes. This mapping into broader categories fosters categorised results and enables a broader perspective when applying clustering and subsequent analyses. Nevertheless, a discrepancy exists between the dataset diagnoses and the ICD-10 codes in the PheWAS catalogue: the former does not contain separating dots, while the latter does. To address this inconsistency, punctuation characters in the ICD-10 codes in the imported phecode catalogue were eliminated. This allowed for automated matching and conversion of the dataset diagnoses into phecodes. Figure 7.3 illustrates this mapping process.

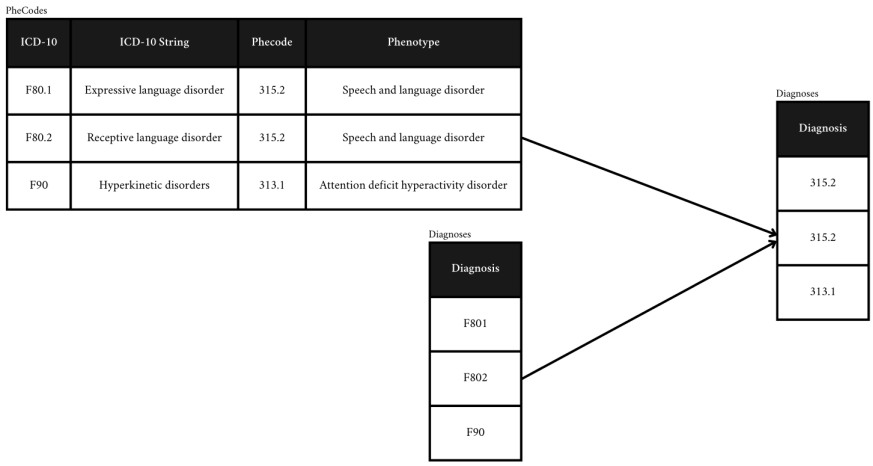


Figure 7.3: Mapping of diagnoses to phecodes by matching identical ICD-10 codes and assigning the corresponding phecode.

Given that the data comprises real medical records, it is prone to errors. Additionally, the phecode system misses the mappings of some diagnoses. Thus, a thorough review of diagnoses that did not receive a code automatically was conducted, and the list of phecodes was supplemented with the missing codes later. This addition enhanced the integrity of the cleaned data. Direktoratet for e-helse [2023] and Wu et al. [2019] were used to interpret, merge and ensure the correctness and quality of the codes added. Additionally, clinicians were consulted to confirm some mappings. Appendix G lists all ICD-10 codes that were appended to the original phecode system. Diagnoses that occurred less than three times in the entire dataset, post-phecode mapping, were eliminated due to their minimal

impact on the results. This removal primarily affected diagnoses coded in axis 4, a less critical axis in CAMHS. Specifically, out of 488 diagnoses excluded, 475 belonged to axis 4 and 13 to axis 1.

Cleaning of Demographics

The inclusion of demographic factors within datasets aids in characterising patient subgroups. Therefore, the dataset deployed for clustering incorporates both age and gender. The gender attribute was pre-processed during the data extraction phase and presented in Section 7.2.3. However, the records with missing gender information were excluded from the experiments. Furthermore, the patient age attribute was grouped into broader age groups to minimise the category count during clustering in this cleaning phase. The MASPC algorithm necessitates one-hot encoding of the demographic information prior to input, implying that each unique gender and age value must be transformed into a separate column for clustering. To optimise the dimensions, the patients' ages were grouped into intervals based on children's developmental stages and the Norwegian school system. Clinicians were consulted to ensure that the defined age groups held clinical significance. Table 7.4 presents the age groupings employed in this study. Patients over the age of 18 were excluded from the dataset.

Table 7.4: Age groupings of patients.

Age Interval (years)	**Age Group**
0-5	Preschooler
6-11	MiddleChildhood
12-18	Teenager

Cleaning of Prescriptions

The study involved the extraction and cleaning of prescription and medication data for each episode. As outlined in Section 2.3, the ATC classification system presents valuable insights into medications' anatomical and therapeutic properties, facilitating the grouping of similar medications into a common category based on their ATC codes.

Nonetheless, certain extracted prescriptions encountered during the process did not include an ATC code but instead featured a trading name. To address this issue, the study assigned ATC codes to these prescriptions by referring to the WHO Collaboration Centre for Drug Statistics Methodology [2022] whenever possible. This method utilised the trade name of frequently occurring prescriptions with

absent ATC codes to identify relevant ATC codes for the missing values. Table 7.5 illustrates the added ATC codes for records initially containing only a trading name. However, certain trading names did not yield an appropriate mapping. Such prescriptions were subsequently eliminated from the dataset and are presented in Table 7.6.

Table 7.5: ATC code mappings based on medication trade names for medications missing ATC codes.

Trading name	**Assigned ATC-code**
Antiepileptika	N03A
Antidepresiva	N06A
Concerta	N06BA04
Dexidrine	N06BA02
Melatonin	N05CH01
Metamina	N06BA02
Nevroleptika	N05A
Sentralstimulerende	N06BA

Table 7.6: Deleted medications due to missing ATC codes.

Trading name	**Number of occurrences**
Annet	1383
Ikke aktuelt	832
Preparat ikke angitt	220
Frebini energy Drink Banan	6
Frebini energy fibre Drink sjokolade	6
Frebini energy fibre Drink vanilje	4
Nutridrink Compact jordbær	2
Nutridrink Multi Fibre jordbær	1

Certain prescriptions posed the challenge of it being difficult to find a corresponding ATC code. To maintain the accuracy of comparisons between medications, prescriptions still retaining a *NaN* or *NULL* value in their ATC code field following the cleaning process were excluded from the dataset. A total of 2,454 prescriptions were removed based on this criterion. This action was necessitated

due to the lack of a dependable method to ascertain the ATC code to which these medications should be mapped.

After verifying that all remaining prescription records contained a valid ATC code, a limit of four characters was imposed on each code. This determination was reached on the basis that a four-character ATC code provides sufficient detail while enabling larger groupings of codes, thereby increasing the frequency of instances within each group. Such an approach enhances the statistical power of the analyses and ensures that the findings are meaningful and interpretable.

Final Dataset

Following the cleaning of all features, sample records are displayed in Table 7.7. It's critical to highlight that before the implementation of the MASPC algorithm, demographic data underwent a transformation into one-hot encoded representations to ensure compatibility with the algorithm's input requirements. It should also be highlighted that all records containing only a single element in their list of diagnoses and medications were excluded, resulting in a reduction of the dataset size from 16,202 to 8,499 episodes. It is important to note that while this exclusion was unnecessary due to the MASPC algorithm automatically removing these records in the MAS phase, it proved valuable for the initial data analysis.

Table 7.7: Example records of the final data. Each record contains gender and age group information as well as a list containing diagnoses and medications related to the episode.

Gender	Age group	Diagnoses and medications
M	Teenager	[Learning disorder, ADHD, N06B]
F	MiddleChildhood	[Anxiety disorder, ADHD, N06A, N06B]

7.3 MASPC Implementation

This section describes the Python-based implementation of the MASPC algorithm employed for clustering in this research. A theoretical explanation of the algorithm can be found in Section 6.3, while the complete code for the implementation is provided in Appendix E. The MASPC algorithm was implemented in a Jupyter Notebook using the Python programming language. The code was adapted from Zhong et al. [2020] and adjusted to fit with the dataset.

As outlined in Section 6.3, the MASPC algorithm's initial stage involves mining MFAs. The FPMax algorithm (explained in Section 3.4.1) is initially applied

to mine MFIs with an appropriate support value. Subsequently, the Apriori algorithm is run as an initial measure for calculating confidence values. Here, the FPMax algorithm ensures the selection of frequent itemsets that are also MFIs. Leveraging the output from both FPMax and Apriori, each MFI's confidence value is computed. If an MFI's confidence meets or surpasses the user-defined *minAc* threshold, it is added to the list of MFIs possessing an all-confidence value equal to or higher than *minAc*. This list of all-confident MFIs is further analysed in a loop. An MFI is included in the list of MFAs if it does not overlap with any itemset already in the MFAs list or if the frequency of the union between the overlapping itemset and the analysed MFI is at least *minOv*. The final list contains all MFAs of diagnoses and ATC codes.

The next phase of the Python implementation involves the PC part of the algorithm described in Section 6.3.2. During this stage, all records are examined, and those whose lists of diagnoses and ATC codes are supersets of one or more MFA are collected in a dataframe. This dataframe, encompassing all episodes of care with records containing one or more MFA, will be used in the following clustering process. After that, hierarchical clustering using average-linkage and cosine similarity is applied to this dataframe to group the records into a user-defined number of k clusters. This process results in the formation of clusters based on the discovered MFAs.

The FPMax and Apriori algorithms used in this study are implemented via SPMF, an open-source Java Data Mining Library specialising in pattern mining [Fournier-Viger et al., 2016]. This library, incorporating various data mining algorithms, is directly integrated into the Python-based implementation of MASPC. The SPMF versions of FPMax and Apriori demand a pre-defined minimum support value and a text file as input. Every line of the input file encompasses all diagnoses and ATC codes related to a particular episode of care, separated by a space. The algorithms return a text file containing the mined frequent itemsets, along with their corresponding support values.

Chapter 8

Experiments and Results

This chapter presents the results obtained by applying the implemented methodology and experimental design. To begin with, an exploratory data analysis (EDA) is undertaken to provide an overview of the experimental dataset. This initial exploration allows for the formulation of hypotheses regarding potential patterns and clustering outcomes that may arise in subsequent analyses. Following the EDA, the results of the application of the MASPC algorithm to the experimental dataset are presented, and some key takeaways from the results are highlighted. This chapter aims to present findings that will be evaluated in a clinical context in Chapter 9. Furthermore, the validity of these findings, considering the project's limitations and technical approaches throughout the research process, is discussed in Chapter 10.

8.1 Exploratory Data Analysis

This section introduces the data utilised in the research via an Exploratory Data Analysis (EDA). An EDA aims to delve into, analyse, and gain insights into the data. When conducting an EDA, one can choose the procedure of choice, but it commonly involves data visualisation using graphs [Morgenthaler, 2009]. Carrying out an EDA on the dataset before clustering provides valuable information and uncovers patterns in the data, providing an understanding of the cohort. The EDA featured in this section aims to present statistics regarding the dataset. It begins by presenting the EDA with key numbers associated with demographics, diagnoses, and ATC codes. Subsequently, it summarises the main takeaways from the EDA and posits some hypotheses about what might be discovered in

the clustering analysis.

The EDA is performed on two datasets; the initial dataset with all extracted episodes (referred to as the *initial dataset*) and the dataset inputted into the MASPC algorithm, which contains only records where the set of diagnoses and ATC codes comprises more than one element (referred to as the *experimental dataset*). It is important to note that in the PC phase of MASPC, only a subset of the experimental dataset is clustered. This subset specifically includes records that contain one or more of the identified MFAs, forming what is later referred to as the *clustering dataset*. However, it is crucial to highlight that the analysis of the entire experimental dataset is conducted to identify these MFAs in the first place.

This EDA initiates with key metrics from the dataset shown in Table 8.1. As indicated in 7.2.3, the data was obtained using two distinct queries. Query 1, employed to extract diagnoses and demographics, yielded 42,798 diagnoses across 16,202 episodes. Query 2, on the other hand, resulted in 53,713 records of prescribed medications spanning 4,489 episodes. It's worth noting that if the same ATC code or diagnosis is repeated within an episode, it is only counted once. These numbers suggest many episodes featuring more than one diagnosis and more than one prescribed ATC code. Moreover, the data shows a significantly higher number of episodes with one or more diagnoses compared to those with prescribed medications, implying that many episodes occur without prescriptions. For experimental purposes, only records with more than one element in the set of diagnoses and ATC codes are included, totalling 8,499 episodes.

Table 8.1: Key numbers from the datasets.

Total number of given diagnoses	27,027
Total number of episodes	16,202
Total number of prescribed medications	7,130
Total number of episodes where medications are prescribed	4,459
Number of episodes in experimental dataset	8,499

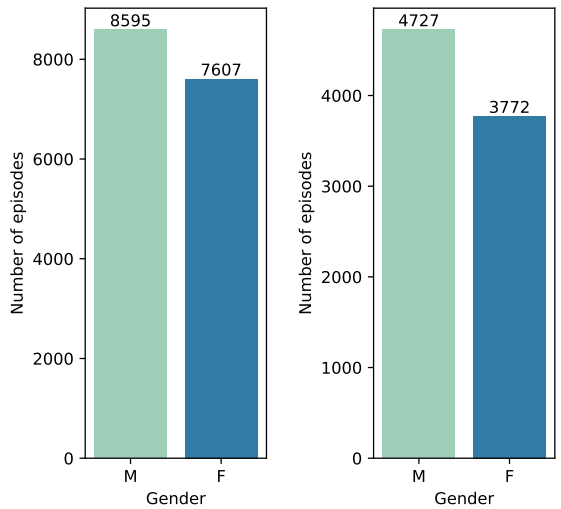
Table 8.2 provides the number of unique categories for each feature in both the initial and experimental datasets. As detailed in section 7.2.4, the gender column allows for two possible values: F (female) and M (male). Likewise, the age group column can take one of three possible values: *Preeschooler*, *Middle Childhood*, and *Teenager*. The initial dataset features 253 unique diagnoses and 57 unique ATC codes, while the experimental dataset comprises 251 unique diagnoses and 55 unique ATC codes. It's worth noting that the number of unique diagnoses and

ATC codes does not significantly decrease in the experimental dataset, indicating that the diversity of diagnoses and ATC codes was preserved even after excluding episodes containing only one diagnosis and no ATC codes.

Table 8.2: Number of unique categories for each feature of the datasets.

Column Name	Initial dataset	Experimental dataset
gender	2	2
age_group	3	3
diagnosis	253	251
atc_code	57	55

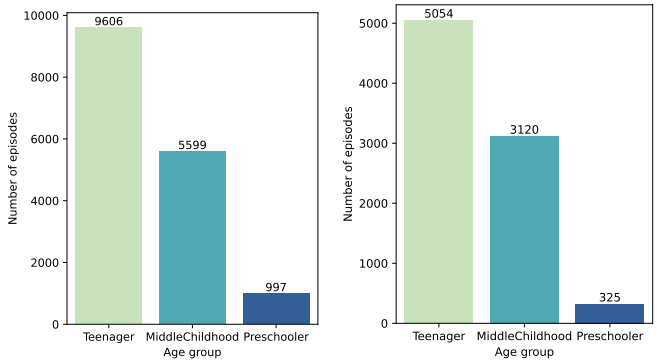
Gender: Figure 8.1 illustrates the gender distribution across both datasets, indicating a higher number of males compared to females. Interestingly, when transitioning from the initial to the experimental dataset, the female count is relatively more reduced than the male count, with the female count reduced by approximately 50% compared to a 44% reduction for males. This trend could suggest that episodes involving males are more likely to present with multiple diagnoses and to involve prescribed medication.



(a) Initial dataset. (b) Experimental dataset.

Figure 8.1: Gender distribution.

Age group: Figure 8.2 shows the distribution of age groups within both datasets. The teenage group is the largest, followed by the middle childhood group, while the preschooler group remains significantly smaller. Upon transitioning to the experimental dataset, the count of preschoolers experiences a more substantial reduction compared to that of teenagers and middle childhood individuals. This trend could imply that preschoolers are often diagnosed with a single condition and are not typically prescribed medication.

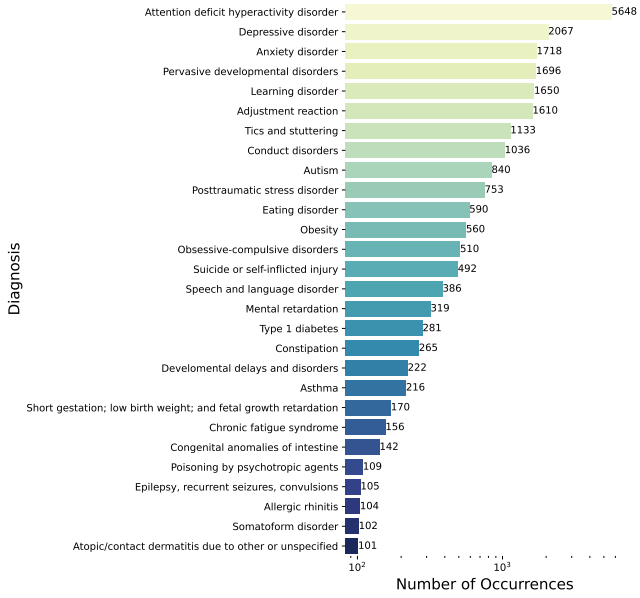


(a) Initital dataset. (b) Experimental dataset.

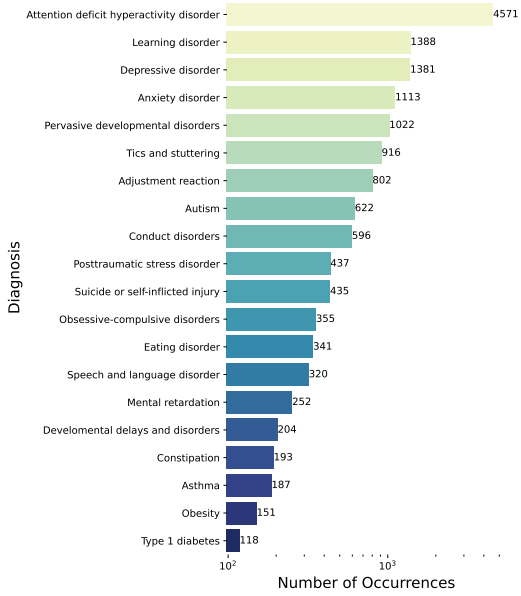
Figure 8.2: Age group distribution. The age is calculated based on the first date the patient was diagnosed in an episode.

Diagnoses: Figure 8.3 illustrates the frequency of each unique phenotype within an episode across both datasets. Only phenotypes occurring in more than 100 episodes are depicted. Notably, if a patient is diagnosed with both F81.0 (specific reading disorder) and F81.1 (specific spelling disorder), these will be collectively considered as the learning disorder phenotype and will count as a single occurrence. From the figure, it is evident that ADHD is the most frequently observed diagnosis in the cohort. Interestingly, while learning disorders are the fifth most frequent diagnosis in the initial dataset, it rises to the second most common diagnosis in the experimental dataset. This shift may imply that many patients receiving medication or diagnosed with multiple disorders also have a learning disorder.

ATC Codes: Figure 8.4 visualises the frequency of each ATC code appearing in an episode. Note that only ATC codes occurring in more than two episodes are included. Remember that the total amount of episodes with prescribed medication is 4,459. Also note that if an episode has multiple medications with the same first four characters of the ATC code, it will be counted as one appearance in the



(a) Initial dataset.



(b) Experimental dataset.

Figure 8.3: Count of occurrences of phenotypes in episodes. The x-axis scale is logarithmic and only phenotypes that appear in more than 100 episodes are shown.

episode. Both datasets maintain a consistent distribution, as every episode of care involving an ATC code is linked to a diagnosed disorder due to the left join operation performed when merging the datasets. As observed, the most frequent ATC codes fall into the N-group, representing medications targeting the nervous system. ATC codes A06B (drugs for constipation) and R06A (antihistamines for systemic use) are also common. The most frequently occurring ATC codes, along with their descriptions, are listed in Table 8.3.

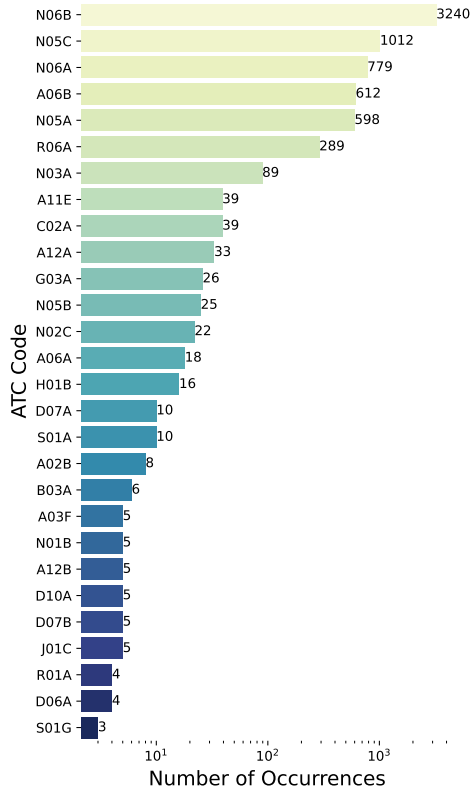


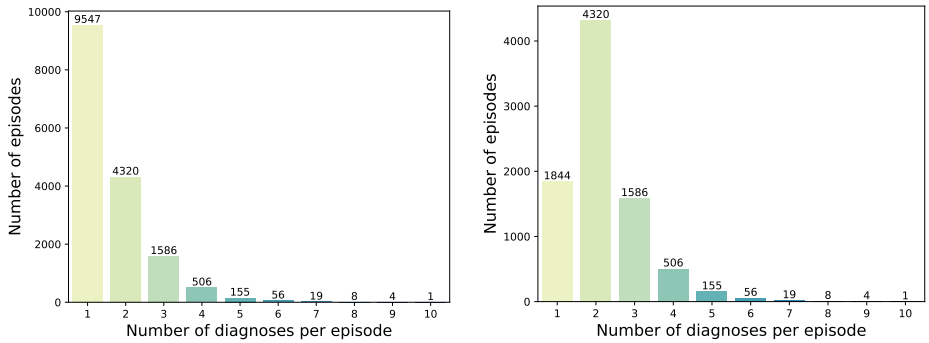
Figure 8.4: Count of occurrences of ATC codes in the initial dataset and the experimental dataset. The x-axis scale is logarithmic, and only ATC codes that appear in more than two episodes are shown.

Diagnosis count distribution: Figure 8.5 displays the distribution of the number of unique diagnoses in each episode. The initial dataset features a large number of episodes involving a single diagnosis. Yet, many patients are receiving two,

Table 8.3: Frequently occurring ATC codes with description.

ATC code	Description
A06B	Drugs for constipation
N05A	Antipsychotics
N05C	Hypnotics and sedatives
N06A	Antidepressants
N06B	Psychostimulants, agents used for ADHD and nootropics

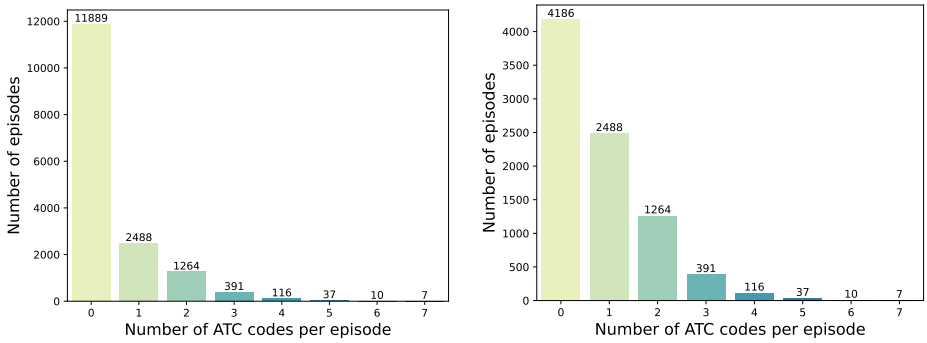
three, or four diagnoses. The episode with the highest count features a patient with ten distinct diagnoses, representing the greatest number of diagnoses in any single episode. It is also apparent that the experimental dataset still includes episodes with only one diagnosis. In these cases, it is known that at least one medication must have been prescribed for these patients, given their inclusion in the dataset.



(a) Initial dataset. (b) Experimental dataset.

Figure 8.5: Distribution of the number of unique diagnoses given in episodes.

ATC code count distribution: Figure 8.6 illustrates the distribution of the number of medications in episodes across both datasets. Most episodes do not involve prescriptions, but the experimental dataset shows a significant reduction in episodes with zero prescriptions. This indicates that many episodes with only one diagnosis do not feature medications. Further, the experimental dataset contains numerous episodes without any prescriptions, suggesting the presence of many patients with two or more distinct diagnoses but without any medication. The figure also indicates that many episodes include one or two different medications.



(a) Initial dataset.

(b) Experimental dataset.

Figure 8.6: Distribution of the number of unique ATC codes prescribed in episodes.

8.1.1 Key Takeaways from EDA

This section aims at summarising the key takeaways from the EDA.

- There are more boys than girls in the cohort.
- Teenagers between the ages 12-18 is the most frequent age group in the cohort with more episodes than the two other age groups in total.
- ADHD is the most frequently occurring diagnosis.
- Learning disorders often co-occur with other diagnoses.
- Medications that affect the nervous system are the most frequently prescribed in the cohort.
- Approximately 41% of the episodes in CAMHS involve patients receiving more than one diagnosis.
- In approximately 27% of CAMHS episodes, patients are prescribed medication.

8.1.2 Hypotheses Derived from EDA

This section aims to state some hypotheses of what will likely be shown in the clustering analysis. Remember that the experimental dataset, not the initial one, is the input dataset of MASPC.

- Given its high prevalence, ADHD will likely be present in many identified patterns and clusters.
- Due to their frequency, ATC codes in the N-group are expected to appear in numerous patterns and clusters.
- There is a likelihood of identifying patterns comprising solely of diagnoses with no associated medication.
- To expose subtle nuances and reveal more intriguing patterns, choosing threshold values that are not excessively high might be necessary, considering the wide variance in the frequency of diagnoses and ATC codes.
- Given the potential influence of demographic information on the clusters, a large number of clusters may be required to characterise distinct patient subgroups.

8.2 Determining Optimal Threshold Values

As described in Section 6.3, MASPC takes four input parameters. Specifically, the MAS component uses the *minSup*, *minAc*, and *minOv* parameters to detect frequently occurring patterns in the input dataset. The PC component uses the k parameter to determine the number of clusters to be generated. This section describes and justifies the approaches employed to determine the threshold values. Its inclusion within this chapter is justified because it constitutes an integral part of executing the MASPC algorithm. By elaborating on the methodologies used to establish these thresholds, the section aims to provide insights into the decision-making process involved in the execution of the algorithm.

As described in section 6.3.1, the choice of the threshold values *minSup*, *minAc*, and *minOv* determines the number of patterns identified in the dataset. Lower values increase the number of patterns, while larger values decrease the number of patterns. The patterns generated using higher thresholds show a less detailed picture of patterns in the dataset but can give a good overview of the most frequent patterns. On the other hand, lower thresholds result in more recognised patterns. However, these patterns may be too infrequent to be of any significance.

One way to determine which threshold values to use is to loop the algorithm using different thresholds and calculate the SI and CI score elaborated in section 6.5. However, as the MASPC algorithm only clusters, and hence also evaluates, records containing one or more frequent patterns, a good CI and SI score might also result from fewer clustered records. By running some experiments, it was seen that higher thresholds and fewer patterns generated better SI and CI scores, which is natural because fewer patterns make it easier to separate the records and create good clusters. However, these results with few patterns and records may not be of any significance to clinicians and medical personnel. As the goal of this thesis is to characterise patient subgroups in CAMHS, preferably of significance for clinicians and CAMHS, the threshold values were decided with the help of a clinician working at the CAMHS clinic at St. Olavs Hospital in Trondheim, Norway.

This method of deciding threshold values differs from the one applied by Zhong et al. [2020], which mainly focuses on each threshold's SI and CI score. It is worth mentioning that they also dealt with datasets with a larger variety of diagnoses, as one of the datasets contained data from an emergency department with 13,521 distinct diagnosis codes and the other de-identified patient data with 558 distinct diagnosis codes. In comparison, the experimental dataset contains 272 unique diagnoses, 108 with less than ten occurrences and 57 unique ATC codes, with 19 with more than ten occurrences. Because CAMHS has a smaller variety of

diagnoses, and the focus is on patterns in a specific area of the medical field, it has been valuable to receive expert feedback. This has helped to obtain results that are interesting and can be useful for CAMHS in the future.

The clinician at the CAMHS clinic at St. Olavs Hospital was shown patterns generated using different threshold values and gave feedback on which set of patterns was more interesting. The clinician was shown patterns generated using low, high and middle high threshold values and provided feedback that the patterns generated using the middle solution were the most interesting. These patterns were generated using $minSup = 0.03$, $minAc = 0.03$ and $minOv = 10$. As most of the patterns contain ADHD, it was discussed to only focus on these patterns. However, the clinician mentioned that there are three main groups of diagnoses in CAMHS; depression, anxiety and ADHD, and as all of these are present in at least one pattern, it is interesting to look at all patterns.

The PC part of MASPC takes in the number of clusters k and generates k clusters. To determine the optimal k-value for producing clusters for records with the selected patterns, PC was executed with varying k-values, and the corresponding SI and CI scores were computed. These scores were used to determine an optimal value for k. The outcomes of this analysis are depicted in Figure 8.7. It is important to mention that k-values lower than ten results in good scores because mostly gender and age groups are taken into account. However, given our specific interest in results focusing more on diagnoses and medications, higher k-values were considered. From the figure, it can be seen that $k = 31$ yields a relatively favourable trade-off among the SI-score, CI-score, and the number of clusters k, with a SI score of 0.18 and a CI score of 117.27. Consequently, a k-value of 31 was chosen for the experiments.

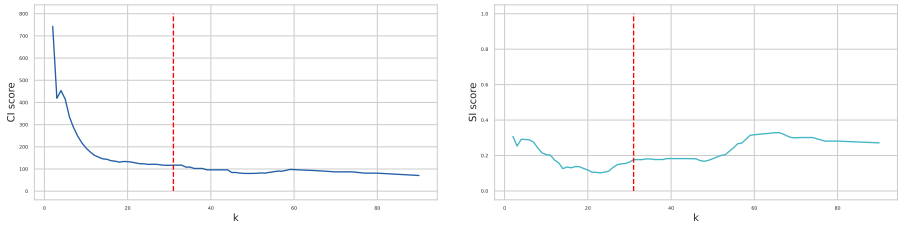


Figure 8.7: SI and CI scores for different k-values. Note that the scales of the y-axis are different. The red lines intersect the chosen k-vale of $k = 31$.

Table 8.4 presents the final threshold values used in the experiments. The criteria

used to determine the optimal values for the *minSup*, *minAc*, *minOv*, and *k* parameters are thoroughly discussed in this section. The selected threshold values are considered the most appropriate for the experimental dataset and have been implemented to obtain the results presented later in this chapter.

Table 8.4: Final threshold values for MASPC.

minSup = 0.03
minAc = 0.03
minOv = 10
k = 31

8.3 MASPC Results

This section presents the results obtained by applying the MASPC algorithm to the experimental dataset with the thresholds discussed in the previous section. Firstly, the patterns identified by the MAS component will be presented. Subsequently, the clustering results derived from the PC component will be shown and visualised in heatmaps, accompanied by key insights extracted from the findings.

8.3.1 Frequent Patterns Detected by MAS

Table 8.5 shows the 13 patterns detected by MAS in the experimental dataset comprising 8,499 records, each with their set of diagnoses and ATC codes containing more than one element. It can be seen that ADHD is present in ten of the patterns and that five of the patterns contain the ATC code N06B. Out of all the patterns, eight of them contain more than one disorder. Seven of the patterns contain one or more ATC codes, and three patterns contain two ATC codes.

8.3.2 Clustering Results of PC

This section presents the clusters and clustering results. After MAS identified MFAs, the PC part was employed to cluster the records. As explained in Section 6.3, only the records containing at least one MFA were included in the clustering step. Initially, the experimental dataset comprised 8,499 records with multiple diagnoses or ATC codes. Applying the thresholds outlined in Section 8.4, approximately 38% (3,240) of the records contained one or more MFAs and were subsequently clustered. This dataset is referred to as the *clustering dataset*. The 5,259 unclustered records were analysed through a basic EDA, but as it did not reveal any significant phenomena, and due to time limitations restricting more

Table 8.5: Patterns generated by MAS using minSup=0.03, minAc=0.03 and minOv=10.

ID	Patterns
0	Depressive disorder, N06A
1	Depressive disorder, Anxiety disorder
2	Anxiety disorder, N06A
3	Attention deficit hyperactivity disorder, N06B, N05C
4	Attention deficit hyperactivity disorder, N06B, A06B
5	Attention deficit hyperactivity disorder, N06B, Learning disorder
6	Attention deficit hyperactivity disorder, N06B, Tics and stuttering
7	Attention deficit hyperactivity disorder, Pervasive developmental disorders
8	Attention deficit hyperactivity disorder, Autism
9	Attention deficit hyperactivity disorder, Anxiety disorder
10	Attention deficit hyperactivity disorder, N06B, N05A
11	Attention deficit hyperactivity disorder, Depressive disorder
12	Attention deficit hyperactivity disorder, Conduct disorders

advanced analysis, it is excluded from this report. The clustering process resulted in the formation of 31 distinct clusters. Figure 8.9, 8.10 and 8.11 summarises the clustering results showing heatmaps of the MFA distribution, gender distribution and age group distribution in the obtained clusters.

Firstly, some statistics of the differences between the experimental dataset and the clustering dataset are presented. Table 8.6 shows the distribution of demographic information of the experimental dataset versus the clustering dataset containing the 38% of records. It can be seen that for most demographic features, somewhere between 35% and 40% were clustered. The middle childhood age group stands out, with 44% of records with patients in this group clustered. Also, the subgroup of boys stands out, with 41 % of the records containing one of the patterns and being clustered.

In general, the clusters separate well on gender and age group. All clusters except for cluster 17 only contain episodes where all patients belong to the same gender. Similarly, all clusters except for cluster 8 contain episodes where all patients in each cluster belong to the same age group.

Size of clusters: Figure 8.8 illustrates a bar plot depicting the number of

Table 8.6: Counts of the experimental dataset and the clustering dataset.

	Experimental dataset	Clustering dataset	Percentage reduced
Female	3772	1318	35 %
Male	4727	1922	41 %
Preschooler	325	116	36 %
Middle childhood	3120	1371	44 %
Teenager	5054	1753	35 %
Total number of records	8499	3240	38 %

episodes in each cluster. The plot demonstrates a notable variation in the record counts across the clusters. Specifically, Cluster 16 comprises the fewest episodes, with a count of n=6, while Cluster 11 has the highest number of episodes, totalling n=445. Moreover, other large clusters are Cluster 7 (n=246), Cluster 20 (n=293), and Cluster 25 (n=206). On the other hand, Cluster 15 (n=8), cluster 16 (n=10) and Cluster 28 (n=13) are smaller clusters.

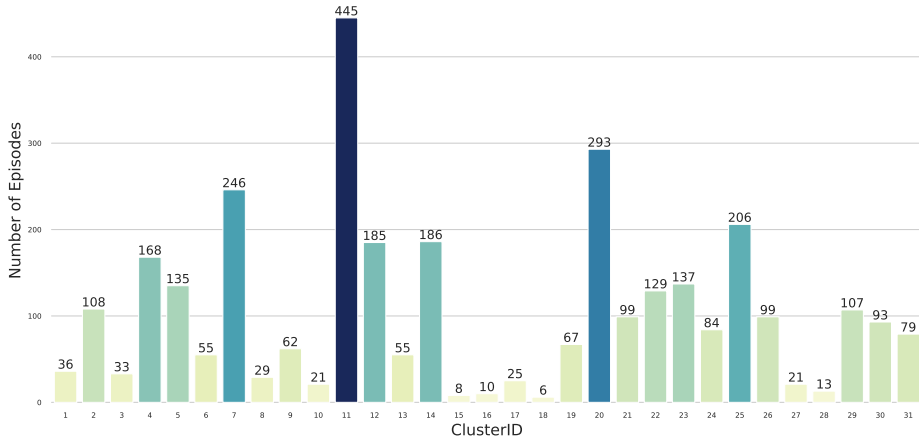


Figure 8.8: Barplot showing the number of episodes in each cluster.

Age group: The heatmap presented in Figure 8.9 reveals that the clusters separate well concerning age groups. Except for Cluster 8, each cluster comprises episodes wherein all patients belong to the same age group. Specifically, 14 clusters are composed of teenagers, 5 of preschoolers, and 12 of patients in the middle childhood age group. Clusters 1-7 and 9-14 exclusively encompass episodes involving teenagers, while Cluster 8 also involves a majority of teenagers. Clusters

15-19 exclusively involve preschoolers, and clusters 20-31 only consist of middle childhood patients. This segregation based on age groups indicates that the clustering algorithm successfully captures the age-related patterns within the dataset.

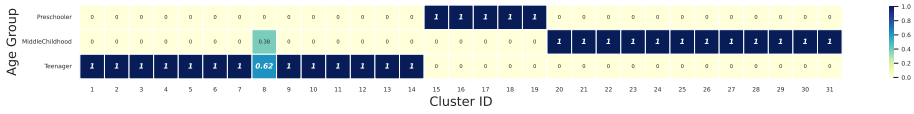


Figure 8.9: Heatmap showing the age group distribution in clusters.

Gender: Figure 8.10 demonstrates that the clusters separate well on gender, as all clusters, except for Cluster 17, consist of patients of the same gender. Notably, there are 17 clusters where the majority of patients are boys, while 14 clusters have a majority of girls. Clusters 8-16 and clusters 27-31 exclusively include episodes involving girls, while clusters 1-7 and clusters 18-26 exclusively involve boys. Additionally, Cluster 17 comprises a majority of boys. The clear differentiation of clusters based on gender indicates that the clustering algorithm effectively captures the gender-related patterns within the dataset.

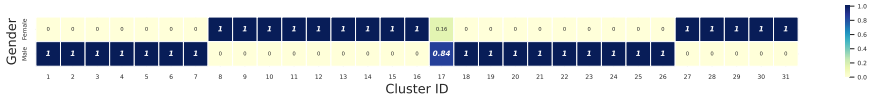


Figure 8.10: Heatmap showing the gender distribution in clusters.

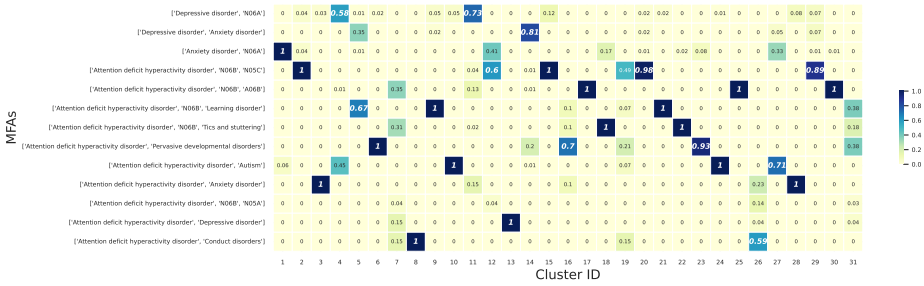


Figure 8.11: Heatmap showing the MFA distribution in the clusters.

Cluster 1: Cluster 1 is relatively small and comprises n=36 episodes and patients with similar demographics, diagnoses and prescriptions. All patients in this cluster are teenage boys diagnosed with anxiety disorder and given antidepressant medications.

Clusters 2, 12, 15, 20 and 29: These clusters comprises n=108, n=185, n=8, n=293 and n=107 yielding a total of 701 episodes. All of these clusters have a majority of patients with ADHD that are prescribed psychostimulants and hypnotics and sedatives. In clusters 2 and 20 (n=401 episodes in total) are all patients boys, with Cluster 2 being only teenagers and Cluster 20 patients in the middle childhood age group. Clusters 12, 15 and 29 (n=300) comprise episodes where the patients are girls in respectively the age groups teenager, preschooler and middle childhood.

Clusters 3 and 28: Clusters 3 (n=33) and 28 (n=13) comprise a total of 46 episodes containing patients with ADHD and anxiety disorder. The patients in Cluster 3 are teenage boys, and the patients in Cluster 28 are middle-childhood girls.

Clusters 4 and 11: Cluster 4 (n=168) and 11 (n=445) comprise a total of 613 episodes, with a majority of patients diagnosed with depressive disorders and given antidepressant drugs. In Cluster 4, 58% of the patients belong to this pattern, while 45% are diagnosed with ADHD and autism. This cluster contains teenage boys. Cluster 11 is the largest of all clusters and contains teenage girls.

Clusters 5, 9 and 21: Clusters 5 (n=135), 9 (n=62) and 21 (n=99) comprise a total of 296 episodes with a majority of patients diagnosed with ADHD and learning disorder that are prescribed psychostimulant drugs. Cluster 5 consists of teenage boys. Here, 35% are diagnosed with depressive disorders and anxiety disorders. Cluster 9 contains teenage girls, and Cluster 21 middle childhood boys. In these two clusters, all patients belong to the same pattern.

Clusters 6, 16 and 23: These clusters have a majority of patients diagnosed with ADHD and pervasive developmental disorders. Cluster 6 (n=55) contains teenage boys, Cluster 16 (n=10) preschooler girls and Cluster 23 (n=137) boys in the middle childhood age group. In total, these clusters comprise a total of 202 episodes.

Clusters 7, 17, 25 and 30: These clusters have a majority of patients diagnosed with ADHD that are prescribed psychostimulant drugs and drugs for constipation. Cluster 7, containing 246 episodes that are teenage boys, has 35% of its episodes belonging to this pattern. This cluster also has 31% of the patients diagnosed with ADHD and tics and stuttering that are prescribed psychostimulants. Cluster 17 (n=25) contains preschoolers, where a majority are boys. Cluster 25 (n=206) consists of middle-childhood boys, and cluster 30 (n=93) middle-childhood girls. In total, these clusters comprise a total of 570 episodes.

Clusters 8 and 26: This group of clusters comprises a total of 128 episodes

where a majority of the patients in each cluster are diagnosed with ADHD and conduct disorders. Cluster 8 (n=29) contains girls, where a majority are teenagers. Cluster 26 contains middle-childhood boys.

Clusters 10, 24 and 27: Clusters 10 (n=21), 24 (n=84) and 27 (n=21) comprise a total of 126 episodes with a majority of patients diagnosed with ADHD and autism. Cluster 10 contains teenage girls, Cluster 24 middle-childhood boys and Cluster 27 middle-childhood girls. In cluster 27, 33% are diagnosed with anxiety disorder and prescribed antidepressants.

Cluster 13: Cluster 13 contains 55 episodes where all patients are diagnosed with ADHD and depressive disorders. All patients in this cluster are teenage girls.

Cluster 14: Cluster 14 contains 186 episodes where 81% are diagnosed with depressive disorders and anxiety disorders. 20% of the patients in this cluster are diagnosed with ADHD and pervasive developmental disorders. All patients in this cluster are teenage girls.

Cluster 18 and 22: This group of clusters comprises a total of 135 episodes where all patients are boys diagnosed with ADHD and tics and stuttering and prescribed psychostimulants. Cluster 18 (n=6) is the smallest cluster and contains preschoolers, and Cluster 22 (n=129) contains patients in the middle childhood age group.

Cluster 19: This cluster contains 67 episodes of preschooler boys. 49% in this cluster are diagnosed with ADHD and prescribed psychostimulants and hypnotics and sedatives.

Cluster 31: Cluster 31 contains 79 episodes of middle-childhood girls. All patients in this group are diagnosed with ADHD, and 38% are additionally diagnosed with a learning disorder and prescribed psychostimulants. Another 38% are diagnosed with pervasive developmental disorders in addition to the ADHD diagnosis.

8.3.3 Main Takeaways

The main takeaways from the clustering results can be summarised as follows:

- Most patients diagnosed with both depressive disorders and anxiety are teenagers (>90%). A majority of these teenagers are girls (ca. 75%).
- A majority of patients diagnosed with ADHD and pervasive developmental disorders are boys (>70%), most in the middle childhood age group (ca.

65%). Relative to the number of preschoolers in the cohort, a relatively large amount (ca. 18%) of these belong to this group.

- Patients diagnosed with both ADHD and autism are usually either teenagers (ca. 50%) or in the middle childhood age group (ca. 50%). Most of these patients are boys (ca. 88%).

- Most patients diagnosed with both ADHD and anxiety disorders are teenagers (ca. 73%). Some are also in the middle-childhood age group (ca. 26%). More girls than boys get these diagnoses when they are teenagers (67% vs 33%), and more boys get these diagnoses when they are in the middle childhood age group (63% vs 37%).

- Patients receiving both an ADHD diagnosis and a depressive disorder diagnosis are usually teenagers (ca. 99%). A majority of these are girls (ca. 60%).

- There is a majority of boys diagnosed with both ADHD and conduct disorders (70%). For both genders with these diagnoses, the majority are diagnosed when they are in their middle childhood (ca. 50%) or teenagers (ca. 50%).

- A majority of the patients diagnosed with ADHD and prescribed both psychostimulants and hypnotics and sedatives are boys (ca. 67%). Most of these boys are in the middle childhood age group (67%). Relative to the number of preschoolers in the cohort, a large amount of the preschoolers (35%) belong to this group.

- Most patients diagnosed with ADHD and getting psychostimulant drugs and drugs for constipation are in the middle childhood age group (ca. 64%). A majority of these are boys (ca. 69%). Out of the total number of preschoolers, a significant amount (21%) of these belong to this group.

- Most patients diagnosed with ADHD that are prescribed psychostimulants and antipsychotic drugs are boys in the middle childhood age group (ca. 99%).

- A majority of patients getting both an ADHD and a learning disorder diagnosis and are prescribed psychostimulant drugs are boys (ca. 67%). Almost all boys in this group are either in their middle childhood (48%) or teenagers (ca. 52%).

- Almost all patients receiving an ADHD and a tics and stuttering diagnosis and prescribed psychostimulants are boys (ca. 94%), most in the middle

childhood age group (ca. 60%). The few girls in this group (ca. 6%) are in the middle childhood age group.

- There are generally more patients who have an ADHD diagnosis and are medicated with psychostimulants and either hypnotics/sedatives or drugs for constipation than the number of patients in the other clusters.
- A majority of men have diagnoses and medications that occur in most patterns, apart from the patterns that include depression and anxiety disorders. These tend to have a predominance of women.
- Preschoolers who have diagnoses and medications that occur in the identified patterns often have ADHD and are prescribed medications with two different ATC codes.

Chapter 9

Results Validation

This chapter discusses and validates the clinical aspects of the findings and discoveries presented in Section 8.3. Throughout this research, clinicians have actively participated, providing input, feedback, and insights into how the characterised subgroups correlate with the context of CAMHS. The clinical input and validation presented in this chapter are derived from the input and feedback received from Hanne Klæboe Greger, a senior physician at CAMHS. On May 16, 2023, a meeting was held with Dr Greger to present the results outlined in Chapter 8, along with the key takeaways highlighted in Section 8.3.3. The meeting involved in-depth discussions regarding the implications of the results and takeaways within the context of CAMHS. These discussions served as the foundation for the validation presented in this chapter.

The chapter begins by discussing the identified patterns and their applicability within CAMHS. This is followed by a discussion of the cluster analysis results, validated against real-world conditions observed in CAMHS. The discussion of patterns and clustering results primarily targets patterns 3 through 12. These patterns warrant special attention due to their encapsulation of either co-occurring morbidities or co-occurring medications. Examining these patterns provides insights into the simultaneous presence of multiple disorders or medications within the CAMHS population. Analysing these co-occurrences can shed light on potential associations, treatment approaches, and the overall complexity of the patient profiles in CAMHS.

9.1 Identified Patterns

The patterns identified and presented in Table 8.5 represent frequent patterns identified by MAS. Each pattern captures a frequently occurring combination of diagnoses and medications in the dataset, mined based on the defined threshold values presented in Section 8.2. According to Dr Greger's feedback, these patterns align well with the reality of CAMHS. The disorders within the patterns are commonly observed in clinical practice, and the ATC codes correspond to frequently prescribed medications. However, Dr Greger did express concerns regarding the broad categorisation of the phenotype labelled as *Pervasive developmental disorders*, which encompasses ICD-10 codes F93, F94, and F98. She pointed out that this categorisation fails to capture the diverse range of diagnoses enclosed within it effectively, and this limitation is discussed further in Section 10.2. Dr Greger proposed that a more suitable name for this phenotype would be *Behavioral disorders, emotional disorders, and disorders in social functioning that occur in childhood* in relation to the CAMHS practice and highlights the F98 group of diagnoses as the likely contributor to records containing pervasive developmental disorders and, thus, the pattern associated with this phenotype. Throughout this thesis, it is important to consider this alternative phenotype description when referring to pervasive developmental disorders.

Alongside the patterns underscoring the prevalence of ADHD within the cohort, the EDA also indicates ADHD as the most frequent disorder in CAMHS, with many patients receiving this diagnosis. This is well known to clinicians in CAMHS. Furthermore, the co-occurrence of ADHD with other disorders is also supported in the literature by [Hansen et al., 2018], [CADDRA, 2020, p.14-15], and [Coghill et al., 2021]. For ADHD treatment, psychostimulants (N06B) are routinely prescribed as they help reduce symptoms and improve overall functioning. These trends correspond with the identified patterns in the dataset, thereby enhancing the alignment between the analysis findings and the existing knowledge in the field.

However, Dr Greger points out that it is surprising that patterns 7, 8, 9, 11, and 12 do not include any medications, particularly considering that patients with these diagnoses typically require medications from the ATC category N06B, commonly associated with these complex conditions. However, the absence of medications in these patterns can be attributed to the fact that they do not meet the user-defined thresholds, indicating that the frequency of medications for these specific combinations is lower compared to the disorders observed in patterns 5 and 6.

9.2 Co-Occurring Morbidities

The results indicate a majority of girls in episodes encompassing patterns involving depression or anxiety in addition to other diagnoses (Patterns 1, 9, and 11). This finding aligns with Dr Greger's expectations, as it is commonly observed that girls receive these diagnoses more frequently. Furthermore, the results highlight that almost exclusively teenagers are diagnosed with both ADHD and depression. Dr Greger underscores the rationale behind this observation, explaining that it is logical as it is uncommon to receive a depression diagnosis before puberty.

The results indicate a higher proportion of boys than girls diagnosed with both ADHD and anxiety during their middle childhood. However, the group receiving these diagnoses when teenagers is larger, and here there is a higher number of girls than boys. This observation aligns with the fact that ADHD is more commonly diagnosed in boys at a younger age. Anxiety is widely recognised as a common comorbid condition with ADHD. Dr Greger highlights that the longer an individual has an ADHD diagnosis, the higher the probability of developing other disorders. Additionally, Dr Greger mentions that the standard practice at the St. Olavs CAMHS clinic in Trondheim has been to delay the coding of an anxiety diagnosis if the patient has already received an ADHD diagnosis. It is mentioned that if a patient is referred to CAMHS for a potential ADHD diagnosis and there are suspicions of anxiety, the usual approach is first to provide an ADHD diagnosis and initiate treatment targeting ADHD symptoms. Subsequently, if the anxiety symptoms persist or worsen, an anxiety diagnosis may be coded and appropriate measures implemented.

The analysis of the results reveals a noteworthy pattern: most patients diagnosed with both ADHD and a pervasive developmental disorder are male individuals in their middle childhood. Dr Greger mentions that the pervasive developmental disorders in this cohort primarily fall within the F98 category, which encompasses patients experiencing difficulties with controlling defecation and urination. It is unsurprising to find a limited number of teenagers in this group, as diagnosing pervasive developmental disorders in teenagers typically requires a higher threshold for this group. Furthermore, the predominantly male composition of this group aligns with Dr Greger's assumption that it may be influenced by a higher incidence of urinary and defecation issues among boys during their initial years of school. Additionally, as mentioned earlier, boys are generally diagnosed with ADHD at an earlier age compared to girls, which further contributes to the gender distribution observed in this particular group.

ADHD and autism are two disorders that often co-occur, which also is seen in the results. Dr Greger emphasises the presence of overlapping symptoms between

ADHD and autism, noting that doctors often remain vigilant for signs of autism in patients with ADHD. Consequently, it is unsurprising that one identified pattern includes both diagnoses. Dr Greger further highlights that a surprisingly large number of patients in this cohort are boys. It is worth noting that the most severe cases of autism typically do not fall within the purview of CAMHS. Instead, the patients receiving such a diagnosis within CAMHS tend to exhibit relatively better social functioning. Moreover, the results align with expectations as they indicate that most patients receiving these diagnoses are either in the middle childhood or teenage age group. This corresponds to the fact that signs of autism tend to become more pronounced when children enter school, leading to increased diagnosis rates during these developmental stages.

The clustering results reveal a predominant presence of males diagnosed with both ADHD and conduct disorders. Dr Greger emphasises the close association between ADHD and conduct disorders, noting that boys often exhibit hyperactivity and difficulties in regulating their impulses. Conduct disorders carry a negative connotation, and physicians exercise caution before assigning this diagnosis, ensuring they have sufficient evidence and certainty. It is customary for preschoolers to exhibit challenging behaviours, but it typically becomes more problematic when they enter school. This is also usually when a patient may receive such diagnoses, explaining the clustering of patients in the middle childhood and teenage groups.

9.3 Morbities and Medications

Pattern 3 reveals a common combination where patients have an ADHD diagnosis, receive psychostimulant drugs, and are prescribed hypnotics and sedatives. The clustering results demonstrate a predominance of boys within this group, primarily in the middle childhood age group. This observation aligns with the fact that boys tend to receive ADHD diagnoses earlier than girls, which may explain the higher representation of boys with this particular combination. Dr Greger explains that hypnotics can refer to sleep medications, and the high frequency of the ATC code N05C (hypnotics and sedatives) is likely due to many patients being prescribed melatonin to address their sleep difficulties. Additionally, it is noteworthy that a relatively significant portion of preschoolers falls within this group. Dr Greger highlights that this finding is not surprising, as parents of children with ADHD often struggle with their children's sleep patterns and seek interventions to improve their sleep quality.

Pattern 4, which showcases the combination of an ADHD diagnosis and the prescription of psychostimulants and drugs for constipation, does not surprise Dr

Greger. In CAMHS, constipation is a common occurrence, and many patients are expected to require these medications for either short-term or long-term management. The fact that most patients with this combination are boys aligns with the observation that boys tend to receive ADHD diagnoses at earlier ages. It is during these earlier ages that constipation issues are also more prevalent.

Pattern 10, characterised by the combination of ADHD, psychostimulant drugs, and antipsychotic drugs, is likely attributed to the usage of the drugs Risperidone and Quetiapine, both classified as antipsychotic drugs. Risperidone is classified as an antipsychotic drug but is often employed for treating conduct disorders, and Quetiapine may be used in low doses as a sleep aid. The clustering results indicate a significant presence of boys within the middle childhood age group who exhibit this pattern. Dr Greger speculates that these boys may struggle with ADHD and conduct disorders as they transition into the school environment. This could explain the higher incidence within this particular age group.

That the majority of patients having both ADHD and learning disorders in addition to being prescribed psychostimulants is also no surprise. Learning disorders are typically identified when children begin their schooling, which aligns with the finding that the majority of patients with both ADHD and learning disorders, who are prescribed psychostimulants, are males in their middle childhood or teenage years. This observation is consistent with the understanding that boys are often diagnosed with ADHD at an earlier age compared to girls. As a result, boys are expected to be more likely to receive additional diagnoses, such as learning disorders, during the same age range.

Dr Greger finds the low number of girls in the clustering results for patients with ADHD, tics, stuttering diagnoses, and prescribed psychostimulants to be unexpected. While it is acknowledged that tics and stuttering diagnoses are more prevalent in boys, the gender disparity within this particular group is notable. Within the context of CAMHS, most of these diagnoses likely correspond to Tourette syndrome, which exhibits a high comorbidity rate with ADHD. Dr Greger highlights that individuals with Tourette syndrome also have a 50% chance of having ADHD. It is not surprising that most patients receiving a Tourette syndrome diagnosis receive the diagnosis during their middle childhood years, as this is when most individuals are diagnosed with the condition.

Chapter 10

Discussion of Methodology and Design

This section discusses the project's methodology and design related to the planning and execution of the project. The discussion begins with an evaluation of the planning of the project, specifically the adherence to the WBS and timeline outlined in sections 7.1.3 and 7.1.4. Next, the cleaning and preprocessing of diagnoses and age groups and their potential errors are addressed, with an emphasis on their potential impact on the results. The subsequent section provides a discussion of the advantages and limitations of the MASPC method. The clinical feedback, input, and validation throughout the project are then discussed. This section focuses on the frequency of communication and the insights and benefits derived from the presentations and meetings conducted. The discussion then moves to address the experimental limitations tied to the project's framework. Finally, the experimental aims are revisited, and the extent to which these have been achieved is assessed.

10.1 WBS and Timeline

This section evaluates the WBS and timeline presented in sections 7.1.3 and 7.1.4.

The WBS was employed to decompose the tasks into smaller components to facilitate project management. Although initially designed for the original scope of the thesis mentioned in Section 6.1, adjustments were made when the scope changed. The WBS was a valuable starting point, providing an overview of the

work and dividing it into distinct WPs. However, it should be mentioned that an enhancement to the WBS could have been the inclusion of a dedicated phase solely for writing the thesis and documenting the process, spanning the entire duration of the project. This addition would have provided a more accurate representation of the overall process, as significant attention was devoted to documenting content along the way to ensure important details were not overlooked. Although this aspect could have been better represented within the WBS, most of the documentation occurred during the latter part of Phase 2 and throughout Phase 3.

In general, the timeline depicted in Figure 7.2 and Table 7.1 was adhered to, with most WPs completed within their designated timeframes. However, it should be acknowledged that the actual progression of the project was not as linear and streamlined as presented. Although WP2.3, preprocessing, was finalised by week 12, subsequent minor refinements and adjustments were made as potential improvements were identified at later stages. Furthermore, while WP1.3, semi-structured literature review, was initially intended to be completed by week 8, the inclusion of new papers in March extended its completion timeline. It is also worth mentioning that the completion of WP3.1, documentation and validation, was delayed a few days due to the unavailability of clinicians to validate and provide feedback on the results. Furthermore, as previously mentioned, it is essential to emphasise that the definition phase, including WP3.2, documentation, was an ongoing aspect throughout the project. Unlike other WPs that were allocated specific timeframes, the documentation process was intertwined with the various stages of the project. By integrating documentation throughout the project, the intent was to maintain a correct record of the research process, results, and findings.

10.2 Data Cleaning and Preprocessing

This section critically evaluates the strengths and limitations of specific data preprocessing techniques employed in the research, focusing on two aspects: the mapping of diagnoses to phenotypes and the grouping of patient ages.

10.2.1 Conversion of ICD-10 Diagnoses to Phenotypes

In the process of converting ICD-10 diagnoses to phenotypes, it became evident that there were some limitations due to missing conversions for certain diagnoses and differing levels of detail in the phenotypes. To address this issue, a solution was sought by interpreting the missing codes, assigning, and sometimes changing appropriate phenotypes. This approach is described in Section 7.2.4, and all

added and changed phecodes conversions can be found in Appendix G. It is important to acknowledge that this approach has limitations as it was undertaken by an individual without medical expertise. It would have been optimal to have this process conducted by someone with medical experience to ensure accuracy and thoroughness. However, due to limited resources, the researcher performed the task, leading to some unavoidable drawbacks. Despite these limitations, every effort was made to ensure that the diagnoses were converted as accurately and comprehensively as possible.

Converting ICD-10 diagnoses to phenotypes in the dataset revealed limitations stemming from the lack of conversions for certain codes and varying levels of detail in the phenotypes. For instance, while most eating disorders were classified as the phecode and phenotype *305.2, eating disorder*, F50.0, Anorexia Nervosa, was classified as *305.21, anorexia nevrosa*. This discrepancy in detail was identified and modified to the phenotype eating disorder through the consultation and approval of a clinician. Nevertheless, there is a possibility that some cases may have gone unnoticed. To address this issue, a potential solution is to shorten the phecodes to a maximum of four characters and assign the corresponding phenotype. Unfortunately, this option was not considered before conducting the experiments. As such, some limitations may have persisted despite the efforts to mitigate them.

During the presentation of the research findings to a clinician, it was conveyed that the phenotype labelled as *Pervasive developmental disorders* encompassing ICD-10 codes F93, F94, and F98 does not effectively capture the diverse range of diagnoses included within it, as the variability among these disorders is too large for such a broad categorisation. Nevertheless, the clinician acknowledged that for practical purposes and to avoid excessive fragmentation of distinct phenotypes, it may still be beneficial and necessary to have certain groups encompassing a range of diagnoses. In light of this discussion, it was proposed that a more suitable name for this phenotype would be *Behavioral disorders, emotional disorders, and disorders in social functioning that occur in childhood* in relation to the CAMHS practice. This alternative description of the phenotype should be considered when referring to pervasive developmental disorders in this thesis.

10.2.2 Age Groups

As explained in Section 7.2.4, the age of patients was grouped into three age groups; preschooler, middle childhood and teenager. This choice was made to align with the school system in Norway, potentially revealing important insights, as certain mental disorders tend to manifest more prominently when children enter school. The selection of these specific age groups was also subject to discussion

and approval by clinicians. However, as will be further discussed in Section 10.3, the evaluation of the results indicates a strong influence of the age groups on the separation of clusters. One potential solution to mitigate this issue would have been to retain the age as an individual feature, encompassing values from 0 to 18, instead of grouping them. This would have resulted in 19 distinct features. While alternative grouping strategies could have potentially revealed different patterns, it proved challenging to identify logical age groupings beyond those employed in this research.

The grouping of ages offers advantages in simplifying the analysis and facilitating the identification of age-related patterns. By grouping patients into age groups, the focus is shifted towards broader age ranges that may exhibit common characteristics. This grouping enhances the interpretability of the results and aids in identifying age-specific patterns and trends. However, there are also limitations associated with this approach. One limitation is the potential loss of individual variability within each age group, as diagnoses and treatment may vary significantly within a given age range.

10.3 MASPC

This section delves into the technical aspects of the clustering results obtained through the MASPC algorithm, specifically evaluating the choice of threshold values and the impact of features on the clustering outcomes. This is discussed to contribute to the understanding of the strengths and limitations of the clustering results.

10.3.1 Threshold Values and Number of Clusters

Section 8.2 addresses the determination of the threshold values *minSup*, *minAc* and *minOv*, which were determined by consulting with a professional rather than by finding thresholds that yielded optimal scores by the validity metrics. This decision was made after experimenting with different thresholds and finding that higher values resulted in better validity scores but did not provide interesting clinical results. Therefore, it was deemed necessary to seek the input of clinicians. In the meeting, the participants discussed and chose threshold values that generated clinically interesting patterns while avoiding excessive patterns.

Figure 8.7, presented in Section 8.2, is the basis for determining this research's optimal number of clusters. The appropriate value for k was determined by examining the SI and CI scores for various k values, as explained in Section 8.2. As observed, lower k values produced better SI and CI scores. However, the resulting

clusters were predominantly separated based on demographic information alone. Given the research goal of characterising subgroups related to patient diagnoses, medication, and demographics, larger k values were necessary to achieve better granularity in the results. Consequently, although the chosen k value did not yield the best scores, it was selected as a compromise between relatively good scores and an appropriate number of clusters.

While these approaches of selecting thresholds and k-value may have led to more accurate and clinically relevant results, it is also important to acknowledge that it can introduce a bias towards what is already known in CAMHS and what may be deemed interesting by a person working in the field. This can potentially hinder the discovery of new patterns or coherences that are not yet known or may be overlooked due to preconceptions. As such, future research efforts may benefit from a more objective and systematic approach to threshold determination, taking into account both the evaluation metrics and the input of medical professionals to ensure the highest level of quality in the data analysis.

10.3.2 Number of Clustered Episodes

As described in Section 6.3, PC exclusively clusters records that exhibit one or more of the identified patterns. Therefore, approximately 38% (3,240 out of 8,499) of the records used in the pattern recognition phase underwent clustering in this study. An advantage of solely clustering records containing an identified pattern is the elimination of noisy data that lacks frequently occurring diagnoses and medications, resulting in a reduction of noise within the dataset. Given the objective of characterising patient subgroups, this approach proves valuable in gaining insights into the broader patient population represented in the dataset. Nevertheless, exploring and clustering larger portions of the data may be desirable depending on the nature of the research. This could be achieved by adjusting the threshold values to encompass a broader range of records, thereby increasing the number of clustered instances.

An alternative approach to increase the number of clustered records is to modify the algorithm to mine frequent itemsets instead of MFIs. This adjustment would involve including more records in the clustering process. For instance, consider Pattern 5, which encompasses ADHD, learning disorder, and N06B. With the current algorithm, patients diagnosed with ADHD and a learning disorder but not receiving medication with the ATC code N06B would be excluded from the clustering dataset. However, given that learning disorders are commonly associated with ADHD as a comorbidity, this limitation should be acknowledged. By mining frequent itemsets rather than MFIs, this limitation would be addressed. Nevertheless, it is important to note that this alternative approach would introduce a

bias towards records containing MFIs, as these would have a higher occurrence of identified patterns within their records because they would be present in all frequent itemsets that are subsets of the MFI.

10.3.3 Impact of Features on Clustering Results

Figure 10.1 displays the SHAP (SHapley Additive exPlanations) summary plot, which illustrates the impact of each cluster feature on the clustering outcome. It is important to note that the classes depicted in the SHAP plot are zero-indexed, but for clarity, they will be referred to as the corresponding clusters they represent. For example, Class 0 corresponds to Cluster 1, and so forth. The SHAP framework, introduced by Lundberg and Lee [2017], offers a means to interpret, evaluate, and explain machine learning predictions. In the context of this research, utilising SHAP values and the accompanying SHAP plot allows for the interpretation of how individual features within the dataset influence the results produced by the machine learning algorithm.

The SHAP plot reveals that the demographic attributes *F*, *MiddleChildhood*, and *Teenager* strongly affect the separation of clusters. Especially, it can be derived from the plot that Cluster 11, the largest cluster, containing female teenagers, is strongly influenced by the presence of the demographic attributes *F* and *Teenager* and Pattern 0. This can also be derived from the heatmaps in figures 8.9, 8.10 and 8.11. The strong impact of demographic attributes can be seen as both an advantage and a disadvantage. On the positive side, it facilitates the characterisation of distinct demographic subgroups and their corresponding diagnoses and medications. However, a disadvantage is that it fails to recognise groups with similar diagnoses and ATC codes across different demographics. Nevertheless, it is possible to derive such comparisons and identify patterns of similar diagnoses and ATC codes by carefully examining the clustering results and observing similarities across clusters.

In conclusion, it can be argued that certain demographic attributes have a disproportionately large influence on the clustering outcome. However, this observation aligns with expectations, considering that records that do not contain any identified patterns are eliminated before clustering. Consequently, it is anticipated that demographic attributes will have a larger impact on the outcome, as all records possess gender and age group information, yielding the presence of two 1s in the five first features, compared to the presence of one or more 1s in the thirteen patterns comprising the rest of the features in the clustering dataset.

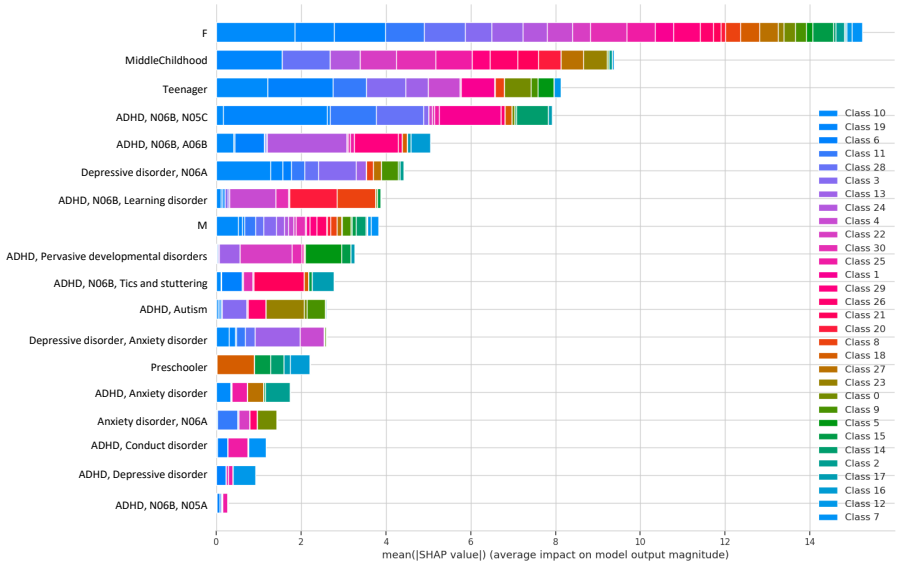


Figure 10.1: SHAP summary plot showing the impact of each feature on the clustering results. The classes depicted are zero-indexed, meaning Class 0 refers to Cluster 1 etc.

10.4 Clinical Validation

This section focuses on the clinical validation of the research, highlighting the involvement of clinicians and professionals throughout the project to ensure its quality, as well as the interpretation and discussion of results. The section provides an overview of the clinical evaluation process, including presentations and meetings. However, the specific discussion and analysis of the clinical results were presented in Chapter 9 and will not be reassessed in this section.

Throughout the project, the IDDEAS team has been available for questions and feedback during their weekly meetings and through email communication. Furthermore, they have been available for additional scheduled meetings as needed. Additionally, on the 14th of April 2023, the data selection, extraction and cleaning process, the choice of algorithm and some initial statistics about the dataset were presented to the team, providing an opportunity for valuable feedback and comments. Subsequently, certain adjustments were implemented based on the feedback received, such as excluding medical records of adults and merging infants, toddlers, and preschoolers into a single age group during the data cleaning

phase. The presentation also played a crucial role in validating and confirming the correctness of the decisions made throughout the project.

On May 16th, 2023, the patterns and clustering results were presented to Dr Hanne Klæboe Greger, a senior physician working in CAMHS. During the meeting, Dr Greger provided an overview of the clustering results, followed by a detailed discussion of each key takeaway and its clinical implications. The meeting was valuable and provided points for the clinical validation in Chapter 9. Furthermore, Dr Greger offered valuable input regarding the clinical application of the prescribed medications and the potential patient profiles associated with specific diagnoses and medications. Feedback related to the phenotype *Pervasive developmental disorders* was also received, as outlined in Section 10.2. In summary, this meeting with Dr Greger provided valuable feedback and established the foundation for discussing the clinical application of the research findings.

The engagement with the IDDEAS team and Dr Greger has been an important aspect of the project, as their medical expertise has enabled them to confirm, provide commentary, and interpret the processes and results. However, due to the multitude of decisions made throughout the project, it is inevitable that not every decision has undergone detailed evaluation by professionals. For instance, the manually added phenotype mappings have not been extensively validated by a professional but rather evaluated at a high-level overview. Ideally, someone with medical expertise should have carried out this manual work. However, due to the time-consuming nature of the mapping process and the sporadic addition of new mappings as deficiencies were identified, it proved challenging to enlist professionals from the field to conduct the mappings. Furthermore, technical knowledge was required to identify missing phenotypes. Although the contact with professionals has been adequate for validating the process and choices, a more thorough validation could have potentially uncovered unknown errors and provided feedback for adjustments to enhance clinical meaningfulness. Nevertheless, considering that the clustering results primarily provide an overview rather than extensive details, it is believed that the choices made in this research have effectively preserved the overarching patterns and insights.

10.5 Experimental Limitations

This section presents and discusses the experimental limitations concerning the project's framework and the overall research process. The limitations relate to the constraints and boundaries within which the project was conducted. By acknowledging and evaluating these limitations, an understanding of the potential constraints and implications on the experimental outcomes can be attained.

10.5.1 Data Basis

For the analysis, data from 16,202 episodes where patients received diagnoses were initially extracted. However, episodes that consisted solely of a single unique diagnosis without any other medications or co-occurring diagnoses were subsequently removed from the dataset. As a result, the dataset was reduced to 8,499 episodes for further analysis. Although having a larger volume of data is generally advantageous and could potentially lead to the identification of new patterns and subgroups, the current number of episodes should still serve as a solid foundation for characterising general subgroups within the cohort. This number of episodes is believed to adequately represent the practices and patients encountered within CAMHS.

The initial dataset of 16,202 episodes was merged with their corresponding prescriptions. Following the data cleaning process, 7,130 prescriptions were identified, associated with 4,459 episodes. Thus, approximately 27% of the episodes were found to have corresponding prescriptions. It is important to note that 2,435 prescriptions were removed during the cleaning stage due to missing information concerning both the trade name and the ATC code. These prescriptions were excluded from the dataset as it was challenging to determine the specific drug they represented without the necessary information. The translated trade names of these excluded prescriptions were labelled as *Other*, *Not relevant*, and *Substance not specified*. Consequently, it became difficult to ascertain the impact of these removed prescriptions on the results, particularly regarding whether they represent frequently occurring medications or a mixture of less significant medications.

These 8,499 records were analysed to identify patterns across the entire dataset based on the user-defined threshold values. This resulted in 13 patterns present in 3,240 of the episodes. Only these episodes were subject to the clustering process. One advantage of excluding episodes without frequent patterns is that the resulting clustering outcome provides clearer insights into the main subgroups within the dataset while minimising noise. Additionally, records that lack a clear natural cluster are excluded, enhancing the overall quality of the clustering results. However, a drawback of this approach is that it may exclude records that could yield more refined clustering outcomes, thereby characterising smaller subgroups within the dataset. Nevertheless, it is possible to achieve a similar outcome by lowering the threshold values, which would result in more frequent patterns and possibly include more records. However, as discussed in Section 8.2, feedback from clinicians indicated that such additional patterns provided excessive detail, thus supporting the decision to maintain the existing threshold values.

In summary, the data utilised in the analysis provides substantial information,

although including more data would always be desirable. It is important to acknowledge that certain records had to be excluded during the cleaning process due to missing fields of information. It is also worth noting that 62% of the data were not included in the clustering process, which may have led to a loss of certain nuances and detailed information within the clustering results. However, this approach effectively removes outliers and noisy data, yielding a comprehensive overview of the main patterns and subgroups of the dataset.

10.5.2 Time and Resources

This section discusses the limitations arising from time and resource constraints within the research process. It elaborates on how these limitations have influenced the outcomes and the choices made during the research. The impact of time constraints on the research outcomes is elaborated upon, highlighting how specific choices were prioritised and trade-offs were made given the available time. Additionally, since the project is interdisciplinary, a discussion is provided on how the reliance on input from medical professionals and their availability has affected the project.

The assigned timeframe for the project was 20 weeks, encompassing the definition, execution, conclusion, and documentation of the research. Given the extensive and time-consuming nature of the research process, it became necessary to prioritise tasks effectively. The main focus was completing essential tasks and generating interesting and applicable results. Furthermore, considerable attention was given to presenting and visualising the findings clearly and comprehensibly, ensuring that individuals without technical expertise could understand them. As a result of these priorities, certain tasks were given higher precedence while others were deprioritised. For instance, conducting multiple experiments, further comparison of different methods, and undertaking a more thorough clinical validation of manually added phenotypes were down-prioritised. Instead, emphasis was placed on delivering meaningful results within the available time frame. Moreover, the delay in changing the research scope and the delayed data access, which was not provided until mid-January, imposed additional constraints on the available time. The acquaintance with the dataset and its associated tools had to be conducted during the master's thesis rather than during the fall specialisation project, further impacting the available time for conducting the research. These factors collectively influenced the prioritisation of tasks and the allocation of time, ensuring that the research could progress within the given timeframe while emphasising the production of valuable and understandable outcomes.

The regular contact and feedback received from clinicians have played a vital role in ensuring the clinical accuracy and interpretability of the data, as well as

facilitating discussions on clinical results. This collaborative aspect of the project has been both helpful and necessary in achieving the research goal. However, it is important to acknowledge that consulting clinicians can be time-consuming, and their availability may not always align with the desired timeline of the project. Due to these factors, there have been instances where waiting for clinical input has resulted in delays in the experimental work. Consequently, to utilise the available time effectively, other tasks, such as documentation, have been prioritised during these waiting periods. While this approach has allowed progress on various fronts, it has also led to some experiments and tasks being deprioritised. The specific tasks affected by these constraints are further elaborated in Section 11.3, providing insights into potential avenues for future research and development.

10.5.3 Problem Definition

In addition to the limitations above, it is worth noting that the project's initial scope was broad and required considerable effort to define a clinically meaningful and focused problem area. Significant time and resources were dedicated during the fall specialisation project and throughout the current semester to carefully delineate the research problem and identify an area of focus that were both clinically relevant and appropriately bounded. This decision-making process was made based on a thorough understanding of the available data, feedback from clinicians, literature reviews and the aim of exploring an area that holds interest and value. By investing time and effort into refining the scope, the research project aimed to ensure that the investigation would yield meaningful insights and contribute to the field.

10.6 Experimental Aims

This section assesses the extent to which the experimental objectives outlined in Section 7.1.1 have been accomplished.

The first experimental aim focuses on the identification of patterns of co-occurring diagnoses and medications in episodes within CAMHS. This aim has been accomplished, and the resulting patterns are presented in Table 8.5. The observed patterns encompass instances where one or more ATC codes accompany disorders and co-occurring disorders. Analysis of the patterns reveals that ADHD frequently co-occurs with other disorders and that patients with ADHD often receive pharmaceutical treatment.

The second aim of this study entails the characterisation of patient subgroups by the co-occurrence of disorders, medication, and demographic attributes. This

objective has been partially achieved by the clustering experiment, which successfully characterises distinct patient subgroups. However, it is important to acknowledge that approximately 62% of the records were not included in any cluster due to the absence of one of the identified patterns. Thus, while the aim can be considered accomplished and patient subgroups have been characterised, it is worth noting that alternative approaches might reveal additional subgroups.

The third aim addresses the feasibility of clustering as a tool to characterise patient subgroups of co-occurring diagnoses and medications. The evaluation of the clinical feasibility and its subsequent discussion were presented in Chapter 9. The assessment of technical feasibility show that applying clustering to clinical diagnostic and medication data is indeed possible, and the results successfully characterise subgroups within the dataset. As described, demographic information seems to affect the clustering results strongly. However, as MASPC first detects patterns and removes noisy records, demographic information was expected to play a significant role when it constitutes five out of eighteen features. In summary, clustering can be seen as a valuable tool to explore clinical data and characterise the main subgroups in the dataset.

The final experimental aim seeks to investigate whether the identified patterns and subgroups can characterise phenomena already established within the context of CAMHS. This analysis and its implications are thoroughly discussed in Chapter 9. Feedback from senior physician Hanne Klæboe Greger affirms that the characterised subgroups of patient demographics, diagnoses and medications aptly characterise patient cohorts within CAMHS.

To summarise, the technical aspect of the experimental aims have been partially achieved. The implementation of MASPC has facilitated the detection of frequently occurring patterns of diagnoses and medications within the records. These identified patterns have subsequently enabled the clustering of records, leading to the characterisation of patient subgroups within the dataset. However, it is important to acknowledge that records lacking frequently occurring patterns were not included in the clustering process, potentially leading to the loss of certain aspects of the data. The clinical validation of the results shows that the identified patient subgroups indeed characterise phenomena in CAMHS and capture patient subgroups of diagnoses and medications in CAMHS.

Chapter 11

Conclusion and Future Work

This chapter aims to conclude the efforts documented in this report by addressing the thesis' goal and research questions and summarise the methodology, experimental design, results, evaluation and discussion. It also addresses the research's contribution to the field. As a final note, potential areas for future research are highlighted to inspire further exploration of the field and methodology.

11.1 Conclusion

In this study, EHRs of patients who received at least one diagnosis in CAMHS were analysed and clustered. The goal of this master's thesis was to analyse co-occurring morbidities and medications of CAMHS patients, with a particular focus on characterising patient profiles and subgroups through cluster analysis. This goal was accomplished by employing the MASPC clustering algorithm on EHR data gathered from the CAMHS clinic at St. Olavs Hospital in Trondheim. The collected data encompassed demographic details, along with diagnostic and medication-related information corresponding to CAMHS episodes of care.

The first research question explore how clustering can be used to analyse diagnoses and medications. As discussed in Section 6.2, successful clustering requires an algorithm that can handle diverse records with a spectrum of diagnoses and medications. These could range from patients with a single diagnosis to those harbouring multiple diagnoses and several medications. The diagnoses and medications were compiled in a list format for each record, enabling the MASPC algorithm to identify frequently occurring patterns. The presence of these iden-

tified patterns was then used in a binary representation to cluster the records. The application of this algorithm yielded 13 unique patterns of co-occurring diagnoses and medications. These patterns subsequently served as a foundation for the arrangement of records into 31 distinct clusters.

Before initiating the clustering experiment, a thorough data selection, extraction, and cleaning process was conducted. A significant part of this process involved identifying which data would be relevant and determining cleaning and preprocessing steps that could boost the performance of the MASPC algorithm on the selected dataset. As a result, patient ages were grouped into distinct age groups, and ICD-10 diagnosis codes were mapped to phenotypes. This mapping reduced the number of unique diagnoses by grouping similar diagnoses. Furthermore, the first four characters of the ATC code of each prescribed medication were utilised to identify and group similar medications. These approaches not only improved the data quality but also helped characterise larger subgroups by focusing on broader patterns rather than granular details during the pattern recognition and clustering analysis.

The second research question investigates how cluster analysis results can characterise patient subgroups according to diagnoses and medications. The clusters provided valuable insights into patient subgroups within the CAMHS population, tied not only to demographic details but also to the patient's diagnoses and medications. These characteristics affirm the potential of clustering techniques in characterising patient subgroups relative to diagnoses and medications and were presented in Chapter 8.

Another part of the project goal was to interpret the results in the context of the Norwegian CAMHS. This process was facilitated by sharing the results and associated statistics with clinicians and soliciting their insights on how these findings relate to their experience within CAMHS. Most of the findings aligned with expectations, including the dominance of female patients in clusters characterised by depression and anxiety disorders. Another expected finding was the earlier diagnosis of ADHD in boys, who are more frequently diagnosed with co-occurring conduct or learning disorders. Several factors could explain this outcome. Firstly, more boys enter CAMHS and receive ADHD diagnoses. Secondly, boys tend to be diagnosed with ADHD at younger ages. Lastly, the prevalence of learning and conduct disorders often increases at younger ages, and the longer a patient has an ADHD diagnosis, the higher the likelihood of additional co-occurring diagnoses.

Nonetheless, certain aspects of the results were surprising. For instance, nearly all the patients in the clusters with an ADHD diagnosis, receiving psychostimulants, and having a comorbid diagnosis of tics and stuttering were boys. While

it was anticipated that boys would form the majority of this group, the fact that they constituted almost 94% of this particular cohort was unexpected to the clinician. Another interesting pattern emerged where some clusters of co-occurring disorders included psychostimulant treatment while others did not, indicating variability in treatment approaches for various diagnostic profiles.

In summary, this study examined how clustering can be used in analysing diagnoses and medications of EHRs of patients in CAMHS. Furthermore, it investigated the potential of these results in characterising patient subgroups within the context of CAMHS. The application of the MASPC algorithm was instrumental in clustering the records into definable subgroups. These subgroups revealed patterns across diagnoses, medications, and demographic traits, characterising both predictable and unexpected subgroups. The research ultimately underscores the feasibility of applying clustering techniques in characterising patient subgroups within CAMHS.

11.2 Contributions

This section highlights the contributions of this research.

This clustering experiment is an important step in understanding co-occurring morbidities and medications within the dataset, thus enriching the knowledge base of the IDDEAS project related to the data at hand. This research is the first to explore this part of the database, characterising various concepts and phenomena in the dataset. By investigating the relationships between different morbidities and pharmaceutical interventions, the research contributes to the understanding of the dataset and lays a foundation for future research within the framework of the IDDEAS project.

Furthermore, the research identifies and characterises clinical phenomena relevant to CAMHS, particularly as they pertain to the practice at the St. Olavs clinic in Trondheim. Moreover, it sheds light on the relationships between co-occurring diagnoses and medication use among CAMHS patients, thus enhancing the comprehension of the complex patient profiles within CAMHS.

As the first study investigating this aspect of the database, the identification of potential future work represents a contribution, elaborated further in Section 11.3. This research opens numerous opportunities for more in-depth exploration of this data, thereby paving the way for future studies. The methods used for data cleaning and preprocessing, along with the implemented methodology, serve as a framework that can guide improvements and inspire further research in this field.

11.3 Future Work

This section explores areas of future work that have emerged during the experiments. These potential areas of future work remain unexplored within this project primarily due to the time constraints and the fact that some topics fell outside the established scope, goals, and research questions defined for this project. However, identifying these areas serves as a valuable starting point for future research and expands the possibilities for further exploration beyond this project's scope. This section presents areas of future work specifically related to the experimental dataset and the scope of this research, as well as areas of future work that encompass the broader dataset available to the IDDEAS project.

Firstly, this section examines areas of future work pertaining to the research conducted within this project. From a technical standpoint, exploring and applying alternative algorithms to the experimental dataset of this research would be valuable. By leveraging different algorithms, each with different strengths and capabilities, additional patient subgroups and intriguing patterns within the cohort may be uncovered. Moreover, as pointed out in Section 8.2, certain thresholds were determined by clinical insights rather than purely validity metrics. A different study, making decisions based solely on scientific metrics, could reveal interesting results. In addition, as deliberated in Section 10.3.2, an increase in the number of clustered episodes, possibly achieved through the mining of frequent itemsets as an alternative to MFAs, could potentially unveil and characterise new patient subgroups. This augmentation could also introduce an interesting feature to the MASPC algorithm.

Furthermore, a comprehensive examination of this research's phenotype mappings and cleaning process merits attention. The relevance and accuracy of the identified patterns may be enhanced by thoroughly evaluating the effectiveness of the existing mappings and considering adjustments to ensure optimal alignment with the specific context of CAMHS. This evaluative process, conducted in close collaboration with clinicians and domain experts, can result in refined groupings that better capture the intricacies and nuances of the diagnoses observed in CAMHS.

In relation to the experimental dataset, it could be interesting to incorporate the psychosocial situations categorised under Axis 5 and the CGAS scores in Axis 6. This inclusion would provide a more comprehensive portrayal of the patient's diagnostic condition and overall functioning. Moreover, incorporating additional demographic information, such as parental situation and care arrangements, may unveil new patterns and correlations. Furthermore, including data on the number of appointments for each episode and the duration of episodes could potentially

characterise novel subgroups within the dataset.

When presenting the experimental findings to the clinician, it was suggested that conducting additional analyses on the demographic information of the subgroups characterised through cluster analysis would be valuable. Furthermore, exploring all the diagnoses and prescriptions within each subgroup, considering that the records in this experiment are clustered solely on diagnoses and medication present in the observed patterns, could provide more in-depth insights into each subgroup. Moreover, a more comprehensive analysis of the unclustered records is recommended to uncover other meaningful patterns and associations.

Secondly, areas of future work related to the potential research avenues associated with the data encapsulated in the IDDEAS dataset is presented. It is essential to recognise and emphasise the potential inherent in the extensive dataset available to IDDEAS. This dataset, comprising authentic patient data, presents a valuable opportunity to extract meaningful insights and gather statistical information about the CAMHS population, CAMHS practices, diagnostic patterns, and medication-related trends. However, it is crucial to approach the handling of this dataset with care, considering its real-world nature. Extensive data cleaning and preprocessing efforts must be considered to ensure that the outcomes obtained are reliable, representative, and of high quality. While these steps may require significant time and effort, they are essential for obtaining accurate and meaningful results from the dataset. Some possible areas of research within the IDDEAS database used are:

- Conduct a more extensive analysis of medications in CAMHS, including additional data on prescription type, dosage, and related regulations. It would be interesting to examine the order of prescriptions in patient trajectories and identify potential patterns. This analysis can provide insights into treatment strategies, medication selection, and areas for improvement in patient care.

- Examine the sequencing of diagnoses given to patients in CAMHS. Specifically, whether certain diagnoses are typically assigned during the first appointment or at later appointments can yield intriguing insights. This analysis can provide valuable information about CAMHS practices and the standard diagnostic protocols associated with each diagnosis and combination of diagnoses. Understanding the patterns and timing of diagnosis assignments can contribute to a better understanding of the diagnostic process and shed light on the usual practices within CAMHS.

- As discussed in Section 6.1, the IDDEAS team has access to text records of journal notes, which, although requiring proper anonymisation before

being added to the HUNT Cloud laboratory, present a wealth of research opportunities. Once these records are added, various NLP techniques and supervised ML approaches can be applied to automatically identify and classify different types of journal notes, patient care details, diagnostics, and treatments. In addition to the focus of drug-related treatments in the current research, this approach enables the exploration of other kinds of treatment, such as psychotherapy, family therapy sessions, and parental training documented in journal notes. This addition can provide a more comprehensive understanding of the treatment processes within CAMHS, expanding the scope of analysis and enhancing insights into diverse CAMHS interventions.

It is important to acknowledge that mental health disorders in children and adolescents are an increasing problem. With the rising prevalence of these issues, gaining a deeper understanding of how treatment planning and implementation can be optimised to enhance efficiency is crucial. By identifying strategies that can save clinicians time, they can allocate resources to assist the growing number of children and adolescents who could benefit from CAMHS support. Additionally, advancements in knowledge regarding mental health disorders can enhance the quality of treatment for patients and subsequently improve their overall quality of life. Continuous research in this field is essential to drive progress and address the evolving needs of this vulnerable population.

Bibliography

CADDRA (2020). *Canadian ADHD Resource Alliance: Canadian ADHD Practice Guidelines*. CADDRA, Toronto ON, 4.1 edition.

Caliñski, T. and Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods*, 3:1–27. doi:https://doi.org/10.1080/03610927408827101.

Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. (1987). A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation. *Journal of Chronic Diseases*, 40(5):373–383. doi:https://doi.org/10.1016/0021-9681(87)90171-8.

Coghill, D., Banaschewski, T., Cortese, S., Asherson, P., Brandeis, D., Buitelaar, J., Daley, D., Danckaerts, M., Dittmann, R. W., Doepfner, M., Ferrin, M., Hollis, C., Holtmann, M., Paramala, S., Sonuga-Barke, E., Soutullo, C., Steinhausen, H. C., Van der Oord, S., Wong, I. C., Zuddas, A., and Simonoff, E. (2021). The Management of ADHD in Children and Adolescents: Bringing Evidence to the Clinic: Perspective from the European ADHD Guidelines Group (EAGG). *European Child & Adolescent Psychiatry*, 10:1–25. doi:https://doi.org/10.1007/s00787-021-01871-x.

DBeaver (2023). Universal Database Tool. https://dbeaver.io/. Accessed 27.05.2023.

De nasjonal forskningsetiske komiteene (2014). Regionale komiteer for medisinsk og helsefagligforskningsetikk (REK). https://www.forskningsetikk.no/om-oss/komiteer-og-utvalg/rek/. Accessed 07.06.2023.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., Basford, M. A., Carrell, D. S., Peissig, P. L., Kho, A. N., Pacheco, J. A., Rasmussen, L. V., Crosslin, D. R., Crane, P. K., Pathak, J., Bielinski, S. J., Pendergrass, S. A.,

Xu, H., Hindorff, L. A., Li, R., Manolio, T. A., Chute, C. G., Chisholm, R. L., Larson, E. B., Jarvik, G. P., Brilliant, M. H., Mccarty, C. A., Kullo, I. J., Haines, J. L., Crawford, D. C., Masys, D. R., and Roden, D. M. (2013). Systematic Comparison of Phenome-Wide Association Study of Electronic Medical Record Rata and Genome-Wide Association Study Data. *Nature Biotechnology 2013 31:12*, 31(12):1102–1111. doi:https://doi.org/10.1038/nbt.2749.

Direktoratet for e-helse (2018). Retningslinjer for koding: Multiaksial klassifikasjon for barn og unge (PHBU). https://www.ehelse.no/kodeverk-og-terminologi/Multiaksial-klassifikasjon-i-psykisk-helsevern-for-barn-og-unge-(PHBU). Accessed 05.03.2023.

Direktoratet for e-helse (2023). ICD-10, Den internasjonale statistiske klassifikasjonen av sykdommer og beslektede helseproblemer. https://finnkode.ehelse.no/#icd10/0/0/0/2622930. Accessed 01.02.2023.

Fournier-Viger, P., Lin, J. C. W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., and Lam, H. T. (2016). The SPMF Open-Source Data Mining Library Version 2. In *Proc. 19th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2016) Part III*, volume 9853 LNCS, pages 36–40. Springer. doi:https://doi.org/10.1007/978-3-319-46131-1_8.

Frades, I. and Matthiesen, R. (2010). Overview on Techniques in Cluster Analysis. *Methods in Molecular Biology (Clifton, N.J.)*, 593:81–107. doi:https://doi.org/10.1007/978-1-60327-194-3_5.

Géron, A. (2019). *Hands-on Machine Learning With Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 3 edition.

Gouda, K. and Zaki, M. (2001). Efficiently Mining Maximal Frequent Itemsets. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 163–170. doi:https://doi.org/10.1109/ICDM.2001.989514.

Grahne, G. and Zhu, J. (2003). High Performance Mining of Maximal Frequent Itemsets.

Hansen, B. H., Oerbeck, B., Skirbekk, B., Éva Petrovski, B., and Kristensen, H. (2018). Neurodevelopmental Disorders: Prevalence and Comorbidity in Children Referred to Mental Health Services. *Nordic Journal of Psychiatry*, 72(4):285–291. doi:https://doi.org/10.1080/08039488.2018.1444087.

HUNT Cloud (2023a). HUNT Workbench. https://docs.hdc.ntnu.no/do-science/hunt-workbench/. Accessed 27.05.2023.

HUNT Cloud (2023b). Welcome to the HUNT Cloud Documentation. https://www.ntnu.edu/mh/huntcloud. Accessed 27.05.2023.

IDDEAS (2022a). Forskningsprojekt IDDEAS. https://www.ntnu.no/rkbu/iddeas#/view/about. Accessed 10.05.2023.

IDDEAS (2022b). IDDEAS - Individual Digital Decision Assist System. https://www.iddeas.no/. Accessed 10.05.2023.

Kashihara, J., Takebayashi, Y., Kunisato, Y., and Ito, M. (2021). Classifying Patients With Depressive and Anxiety Disorders According to Symptom Network Structures: A Gaussian Graphical Mixture Model-Based Clustering. *PLOS ONE*, 16(9):1–18. doi:https://doi.org/10.1371/journal.pone.0256902.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. doi:https://doi.org/10.48550/arXiv.1705.07874.

Malt, U. and Braut, G. S. (2022). ICD-10. https://sml.snl.no/ICD-10. Accessed 15.05.2023.

Morgenthaler, S. (2009). Exploratory Data Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):33–44. doi:https://doi.org/10.1002/wics.2.

Nidheesh, N., Nazeer, K. A., and Ameer, P. M. (2020). A Hierarchical Clustering Algorithm Based on Silhouette Index for Cancer Subtype Discovery from Genomic Data. *Neural Computing and Applications*, 32(15):11459–11476. doi:https://doi.org/10.1007/s00521-019-04636-5.

Nitin, R., Shaw, D. M., Rocha, D. B., Walters, Courtney E., J., Chabris, C. F., Camarata, S. M., Gordon, R. L., and Below, J. E. (2022). Association of Developmental Language Disorder With Comorbid Developmental Conditions Using Algorithmic Phenotyping. *JAMA Network Open*, 5(12):e2248060–e2248060. doi:https://doi.org/10.1001/jamanetworkopen.2022.48060.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20:53–65. doi:https://doi.org/10.1016/0377-0427(87)90125-7.

Slaby, I., Hain, H. S., Abrams, D., Mentch, F. D., Glessner, J. T., Sleiman, P. M., and Hakonarson, H. (2022). An Electronic Health Record (EHR) Phenotype Algorithm to Identify Patients With Attention Deficit Hyperactivity Disorders (ADHD) and Psychiatric Comorbidities. *Journal of Neurodevelopmental Disorders*, 14. doi:https://doi.org/10.1186/s11689-022-09447-9.

Solheim, F. S. (2022). Characterising Patients Referred on Suspicion of ADHD and Behavioral Difficulties: An Exploratory Cluster Analysis of Norwegian Electronic Health Records. Master's thesis, Norwegian University of Science and Technology, NTNU.

Sosial- og helsedirektoratet (1999). *ICD-10 Psykiske lidelser og atferdsforstyrrelser, kliniske beskrivelser og diagnostiske retningslinjer*. Gyldendal akademisk.

Stawicki, S. P., Kalra, S., Jones, C., Justiniano, C. F., Papadimos, T. J., Galwankar, S. C., Pappada, S. M., Feeney, J. J., and Evans, D. C. (2015). Comorbidity Polypharmacy Score and its Clinical Utility: A pragmatic Practitioner's Perspective. *Journal of Emergencies, Trauma, and Shock*, 8:224. doi:https://doi.org/10.4103/0974-2700.161658.

Swain, S., Coupland, C., Strauss, V., Mallen, C., Kuo, C., Sarmanova, A., Bierma-Zeinstra, S., Englund, M., Prieto-Alhambra, D., Doherty, M., and Zhang, W. (2022). Clustering of Comorbidities and Associated Outcomes in People With Osteoarthritis - A UK Clinical Practice Research Datalink study. *Osteoarthritis and Cartilage*, 30(5):702–713. doi:https://doi.org/10.1016/j.joca.2021.12.013.

Walters, C. E., Nitin, R., Margulis, K., Boorom, O., Gustavson, D. E., Bush, C. T., Davis, L. K., Below, J. E., Cox, N. J., Camarata, S. M., and Gordon, R. L. (2020). Automated Phenotyping Tool for Identifying Developmental Language Disorder Cases in Health Systems Data (APT-DLD): A New Research Algorithm for Deployment in Large-Scale Electronic Health Record Systems. *Journal of Speech, Language, and Hearing Research*, 63(9):3019–3035. doi:https://pubs.asha.org/doi/pdf/10.1044/2020_JSLHR-19-00397.

Wartelle, A., Mourad-Chehade, F., Yalaoui, F., Chrusciel, J., Laplanche, D., and Sanchez, S. (2021). Clustering of a Health Dataset Using Diagnosis Co-Occurrences. *Applied Sciences 2021*, 11(5):2373. doi:https://doi.org/10.3390/app11052373.

WHO Collaboration Centre for Drug Statistics Methodology (2022). ATC - Structure and Principles. https://www.whocc.no/atc/structure_and_principles/. Accessed 28.04.2023.

Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J., Theodoratou, E., and Wei, W.-Q. (2019). Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Medical Informatics*, 7(4):e14325. doi:https://doi.org/10.2196/14325.

Yang, K.-W. K., Paris, C. F., Gorman, K. T., Rattsev, I., Yoo, R. H., Chen, Y., Desman, J. M., Wei, T. Y., Greenstein, J. L., Taylor, C. O., and Ray, S. C. (2023). Factors Associated With Resistance to SARS-CoV-2 Infection Discovered Using Large-Scale Medical Record Data and Machine Learning. *PLOS ONE*, 18(2):1–14. doi:https://doi.org/10.1371/journal.pone.0278466.

Zaki, M. J. and Meira, W. (2020). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press, 2 edition.

Zhong, H., Loukides, G., and Gwadera, R. (2020). Clustering Datasets With Demographics and Diagnosis Codes. *Journal of Biomedical Informatics*, 102:103360. doi:https://doi.org/10.1016/j.jbi.2019.103360.

Zhong, H., Loukides, G., and Pissis, S. P. (2022). Clustering Demographics and Sequences of Diagnosis Codes. *IEEE Journal of Biomedical and Health Informatics*, 26(5):2351–2359. doi:https://doi.org/10.1109/JBHI.2021.3129461.

Appendices

Appendix A

Abbreviations

Table A.1: Abbreviations.

Abbreviation	**Explanation**
ADHD	Attention Deficit Hyperactivity Disorder
AI	Artificial Intelligence
APT-DLD	Automated Phenotyping Tool for Identifying Developmental Language Disorder
ASPJ	Average Sum of Pairwise Jaccard distance
ASPWE	Average Sum of Pairwise Weighted Edit distance
ASPLCS	Average Sum of Longest Common Subsequence
ATC	Anatomical Therapeutic Chemical
CAMHS	Child and Adolescent Mental Health Services
CCI	Charlson Comorbidity Index
CDSS	Clinical Decision Support System
CHOP	Children's Hospital of Philadelphia
CI	Calinski-Harabasz Index
CPS	Comorbidity Polypharmacy Score
DLD	Developmental Language Disorder
DDSCA	Demographics and Diagnosis Sequences Clustering Algorithm
DPIA	Data Protection Impact Assessment
DSM-5	Diagnostic and Statistical Manual of Mental Disorders 5
EHR	Electronic Health Record
FP	Frequent Pattern
GDPR	General Data Protection Regulation
HAC	Hierarchical Agglomerative Clustering
ICD	International Classification of Diseases
ICD-10	International Classification of Diseases, 10th Revision
IDDEAS	Individualised Digital DEcision Assist System
LCA	Latent Class Analysis
MAS	Maximal-frequent All-confident pattern Selection
MASPC	Maximal-frequent All-confident pattern Selection and Pattern-based Clustering
MFA	Maximal-Frequent All-confident itemset
MFI	Maximal Frequent Itemset
ML	Machine Learning
NDA	Non-Disclosure Agreement
NLP	Natural Language Processing
NTNU	Norwegian University of Science and Technology
PC	Pattern-based Clustering
PheWAS	Phenome-Wide Association Studies
PPV	Positive Predictive Value
REK	Regional Committees for Medical and Health Research Ethics
SI	Silhouette Index
WHO	World Health Organisation
WP	Work Package
WBS	Work Breakdown Structure

Appendix B

PSQL Queries for Data Extraction

B.0.1 PSQL Query Extracting Diagnoses and Demographics

```
select
  pasient.nr as patient,
  diagnose.sak as episode_id,
  case
    when pasient.kjonn = '1' then 'F'
    when pasient.kjonn = '2' then 'M'
    else '0'
  end as gender,
  diagnose.akse as axis,
  case
    when diagnose.diagnose = '00' and akse = 1 then 'F00'
    when diagnose.diagnose = '99' and akse = 1 then 'F99'
    when diagnose.diagnose = '5' and akse = 3 then 'F70'
    when diagnose.diagnose = '6' and akse = 3 then 'F71'
    when diagnose.diagnose = '7' and akse = 3 then 'F72'
    when diagnose.diagnose = '8' and akse = 3 then 'F73'
    when diagnose.diagnose = '9' and akse = 3 then 'F79'
    else diagnose.diagnose
  end as diagnosis,
  case
    when diagnose.dato is null then extract(year from age(diagnose.
    endrdato, pasient.fdt))
    else extract(year from age(diagnose.dato, pasient.fdt))
  end as age_patient
from
```

```
25   diagnose left join pasient on diagnose.pasient = pasient.nr
26 where
27   diagnose.diagnose not like '%Z%' and diagnose.diagnose not like '%
       R%' and
28   (
29   (diagnose.akse = 1 and diagnose.diagnose != '999' and diagnose.
       diagnose != '000' and diagnose.diagnose != '1000' and diagnose.
       diagnose != '1999') or
30   (diagnose.akse = 2 and diagnose.diagnose != '999' and diagnose.
       diagnose != '000' and diagnose.diagnose != '2000' and diagnose.
       diagnose != '2999') or
31   (diagnose.akse = 3 and diagnose.diagnose != '30' and diagnose.
       diagnose != '39' and diagnose.diagnose != '99'
32   and diagnose.diagnose != '3999' and diagnose.diagnose != '3000'
       and diagnose.diagnose != '1'
33   and diagnose.diagnose != '2' and diagnose.diagnose != '3' and
       diagnose.diagnose != '4'
34   ) or
35   (diagnose.akse = 4 and diagnose.diagnose not like '%99%' and
       diagnose.diagnose not like '%00%')
36   )
37 order by episode_id, age_patient;
```

Listing B.1: PostgreSQL query for extraction of diagnoses and demographics.

B.0.2 PSQL Query Extracting Prescriptions

```
select
  forordning.saknr as episode_id,
  forordning.forordning as regulation,
  resept.resepttype as prescription_type,
  preparat.handelsnavn as trade_name,
  preparat.atckode as atc_code,
  preparat.atcnavn as atc_name
from forordning
left join preparat on forordning.preparatid = preparat.id
left join resept on forordning.nr = resept.forordningnr
order by forordning.saknr;
```

Listing B.2: PostgreSQL query for extraction of prescriptions.

Appendix C

Data Cleaning and Preprocessing

```
# Imports
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import pickle as pkl
import seaborn as sns
from sklearn.preprocessing import MultiLabelBinarizer

phecodes_df = pd.read_csv(
    "/home/iascheft/workbench/phecodes/phecode_icd10.csv", decimal=','
    )
diagnoses_df = pd.read_csv(
    "/home/iascheft/workbench/data/diagnoses.csv", decimal=',')
prescriptions_df = pd.read_csv(
    "/home/iascheft/workbench/data/prescriptionsEachSak.csv")


# ----------------------------------------
# Diagnoses cleaning
# ----------------------------------------


# Remove punctuation from icd-codes to match data
for i in range(len(phecodes_df["ICD10"])):
    phecodes_df.at[i, "ICD10"] = phecodes_df["ICD10"][i].replace("."
    , "")


```

```
# Convert phecodes df to dictionary
phecodes_dict = dict([(i, [x, y, z]) for i, x, y, z in zip(
    phecodes_df['ICD10'],
                     phecodes_df['PheCode'], phecodes_df['Phenotype'
    ], phecodes_df['Excl. Phenotypes'])])

# Add missing codes to dictionary
phecodes_dict["B209"] = [
    071.1, 'HIV infection, symptomatic', 'Viral infection']

phecodes_dict["E90"] = [277, 'Other disorders of metabolism',
                        'Other/unspecified disorders of metabolism']


phecodes_dict["F068"] = [291.4, 'Specific nonpsychotic mental
    disorders due to brain damage',
                         'dementia and related cognitive disorders/
    symptoms']
phecodes_dict["F069"] = [291.4, 'Specific nonpsychotic mental
    disorders due to brain damage',
                         'dementia and related cognitive disorders/
    symptoms']

phecodes_dict["F18"] = [290.3, 'Other persistent mental disorders
    due to conditions classified elsewhere',
                        'dementia and related cognitive disorders/
    symptoms']
phecodes_dict["F180"] = [290.3, 'Other persistent mental disorders
    due to conditions classified elsewhere',
                         'dementia and related cognitive disorders/
    symptoms']

phecodes_dict["F28"] = [295.3, 'Psychosis', 'psychological disorders
    ']

phecodes_dict["F32"] = [
    296.22, 'Depressive disorder', 'psychological disorders']
phecodes_dict["F320"] = [
    296.22, 'Depressive disorder', 'psychological disorders']
phecodes_dict["F321"] = [
    296.22, 'Depressive disorder', 'psychological disorders']
phecodes_dict["F322"] = [
    296.22, 'Depressive disorder', 'psychological disorders']
phecodes_dict["F323"] = [
    296.22, 'Depressive disorder', 'psychological disorders']
phecodes_dict["F329"] = [
    296.22, 'Depressive disorder', 'psychological disorders']
phecodes_dict["F339"] = [
    296.22, 'Depressive disorder', 'psychological disorders']

phecodes_dict["F204"] = [
    296.22, 'Depressive disorder', 'psychological disorders']
```

```
68 phecodes_dict["F328"] = [
69     296.22, 'Depressive disorder', 'psychological disorders']
70 phecodes_dict["F330"] = [
71     296.22, 'Depressive disorder', 'psychological disorders']
72 phecodes_dict["F331"] = [
73     296.22, 'Depressive disorder', 'psychological disorders']
74 phecodes_dict["F332"] = [
75     296.22, 'Depressive disorder', 'psychological disorders']
76 phecodes_dict["F333"] = [
77     296.22, 'Depressive disorder', 'psychological disorders']
78 phecodes_dict["F334"] = [
79     296.22, 'Depressive disorder', 'psychological disorders']
80 phecodes_dict["F338"] = [
81     296.22, 'Depressive disorder', 'psychological disorders']
82
83 phecodes_dict["F403"] = [300.13, 'Phobia', 'psychological disorders'
       ]
84
85 phecodes_dict["F064"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
86 phecodes_dict["F40"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
87 phecodes_dict["F400"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
88 phecodes_dict["F401"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
89 phecodes_dict["F402"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
90 phecodes_dict["F408"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
91 phecodes_dict["F409"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
92 phecodes_dict["F410"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
93 phecodes_dict["F411"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
94 phecodes_dict["F412"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
95 phecodes_dict["F413"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
96 phecodes_dict["F418"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
97 phecodes_dict["F419"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
98 phecodes_dict["F606"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
99 phecodes_dict["F930"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
100 phecodes_dict["F931"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
101 phecodes_dict["F932"] = [300.1, 'Anxiety disorder', 'psychological
       disorders']
```

```
102
103 phecodes_dict["F500"] = [305.2, 'Eating disorder', 'psychological
        disorders']
104 phecodes_dict["F504"] = [305.2, 'Eating disorder', 'psychological
        disorders']
105 phecodes_dict["F518"] = [327, 'Sleep disorders', 'Sleep disorders']
106 phecodes_dict["F552"] = [
107     316, 'Substance addiction and disorders', 'Substance-related
        disorders']
108 phecodes_dict["F555"] = [
109     316, 'Substance addiction and disorders', 'Substance-related
        disorders']
110 phecodes_dict["F558"] = [
111     316, 'Substance addiction and disorders', 'Substance-related
        disorders']
112
113 phecodes_dict["F623"] = [
114     306, 'Other mental disorder', 'psychological disorders']
115 phecodes_dict["F638"] = [312.3, 'Impulse control disorder',
116                          'Developmental/behavioral disorders']
117 phecodes_dict["F629"] = [
118     306, 'Other mental disorder', 'psychological disorders']
119
120 phecodes_dict["F7"] = [315.3, 'Mental retardation',
121                        'Developmental/behavioral disorders']
122 phecodes_dict["F710"] = [315.3, 'Mental retardation',
123                          'Developmental/behavioral disorders']
124 phecodes_dict["F718"] = [315.3, 'Mental retardation',
125                          'Developmental/behavioral disorders']
126 phecodes_dict["F780"] = [315.3, 'Mental retardation',
127                          'Developmental/behavioral disorders']
128 phecodes_dict["F791"] = [315.3, 'Mental retardation',
129                          'Developmental/behavioral disorders']
130 phecodes_dict["F798"] = [315.3, 'Mental retardation',
131                          'Developmental/behavioral disorders']
132
133 phecodes_dict["F81"] = [315.1, 'Learning disorder',
134                         'Developmental/behavioral disorders']
135 phecodes_dict["F813"] = [315.1, 'Learning disorder',
136                          'Developmental/behavioral disorders']
137 phecodes_dict["F83"] = [315.1, 'Learning disorder',
138                         'Developmental/behavioral disorders']
139
140 phecodes_dict["F900"] = [313.1, 'Attention deficit hyperactivity
        disorder',
141                          'Developmental/behavioral disorders']
142 phecodes_dict["F908"] = [313.1, 'Attention deficit hyperactivity
        disorder',
143                          'Developmental/behavioral disorders']
144 phecodes_dict["F928"] = [312, 'Conduct disorders',
145                          'Developmental/behavioral disorders']
146 phecodes_dict["F933"] = [313, 'Pervasive developmental disorders',
```

```
147                          'Developmental/behavioral disorders']
148 phecodes_dict["F98"] = [313, 'Pervasive developmental disorders',
149                          'Developmental/behavioral disorders']
150 phecodes_dict["F982"] = [305.2, 'Eating disorder', 'psychological
    disorders']
151
152 phecodes_dict["G40."] = [345, 'Epilepsy, recurrent seizures,
    convulsions',
153                          'hereditary/degenerative nervous conditions
    ; other diseases of CNS']
154
155 phecodes_dict["H50"] = [
156     378.1, 'Strabismus (not specified as paralytic)', 'other eye
    disorders']
157 phecodes_dict["H549"] = [
158     367.9, 'Blindness and low vision', 'Blindness and low vision']
159 phecodes_dict["H91"] = [389, 'Hearing loss',
160                          'Hearing loss and related disorders']
161
162 phecodes_dict["J951"] = [
163     519.2, 'Respiratory complications', 'tracheostomy complications'
    ]
164
165 phecodes_dict["K523"] = [555.2, 'Ulcerative colitis',
166                          'noninfective gastrointestinal disorders']
167 phecodes_dict["K649"] = [455, 'Hemorrhoids',
168                          'diseases of veins and lymphatics']
169 phecodes_dict["K720"] = [
170     571.8, 'Liver abscess and sequelae of chronic liver disease', '
    Liver disease']
171 phecodes_dict["K858"] = [577.1, 'Acute pancreatitis', 'pancreatic
    disorders']
172
173 phecodes_dict["L65"] = [704.1, 'Alopecia', 'Diseases of hair and
    nails']
174
175 phecodes_dict["M090"] = [714.2, 'Juvenile rheumatoid arthritis',
176                          'Autoimmune arthritis and psoriasis']
177 phecodes_dict["M091"] = [714.2, 'Juvenile rheumatoid arthritis',
178                          'Autoimmune arthritis and psoriasis']
179 phecodes_dict["M248"] = [742.9, 'Other derangement of joint',
180                          'other non-traumatic joint disorders']
181 phecodes_dict["M609"] = [313.1, 'Myopathy',
182                          'Disorders of the peripheral nervous system
    ']
183 phecodes_dict["M726"] = [
184     727, 'Symptoms of the muscles', 'Other muscular symptoms']
185 phecodes_dict["M828"] = [743.1, 'Osteoporosis NOS',
186                          'osteopenia, osteoporosis, pathological
    fractures']
187
188 phecodes_dict["N038"] = [
```

```
189     580.14, 'Chronic glomerulonephritis, NOS', 'diseases of kidney
        and ureters']
190
191 phecodes_dict["O04"] = [
192     634, 'Miscarriage; stillbirth', 'complications of pregnancy']
193 phecodes_dict["O94"] = [
194     676, 'Other disorders of the breast associated with childbirth
        and disorders of lactation', 'COMPLICATIONS OF THE PUERPERIUM']
195
196 phecodes_dict["P043"] = [658, 'Maternal complication of pregnancy
        affecting fetus or newborn',
197                          'Maternal complication of pregnancy
        affecting fetus or newborn']
198 phecodes_dict["P044"] = [658, 'Maternal complication of pregnancy
        affecting fetus or newborn',
199                          'Maternal complication of pregnancy
        affecting fetus or newborn']
200 phecodes_dict["P070"] = [637, 'Short gestation; low birth weight;
        and fetal growth retardation',
201                          'miscarriage, early labor, hemorrhage']
202 phecodes_dict["P071"] = [637, 'Short gestation; low birth weight;
        and fetal growth retardation',
203                          'miscarriage, early labor, hemorrhage']
204
205 phecodes_dict["Q044"] = [752.2, 'Other specified congenital
        anomalies of nervous system',
206                          'congenital anomalies of nervous system,
        spine']
207 phecodes_dict["Q315"] = [748, 'Anomalies of respiratory system,
        congenital',
208                          'congenital anomalies of respiratory system
        , face and neck']
209 phecodes_dict["Q64"] = [751.2, 'Congenital anomalies of urinary
        system',
210                         'congenital anomalies of gi, urinary tract']
211 phecodes_dict["Q900"] = [758.1, 'Chromosomal anomalies',
212                          'All other congenital anomalies']
213 phecodes_dict["Q914"] = [758.1, 'Chromosomal anomalies',
214                          'All other congenital anomalies']
215
216 phecodes_dict["S327"] = [1009, 'Injury, NOS', None]
217 phecodes_dict["S361"] = [1008, 'Crushing or internal injury to
        organs', None]
218
219 phecodes_dict["T012"] = [
220     870, 'Open wounds of head; neck; and trunk', 'Open wound']
221 phecodes_dict["T141"] = [1009, 'Injury, NOS', None]
222 phecodes_dict["T142"] = [1009, 'Injury, NOS', None]
223 phecodes_dict["T4n"] = [969, 'Poisoning by psychotropic agents',
224                         'Poisoning By Drugs, Medicinal And
        Biological Substances']
225 phecodes_dict["T740"] = [1015, 'Effects of other external causes',
```

```
    None]
phecodes_dict["T748"] = [1015, 'Effects of other external causes',
    None]
phecodes_dict["T742"] = [1015, 'Effects of other external causes',
    None]

phecodes_dict["X4n"] = [981, 'Toxic effect of (non-ethyl) alcohol
    and petroleum and other solvents',
                        'Toxic effects of substances chiefly
    nonmedicinal as to source']

phecodes_dict["X6n"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X61"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X60"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X6nx"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X6n0"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X6n1"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X6n2"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X6n4"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X6n5"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X6n6"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X6n8"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']
phecodes_dict["X6n9"] = [
    297.2, 'Suicide or self-inflicted injury', 'psychological
    disorders']

phecodes_dict["Y912"] = [981, 'Toxic effect of (non-ethyl) alcohol
    and petroleum and other solvents',
                         'Toxic effects of substances chiefly
    nonmedicinal as to source']
```

```
phecodes_dict["Y913"] = [981, 'Toxic effect of (non-ethyl) alcohol and petroleum and other solvents',
                         'Toxic effects of substances chiefly nonmedicinal as to source']

pkl.dump(phecodes_dict, open(
    '/home/iascheft/workbench/clustering/maspc/pickleDump/phecodes_dict.pkl', 'wb'), protocol=4)


def replace_icd_with_phenotype(df):
    col = "diagnosis"
    for i in df.index:
        if len(df[col][i]) > 4:
            # Shorten codes that are too long
            df.at[i, col] = df[col][i][:4]
        if df[col][i] in phecodes_dict:
            # Replace ICD-code with phenotype
            icd_code = df.at[i, col]
            df.at[i, col] = phecodes_dict[icd_code][0]
        else:
            df.drop(index=i, inplace=True)


replace_icd_with_phenotype(diagnoses_df)

# Drop rows with null
diagnoses_df = diagnoses_df.dropna(subset=['episode_id'])
diagnoses_df = diagnoses_df.dropna(subset=['gender'])
diagnoses_df = diagnoses_df.dropna(subset=['diagnosis'])
diagnoses_df = diagnoses_df.dropna(subset=['age_patient'])

# Drop diagnoses with less than 3 occurrences
diagnoses_df = diagnoses_df.groupby("diagnosis").filter(lambda x: len(x) > 3)

# df containing list of diagnoses
diagnoses_list_df = diagnoses_df.drop(["patient", "axis"], axis=1)

diagnoses_list_df['diagnosis'] = diagnoses_list_df['diagnosis'].astype(str)
diagnoses_list_df.info()

diagnoses_list_df = diagnoses_list_df.groupby('episode_id').agg(
    {'diagnosis': lambda x: list(set(x)), 'gender': min, 'age_patient': min})

diagnoses_list_df["age_patient"] = diagnoses_list_df["age_patient"].astype(int)

# ----------------------------------------
# Demographics cleaning
```

```
# ----------------------------------------

# Add age group column
diagnoses_list_df.loc[diagnoses_list_df["age_patient"].between(0, 5), [
    "age_group"]] = "Preschooler"
diagnoses_list_df.loc[diagnoses_list_df["age_patient"].between(
    6, 11), ["age_group"]] = "MiddleChildhood"
diagnoses_list_df.loc[diagnoses_list_df["age_patient"].between(12, 18), [
    "age_group"]] = "Teenager"
diagnoses_list_df.loc[diagnoses_list_df["age_patient"].between(19, 60), [
    "age_group"]] = "Adult"

diagnoses_list_df = diagnoses_list_df[["diagnosis", "gender", "age_group"]]
diagnoses_list_df = diagnoses_list_df.rename(
    columns={"diagnosis": "diagnoses"})

# Remove dots in diagnoses
diagnoses_list_df["diagnoses"] = [str(x).replace(
    '.', '') for x in diagnoses_list_df["diagnoses"]]

# Drop records with no gender info
diagnoses_list_df = diagnoses_list_df.drop(
    diagnoses_list_df[diagnoses_list_df['gender'] == "0"].index)

# Drop adults, because not that many and CAMHS are mainly for children and adolescents
diagnoses_list_df = diagnoses_list_df.drop(
    diagnoses_list_df[diagnoses_list_df['age_group'] == "Adult"].index)

diagnoses_list_df = diagnoses_list_df.dropna(subset=['age_group'])

diagnoses_list_df.to_csv(
    "/home/iascheft/workbench/data/cleaned_data/diagnoses_listed.csv")

# ----------------------------------------
# Prescriptions cleaning
# ----------------------------------------


# Remove spacing in atc codes
prescriptions_df["atc_code"] = [str(x).replace(
    ' ', '') for x in prescriptions_df["atc_code"]]


def update_atc_code(df):
    df.atc_code = np.where(((df.atc_code == 'nan') & (
```

```
        df.trade_name.str.contains('Antiepileptika'))), 'N03A', df.
    atc_code)
    df.atc_code = np.where(((df.atc_code == 'nan') & (
        df.trade_name.str.contains('Antidepresiva'))), 'N06A', df.
    atc_code)
    df.atc_code = np.where(((df.atc_code == 'nan') & (
        df.trade_name == 'Concerta')), 'N06BA04', df.atc_code)
    df.atc_code = np.where(((df.atc_code == 'nan') & (
        df.trade_name == 'Dexidrine')), 'N06BA02', df.atc_code)
    df.atc_code = np.where(((df.atc_code == 'nan') & (
        df.trade_name == 'Melatonin')), 'N05CH01', df.atc_code)
    df.atc_code = np.where(((df.atc_code == 'nan') & (
        df.trade_name == 'Metamina')), 'N06BA02', df.atc_code)
    df.atc_code = np.where(((df.atc_code == 'nan') & (
        df.trade_name.str.contains('Nevroleptika'))), 'N05A', df.
    atc_code)
    df.atc_code = np.where(((df.atc_code == 'nan') & (
        df.trade_name.str.contains('Sentralstimulerende'))), 'N06BA'
    , df.atc_code)


update_atc_code(prescriptions_df)

# Cut atc codes at 4 letters
atc_df = prescriptions_df[["episode_id", "atc_code"]].copy()
atc_df["atc_code"] = [x[0:4] for x in atc_df["atc_code"]]

# Replace "O" with 0, because one instance is N06B and not N06B
atc_df["atc_code"] = [str(x).replace('O', '0') for x in atc_df["
    atc_code"]]

# Delete records that still have "nan" as ATC code
atc_df = atc_df.drop(atc_df[atc_df['atc_code'] == "nan"].index)

# Make a dictionary to be used for converting letters in atc codes
    to numbers.
# The dictionary is needed because FPMax and Apriori only accepts
    numbers
alphabet_dict = {}
for i in range(26):
    letter = chr(i + ord('A'))
    if i < 9:
        alphabet_dict[letter] = "0" + str(i + 1)
    else:
        alphabet_dict[letter] = str(i + 1)

pkl.dump(alphabet_dict, open(
    '/home/iascheft/workbench/clustering/maspc/pickleDump/
    alphabet_dict.pkl', 'wb'), protocol=4)

# Convert atc codes to numbers. Adding "999" in the beginning of all
     codes to be able to separate the codes from diagnosid codes
```

```
for i in atc_df.index:
    atc = atc_df["atc_code"][i]
    atc_string = "999" + alphabet_dict[atc[0]
                                       ] + atc[1:3] + alphabet_dict[
    atc[3]]
    atc_df.loc[i, "atc_code"] = atc_string

atc_list_df = atc_df.copy()
atc_list_df = atc_list_df.groupby('episode_id').agg(
    {'atc_code': lambda x: list(set(x))})

atc_list_df.to_csv("/home/iascheft/workbench/data/cleaned_data/
    atc_listed.csv")

# ----------------------------------------
# Merging the cleaned data into one df, making it ready for analysis
# ----------------------------------------

atc_df = pd.read_csv(
    "/home/iascheft/workbench/data/cleaned_data/atc_listed.csv")
diagnoses_df = pd.read_csv(
    "/home/iascheft/workbench/data/cleaned_data/diagnoses_listed.csv
    ")

atc_df.rename(columns={'atc_code': 'atc_codes'}, inplace=True)

df = pd.merge(diagnoses_df, atc_df, on="episode_id", how="left")

# Fill rows with nan with empty list
df["atc_codes"] = df["atc_codes"].replace(
    [np.inf, -np.inf], np.nan).fillna("0")
for i in df.index:
    if df["atc_codes"][i] == "0":
        df.at[i, "atc_codes"] = []

df = df[["episode_id", "gender", "age_group", "diagnoses", "
    atc_codes"]].copy()

# Merge diagnosis and atc columns
df['diag_atc'] = df.apply(lambda row: eval(
    str(row['diagnoses'])) + eval(str(row['atc_codes'])), axis=1)

df = df[["episode_id", "gender", "age_group", "diagnoses", "
    atc_codes"]].copy()

df.to_csv(
    "/home/iascheft/workbench/data/cleaned_data/
    diagnoses_atc_episode.csv", index=False)
```

Listing C.1: Cleaning of diagnoses and demographics and prescriptions in Pyhton.

Appendix D

EDA

```
# Imports
import matplotlib.pyplot as plt
import seaborn as sns


# ----------------------------------------
# Bar chart showing diagnosis count. If a diagnosis occur more than once in an episode it is only counted once.
# ----------------------------------------

plt.figure(figsize=(5, 10))

unique_diag_episode_df = diagnoses_df.groupby(['episode_id', 'diagnosis']).count().reset_index()
diag_counts = unique_diag_episode_df.diagnosis.value_counts().loc[lambda x: x > 200]

ax = sns.barplot(x=diag_counts, y=diag_counts.index, palette="YlGnBu")

ax.bar_label(ax.containers[0])
ax.set_xscale("log")

ax.set_xlabel("Number of Occurrences")
ax.set_ylabel("Diagnosis")

sns.despine(left=True, bottom=True)


plt.savefig('/home/iascheft/workbench/clustering/maspc/Diagrams/DatasetChapter/diagnosis_count_ICD-10_BARPLOT.pdf', bbox_inches=
```

```
    "tight")
plt.show()


# ----------------------------------------
# Bar chart showing ATC count. If an ATC code occur more than oce in
    an episode it is only counted once.

# ----------------------------------------

plt.figure(figsize=(5, 10))

unique_atc_episode_df = prescriptions_df.groupby(
    ['episode_id', 'atc_code']).count().reset_index()
atc_counts = unique_atc_episode_df.atc_code.value_counts(
).loc[lambda x: x > 20]

ax = sns.barplot(x=atc_counts, y=atc_counts.index, palette="YlGnBu")

ax.bar_label(ax.containers[0])
ax.set_xscale("log")

ax.set_xlabel("Number of Occurrences")
ax.set_ylabel("ATC code")

sns.despine(left=True, bottom=True)


plt.savefig('/home/iascheft/workbench/clustering/maspc/Diagrams/
    DatasetChapter/atc_count_notCut_BARPLOT.pdf', bbox_inches="tight
    ")
plt.show()


# ----------------------------------------
# Plot: gender distribution
# ----------------------------------------

plt.figure(figsize=(2, 4))

count = diagnoses_list_df.gender.value_counts()

ax = sns.barplot(x=count.index, y=count, palette="YlGnBu")

ax.bar_label(ax.containers[0])

ax.set_xlabel("Gender")
ax.set_ylabel("Number of episodes")

plt.savefig('/home/iascheft/workbench/clustering/maspc/Diagrams/
    DatasetChapter/gender_dataset_dist_BARPLOT.pdf', bbox_inches="
    tight")
```

```
75
76
77 # ----------------------------------------
78 # Plot: age distribution.
79 # ----------------------------------------
80
81 plt.figure(figsize=(10, 4))
82
83 count = diagnoses_list_df.age_patient.value_counts()
84 count = count.loc[lambda x: x.index > -1]
85 count = count.loc[lambda x: x.index < 19]
86
87
88 ax = sns.barplot(x=count.index, y=count, palette="YlGnBu")
89
90 ax.bar_label(ax.containers[0])
91
92 ax.set_xlabel("Age")
93 ax.set_ylabel("Number of episodes")
94
95 plt.savefig('/home/iascheft/workbench/clustering/maspc/Diagrams/
    DatasetChapter/age_dist_BARPLOT.pdf', bbox_inches="tight")
```

Listing D.1: Python code for EDA Plots. The dataframes used are the ones defined and cleaned in the previous appendix showing the cleaning process.

Appendix E

MASPC Implementation

```
1
2  # Code inspired from https://bitbucket.org/EHR_Clustering/maspc/src/
       master/
3
4  # Imports
5  import subprocess
6
7  import pandas as pd
8  from scipy.cluster.hierarchy import fcluster
9  from scipy.cluster.hierarchy import linkage
10 from sklearn import metrics
11 from sklearn.preprocessing import LabelBinarizer
12 from sklearn.preprocessing import OneHotEncoder
13
14 # Data input
15 df = pd.read_csv(
16     "/home/iascheft/workbench/data/cleaned_data/
       diagnoses_atc_episode.csv")
17
18
19 # global variables
20 clusteredData = pd.DataFrame()  # emptyDF. Clustered data will be
       added here
21 linkage_result = ""  # Variable to be used for drawing dendogram
22 mfas = []
23
24 # Extract demographic information
25 demographic = df[["episode_id", "gender", "age_group"]]
26
27 # Binarize and one-hot encode demographic input
28 lb = LabelBinarizer()
```

```
ohe = OneHotEncoder()

demographic = demographic.join(pd.DataFrame(lb.fit_transform(
    demographic['age_group']),
                                            columns=lb.classes_,
                                            index=demographic.index)
    )

# one-hot encode the 'gender' column and create new columns
one_hot = pd.get_dummies(df['gender'])
one_hot = one_hot.astype(int)
one_hot.columns = ['F', 'M']

# concatenate the original dataframe with the one-hot encoded
    columns
demographic = pd.concat([demographic, one_hot], axis=1)

demographic = demographic.drop(["gender", "age_group"], axis=1)

# Read diagnosisCodes.txt as input for diagnosis codes
diagnosis = open('diagnosesCodesAndAtc.txt', 'r')
diagnosisCodes = [line[:-2].split(' ') for line in diagnosis.
    readlines()]

# ----------------------------------------

# Define Apriori algorithm
# The Apriori algorithm used at here is built upon SPMF (http://www.
    philippe-fournier-viger.com/spmf/)
# Please download spmf.jar from its website before you run Apriori
    algorithm
# Reference of Apriori: https://en.wikipedia.org/wiki/
    Apriori_algorithm
# Input of Apriori is self._input = "***.txt", which includes each
    patients' diagnosis codes
# Output of Apriori is self._output = "***.txt"


class Apriori():

    def __init__(self):
        self._executable = "spmf.jar"
        self._input = "diagnosesCodesAndAtc.txt"
        self._output = "Apriori_output.txt"

    def run(self, min_supp):
        subprocess.call(["java", "-Xmx512m", "-jar", self.
    _executable,
                         "run", "Apriori", self._input, self._output,
     str(min_supp)])

    def encode_input(self, data):
```

```
71          pass
72
73      def decode_output(self):
74          # read
75          lines = []
76          try:
77              with open(self._output, "rU") as f:
78                  lines = f.readlines()
79          except:
80              print("read_output error")
81
82          # decode
83          patterns = []
84          for line in lines:
85              line = line.strip()
86              patterns.append(line.split())
87
88          return patterns
89
90  # ----------------------------------------
91
92  # Define FPMax algorithm
93  # FPMax Algorithm can return Frequent Maximal Itemsets
94  # Reference of FPMax Algorithm: Grahne, G., & Zhu, J. (2003, May).
        High performance mining of maximal frequent itemsets.
95  # Input of FPMax is self._input = "***.txt", which includes each
        patients' diagnosis codes
96  # Output of FPMax is self._output = "***.txt"
97
98
99  class FPMax():
100
101     def __init__(self):
102         self._executable = "spmf.jar"
103         self._input = "diagnosesCodesAndAtc.txt"
104         self._output = "FPMax_output.txt"
105
106     def run(self, min_supp):
107         subprocess.call(["java", "-Xmx512m", "-jar", self.
        _executable,
108                          "run", "FPMax", self._input, self._output,
        str(min_supp)])
109
110     def encode_input(self, data):
111         pass
112
113     def decode_output(self):
114         # read
115         lines = []
116         try:
117             with open(self._output, "rU") as f:
118                 lines = f.readlines()
```

```
119         except:
120             print("read_output error")
121 
122         # decode
123         patterns = []
124         for line in lines:
125             line = line.strip()
126             patterns.append(line.split())
127 
128         return patterns
129 
130 # ----------------------------------------
131 
132 
133 def allconfidence(list_1, list_max):
134     # Compute All_confidence of an itemset
135     b = []
136     for i in list_max[:len(list_max)-2]:
137         for j in list_1:
138             if i == j[0]:
139                 b.append(int(j[2]))
140     return int(list_max[-1])/max(b)
141 
142 
143 def get_all_allconfidence(list_1, list_all_max, threshhold):
144     # Input a list of MFIs
145     # Return MFIs whose All_confidence is above minAc
146     all_max = []
147     for i in list_all_max:
148         if allconfidence(list_1, i) >= threshhold:
149             i[-1] = allconfidence(list_1, i)
150             all_max.append(i)
151     return all_max
152 
153 # ----------------------------------------
154 
155 # MASPC algorithm
156 
157 
158 class MASPC():
159 
160     def MAS(self, minSup, minAc, minOv):
161         # Run FPMax to get MFI
162         fpmax = FPMax()
163         fpmax.encode_input([])
164         fpmax.run(minSup)
165 
166         # Running Apriori is a preparatory step for getting MFA
167         apriori = Apriori()
168         apriori.encode_input([])
169         apriori.run(minSup)
170 
```

```
171         list_1 = []
172         for i in apriori.decode_output():
173             if len(i) == 3:
174                 list_1.append(i)
175         # Get MFA
176         all_con = get_all_allconfidence(list_1, fpmax.decode_output
    (), minAc)
177         all_con.sort(key=lambda x: x[-1], reverse=True)
178 
179         all_con_withoutSUP = []
180         for i in all_con:
181             all_con_withoutSUP.append([x for x in i[:len(i)-2]])
182 
183         all_con_target = []
184         for i in all_con_withoutSUP:
185             flag = 0
186             for j in all_con_target:
187                 if (set(i) & set(j) != set()):
188                     number = 0
189                     for k in diagnosisCodes:
190                         if ((set(k) & (set(i) | set(j))) == (set(i)
    | set(j))):
191                             number = number + 1
192                     if number <= minOv:
193                         flag = 1
194                         break
195             if flag == 0:
196                 all_con_target.append(i)
197 
198         all_con_target_without1 = []
199 
200         for i in all_con_target:
201             if len(i) != 1:
202                 all_con_target_without1.append(i)
203 
204         # save MFAs
205         self.MFAs = all_con_target_without1
206         global mfas
207         mfas = self.MFAs
208 
209 
210 class MASPC(MASPC):
211     def PC(self, k, method, metric):
212         w, h = len(self.MFAs), len(diagnosisCodes)
213         all_con_tables_without1 = [[0 for x in range(w)] for y in
    range(h)]
214 
215         # project maximum set of independent frequnet patterns
216         for i, j in enumerate(diagnosisCodes):
217             temp = set(j)
218 
219             l = len(temp)
```

```
220             for a, b in enumerate(self.MFAs):
221                 while (set(b) <= temp):
222                     temp = temp.difference(set(b))
223 
224                     all_con_tables_without1[i][a] += 1
225 
226         # build a dataframe
227         all_con_part_2_without1 = pd.DataFrame(
    all_con_tables_without1, columns=[
228                                                str(sublist) for
    sublist in self.MFAs])
229         all_con_final_t_without1 = demographic.join(
    all_con_part_2_without1)
230 
231         # delete the data that not be subscribed
232         all_con_delete_without1 = [sum(i) for i in
    all_con_tables_without1]
233         all_con_delete_idex_without1 = [
234             i for i, e in enumerate(all_con_delete_without1) if e ==
     0]
235 
236         all_con_final_t_without1.drop(
237             all_con_delete_idex_without1, inplace=True)
238 
239         self.binaryData = all_con_final_t_without1
240 
241         # do clustering
242         all_con_cos_ave_without1 = linkage(all_con_final_t_without1.
    drop(
243             "episode_id", axis=1).values, method, metric)
244         self.ClusterResult = fcluster(
245             all_con_cos_ave_without1, k, criterion='maxclust')
246 
247         global linkage_result
248         linkage_result = all_con_cos_ave_without1
249 
250         global clusteredData
251         clusteredData = all_con_final_t_without1
252 
253 # ----------------------------------------
254 
255 # Run MASPC
256 # Input parameters: minSup=0.03, minAc=0.03, minOv=10, k=31
257 # method='average' and metric='cosine' are parameters for
    agglomerative average-linkage hierarchical clustering
258 
259 
260 k = 31
261 
262 if __name__ == "__main__":
263     maspc = MASPC()
264     maspc.MAS(minSup=0.03, minAc=0.03, minOv=10)
```

```
265     maspc.PC(k=k, method='average', metric='cosine')
266
267 # Add label to binary representation
268 maspc.binaryData['label'] = maspc.ClusterResult
269
270 # ----------------------------------------
271
272 # Preparation for SI and CI calculation
273
274 # Get all unique diagnosis codes and build a binary representation
    for evaluation
275 allUniqueCodes = []
276
277 for i in diagnosisCodes:
278     for j in i:
279         allUniqueCodes.append(j)
280 allUniqueCodes = list(set(allUniqueCodes))
281 new_list = [allUniqueCodes[i:i+1] for i in range(0, len(
    allUniqueCodes), 1)]
282
283
284 w, h = len(allUniqueCodes), len(diagnosisCodes)
285 atables = [[0 for x in range(w)] for y in range(h)]
286
287 # project maximum set of independent frequnet patterns
288 for i, j in enumerate(diagnosisCodes):
289     temp = set(j)
290     for a, b in enumerate(new_list):
291         while (set(b) <= temp):
292             temp = temp.difference(set(b))
293             atables[i][a] += 1
294
295 diga_codes = pd.DataFrame(
296     atables, columns=[str(sublist) for sublist in allUniqueCodes])
297
298
299 # Keep only demographic info of clustered records
300 demographicClusteredData = clusteredData.iloc[:, 0:6]
301
302 temp_df = pd.merge(df, diga_codes, left_index=True, right_index=True
    )
303 temp_df = temp_df.drop(
304     ["episode_id", "gender", "age_group", "diagnoses", "atc_codes"],
     axis=1)
305
306 # eval_df for calculation of SI and CI. Info about only the records
    that are clustered
307 eval_df = pd.merge(demographicClusteredData, temp_df,
308                    left_index=True, right_index=True)
309
310 # Binary representation for evaluation
311 testdata = pd.concat([demographic, diga_codes], axis=1, sort=False)
```

```
312 testdata.head()
313
314 # ----------------------------------------
315
316 # SI and CI calculation
317
318 print('CI: ', metrics.calinski_harabasz_score(eval_df.drop(
319     "episode_id", axis=1).values, maspc.ClusterResult.tolist()))
320 print('SI: ', metrics.silhouette_score(eval_df.drop("episode_id",
321       axis=1).values, maspc.ClusterResult.tolist(), metric='cosine')
    )
```

Listing E.1: Python implementation of MASPC and calculation of SI and CI evaluation metrics.

Appendix F

Clustering Result Plots

```
# Imports
import pickle as pkl
import re

import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
import shap
from lightgbm import LGBMClassifier


heat_map_AGEGROUP = np.zeros([3, k])
heat_map_GENDER = np.zeros([2, k])
heat_map_MFAs = np.zeros([len(mfas), k])

barplot_GENDER = np.zeros([2, k])

num_episodes_each_cluster_list = []

# Calculate values for plots
for i in range(0, k):
    group_df = maspc.binaryData.groupby(['label']).get_group(i+1)
    num_episodes_in_group = len(
        maspc.binaryData.groupby(['label']).get_group(i + 1))
    num_episodes_each_cluster_list.append(num_episodes_in_group)

    # Calculate gender distribution in number
    barplot_GENDER[0][i] = len(group_df[group_df["F"] == 1])
    barplot_GENDER[1][i] = len(group_df[group_df["M"] == 1])

```

```
    # Calculating the percentage of each gender in each group, 0 =
    Female, 1 = Male
    heat_map_GENDER[0][i] = len(
        group_df[group_df["F"] == 1])/num_episodes_in_group
    heat_map_GENDER[1][i] = len(
        group_df[group_df["M"] == 1])/num_episodes_in_group

    # Calculating the percentage of age groups in each group
    heat_map_AGEGROUP[0][i] = len(
        group_df[group_df["Preschooler"] == 1])/
    num_episodes_in_group
    heat_map_AGEGROUP[1][i] = len(
        group_df[group_df["MiddleChildhood"] == 1])/
    num_episodes_in_group
    heat_map_AGEGROUP[2][i] = len(
        group_df[group_df["Teenager"] == 1])/num_episodes_in_group

    for mfa in range(len(mfas)):
        heat_map_MFAs[mfa][i] = len(
            group_df[group_df.iloc[:, 6 + mfa] == 1])/
    num_episodes_in_group

# Decode labels to text, so that it is not only numbers

mfa_text = clusteredData.columns[6:len(mfas)+6]
plot_labels = []

with open('/home/iascheft/workbench/clustering/maspc/pickleDump/
    alphabet_dict.pkl', 'rb') as a_dict:
    alphabet_dict = pkl.load(a_dict)

reverse_alphabet_dict = {}
for i in range(26):
    letter = chr(i + ord('A'))
    if i < 9:
        reverse_alphabet_dict["0" + str(i+1)] = letter
    else:
        reverse_alphabet_dict[str(i+1)] = letter

with open('/home/iascheft/workbench/clustering/maspc/pickleDump/
    phecodes_dict.pkl', 'rb') as p_dict:
    phecodes_dict = pkl.load(p_dict)

reverse_phecodes_dict = {}

for item in phecodes_dict.values():
    reverse_phecodes_dict[str(item[0]).replace(".", "")] = item[1]

for labels in mfa_text:
    labels_list = labels.replace("[", "").replace("]", "").replace(
        "'", "").replace(" ", "").split(",")
    one_mfa_labels = []
```

```
78      for label in labels_list:
79          if label[0:3] == "999":
80              atc = label[3:]
81              atc_string = reverse_alphabet_dict[atc[0:2]] + \
82                  atc[2:4] + reverse_alphabet_dict[atc[4:]]
83              one_mfa_labels.append(atc_string)
84          else:
85              diagnosis = reverse_phecodes_dict[label]
86              one_mfa_labels.append(diagnosis)
87      plot_labels.append(one_mfa_labels)
88
89
90  # Create a color scale for bar plot so that large values have
        similar colors
91  # Inspired from here: https://stackoverflow.com/a/60917129
92  def colors_from_values(values, palette_name):
93      # normalize the values to range [0, 1]
94      normalized = (values - min(values)) / (max(values) - min(values)
        )
95      # convert to indices
96      indices = np.round(normalized * (len(values) - 1)).astype(np.
        int32)
97      # use the indices to get the colors
98      palette = sns.color_palette(palette_name, len(values))
99      return np.array(palette).take(indices, axis=0)
100
101
102 # Bar chart showing number of episodes in each cluster
103 plt.figure(figsize=(15, 7))
104
105 x_vals = np.arange(1, k+1)
106 y_vals = num_episodes_each_cluster_list
107
108 ax = sns.barplot(x=x_vals, y=y_vals, palette=colors_from_values(
109     pd.Series(y_vals), "YlGnBu"), )
110
111 ax.set_ylabel("Number of Episodes")
112 ax.set_xlabel("ClusterID")
113 ax.bar_label(ax.containers[0])
114 sns.despine(left=True, bottom=True)
115
116 plt.savefig('/home/iascheft/workbench/clustering/maspc/Diagrams/
        BarPlots/numEpisodesClusterBARPLOT.pdf', bbox_inches="tight")
117
118
119 # Heatmap for age group
120
121 plt.figure(3, figsize=(20, 5))
122 plt.rc('xtick', labelsize=7)
123 plt.rc('ytick', labelsize=7)
124 plt.rc('axes', labelsize=10, linewidth=1)
125
```

```
y_axis_labels = ["Preschooler", "MiddleChildhood", "Teenager"]
x_axis_labels = list(range(1, k+1))
ax = sns.heatmap(np.round(heat_map_AGEGROUP, 2), cmap="YlGnBu",
    square=True, annot=True, fmt=".2g", annot_kws={"fontsize": 7},
                 cbar=True, cbar_kws={"shrink": .3}, linewidths=.1,
    xticklabels=x_axis_labels, yticklabels=y_axis_labels)


plt.xlabel('Cluster ID', fontweight='bold')
plt.ylabel('Age Group', fontweight='bold')

for text in ax.texts:
    text.set_size(6)
    if float(text.get_text()) >= 0.5:
        text.set_size(8.5)
        text.set_weight('bold')
        text.set_style('italic')

plt.savefig('/home/iascheft/workbench/clustering/maspc/Diagrams/
    Heatmaps/ageGroupHEATMAP.pdf',
            bbox_inches="tight")


# Heatmap for gender

plt.figure(figsize=(20, 5))
plt.rc('xtick', labelsize=7)
plt.rc('ytick', labelsize=7)
plt.rc('axes', labelsize=10, linewidth=1)

y_axis_labels = ["Female", "Male"]
x_axis_labels = list(range(1, k+1))
ax = sns.heatmap(np.round(heat_map_GENDER, 2), cmap="YlGnBu", square
    =True, annot=True, fmt=".2g", annot_kws={"fontsize": 7},
                 cbar=True, cbar_kws={"shrink": .3}, linewidths=.1,
    xticklabels=x_axis_labels, yticklabels=y_axis_labels)


plt.xlabel('Cluster ID', fontweight='bold')
plt.ylabel('Gender', fontweight='bold')

for text in ax.texts:
    text.set_size(6)
    if float(text.get_text()) >= 0.5:
        text.set_size(8.5)
        text.set_weight('bold')
        text.set_style('italic')

plt.savefig('/home/iascheft/workbench/clustering/maspc/Diagrams/
    Heatmaps/genderHEATMAP.pdf',
            bbox_inches="tight")

```

```
# Heatmap for patterns
plt.figure(figsize=(15, 15))
plt.rc('xtick', labelsize=7)
plt.rc('ytick', labelsize=7)
plt.rc('axes', labelsize=10, linewidth=1)

y_axis_labels = clusteredData.columns[6:len(mfas)+6]
x_axis_labels = list(range(1, k+1))
ax = sns.heatmap(np.round(heat_map_MFAs, 2), cmap="YlGnBu", square=True, annot=True, fmt=".2g", annot_kws={"fontsize": 7},
                 cbar=True, cbar_kws={"shrink": .1}, linewidths=.1, xticklabels=x_axis_labels, yticklabels=plot_labels)

plt.xlabel('Cluster ID', fontweight='bold')
plt.ylabel('MFAs', fontweight='bold')

for text in ax.texts:
    text.set_size(6)
    if float(text.get_text()) >= 0.5:
        text.set_size(8.5)
        text.set_weight('bold')
        text.set_style('italic')

plt.savefig('/home/iascheft/workbench/clustering/maspc/Diagrams/Heatmaps/mfaHEATMAP.pdf',
            bbox_inches="tight")


# SHAP Summaryplot showing how much each feature affect the results
# https://towardsdatascience.com/how-to-make-clustering-explainable-1582390476cc

y = maspc.binaryData["label"]
shap_df = maspc.binaryData.drop(["label"], axis=1)
shap_df = shap_df.rename(columns=lambda x: re.sub('[^A-Za-z0-9_]+', ' ', x))
clf = LGBMClassifier(objective="binary")
clf.fit(shap_df, y)

explainer = shap.TreeExplainer(clf)
shap_values = explainer.shap_values(shap_df)

# summarize the effects of all the features
plt.figure(figsize=(15, 7))
shap.summary_plot(shap_values, shap_df, plot_type="bar",
                  plot_size=(15, 10), show=False)
plt.savefig('/home/iascheft/workbench/clustering/maspc/Diagrams/ShapPlots/shapBARPLOT.pdf',
            bbox_inches="tight")
```

Listing F.1: Pyhton code for creating plots visualising clustering results and SHAP plot.

Appendix G

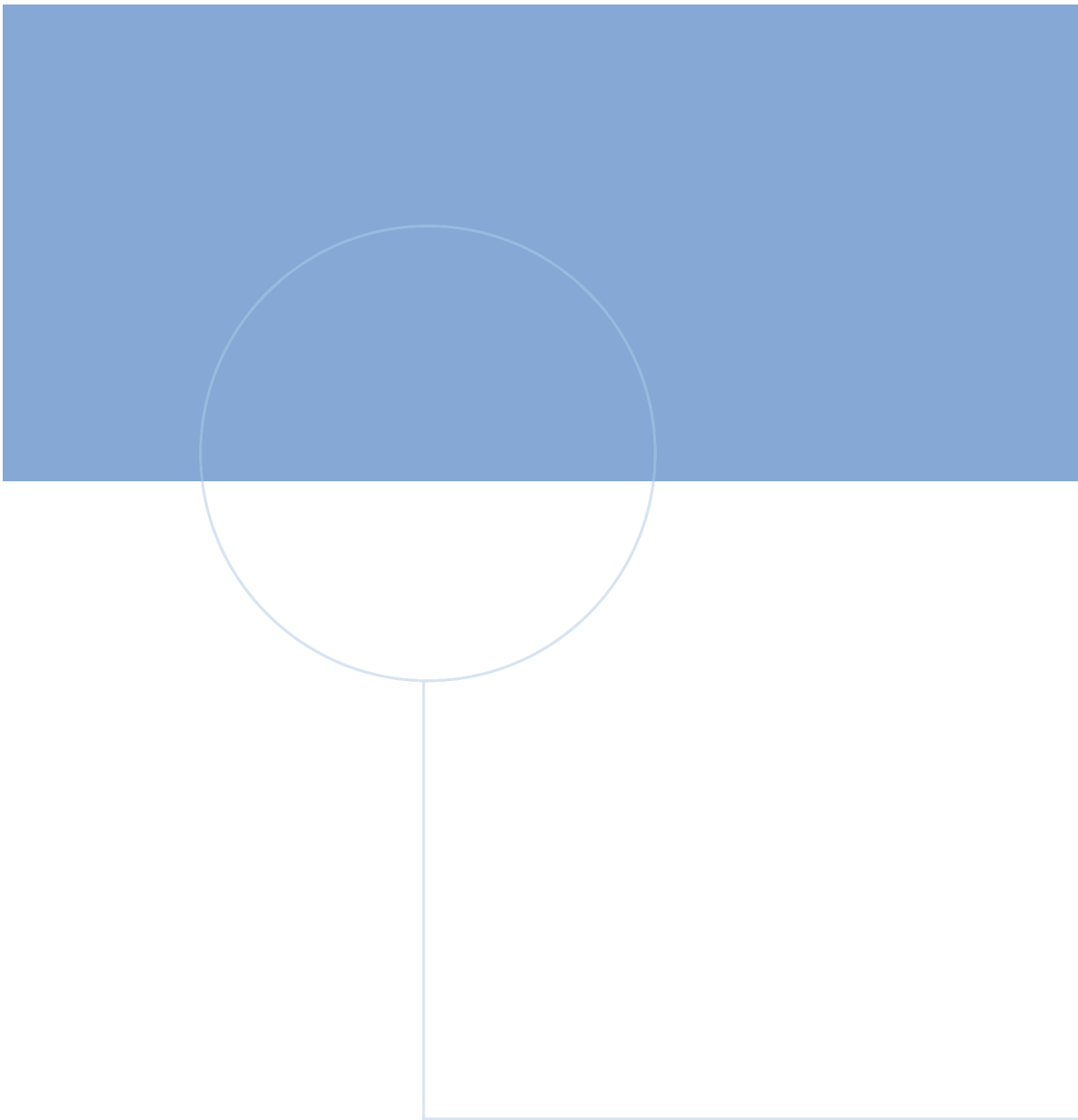
Manual Phecode and Phenotype Mappings

ICD-10 diagnosis	Phecode	Phenotype	Comment
B209	071.1	HIV infection, symptomatic	
E90	277	Other disorders of metabolism	
F068	291.4	Specific nonpsychotic mental disorders due to brain damage	
F069	291.4	Specific nonpsychotic mental disorders due to brain damage	
F18	290.3	Other persistent mental disorders due to conditions...	
F180	290.3	Other persistent mental disorders due to conditions...	
F28	295.3	Psychosis	
F32	296.22	Depressive disorder	Original phenotype "Major depressive disorder", changed to "Depressive disorder" to include more diagnoses.
F320	296.22	Depressive disorder	Original phenotype "Major depressive disorder", changed to "Depressive disorder" to include more diagnoses.
F321	296.22	Depressive disorder	Original phenotype "Major depressive disorder", changed to "Depressive disorder" to include more diagnoses.
F322	296.22	Depressive disorder	Original phenotype "Major depressive disorder", changed to "Depressive disorder" to include more diagnoses.
F323	296.22	Depressive disorder	Original phenotype "Major depressive disorder", changed to "Depressive disorder" to include more diagnoses.
F329	296.22	Depressive disorder	Original phenotype "Major depressive disorder", changed to "Depressive disorder" to include more diagnoses.
F339	296.22	Depressive disorder	Original phenotype "Major depressive disorder", changed to "Depressive disorder" to include more diagnoses.
F204	296.22	Depressive disorder	Did not have a phenotype, but added to "Depressive disorder"
F328	296.22	Depressive disorder	Did not have a phenotype, but added to "Depressive disorder"
F330	296.22	Depressive disorder	Did not have a phenotype, but added to "Depressive disorder"
F331	296.22	Depressive disorder	Did not have a phenotype, but added to "Depressive disorder"

F332	296.22	Depressive disorder	Did not have a phenotype, but added to "Depressive disorder"
F333	296.22	Depressive disorder	Did not have a phenotype, but added to "Depressive disorder"
F334	296.22	Depressive disorder	Did not have a phenotype, but added to "Depressive disorder"
F338	296.22	Depressive disorder	Did not have a phenotype, but added to "Depressive disorder"
F403	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F064	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F40	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F400	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F401	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F402	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F408	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F409	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F410	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F411	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F412	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F413	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F418	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F419	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F606	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F930	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F931	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F932	300.1	Anxiety disorder	Changed by input from clinician to merge all anxiety diagnoses into one category.
F500	305.2	Eating disorder	Changed from "Anorexia nevrosa" to "Eating disorder"
F504	305.2	Eating disorder	
F518	327	Sleep disorders	
F552	316	Substance addiction and disorders	
F555	316	Substance addiction and disorders	
F558	316	Substance addiction and disorders	
F623	306	Other mental disorder	
F638	312.3	Impulse control disorder	
F629	306	Other mental disorder	
F7	315.3	Mental retardation	
F710	315.3	Mental retardation	
F718	315.3	Mental retardation	
F780	315.3	Mental retardation	
F791	315.3	Mental retardation	
F798	315.3	Mental retardation	
F81	315.1	Learning disorder	
F813	315.1	Learning disorder	
F83	315.1	Learning disorder	
F900	313.1	Attention deficit hyperactivity disorder	
F908	313.1	Attention deficit hyperactivity disorder	
F928	312	Conduct disorders	

F933	313	Pervasive developmental disorders	
F98	313	Pervasive developmental disorders	
F982	305.2	Eating disorder	
G40.	345	Epilepsy, recurrent seizures, convulsions	
H50	378.1	Strabismus (not specified as paralytic)	
H549	367.9	Blindness and low vision	
H91	389	Hearing loss	
J951	519.2	Respiratory complications	
K523	555.2	Ulcerative colitis	
K649	455	Hemorrhoids	
K720	571.8	Liver abscess and sequelae of chronic liver disease	
K858	577.1	Acute pancreatitis	
L65	704.1	Alopecia	
M090	714.2	Juvenile rheumatoid arthritis	
M091	714.2	Juvenile rheumatoid arthritis	
M248	742.9	Other derangement of joint	
M609	313.1	Myopathy	
M726	727	Symptoms of the muscles	
M828	743.1	Osteoporosis NOS	
N038	580.14	Chronic glomerulonephritis, NOS	
O04	634	Miscarriage; stillbirth	
O94	676	Other disorders of the breast associated with childbirth and disorders of lactation	
P043	658	Maternal complication of pregnancy affecting fetus or newborn	
P044	658	Maternal complication of pregnancy affecting fetus or newborn	
P070	637	Short gestation; low birth weight; and fetal growth retardation	
P071	637	Short gestation; low birth weight; and fetal growth retardation	
Q044	752.2	Other specified congenital anomalies of nervous system	
Q315	748	Anomalies of respiratory system, congenital	
Q64	751.2	Conqenital anomalies of urinary system	
Q900	758.1	Chromosomal anomalies	
Q914	758.1	Chromosomal anomalies	
S327	1009	Injury, NOS	
S361	1008	Crushing or internal injury to organs	
T012	870	Open wounds of head; neck; and trunk	
T141	1009	Injury, NOS	
T142	1009	Injury, NOS	
T4n	969	Poisoning by psychotropic agents	
T740	1015	Effects of other external causes	
T748	1015	Effects of other external causes	
T742	1015	Effects of other external causes	
X4n	981	Toxic effect of (non-ethyl) alcohol and petroleum and other solvents	
X6n	297.2	Suicide or self-inflicted injury	
X61	297.2	Suicide or self-inflicted injury	
X60	297.2	Suicide or self-inflicted injury	
X6nx	297.2	Suicide or self-inflicted injury	
X6n0	297.2	Suicide or self-inflicted injury	
X6n1	297.2	Suicide or self-inflicted injury	
X6n2	297.2	Suicide or self-inflicted injury	
X6n4	297.2	Suicide or self-inflicted injury	
X6n5	297.2	Suicide or self-inflicted injury	
X6n6	297.2	Suicide or self-inflicted injury	
X6n8	297.2	Suicide or self-inflicted injury	
X6n9	297.2	Suicide or self-inflicted injury	

Y912	981	Toxic effect of (non-ethyl) alcohol and petroleum and other solvents	
Y913	981	Toxic effect of (non-ethyl) alcohol and petroleum and other solvents	



NTNU

Norwegian University of
Science and Technology