Master's thesis

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

Frida Sofie Solheim

Characterising Patients Referred on Suspicion of ADHD and Behavioral Difficulties

An Exploratory Cluster Analysis of Norwegian Electronic Health Records

Master's thesis in Computer Science
Supervisor: Øystein Nytrø
June 2022

NTNU
Kunnskap for en bedre verden

IDDEAS

Frida Sofie Solheim

Characterising Patients Referred on Suspicion of ADHD and Behavioral Difficulties

An Exploratory Cluster Analysis of Norwegian Electronic Health Records

Master's thesis in Computer Science
Supervisor: Øystein Nytrø
June 2022

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

NTNU

Kunnskap for en bedre verden

Abstract

Hyperkinetic disorders is a group of disorders associated with impaired attention and overactivity. They are some of the most common childhood psychiatric disorders and usually arise in the first five years of life. Early assessment and intervention is necessary to mitigate the impact on the child and its family. To better understand the trajectories, we need to better understand the patients.

This thesis investigates and clusters data from the Child and Adolescent Mental Health Services (CAMHS) clinic at St. Olavs Hospital in Trondheim, specifically for patients with relevance to hyperkinetic disorders in their very first referral period. This is done with the intention of identifying characteristics and latent subgroups of patients, as well as uncovering natural patterns and phenomena in data from electronic health records from the last 20 years.

Clustering was used to probe health data of 4,201 patients with relevance to hyperkinetic disorders, either by referral reason or diagnose. Based on age, gender, and 10 variables connected to their first referral period to CAMHS in Norway, six clusters were generated. The clusters were able to capture important aspects and differences in the referral process, identified patient profiles related to gender and rejection rates, as well as unanticipated referral and diagnostic phenomena. As part of that process, an assessment has also been made of the utility of clustering as methodology for analysing patient trajectories.

A significant aspect of the research is to understand the information and results in the relevant context of clinical psychiatry. This is ensured by analysing, building comprehension and engaging in dialogue with clinics and professionals that understand the data that is used and the results that have emerged. This has enabled the research to build comprehension beyond the limitations of only an isolated data analysis.

This thesis is among the first to analyse, cluster and document data that confirms the clinical reality in Norway. The results also confirm the feasibility of working with clinical data, despite clinical data commonly being deficient and prone to human error. This study proves that research on datasets like this can produce accurate results that are aligned and consistent with what can be found internationally, and sheds light on what can be seen in clinical practice within our own borders.

The study makes an introductory foundation for future research on the area, by uncovering interesting and important phenomena that occur during referral and treatment, assess and confirm the potential of clinical data, and compose novel documentation of CAMHS patient trajectories.

Sammendrag

Hyperkinetiske lidelser er en gruppe lidelser assosiert med nedsatt oppmerksomhet og overaktivitet. De er noen av de mest vanlige barndomspsykiatriske lidelsene og oppstår vanligvis i løpet av barnets fem første leveår. Tidlig vurdering og intervensjon er nødvendig for å dempe innvirkningen på barnet og dets familie. For bedre å forstå forløpene, må vi bedre forstå pasientene.

Denne avhandlingen undersøker og klynger data fra Barne- og Ungdomspsykiatrisk Poliklinikk (BUP) ved St. Olavs Hospital, spesifikt for pasienter med tilknytning til hyperkinetiske lidelser i deres aller første henvisningsperiode. Dette gjøres med hensikt om å identifisere egenskaper og latente subgrupper av pasienter, samt avduke naturlige mønstre og fenomen i data fra elektroniske helsejournaler fra de siste 20 årene.

Klynging ble brukt til å undersøke helsedata til 4,201 pasienter med tilknytning til hyperkinetiske lidelser, enten ved henvisningsårsak eller diagnose. Basert på alder, kjønn og 10 andre variabler knyttet til deres første henvisningsperiode til BUP, ble seks klynger generert. Klyngene var i stand til å fange viktige aspekter og forskjeller ved henvisningsprosessen, og kunne identifisere pasientprofiler relatert til kjønn, avslagsprosenter, samt overraskende fenomen relatert til henvisning og diagnostikk. Som en del av den prosessen, er det også foretatt en vurdering av egnetheten til klynging som metodikk for å analysere pasientforløp.

Et vesentlig aspekt ved denne forskningen er å forstå informasjonen og resultatene i den relevante konteksten av klinisk psykiatri. Dette sikres ved å analysere, bygge forståelse og gå i dialog med klinikker og fagpersoner som forstår dataene vi bruker og resultatene som har fremkommet underveis. Dette har gjort det mulig for forskningen å bygge forståelse utover begrensningene til kun en isolert dataanalyse.

Denne avhandlingen er blant de første som analyserer, klynger og dokumenterer data som bekrefter den kliniske virkeligheten i Norge. Arbeidet bekrefter potensialet og muligheten for å forske på kliniske data, til tross for at kliniske data tidvis er svært mangelfulle og utsatt for menneskelige feil. Studien beviser at forskning på datasett som dette kan gi nøyaktige resultater som er på linje og konsistente med det som finnes internasjonalt, og belyser det man kan se innenfor våre egne grenser hva angår klinisk praksis.

Studien skaper et innledende grunnlag for fremtidig forskning på området, ved å avdekke interessante og viktige fenomen som opptrer i sammenheng med henvisning og behandling, bekrefte mulighetsrommet og potensialet til kliniske data, og sammenstille ny dokumentasjon av pasientforløp i BUP.

Preface

The research for this Master's thesis was conducted and written in the Spring of 2022 at the Norwegian University of Science and Technology in Trondheim. The project was carried out in collaboration with the IDDEAS project, an interdisciplinary and international research group. The project was supervised by Øystein Nytrø, Department of Computer Science.

The study aims to analyse electronic health records from the CAMHS clinic at St. Olavs Hospital, and identify and discuss patient characteristics and subgroups of children and adolescents with relevance to hyperkinetic disorders. This also includes an assessment of the feasibility of working and conducting research with clinical data, and clustering as a tool for this task.

In order to build comprehension of data and results in the context of Norwegian clinical psychiatry, these are interpreted in collaboration with professionals. For the research to be valuable, this ongoing dialogue has been devoted a lot of time and resources in the project. Findings have been interpreted and discussed with a number of professionals, including psychologist specialists, computer scientists, researchers and developers from both the IDDEAS project and CAMHS at St. Olavs Hospital in Trondheim. This ensured academic grounding, a clarification of findings, and a confirmation of where the work is located in the research area.

I would like to take the opportunity to thank my supervisor, Øystein Nytrø, for his help, guidance and contribution to this project, for encouraging my research, and for challenging me along the way. The help of Nytrø was essential for navigating the unfamiliar intersection between technology and clinical medicine.

In addition to Nytrø, several resources in the IDDEAS project greatly contributed to the research. I would like to thank the IDDEAS team for being essential in the implementation of this project, and the interpretation of research results. I especially want to give my thanks to Odd Sverre Westbye, for providing valuable input and putting me in touch with useful resource persons along the way.

Furthermore, I want to express my gratitude to the professionals at CAMHS St. Olavs Hospital, particularly psychologist specialist Jostein Arntzen for the essential contribution to the analysis of research results and comprehension of clinical diagnosis today.

Lastly, I would like to thank the HUNT Cloud team for all the indispensable help given to me during my endeavours with their cloud services, and when technology was a bit challenging even for the data scientist.

The contributions of these resourceful professionals have been significant for the work with my Master's thesis.

Frida Sofie Solheim
Trondheim, 1st June 2022

Contents

List of Figures

List of Tables

1. Introduction

The aim of this chapter is to give an introduction to the project, the motivation for the research and the structure of the thesis.

As the title suggest, the project is concerned with the clustering and exploratory analysis of electronic health records of patients with relevance to hyperkinetic disorders. Patients included in this study are both those referred due to suspicion of hyperkinetic disorder (ADHD) or behavioral difficulties, and those that are diagnosed with a hyperkinetic disease in their first referral period. The scope is limited to their first encounter with the Child and Adolescent Mental Health Services, further referenced as CAMHS. The analytical focus is from the time of making the referral to the assessment of the referral. The aim is to identify and characterise potential patient profiles and latent subgroups, by clustering and analysing data of patients referred to a Norwegian CAMHS clinic. Due to little research previously done on the data, the ambition is to probe the electronic health records in an exploratory manner in order to identify phenomena and patterns of interest. Consequently, this may improve the comprehension of patients referred to CAMHS, and hopefully aid in the long-term goal of improving clinical diagnosis and care in Norway.

In addition to giving insight into the background and motivation for the research on patient profiles and subgroups in CAMHS, the chapter also presents the goals and research questions that are to be explored throughout the project. Lastly, the research methodology as well as an overview of the next chapters are provided to the reader.

1.1. Background and Motivation

Hyperkinetic disorders, i.e. diagnoses in the F90-group in ICD-10 (World Health Organization, 1992a), is a group of disorders associated with impaired attention and overactivity. They are some of the most common childhood psychiatric disorders and usually arise in the first five years of life. According to psychologists, it is also the largest cause and diagnostic group in CAMHS in Norway (meeting with Jostein Arntzen, 11.05.22). Hyperkinetic children are often reckless, impulsive, exposed to accidents and disregarding of social rules, which often leads to disciplinary difficulties. Due to this, they may be unpopular with other children and can experience social isolation (Helsebiblioteket, 2022). To mitigate the impact of a hyperkinetic disorder on the child, the family and the local

community, it is necessary to ensure early assessment and intervention.

The IDDEAS team is an interdisciplinary group consisting of developers, researchers, health informatics specialists and clinicians. IDDEAS, which is short for Individualised Digital Decision Assist System, is a clinical decision support system that seeks to improve patient care by providing data driven and evidence based guidelines to healthcare professionals. It is specifically designed for use in mental health services for children and adolescents in Norway. The aim is to improve timing and precision of decision making, reduce the extent of misdiagnosis, and increase patient contact efficiency. According to the IDDEAS team, the first version of IDDEAS focuses on preventive treatment, early diagnosis and intervention, as well as the treatment and care of hyperkinetic disorders like ADHD (IDDEAS, 2021a).

For my Master's thesis, I have collaborated with the IDDEAS team with the aim of identifying patient characteristics and latent subgroups among a selection of patients, investigate the feasibility of conducting research on the basis of clinical data, and assessing clustering as a tool for performing an exploratory analysis of electronic health records of patients referred to CAMHS. The aim is to probe the clinical data we have at hand, assess its potential, investigate the adequacy of using clustering for analysing patient profiles and subgroups, and interpret and discuss the identified phenomena. A subset of the data available, derived from electronic health records from the CAMHS clinic at St. Olavs Hospital in Trondheim, is the basis for the analysis.

Patients of interest are those that are either referred on the suspicion of ADHD or behavioral difficulties, or are diagnosed with a hyperkinetic disorder. Some of the referred patients receive a diagnose, and some do not. Both are equally included, regardless if the problem was formally characterised by CAMHS. A thorough reasoning for this selection can be found in section 2.1.3. To limit the scope of the research, we only investigate the very first referral period to CAMHS of every unique patient, disregarding the history of any later referral periods. Data from their electronic health records are carefully selected, clustered and analysed. This is done on the basis of a variety of observational data, including demographics (gender and age) and clinical data related to their first referral period. An unsupervised clustering technique is used for exploratory analysis in order to probe underlying patterns within the dataset. Using clinical data of 4,201 children and adolescents referred to and assessed by CAMHS in Norway, this enables the identification of clinical and trajectorial patterns and possibly novel insight from the clusters.

The K-Prototype algorithm has been chosen as clustering tool for this task. A part of this thesis is to investigate and assess if clustering, secondly K-Prototype, is a useful and beneficial tool for clinical research on patient trajectories. The task is, however, not to use K-Prototype as means to an end regarding clustering patients in order to obtain concrete results with an expected utility value. The purpose of this project is to attempt to find clusters of patients that can be meaningful to professionals, use these to identify patient profiles, and assess whether clustering can be useful for such an undertaking. The

research is as much about measuring the utility and meaningfulness of clustering in a clinical perspective, as discovering patient characteristics and subgroups in the context of clinical psychiatry.

An essential aspect of this process is to evaluate information and results in their relevant context. Initially, this includes comprehensive literature-, documentation-, and systems archaeology in order to assemble sources and information on the domain, of which are scattered around numerous platforms. This is a prerequisite for ensuring the sufficient domain knowledge prior to analysis. Moreover, the majority of the research is concerned with understanding the information we uncover and work with in context of clinical psychiatry, by analysing, building comprehension, and engaging in dialogue with actual clinics, people and professionals that understand the data I am using. Several meetings and presentations with professionals were held to facilitate the interpretation of clinical phenomena identified in the research, to understand if the findings are meaningful, understandable, and reflects clinical practice today. Their input also supports the choice of direction for future work.

Clinical practice and diagnosis will vary between countries and local regions. Although many countries follow the guidelines for diagnosis according to ICD-10, referral and clinical patterns also vary and change. Thus, research and findings in international studies may not apply to the clinical reality in Norway. Consequently, this project focuses on clinical practice within our own borders, of which the continuous dialogue with Norwegian clinics and professionals is essential for analysing results in context of Norwegian psychiatry. However, international research has been used as assistance in identifying relevant phenomena, and as a basis for comparing patterns of clinical practice.

The ambition is for the research to chart the potential and limitations of the data at hand, provide better understanding of children and adolescents that are referred to CAMHS, facilitate more future research, and make a small contribution to enhancing medical diagnosis and care in Norway.

1.2. Goal and Research Questions

In light of the background and motivation presented in the former section, this section presents the goal and research questions that have been defined for this project.

The overall goal of the project is:

Goal *To analyse electronic health records of patients with relation to hyperkinetic disorders in Child and Adolescent Mental Health, investigate if patient profiles and subgroups can be identified by cluster analysis, and interpret phenomena in the context of Norwegian psychiatry.*

The objective is to analyse and interpret electronic health records of patients referred to a Norwegian CAMHS clinic. This is done by carrying out several iterations of exploratory data analysis, performing a clustering experiment in which interesting profiles and subgroups may be identified, and by discussing and interpreting findings and phenomena with professionals. The ambition is to better understand the patient situations and referral trajectories of patients either referred for or diagnosed with a hyperkinetic disorder. We want to explore whether users of CAMHS can be grouped by who they are, their situation, their medical history and their comorbidities. It should also be investigated if clustering can yield interesting results, is relevant for clinical analysis, and is applicable to similar projects or future work on the area. Lastly, whether the findings are relevant in the research field and in the context of Norwegian psychiatry must be evaluated, of which discussing and interpreting the findings in collaboration with professionals is a key part of the process.

Moreover, in the way the following research questions are defined, they aim to answer both the utility of analysing electronic health records, as well as assessing the value of clustering clinical data for the discovery of patterns. The research questions are:

Research question 1 *What is the utility of an analysis of health record data of patients with relation to hyperkinetic disorders, and can it be used to identify phenomena of interest in the context of Norwegian psychiatry?*

Utility emphasises both the feasibility of analysing electronic health records, as well as its ability to yield results of interest in the context of patients with relation to hyperkinetic disorders and their trajectories. *In the context of Norwegian psychiatry* emphasises the degree to which an analysis is useful for CAMHS, clinical professionals and current research in improving the comprehension of what characterises these patients, and whether novel findings on clinical phenomena can be highlighted. The latter part of the research question is especially concerned with establishing dialogue with clinics and professionals, in order to properly discuss and interpret the findings.

Research question 2 *How meaningful is a clustering of clinical data as a tool for the discovery of patient profiles, subgroups and referral patterns in CAMHS?*

The word *meaningful* is used to emphasise whether it is relevant to apply clustering to clinical data like electronic health records, if this process can aid in the identification of valuable clinical phenomena and patient subgroups, and if clustering as a tool is at all useful in the context of clinical analysis. This also highlights the need for a thorough assessment of clustering in order to map its strengths, weaknesses and prerequisites.

1.3. Research Method

This section aims to describe the research methodology applied in this research, and why it has been chosen. The stages of the research are briefly presented, before being thoroughly explored in chapter 6: Experiment.

To address the defined goal and research questions, the research method of choice is firstly to conduct an experiment with the clustering algorithm of choice. This is done by extracting a relevant selection of the overall cohort and conducting an exploratory cluster analysis in order to discover patient characteristics, patient subgroups and referral period patterns. This requires careful extraction of relevant data, precise mapping between clinical codes and descriptions, and a preprocessing of the dataset in order to prepare it for clustering. In order to not be repetitive, the detailed experimental steps can be found in section 6.1.

Secondly, the aim is to analyse, interpret and discuss the findings and results, if any, with clinicians and CAMHS professionals. The ambition is for this process to facilitate an evaluation of the research results, clustering as methodology for assessing the similarity of patient situations, the feasibility of conducting research on clinical data, and if relevant and useful results can be produced. The dialogue with professionals is essential to the comprehension of information in the context of clinical practice, treatment and care, and a significant part of the research.

1.4. Thesis Structure

This sections provides an overview of the thesis structure.

Following this introductory chapter, a thorough theoretical introduction is given in chapter 2, which presents the background theory. This chapter aims to give the reader the necessary theoretical comprehension of clinical diagnostics, process and care. It also explains the time scope of the data selection in the upcoming experiment.

After the theoretical introduction, chapter 3 describes the concept of clustering and how it is applied in this project. It also elaborates on the requirements for this specific application, and presents the chosen clustering technique.

Chapter 4 presents any related work on the area, and where my research is situated in the field.

Chapter 5 thoroughly explains the cohort, both the overall cohort and our specific selection, the environments in which the data has been processed, and how sensitive data has been managed in accordance with security regulations and agreements.

Following this, the experiment and results are presented in very much detail in chapter 6. This includes the experimental plan, setup, the exploratory data analysis, and the execution of the experiment itself. Lastly, the clustering results are presented at the end of the chapter.

The results, including the methodology and process, are evaluated in chapter 7. This also includes the clinical evaluation, of which the results were presented, interpreted and discussed with a panel of professionals.

The next chapter, chapter 8 presents the discussion of the results in light of recent research, knowledge, the clinical evaluation, and the aim of the project. Important information and inputs from the meeting with CAMHS are also included. At the end of the chapter, the research questions are revisited and answered.

Finally, the research is concluded in chapter 9, which also includes a summary of the contributions and a presentation of future work.

2. Background Theory

Due to the multidisciplinary nature of this research project, it is necessary to increase domain knowledge in several areas. The aim of this chapter is to present the theory that is necessary for the comprehension of CAMHS process and procedures in Norway. Theoretical aspects related to computer science and machine learning are presented in the subsequent chapter 3.

As briefly mentioned, a significant part of the research is to understand the information in its relevant context. For the data analysis and clustering results to be valuable, they need to be interpreted and understood in connection with clinical practice, treatment and care. The prerequisite for such an undertaking is firstly to assemble, interpret and be acquainted with the available domain knowledge. This chapter focuses on this comprehension, and is also the result of comprehensive literature-, documentation-, and systems archaeology in order to assemble sources and information on the domain.

Section 2.1 is concerned with clinical diagnostics in relation to international diagnostic guidelines like ICD-10 and the topic of hyperkinetic disorders. In section 2.2, we scope in on CAMHS in Norway and the available documentation on the aspects of clinical care within our own borders.

It is suggested to visit appendix A for translations made from Norwegian to English for any referring instances or referral reasons mentioned in this chapter, some of which are unique in Norway and do not follow any international definition.

2.1. Clinical Diagnostics

This section aims to give an introduction to diagnostics in CAMHS, the system it is based on, and hyperkinetic disorders, which is the target disorder group for this project. This is done in the context of clinical assessment and care, in order to build familiarity with the symptoms and complications, as well as related clinical codings which is necessary for the data analysis. In order to know what specific data to look into when performing experiments with the dataset, it is necessary to elaborate on which data that represents the diseases we want to investigate, what codings that are relevant to the diagnose we investigate, and which restrictions we want to set to reduce the scope.

2.1.1. ICD-10

The International Classification of Diseases (ICD) is an international, collective classification of diseases that WHO wants all membering countries to use. According to WHO (World Health Organization, 2022a), ICD:

- *allows the systematic recording, analysis, interpretation and comparison of mortality and morbidity data collected in different countries or regions and at different times*
- *ensures semantic interoperability and reusability of recorded data for the different use cases beyond mere health statistics, including decision support, resource allocation, reimbursement, guidelines and more.*

In short, ICD contains disease codes, and CAMHS takes use of these ICD-codes when associating a diagnosis to a patient. For example, if a patient is examined at CAMHS and they find that the symptoms meet the criteria for Oppositional defiant disorder, they will code F913 (World Health Organization, 1992b).

However, In 2018, WHO published the 11th revision of the International Classification of Diseases (ICD-11). According to WHO, this revision *better reflects advances in science and medicine, aligning classification with the latest knowledge of disease treatment and prevention. There is more meaningful clinical content than ICD–10* (World Health Organization, 2022a). Data in this project is based on diagnostic codes and guidelines from ICD-10, as they were recorded prior to the newest revision.

There are several key differences from ICD-10 to ICD-11. There are especially two aspects to highlight, as these are relevant to upcoming challenges of working with clinical data. Firstly, ICD-11 is supposed to be a flexible system that eliminates the need for local variants. All kinds of clinical detail can be documented, which essentially means more detailed patient trajectories. Secondly, its coding is simplified, which enables seamless integration into already established clinical routine. It is also stated that correct use of ICD requires less training with ICD-11, as well as less time for coding (World Health Organization, 2022b).

These changes from ICD-10 to ICD-11, as well as the fact that the data used for this research is based on ICD-10, is important to keep in mind as we shall revisit this topic when evaluating our experimental results in chapter 8.

2.1.2. Multi-Axial Classification System

The multi-axial classification of Child and Adolescent Psychiatric Disorders system was produced to be used with ICD-10. It has been in use in Norway since 2008 (Direktoratet for e-helse, 2022), and a major advantage of the system is that composite conditions can

be described using the axis system. Every group of disorders belong to one axis, and if a set of criteria specific to an individual diagnosis are met, the illness can be coded on its associated axis.

The six axes are:

1. Clinical psychiatric syndrome
2. Specific disorders of psychological development
3. Intellectual level
4. Co-existent medical conditions
5. Associated abnormal psychosocial situations
6. Global assessment of disability

CAMHS clinics need to follow the guidelines in ICD-10 for coding of illnesses and health related issues when describing a condition. An updated and complete list of the different codes, as well as more detailed guidelines for each axis can be found at Direktoratet for e-helse. Hyperkinetic disorders, which will be further presented in the next section, are coded on axis 1. This essentially also means that when selecting data for the experiment, any code - or lack of - on the first axis during a referral period is of relevance.

2.1.3. Hyperkinetic Disorders

According to the ICD-10 Classification of Mental and Behavioural Disorders (World Health Organization, 1992a), hyperkinetic disorders are any disorder in the F90-group, and is classified on axis 1 in the multi-axial classification system.

According to Helsebiblioteket (2022), hyperkinetic disorders usually arise in the first five years of life. The cardinal features are said to be impaired attention and overactivity, as well as being characterised by frequently changing from one task to another, leaving tasks unfinished, and being distracted by new ones. Overactivity manifests in restlessness, especially in situations where the child is expected to be calm. These behavioral characteristics are most prominent in organised and structured situations that would normally require a certain degree of self-control. The excessiveness of both impaired attention and overactivity should be assessed in comparison to the child's age and IQ, and to other children of similar age (Helsebiblioteket, 2022).

Children that are hyperkinetic are often reckless and impulsive, exposed to accidents, and flouting of social rules like interrupting the activities of others or having trouble in

waiting turns. This often leads to more disciplinary difficulties. Due to this, they may be unpopular with other children and can experience social isolation. It is also usual to see cognitive disturbances, and specific motor and language development disorders are frequent. Moreover, secondary complications are dyssocial behavior and low self esteem (World Health Organization, 1992a).

In WHO's disease classification system ICD-10 here are four codes in the F90-group:

- F900 Disturbance of activity and attention
- F901 Attention Deficit Hyperactivity Disorder
- F908 Other hyperkinetic disorders
- F909 Hyperkinetic disorder, unspecified

A major challenge in diagnosis is the differentiation from other disorders, in this case especially conduct disorder (attention deficit). If the criteria of hyperkinetic disorder is met, it is diagnosed with priority over conduct disorder. When features of both hyperactivity and conduct disorder are present, and the hyperactivity is pervasive and severe, F901: *Attention Deficit Hyperactivity Disorder* should be the diagnosis (World Health Organization, 1992a).

This means that the code should be F900 when the overall criteria for a hyperkinetic disorder (the F90-group) are met, and those for conduct disorders/attention deficit (the F91-group) are not. Furthermore, F900 then includes *attention deficit disorder or syndrome with hyperactivity* and *attention deficit hyperactivity disorder*, but consequently excludes hyperkinetic disorder associated with conduct disorder. For the latter, F901: Attention Deficit Hyperactivity Disorder (also commonly referred to as Hyperkinetic Conduct Disorder) should be coded instead, i.e. when both the overall criteria for hyperkinetic disorders (The F90-group) and the overall criteria for conduct disorders (The F91-group) are met.

When there is a lack of differentiation between F900 and F901, but the overall criteria for F90-group are fulfilled, F909: *Hyperkinetic disorder, unspecified* can be coded, but is not recommended to be used.

The F91-group, behavioral disorders, is another class that to some degree have similar features to F90. However, the focus remains on the diseases in the class of hyperkinetic disorders, and not on behavioral disorders.

Regardless of any disorders or diagnoses, not every patient is diagnosed with a disorder in the F90-group, even though they were referred on the suspicion of it. For research purposes and composite understanding of the trajectories, it is important to be familiar with the associated referral reasons. Table 2.1 presents the four most commonly used

codes in cases with relation to hyperkinetic disorders. These were identified through initial data analysis of the cohort when looking into patients given a diagnose in the F90-group, and their most associated referral reasons. A thorough elaboration on this based on the actual data can be found in section 5.1.2.

New code	Map	Old code	Map
3	Suspicion of defiance/conduct disorder	29	Behavioral difficulties
4	Suspicion of hyperkinetic disorder (ADHD)	30	Hyperactivity/concentration difficulties

Table 2.1.: Relevant referral reasons for hyperkinetic disorders.

Due to the change of Norwegian referral reasons in 2009/2010 (meeting between CAMHS and IDDEAS, 18.11.2021) some referral reasons were able to be mapped directly to new ones, and some were not. For the upcoming experiment, an attempt was made to map old reasons to new reasons, and can be found in appendix A.

Other referral reasons frequently seen in connection to hyperkinetic disorders, are 16: *other reasons*, 10: *suspicion of depression* and 7: *suspicion of anxiety*. This is confirmed by own research, see section 6.5.

To conclude, both the referral reasons above and all of the diagnoses belonging to the F90-group are highly relevant to this research, and will equally be included in the data selection.

2.2. Child and Adolescent Mental Health Services

This section provides an introduction to the Child and Adolescent Mental Health Services (CAMHS) in Norway. This includes CAMHS as an organisation, its processes, and its clinical routines and systems. The purpose is to better understand the aspects we investigate in the upcoming experiment; the referral process, the assessment process, the different patient trajectories, and the data behind them.

BUPdata, an electronic health record system, was developed and previously used by CAMHS in Norway. BUPdata was in use for around 30 years before it was completely discontinued around 2019, when CAMHS was integrated with the Specialist Health Service and BUPdata was replaced (Koochakpour et al., 2022). The data used in this

research project is based on health record data from the BUPdata system.

In order to be acquainted with terms defined in the BUPdata system that are used in this section, some useful clinical definitions are firstly presented in section 2.2.1. These are helpful both for understanding clinical practice in Norway, and for the identification of clinical registrations to use in the experiment.

2.2.1. Terms and Definitions

Before moving on to CAMHS, we highlight some relevant terms retrieved from the BUPdata Kodebok (Norsk forening for Barne- og ungdomspsykiatriske institusjoner and Hiadata AS, 1999). These terms, including their definitions, are helpful for building comprehension of the different clinical phrases used in this research.

- Patient: One that is referred to CAMHS and admitted to treatment.
- Referral period: The time from reception of referral for one and the same illness for assessment, treatment, rehabilitation and follow-up is completed, and no new contacts are agreed. A patient can have several referral periods within a health institution if several illnesses are present. A referral period may include several series of outpatient treatment.
- Episode: An episode, often referred to as stay, is a time period of which a patient receives health care by one and the same place of treatment for one and the same health issue. It can either be a 24-hour stay, a day stay, an outpatient contact, or indirect contact. The term episode defines the single contact, and the length of each episode may vary.
- Contact: Contact is defined as uninterrupted interaction between patient and health personnel where the patient receives health care, within a stay/episode.
- Unit or care unit: Unit and care unit are used interchangeably. A CAMHS clinic can consist of one or more units. Locally, a unit can be denoted as department, outpatient clinic, post or similar. Every unit is associated with a unit type referred to as level of care; outpatient clinic, day or 24-hour.
- Team: A team is a working group within a unit or across several units. They are usually composed on the basis of geographical area of responsibility, by patient issues and/or treatment method.

2.2.2. Referral Process

As elaborated on in chapter 1: Introduction, we scope in on data concerning the first referral period or time of contact for patients with relevance to hyperkinetic disorders, as well as the early stages of assessment. All these aspects are described in this section, which presents some of the available documentation on the referral process in Norway.

According to Helsedirektoratet (2020), if there is suspicion of a hyperkinetic disorder, parents or guardians can contact the kindergarten, school, health nurse, or the general physician (also referred to as GP). It may also be the other way around, if suspicion arises in arenas outside the home. Together they assess whether there is reason to move on to municipal services, e.g. the GP or the Educational Psychological Service, who in turn assess the need for reaching out to the Specialist Health Service. The municipal services can refer children and adolescents to CAMHS, when one or more symptoms of severe mental illness are present. Based on the referral, CAMHS will assess whether the child or adolescent has the right to receive care in the Specialist Health Service (Helsedirektoratet, 2020).

Note that the Help Service for Children and Young People some places often is a collective designation for instances like the Health Station, The Child Welfare Service, Educational Psychological service, and school and kindergarten. When parents or other care takers have concerns or want to reach out for help regarding their child, they may get in touch with the Help Service, which constitutes of several instances like the ones mentioned above. This is important to keep in mind when numbers are assessed in chapter 6.

The most common practice is for the referral to be made by the GP, but others like the child welfare manager may also make the referral. In case of the latter, it is advised that it is coordinated with other instances, e.g. the GP, to make sure the relevant and necessary information is included in the referral. It is not regulated by law which actors that can refer to the Specialist Health Service.

When making a referral, the referral includes relevant information for further assessment. This is any necessary patient information, referral date, information about referring instance and actors involved, up to three reasons for the referral regarding the child, up to three reasons for the referral regarding the child's environment, legal basis for the referral, the Child Welfare Services' involvement in the case, what kind of case type, and so on. Together with personal information on the patient, the referral is the basis for assessment upon arrival at CAMHS.

Figure 2.1 illustrates some of the possible pathways of a referral to CAMHS. This is a non-standardised process and as mentioned, it is not regulated by law which actors can make the referral. Thus it may look different across regions and municipalities, and for every individual case. In some cases, a CAMHS clinic is a part of the pediatrics

clinic in the somatic hospital, e.g. at St. Olav's University Hospital. They have their own outpatient clinic in the pediatric clinic, and only accept referrals from its associated clinic. This is relevant because it means that even though a large majority of referrals nationally come from municipal services like the GP or the Child Welfare Services, some regions may also have a considerable amount coming from somatic hospitals.

When a referral is received, it is assessed to one of four categories: *Accepted*, *rejection due to capacity*, *rejection for professional reasons*, and *assessment so far*. Note that *rejection due to capacity* has been excluded from figure 2.1 as it is no longer in use as a reason for rejection (meeting between CAMHS and IDDEAS, 18.11.21).

If a referral is accepted and the child or adolescent is given the right to patient care in the Specialist Health Service, an individual deadline for health care to be initiated is also provided (Helsedirektoratet, 2020). If a referral is still in assessment, it means it is uncertain at this time if the case fulfills the requirements for the right to patient care, but it is also described in such a way that it is not yet to be rejected. Missing information should not automatically give a rejection, and CAMHS will investigate the case by asking for more details from the referring actor.

In the case of the Specialist Health Service declining the right to health care, CAMHS will provide a reasoning for the decline, as well as recommended measures to take care of the child. This is usually based on an assessment by CAMHS that the symptoms of the patient do not meet the criteria in the regulation for priority health care, or the Prioritisation guide for CAMHS (Helsedirektoratet, 2020). Furthermore, an important part of this is that if CAMHS can assess that other services like the municipal health services can accommodate the needs and follow-up of the child, this can be sufficient for a decline. This is usually based on a professional assessment of needs and local health resources. According to conversations between the IDDEAS project and CAMHS (meeting between IDDEAS and CAMHS, 18.11.2021), the main reason for referrals being rejected today, is due to too few interventions in the municipality by the referrals.

Table 5.1 in section 5.1.1 recaps the assessments made of all incoming referrals in CAMHS for the given dataset. As can be seen from the table, 85.80% were accepted, 0.21% were rejected due to professional reasons, and 10.48% were rejected due to capacity. However, the latter is no longer in use, as it is not an acceptable reason for rejection. This can actually be considered to be a rejection due to professional reasons, not due to capacity issues. However, if a rejection is made, we do not know the details of the professional reasons, e.g. that the municipality have not enforced enough interventions prior to the referral, if the referral is missing information, and so on. These are documented by text in records not available to the project at the time during work with this thesis. However, at the meeting between CAMHS and IDDEAS 18.11.2021, it was stated by CAMHS representatives that too few interventions in the municipality by the referring actor is the primary reason for rejection today.

2.2.3. Assessment, Stays and Progress Time

This section briefly discusses the assessment of the patient after a referral is accepted, the recommended progress time for assessments as recommended by the Norwegian Directorate of Health, and related stays a patient may have at CAMHS.

In 2019, the Norwegian Directorate of Health implemented package procedures for mental health and substance abuse (Helsedirektoratet, 2020). The aim was to warrant predictable progress time without professionally unfounded waiting time, ensure that patients experience a well organised and complete process, and ensure a nationally equal offer (Helsedirektoratet, 2021).

Once a referral reaches CAMHS and the referral is assessed to be accepted, the progress time is initiated. Progress time is evaluated from clinical decision (assessment) to the first evaluation. Assessment in the package procedure is divided into basis and extended assessment, each with a recommended progress time of 42 calendar days (Helsedirektoratet, 2021). Every patient will receive an offer of basis assessment, before the need for an extended assessment is decided.

A stated goal is for at least 80% of the patients that are in a package procedure for assessment and treatment of the patient to have completed assessment of the patient within the recommended progress time (Helsedirektoratet, 2021). In the second tertial of 2021, 52% of patients were assessed within the recommended progress time (including both basis and extended assessment of the patient). There are regional variations, but none of the regional health authorities reached the goal of at least 80% completion within the recommended progress time.

Regardless of assessment outcome and progress time, every patient that has been referred to CAMHS are registered with at least one stay/episode. That includes rejected patients; even these have one registered episode for every referral received by CAMHS. These registrations are used to count and compare the number of stays for all patients in the upcoming experiment.

Furthermore, for every episode in a given referral period, there can be several contacts. Every contact is registered with a type, e.g. assessment, treatment or control inspection. The place where an activity was conducted may also be registered, e.g. at the health institution, at home with the patient, or at an external institution, to mention some. There may also be registrations on whom participated in the different activities. All these variables may provide useful information when looking into the different patient trajectories for patients referred with suspicion of hyperkinetic disorders.

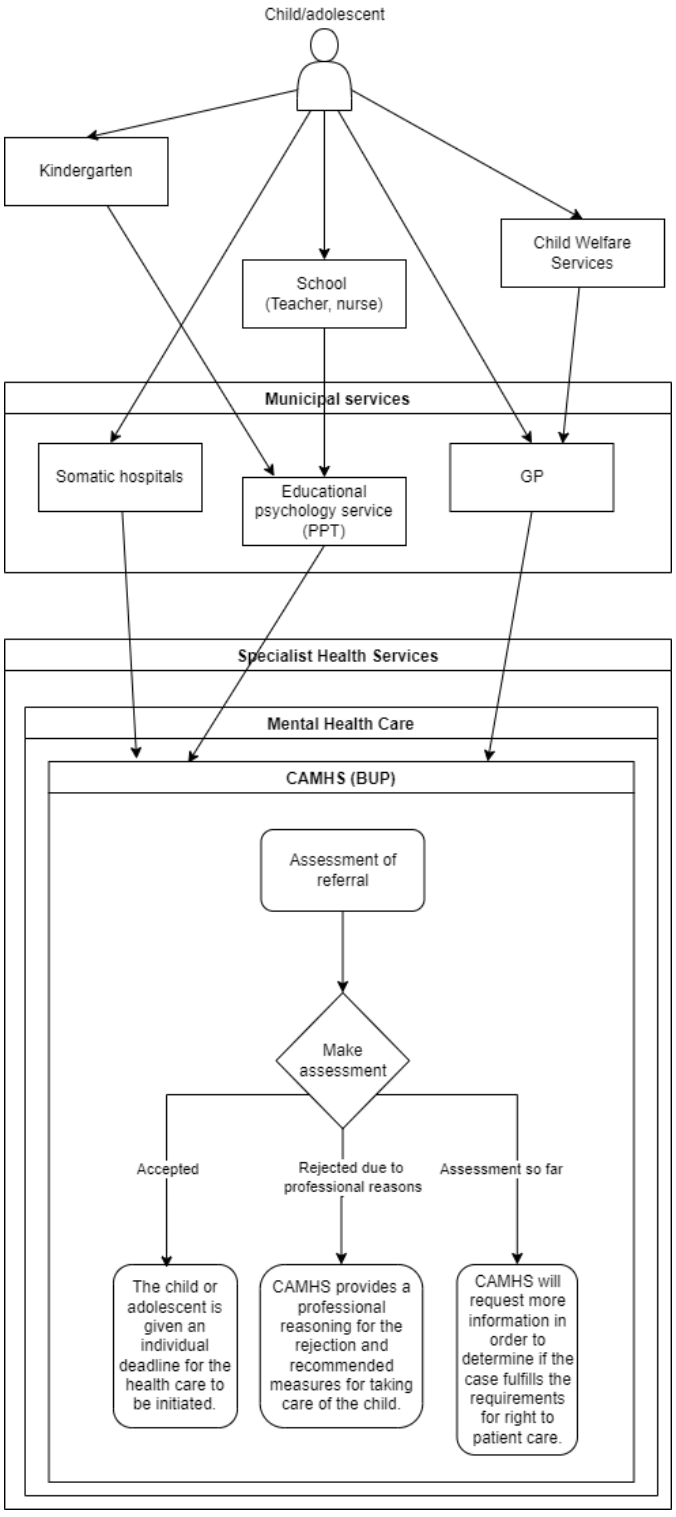


Figure 2.1.: Example of different pathways in the referral process.

3. Clustering Methodology

This chapter presents clustering, an unsupervised machine learning technique used as a key tool in the work of identifying patterns and interesting subgroups of patients in the datasets. Furthermore, we look into the requirements when choosing an algorithm, the specific clustering method applied in the upcoming experiment, and the rationale for using it. The aim is to concisely ensure the minimum knowledge required within the field of computer science to understand the subsequent research, and to provide some important reflections on aspects of the CAMHS dataset that must also be accounted for in future research.

3.1. Machine Learning

In order to build acquaintance with clustering, we start with a brief introduction to machine learning. Machine learning is a branch within computer science and Artificial Intelligence (AI). It is the science of training machines to analyse and learn from data, by imitating the ways humans learn (IBM, 2020a). This gradually improves its accuracy. By using statistical methods, algorithms can be trained to identify patterns, make classifications, or even predict new cases. These key insights provide data understanding and drive decision making.

Machine learning is progressively playing a major role in healthcare and medicine. In the paper by Sidey-Gibbons and Sidey-Gibbons (2019), they highlight several applications of machine learning in medicine. It is used as diagnostic tools, for patient monitoring, identification of latent phenotypes, for predicting new outcomes, or to identify patients in risk groups, to mention some. As providers increasingly employ electronic health records, machine learning techniques offer a huge potential for enhancing medical research and clinical care.

Furthermore, Sidey-Gibbons and Sidey-Gibbons (2019) address two areas which may benefit from the application of machine learning techniques in the medical field, namely diagnosis and outcome prediction. These are relevant to our long-term project aims. There are numerous studies that support the accuracy of using machine learning in medicine for such purposes, for example a study on cancer prediction from Oslo University Hospital on predicting clinical outcomes of colorectal cancer based on patient clinical

and laboratory data (Oslo University Hospital, 2018), or a study from Washington on clustering of patients with overactive bladder syndrome to better facilitate treatments (Gross et al., 2021). The former used a classification technique to predict the time to disease recurrence for new patients. By using cross-validation technique on the results, they found that it correctly labeled 76% of the patients. Of these 35 correctly classified cases, 21 of the 28 long-recurrence patients were correctly classified (75%), and 14 of 18 short-recurrence patients were correctly classified (78%) (Oslo University Hospital, 2018). They concluded that the classification provided results of high accuracy. In the latter, they were able to identify two clusters of overactive bladder-patients: a urinary cluster and a systemic cluster. Furthermore, they were able to identify strong characteristics of comorbidities associated with each cluster, which essentially improved the understanding of pathophysiology of overactive bladder subtypes (Gross et al., 2021). These studies are among several that confirm the feasibility of applying machine learning techniques in the medical field and the accuracy of results that can be produced.

3.2. Clustering

The implementation of this project applies clustering as means for data analysis. Clustering is a form of unsupervised machine learning, of which the aim is to analyse and identify structures or subgroups within the data. As opposed to supervised machine learning, it is done by using unlabeled datasets (IBM, 2020b).

In supervised machine learning, labeled datasets are used to train algorithms to classify data or predict outcomes accurately. A real-world example can be the filtering of spam in an inbox, by training the algorithm with pre-labeled mail (spam or not spam) in order to automatically classify new, incoming mail to the spam folder. Unsupervised machine learning do not use labeled data like this, which requires the algorithm to identify any underlying patterns or structures in the data.

The use of clustering algorithms are often the first step in machine learning. The ability clustering has to discover similarities and differences in data makes it the ideal solution for topics like exploratory data analysis and patient segmentation, which essentially is what this study is interested in. Unsupervised machine learning has been integrated into this project because of its strengths in identifying patterns and structures, revealing key data insights at an early stage in a project, enabling process automation which saves time and resources, grouping data so that patient subgroups may be identified, and providing decision support with respect to the different patient groups.

3.2.1. Clustering Algorithms

A clustering algorithm classifies a set of given data points into a specific group. When the choice of algorithm is suitable and scales well to the dataset, this means that data points that have been assigned to the same group should have similar properties, and data points that are not in the same group should have different properties.

In order to group similar data points, similar examples of data need to be identified. A typical clustering algorithm uses a similarity measure to compare different data points. In most applications of clustering, the similarity measure is based on distance functions like e.g. Euclidean distance, Manhattan distance and Cosine similarity (Irani et al., 2016), or Pearson correlation and Mahalanobis distance (Xu and Tian, 2015). The clusters are formed so that any two objects within a cluster have a minimum distance value, and objects across different clusters have a maximum distance value.

We will not go into further detail on the different categories of clustering, but the choice of clustering technique is largely dependent upon the kind of data one has, i.e. dimensionality, types (categorical or numerical) and size. At the start of this research project, netDx, a patient classifier for building patient similarity networks, was initially chosen. The discard of netDx as algorithm of choice, is thoroughly discussed in section 7.4.1.

3.3. Requirements for Clustering Technique

In this section, we discuss the requirements and prerequisites for choosing a cluster algorithm for the analysis and identification of patient characteristics and latent subgroups in the specific context of clinical data. When working with clinical data, there are some aspects to consider. This reasoning will support the decision of which clustering algorithm to use in the upcoming experiment, as well as provide important aspects to consider in future research.

3.3.1. Scalability

Even though many algorithms work well on smaller datasets, a large database can contain millions of objects. In the context of this project, the initial selected dataset is 4201 rows, after careful selection from a database with over 3 million objects. In order to support such a potential, a clustering algorithm should be scalable to sizes beyond the initial experiment of this project.

3.3.2. Mixed Datatypes

Many applications of real-world clustering examples need to handle the presence of different datatypes in the datasets, like numerical values (also referred to as continuous), binary values or categorical values. Most clustering algorithms efficiently handle single types of attributes, but later developments have made it possible to cluster objects of different types by enforcing different similarity measures dependent on the type of attribute. In this project, the dataset includes both numerical and categorical values, which needs to be accounted for.

3.3.3. Noisy Data

Datasets collected from real-world situations more often than not contain outliers, missing data, or erroneous data, all three of which are present in our dataset. The records are manually registered clinical codings, which makes them prone to human errors like mistyping and inaccurate coding, and the same observations in patients may even be coded differently from one clinic to another. Preprocessing and cleaning may be used to mitigate some of these errors, but not all. Thus, the clustering algorithm must not be too sensitive to outliers, as these may well be present in our manually recorded datasets.

3.3.4. High Dimensionality

It is evident that computing on lower-dimensional data is a less expensive computational task than higher-dimensional data. Most clustering algorithms handle two or three dimensions easily, but having anything from a tens to thousands of dimensions, the algorithm should be chosen with the number of dimensions in mind. Our initial data selection has 12 columns, which should be a a sensible dimensionality to handle. However, later experiments do not exclude the possibility of as much as 30 columns related to patient trajectories, of which an algorithm should be able to handle.

3.3.5. Similarity Measure and Transformation Potential

As briefly mentioned, a similarity measure is a measure used to compare different data points. We want to have similarity measures that are suitable for our data types, and applicable to their transformation potential. Datasets with numerical values can use algorithms with similarity measures like Euclidean or Manhattan distance (Irani et al., 2016). Given a dataset with categorical data, the sample space is discrete, and doesn't have a natural origin. Due to this, measures like Euclidean distance is not meaningful to

use. An option is to use a similarity measure that is applicable to categorical values, like Hamming distance (Huang, 1997), or to transform the categorical values to binary values.

For datasets with both categorical and numerical values, there are several alternatives. One alternative is to use an algorithm like K-Means, which is suitable for numerical values, with a binary transformation technique called one-hot encoding (Scikit-learn, 2022a). One-hot encoding enables the appliance of Euclidean distance by binarising the categorical data. An example of one-hot encoding in the context of our own dataset, is the possible transformation of the column named *custody*, which takes on the values *Mother and father*, *Mother*, *Father* and *Other*. These column values could be transformed to their own columns, of which each new column, e.g. *Mother and father* could either have the value 1 when it is true that both mother and father has custody of the child, and 0 when it is false.

However, in comparison to our dataset, each categorical column takes on anything from 2 to 56 unique values, and a binary transformation would unreasonably increase the dimensionality of the dataset. As the dimensions increase, the distances between different data points tend to be closer together. This is a problem when using algorithms that deal with distance-based metrics. If one has the option, one can either reduce the number of similar features (i.e. columns), join similar features that can make sense even when merged to only one column (e.g. *relation to mother* and *relation to father* could be joined and renamed to *parental relation*), or aim for a dimensionality reduction method.

However, having quite a few columns in our initial dataset, dimensionality reduction may not be necessary if the categorical values can be evaluated as they are without the need for transformation. In that case, a suitable algorithm would be one that can handle both categorical and numerical values, with sufficient similarity measures.

3.4. Choice of Clustering Technique

In light of the requirements discussed in the previous section, this excludes a number of algorithms, as well as it highlights the prerequisites of the ones that are relevant.

If the dataset solely consisted of numeric values, K-means would be a good choice. Firstly, it is easy to implement, but can even scale well to large data sets, which makes it a good starting point. It is also one of the most popular clustering algorithms. On the other hand, if we only had categorical values, K-modes would be a good choice. It defines clusters based on the number of matching categories between data points, as opposed to K-means, which clusters numerical data based on Euclidean distance (Rodriguez et al., 2019).

A lesser known, but powerful sibling of the two, is the K-prototype algorithm. Ac-

cording to the paper on K-prototype by Huang (1997), it offers the advantage of being able to handle mixed data types, by providing a sufficient similarity measure for both numerical and categorical data. It unites K-means and K-modes by measuring distance between numerical features using Euclidean distance, and the distance between categorical features using the number of matching categories. This means that there is no need for transforming categorical values to numerical values. It scales well to larger datasets, and is not too sensitive to outliers. For eager readers, the cost and similarity function of K-prototype can be found in the paper by Huang (1997), section 2.

Several experiments and studies conclude that K-prototype is just as good or even better than using K-means with one-hot encoding (Ruberts, 2020). Another such study is the paper by Irani et al. (2022), which is also elaborated on in section 4. In their study, they showed great success in using the K-Prototype algorithm for better understanding the different clinical phenotypes across the disease spectrum in patients with COVID-19. Their data constituted of observational data, including demographics (age and gender), and 20 basic laboratory tests, and a total of 7606 patients. Our initial data selection is both lower in dimension and size.

The documentation of successful implementations of the K-prototype provides a good basis for choosing it as a cluster technique, and it is reasonable to assume that K-prototype can accommodate the requirements for our dataset from the former section.

3.4.1. K-Prototype

K-prototype was developed by Zhexue Huang in 1997. The aim was to develop a clustering algorithm that is based on the K-means paradigm but removes the numeric data limitation, whilst still preserving its strength and efficiency (Huang, 1997).

As mentioned in the former section, K-prototype is essentially a cross between the K-means algorithm and the K-modes algorithm. K-means clusters data using Euclidean distance. K-modes clusters categorical data based off the number of matching categories between data points. Thus, the K-prototype uses a distance measure that mixes the Hamming distance for categorical features and the Euclidean distance for numeric features. The K-prototype algorithm clusters objects with numeric and categorical attributes in a way similar to K-means, but because the objects are clustered against k number of prototypes instead of k means of clusters, it is called K-prototype (Huang, 1997).

The mathematical preliminaries of K-prototype is not emphasised in this thesis, but we briefly describe the steps in the algorithm. For readers interested in the mathematical details of the clustering algorithm, they are suggested to read the paper on K-prototype by Huang (1997), from which the following steps are retrieved from.

1. Select k initial prototypes from a data set X, one for each cluster.

2. Allocate each object in X to a cluster whose prototype is the nearest to it. Update the prototype of the cluster after each allocation.

3. After all objects have been allocated to a cluster, retest the similarity of objects against the current prototypes. If an object is found such that its nearest prototype belongs to another cluster rather than its current one, reallocate the object to that cluster and update the prototypes of both clusters.

4. Repeat step 3 until no object has changed clusters after a full cycle test of X (Huang, 1997).

Based on the elaborations in this chapter, K-prototype is an adequate algorithm for the purpose of this thesis, with good documentation and several similar studies to support its use (Irani et al., 2022). It allows us not having to one-hot encode all categorical columns, and still benefit from the efficiency K-means has on large datasets. Given its popularity, the extensiveness of its documentation and own previous experience with clustering, K-prototype is a safe, reliable and well-proven choice.

As discussed in section 1, a target is to assess whether clustering is an adequate tool for characterising and identifying subgroups of patients referred to CAMHS. In order to see the maximal effect of even minor adjustments to the dataset, columns, or number of clusters, a choice has been made to only focus on one clustering technique. If one were to change between several algorithms, one might miss out on the subtle nuances that are important to be familiar with when working with clinical data. Additionally, this is not a thesis exploring the best suitable clustering method to apply to clinical data; it is primarily a thesis investigating how clustering can be used to probe latent subgroups of patients in CAMHS, and secondly on clustering as tool for analysing data from electronic health records. Thus the exploration of several techniques with the same dataset is a matter of future work, as described in section 9.3.

4. Related Work

This chapter elaborates on some of the previous research in the field of mental health in children and adolescents. As this research project is primarily concerned with the identification of patient subgroups in the context of Norwegian clinical psychiatry, we only address research on hyperkinetic disorders in children and adolescents, and not on the appliance of clustering in medical research.

Upon entering this project, it was with the comprehension that the clinical field of mental health in children and adolescents in Norway is rather unexplored, and its research feasibility uncertain. At the beginning of the collaboration with the IDDEAS project, it was clarified that we do not necessarily know what we are looking for, or if something can be found. Little to no experiments have been conducted on the available CAMHS dataset prior to this project, so the potential of the data was uncharted. Furthermore, clinical practice differs between nationalities, counties, and even local clinics, so previous research results may not be applicable outside the scope of the region the research was based on, and to the clinical reality in Norway.

There are few published papers that touch on the composite study of characterising patients with relation to hyperkinetic disorders by a clustering of referral period trajectories. However, research can be found regarding isolated subjects like hyperkinetic disorders, referral situations, rejection rates and gender differences in CAMHS. The majority of these studies are foreign. Even though the actual clinical practice differs, these studies are relevant in the context of guiding own direction of work and comparing research results. By looking into what other studies have examined, these can help to identify what to look for, and how to find it. The aim is to be able to identify phenomena to look out for when first starting to probe our own data, and review the findings in the context of previous research results on the area.

Several foreign studies indicate the potential of interesting findings regarding referring instance, referral reasons and referral assessment. A Scottish study by Smith et al. (2017) strongly suggests that referring actor is a variable with a very high impact on the referral. In their study, they found that the odds of the referral being rejected by CAMHS were significantly higher if referred by teachers. It also found that children and adolescents with emotional and behavioral difficulties were more likely to be rejected. Children and adolescents referred for hyperactivity/inattention also had significantly longer waiting times. A more recent Danish study by Hansen et al. (2021) concludes that referrals from

general physicians (GPs) were more associated with increased risk of rejection. This may indicate that young people with emotional and behavioral difficulties like hyperactivity and inattention are more likely to be rejected and to expect longer waiting times, but the referring actor with greatest risk of rejection varies from different CAMHS. Observations like these are something the upcoming experiment will look into.

Additionally, looking into internationally identified gender-based differences may yield insights of importance. An Italian study looked into gender-related clinical characteristics in children and adolescents with ADHD, and highlights that boys with ADHD are more likely to be referred for clinical assessments due to a higher prevalence of externalising symptoms (Rossi et al., 2022). Boys would have more disruptive and externalising symptoms than girls, which leads to earlier diagnostic evaluations. Boys showed higher impulsivity, while girls displayed higher levels of inattention. The study by Rossi et al. (2022) also states that available literature generally supports a higher prevalence rate of internalising disorders like anxiety and worry in girls, and a higher prevalence of externalising disorders like conduct disorder and oppositional defiant disorder, as well as symptoms like aggression and rule-breaking in boys.

Furthermore, when evaluating the severity of ADHD symptoms, the girls had more symptom severity than boys. Rossi et al. (2022) explains this as a possible referral bias, as it is possible that only the most severe girls were referred for early assessment and diagnosis. This may be because internalising disorders and inattentive aspects were generally harder to detect and less disturbing in the classroom or at home. Another study from the UK by Young et al. (2020) disagrees about symptom severity being greater in girls. They state that symptom severity may be lower in boys than in girls, particularly for hyperactive-impulsive symptoms. These symptoms may also become more obvious later in females.

However, the study by Young et al. (2020) supports the assertion that girls have more internalising disorders, and that low mood, emotional liability or anxiety may be especially common in females. It also addresses the increasing recognition that girls with ADHD show a modified set of behaviours, symptoms and comorbidities compared to boys, which also makes them less likely to be identified and referred for assessment.

Young et al. (2020) also highlight that psychiatric comorbidity is very common. This may complicate identification and treatment of hyperkinetic disorders. In children with ADHD this includes conduct disorder, oppositional defiant disorder, disruptive mood dysregulation disorder, autism spectrum disorder, developmental coordination disorder, tic disorders, anxiety and depressive disorders, reading disorders, and learning and language disorders. In adults, where comorbidity is also very common, this includes Autism spectrum disorder, anxiety and depressive disorders, bipolar disorder, eating disorders, obsessive compulsive disorder, substance use disorders, personality disorders, and impulse control disorders. Young et al. (2020) convey that the key message is not to discount hyperkinetic disorders like ADHD in females because they do not display the

behavioural problems commonly associated with the same disorder in males.

This indicates that when looking into gender-related differences, referral reasons as well as any diagnose are important variables in the data. Furthermore, internalising and externalising symptoms reported as referral reasons may have a clear separation between the genders, and less obvious symptoms like depression and anxiety may be just as interesting. One should also look into patient trajectories of boys and girls separately in order to identify possible referral patterns that differentiates between genders.

Looking within our own borders, there are rather few publications of studies that investigate and confirm the clinical reality in Norway, and similar research to my own is sparse. However, there are a few that touch on the subject of hyperkinetic disorders, and the prevalence of mental illnesses.

In their paper from 2019, Surén et al. (2018) elaborate on the diagnosis of hyperkinetic disorders among children in Norway. They state that hyperkinetic disorders are some of the most frequently used psychiatric diagnoses among children and adolescents in Norway, and also highlight that it has been shown that prevalence of the diagnoses in the F90-group greatly varies between counties.

In their study, they estimated the number of children with hyperkinetic disorders using patient data from the Norwegian Patient Registry, and also reviewed medical records from specialist mental health services for children and adolescents. Their study is concerned with the exact same diagnoses as in this project; Hyperkinetic disorders were defined as one or more entries in the medical record of the diagnostic codes F900: *Disturbance of activity and attention*, F901: *Hyperkinetic conduct disorder*, F908: *Other hyperkinetic disorders*, and F909: *Hyperkinetic disorder, unspecified*.

They essentially found that at the age of 12, 5.4% of boys and 2.1% of girls in Norway had been diagnosed with hyperkinetic disorder by the specialist health services. They also highlight the geographical variety between counties for children with this diagnosis, of which they believe the most likely explanation is regional differences in diagnostic practice.

From the medical record review, their findings indicate that hyperkinetic disorders are often less well documented than other chronic conditions. In their review of medical records, they also assessed to what degree diagnoses are reliably documented. They found that in about half of the cases, diagnose was not properly recorded. Surén et al. (2018) they conclude that there is a need to review the national guideline for evaluation and diagnostics, how it is followed in practice, and the requirements for medical record keeping.

Another article, this one only by Surén (2018), elaborates on the increased prevalence of mental illnesses in girls in Norway. Especially of interest is the number of teenage girls in the age 15-17 that are treated in the Specialist Health Service. This increase is

specifically for the case of depression, anxiety, adjustment disorders and eating disorders. A doubling of antidepressant use is reported, and teenage girls also report more mental ailments than before. However, the article concludes that we do not know whether this is due to a real increase in incidence, or any causes of it. However, Surén (2018) also emphasise that we need more information about mental illness in Norwegian youth, and suggest using national health registries and population surveys to study risk factors for mental disorders, how the disorders progress and what consequences they have in adulthood.

To summarise, international research indicate a variation in referring actors with greatest risk of rejection, but agree that the impact of the most rejected actor is significant. They also found that young people with emotional and behavioral problems can expect longer waiting times, and are more likely to be rejected. Several studies indicate that gender differences in hyperkinetic disorders are prominent, and that hyperkinetic disorders suffer high degree of comorbidity. Norwegian research on both the prevalence of hyperkinetic disorders and the development of prevalence of mental illnesses in children and adolescents has gradually taken hold and is under development. Norwegian studies also address the need for data-driven research on the area, and the need for improved medical record keeping. Future research may aid in improving journal keeping quality, and explain some of the phenomenon that have been seen in clinics nationally.

5. Data

This chapter presents the data that has been used in the research, the environments it has been processed in, which authorisations and agreements that are prerequisites for the project, and how the handling of sensitive data has been taken care of. For the data, the aim is to provide a description of the data as well as some descriptive statistics. We present this both in the context of the entire cohort, and for the subset of the cohort that has been selected for the subsequent experiment. A reasoning for the data basis in the upcoming experiment is also provided. The actual work done on the dataset, i.e. the specific selection of data as well as the process of cleaning and preprocessing, is presented in section 6.

The IDDEAS project takes use of large datasets derived from interdisciplinary patient journals, specifically from patients referred to the CAMHS clinic at St. Olavs Hospital in Trondheim (IDDEAS, 2021b). Data has been collected over the past ten years by the Norwegian Association for Child and Adolescent Psychiatric Institutions, NFBUI (IDDEAS, 2021a). The overall cohort includes all referrals made to CAMHS at St. Olavs Hospital, of which some referrals are from as early as 1982 up to 2018. When selecting the patients to include in this research project, we have extracted patients specifically with relevance to hyperkinetic disorders, and the subset includes referrals made between 01.01.1992 and 05.03.2018.

5.1. Description of Cohort

In this section, the cohort is briefly described. The purpose is to gain acquaintance with the composition of the cohort, as well as some important distributions. We firstly describe the overall cohort, which is important for later comparison with our subset of patients. Secondly we look into specific numbers for hyperkinetic disorders, both as reason for referral and as diagnosis. Some general discussion and analysis of the data is provided in order to gain understanding of the numbers.

5.1.1. General Description

There is a total of 22,643 distinct patients in the entire cohort. The earliest referral was made 07.07.1982 and the most recent referral was made 01.07.2018, as well as the latest referral period being closed as late as 03.07.2019. There are 30,938 distinct cases (i.e. referral periods) in the database, 41,411 distinct registered stays, and as much as 1,840,045 distinct journal registrations.

The data that is available covers several relevant aspects related to CAMHS patients. It covers the time from which a referral is received at CAMHS, to the case, i.e. referral period, is closed. In other words, it is possible to find data related to patient situation at the time of referral, the referring instance, the reasons for which the referral was made, the assessment, any related stays, contacts and activities, as well as any professionals involved. Each patient's electronic health record includes patient information like age, gender, relation to parents, language at home, custody, care situation and ethnicity.

The data does not include instantly identifiable information like name, address and phone number, but is defined as *potentially re-identifiable health data*, which will further be elaborated on in section 5.3.2. Data does not include more information about the involved parties before or after a referral period other than what can be derived from the clinical codings in each referral, stay, journal or patient table, or any subsequent related tables connected to these by a foreign key. [1]

Gender: The gender distribution of the cohort is quite even, and there are slightly more males than females. The gender distribution of the 22,643 unique patients can be seen in figure 5.1.

Age: CAMHS is primarily a service for children and adolescents in the age 0-18. When looking into the age distribution, age of a patient can be found by calculating the difference between birth date and the date of which a referral was made. By further disregarding records with negative age or age above 84, we get an age interval between 0 to 84. Patients over the age of 18, which are outside the target group of CAMHS, can also be disregarded, and we get an age distribution as shown in figure 5.2.

Note that patients may be in the CAMHS system for several years after the age of 18 due to treatment and transfer processes, which may be the reason for the number of patients over the age of 18 in the dataset. As can be noticed, the curve is higher on the right side, and peaks around the age of 14-16. Patients can have several referral periods, and receive several diagnoses. As the age was calculated at the time of referring the patient, this could e.g. indicate that some patients are not referred until this age, or received their latest referral at this time. It is also conceivable that since CAMHS is a service for patients in the age 0-18, CAMHS will accumulate more and more patients in

[1] Foreign keys in SQL link data in one table to data in another table in the same database, and are used to define relationships between tables. (Custer, 2021)

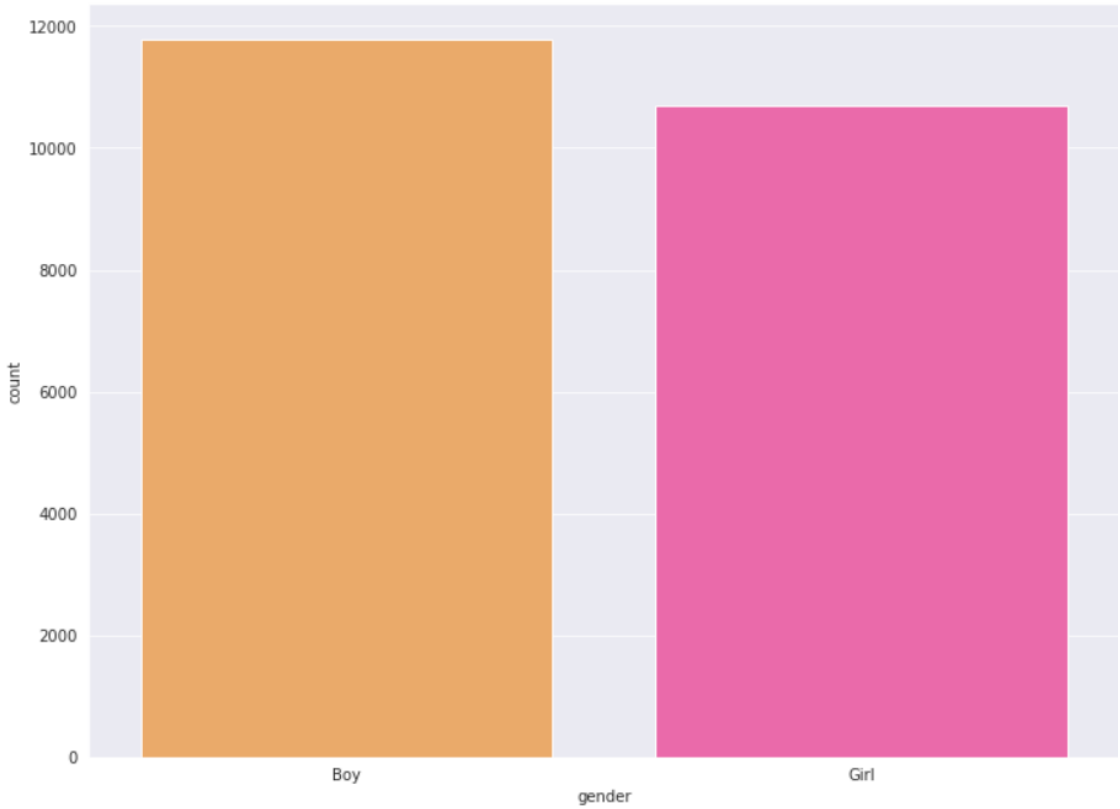


Figure 5.1.: Gender distribution in the cohort.

the higher age ranges, as both new patients come in and patients grow older.

If we specifically look into the age distribution when the condition is to only include the first referral period of every patient, the distribution looks like figure 5.3. Notice that the difference between the two figures is very small, and that the age distribution is quite similar regardless if we look into only the first referral, or both the first and any subsequent referrals. There are slightly more referrals made at the age of 0 and slightly less referrals at the age of 18 in figure 5.3.

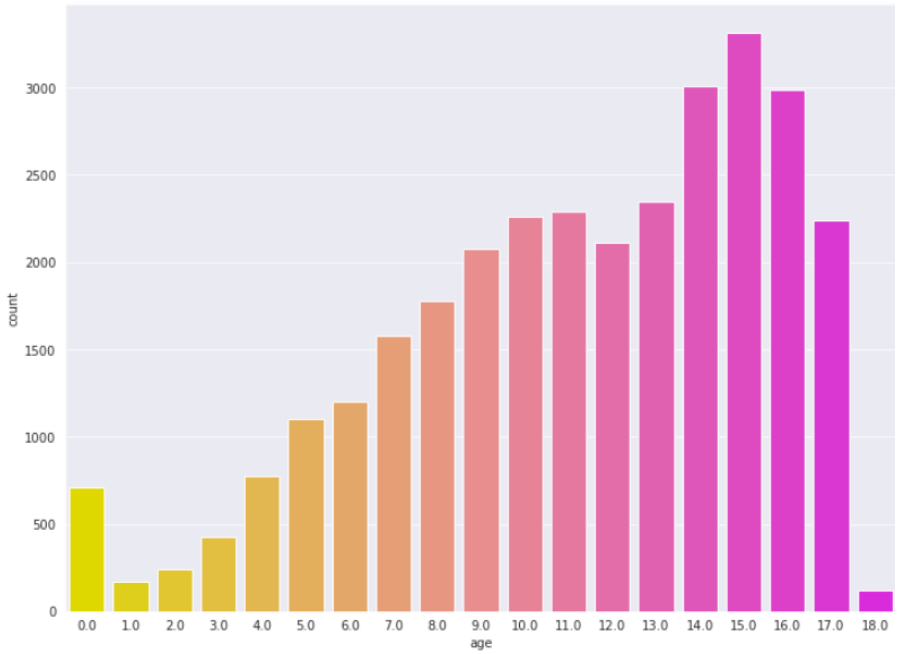


Figure 5.2.: Age distribution in the cohort.

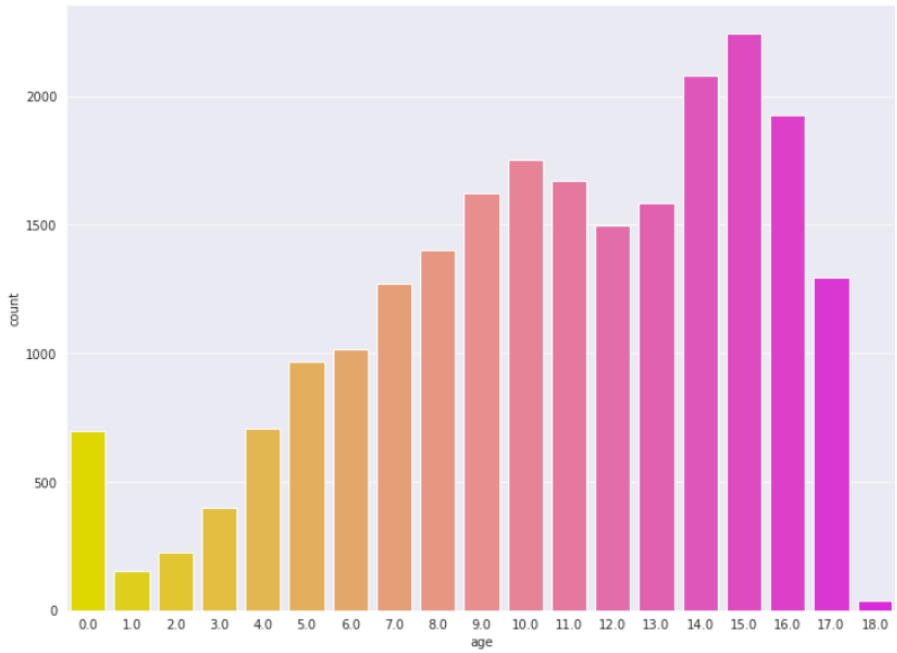


Figure 5.3.: Age distribution in the cohort for patients in their first referral period.

Assessments: When a referral is made and sent to CAMHS, the local clinic will assess the referral, and the outcome of the assessment is represented by a code from 1-4. The total number of assessments made is 30,938, same as the number of referrals. Table 5.1 presents the frequency of each assessment outcome. Note that code 2, *Rejection due to capacity*, has a significantly lower frequency than the others. According to a meeting between CAMHS and IDDEAS 18.11.2021, this is due to *Rejection due to capacity* no longer being used, and rejections today are in fact only due to professional reasons.

Code	**Map**	**Count**	**%**
0	null	917	2.96%
1	Accepted	26,545	85.80%
2	Rejection due to capacity	65	0.21%
3	Rejection for professional reasons	3,243	10.48%
4	Assessment so far	168	0.54%
	In total	30,938	

Table 5.1.: Assessment outcomes.

5.1.2. Data on Hyperkinetic Disorders

For this project, patients of interest are those with relation to hyperkinetic disorders, either by referral reason or to the diagnostic group F90: Hyperkinetic disorders. The topic of hyperkinetic disorders is thoroughly elaborated on in section 2.1.3, and it is advised to revisit that section for more information. For now, it is sufficient to know that cases with relation to hyperkinetic disorders can be identified by looking at either referral reasons, diagnoses on axis 1, or both. This section is concerned with the actual data in the context of hyperkinetic disorders, and also aims to provide reasoning for the data basis when selecting patients for the upcoming experiment.

Hyperkinetic Disorders as Referral Reason

When the referring actor is making a referral, he or she can list up to three reasons for referral, recorded as codes. In the database, these can be identified by looking at *sak.henvgrunnb1, sak.henvgrunnb2* and *sak.henvgrunnb3*. Note that *sak.henvgrunnm1-3* are related to the child's environment, and not the child itself, and is not included as a selection criteria. There are 36 referral reason codes in use in our data, including both the old and new reasons for referral. As pointed out in section 2.1.3, there are some referral reason codes commonly associated with hyperkinetic disorders. For the purpose of the upcoming experiment, we use referral reasons 3, 4, 29 and 30. These were used based on two arguments.

Firstly, in order to determine the most commonly recorded referral reasons, all cases given a diagnose in the F90-group were extracted from the database, before counting the most frequently used referral reasons used in each of these cases. This process found 3, 4, 29 and 30, in addition to 16: *other reasons* to be most frequently recorded in cases where a hyperkinetic disorder was diagnosed. It is important to highlight that when a diagnose in the F90-group was given, the most used first referral reason was predominately either 3, 4, 29 or 30, and it is reasonable to refer to these patients as *referred on suspicion of ADHD and behavioral difficulties*, as defined in the title. The count of first referral reason codes can be found in table 5.2. For the second referral reason, if any referral reason was coded at all, it was either 3, 4, 29 or 30, in addition to common comorbidities like *learning difficulties* and *suspicion of depression*. Regarding the third referral reason, it was either *learning difficulties*, *language and speech difficulties* or *suspicion of depression*, or none at all. Referral reason 16: *other reasons* was also frequently recorded. The common use of referral reason *other reasons* will also be discussed later in this thesis.

It can be expected to see comorbidities like these recorded as either second or third referral reason. As briefly mentioned in chapter 4, hyperkinetic disorders are mental illnesses with a high degree of comorbidity, i.e. the presence of one or more additional conditions (Young et al., 2020) (World Health Organization, 1992a). Thus, the main indicators or suspected illnesses are frequently recorded as first referral reason due to often being the primary concern for making the referral, and the second or third referral reason are often symptoms of common comorbidities, rather than hyperkinetic disorders.

Secondly, the decision on which referral reasons to use as basis for patient selection is based on dialogue between CAMHS and IDDEAS (Meeting between CAMHS and IDDEAS, 18.11.21), in which it was stated that code 3 and 4 are mostly used when there are symptoms relevant to hyperkinetic disorders. Referral reason 29 and 30, which are from the old set of referral reasons prior to 2009/2010, were mapped to their most similar counterparts in the new set of referral reasons. This essentially means that even though some patients do not receive an F90-diagnose in their first referral period, it has been confirmed by CAMHS that these referral reasons are commonly used when there is

Referral reason	Map	Count
4	Suspicion of hyperkinetic disorder (ADHD)	1,670
30	Hyperactivity/concentration difficulties	613
16	Other reasons	465
29	Behavioral difficulties	464
3	Suspicion of defiance/conduct disorder	268
10	Suspicion of depression	174
20	Not recorded by referring instance	114

Table 5.2.: Most frequently used first referral reason codes in cases where a diagnose in the F90-group is given.

suspicion of a hyperkinetic disorder, and we can use these as basis for identifying patients of interest.

It is important to mention that in most cases, the referrals only had a first reason recorded. Of the total of 30,938 cases, there were 30,401 first reasons registered, 11,947 second reasons registered, and only 3,478 third reasons registered. Regardless if a diagnose was registered during a referral period or not, the number of referrals made with the use of either 3, 4, 29 or 30 as first, second or third reason for referral can be found. For the total number of 30,938 referrals made, the occurrence of one of these codes appeared as the first reason for referral in 9,175 cases, second reason in 3,078 cases, and third reason in 624 cases.

Lastly, an important reflection to make is the choice of whether to include the third reason for referral in further analysis. Only 3,478 of 30,401 referrals have a third referral reason, which is a registration rate of 11.2%. However, looking into the dataset, one can find the different combinations of first, second and third reason for referral, given that the third reason is 3, 4, 29 or 30. This showed that most frequently, the first and second

reasons were either one of the other four reasons related to hyperkinetic disorders, or reasons that are commonly seen in combinations with these, e.g. *suspicion of depression*, *suspicion of anxiety*, or *learning difficulties*. This indicates that even though 3, 4, 29 or 30 are only registered as the last referral reason, they may still convey relevant information. It should thus be assessed whether to include the third referral reason in chapter 6, when the specific data selection is made. Additionally, it is important to keep in mind that some studies report that symptoms like anxiety and depression are more commonly seen in females than in males even when the patient is diagnosed with a mental illness like ADHD (Young et al., 2020). Due to this, the second and third referral reasons can be important identifiers for hyperkinetic disorders in groups of patients that are by bias not commonly associated with hyperkinetic disorders, like for example girls.

Hyperkinetic Disorders as Diagnose

Before closing our introduction to the data, we briefly visit the topic of hyperkinetic disorders as diagnose. As hyperkinetic disorders are coded on axis 1 in the multi-axial classification system, these can be identified as *sak.icd1* in the database[2]. Of a total of 28,461 registrations on axis 1 among the total number of cases, there were 3,831 registrations of F900, 509 of F901, 155 of F908 and 33 of F909. In the identification of patients with relevance to hyperkinetic disorders, these diagnostic codes are all equally important, as described in section 2.1.3.

Furthermore, F900 is the third most frequently recorded registration on axis 1, regardless of diagnostic group and referral reason. It is actually also the most commonly recorded diagnose, as the preceding registrations are not formally defined as diagnoses. The most frequent registrations are firstly no diagnose, i.e. the patients do not receive any diagnose during their referral period (5,304), secondly Z032, an observational code only used as an additional, and not diagnostic, code (4,037), and then F900 (3,831), classified as *Disturbance of activity and attention*. This is also consistent with hyperkinetic disorders being both the largest diagnostic group and reason for referral in Norway, according to psychologist specialist Jostein Arntzen (meeting with Jostein Arntzen, 11.05.2022).

Either being referred with one of the four referral reasons discussed in this section, or having a diagnose in the F90-group, i.e. F900, F901, F908 and F908, are the criteria when selecting patients in their first referral period in the later experiment.

[2]Note that registrations on axis 1 for each patient can be found in several tables in the database, but we mostly use the table named *sak* in this project. This field may also be referenced as ICD1 in further analysis.

5.2. Environments

This section presents the different environments that have been used for working with the data during this research. Most importantly is HUNT Cloud, through which the IDDEAS project has access to data and tools.

HUNT Cloud is a scientific computing environment located at NTNU in Norway. According to their own website, HUNT CLOUD focuses on technologies at scale in scientific computing on sensitive data, and provides structure and services for handling and processing huge data quantities (HUNT Cloud, 2022b). Solutions are provided through Cloud services which enable the user to view, process and retrieve results without doing the computations locally. That way, the sensitive data never leave the cloud, and heavy computations are independent of the user's personal computer. The personal computer simply displays the processes and results happening on the HUNT Cloud service side. More importantly, all handling of data is in accordance with security regulations and privacy rules.

As a member of the IDDEAS project, one is also granted access to the IDDEAS digital laboratory, which includes access to all CAMHS data and file management tools. As a lab user, one can request access to several useful tools like the Workbench. Workbench provides smooth access to modern data science tools such as Jupyter Notebooks, Python, RStudio, R, Stata notebook or MATLAB (HUNT Cloud, 2022a).

Workbench was used for the purpose of downloading Python packages and libraries, data preprocessing, exploratory data analysis, data visualisation, clustering, and result analysis. Most importantly was Jupyter Notebook, which has been included in its entirety in appendix D. MobaXterm, an application that simplifies SSH connections to the lab from a local Windows machine, was used primarily for file management. X2Go Client, software that enables instant access to graphical tools in the lab (HUNT Cloud, 2022a), was used for facilitating rapid launch of DBeaver, a database management tool (DBeaver, 2022), and SPSS, a statistical tool by IBM (IBM, 2019). Dbeaver is free and suitable for managing our PostgreSQL-database, and was used to make data selections (queries) and retrieve the dataset. SPSS is licensed software, of which the license was provided by NTNU. In this project, SPSS version 27 was used for descriptive statistical analysis of numerical and categorical data.

5.3. Data Approval and Agreements

Data used for research purposes must be handled respectfully, and the research must comply with the data user agreements. This section is concerned with the authorisation and approvals given for the research, the handling and processing of sensitive data, what

it includes, and how handling of sensitive data is carried out in practice throughout this research to ensure that the work is in compliance with data requirements.

5.3.1. Legal Project Approval

All Norwegian research in health and and medicine requires prior approval from the Regional Committee for Medical and Health Research Statistics (REC), and the approval must be available before initiation of a project (De nasjonale forskningsetiske komiteene, 2014).

The IDDEAS project was reviewed by REC 09.10.2021 (Case 2018/2186). REC made the following decision:

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.

REC justified its decision on the grounds that if the project was successful, it will be of significant interest to society.

The project was granted access to data by application to the regional health authority (St. Olavs Hospital, formal owners of health data from CAMHS patients), upon completion of a Data Protection Impact Assessment (DPIA) and a risk- and vulnerability analysis, in addition to the recommendation by REC. An extensive, time consuming and resource-intensive process.

5.3.2. Agreements

Prior to being admitted in the IDDEAS project, a non-disclosure agreement (NDA) had to be signed. The IDDEAS project receives potentially re-identifiable health data, which means that identification is possible if one has knowledge of personal characteristics of a given identity. The statement made by signing the NDA is as strict as what the health personnel are subject to.

Additionally, a user agreement for HUNT Cloud services had to be signed in order to become a lab user and gain technical privileges, before gaining access to the digital project laboratory. All connections to the lab is made through HUNT VPN with a rotating verification code on a personal phone as key. Login to lab is made by the use of a SSH-tunnel.

All data management, viewing, extraction, processing, visualisation, analysis and statistical calculations are strictly conducted inside the HUNT Cloud environment, and raw data never leaves this environment.

5.3.3. Data Classification

According to the HUNT Cloud security section, there are two levels to their data classification (HUNT Cloud, 2022c).

- **Sensitive data:** Research data that can indirectly identify research participants. E.g. Individual level data such as phenotype data and genotype data.
- **Internal data:** Research data that can not identify research participants. E.g. Summary statistics, figures, computer code, non-human data, and encrypted sensitive data.

Immediately identifiable health information can not be stored in labs by default, like names of research participants, personal identification numbers, phone numbers, address information, and so on. These must be stored outside the HUNT Cloud system, and hence are not available information through the IDDEAS lab. Storage volumes are classified as sensitive by default, and may only be declassified to internal in agreement with the respective data controllers and lab owners (HUNT Cloud, 2022c).

5.3.4. Traceability of Research Results

Research results in this project visualise and summarise the general characteristics of the data, without compromising personal information or present the information in such a manner that it is possible to recompose the sensitive information and identify personal characteristics.

In processing the data, identifiable information like birth date and referral date have been removed. Dates have been used in the calculation of age and for the selection of the first referral period of every patient, without giving reference to a specific point in time. Patient situation, i.e. situation prior to referral, referral situation and assessment situation are never presented linearly or composed, but as a frequency or mean value in an overall group.

The declassification from sensitive to internal classification of the data derived from the experiment, like cluster results, summary visualisations and cluster descriptions, has been consolidated and allowed by the respective lab owner. Due to this, the sharing and distribution of research results will not compromise sensitive information, and research results are confirmed to not be traceable.

6. Experiment and Results

This chapter thoroughly presents the experiment that has been conducted, a key component of the research. This includes the implementation plan, the process of cleaning and preprocessing the data, the exploratory data analysis (from now on referred to as EDA), the clustering process, as well as any reiterations of these steps in order to perform the experiment according to the initial vision.

Section 6.1 covers the aim of the experiment, and a brief overview of the experimental plan. This also includes a tentative time frame. Section 6.2 presents the complete, detailed experimental setup in a manner that enables reproducibility, as well as an EDA. The complete EDA includes pre-analysis statistics for the dataset, as well as a pre-clustering data analysis. After this section, section 6.4 is concerned with the experimental execution, in which the optimal number of clusters is determined.

Lastly, the experimental results will be presented towards the end of this chapter, in section 6.5. The evaluation of the experiment is presented in chapter 7, and the discussion itself will be presented in chapter 8.

All steps of the experiment are thoroughly described with the aim of reproducibility and the ambition that initial work with clustering clinical datasets can be improved and continued in the future, and that any research limitations can be identified and mitigated. Working with clinical data is exciting, but tedious work, and prone to bias, misinterpretation, and faulty data selection. Some of the countermeasures for these challenges are described in section 6.2.

6.1. Experimental Plan

The target for this project is to conduct one experiment with the selected dataset. This has been chosen on the basis of feasibility, implementation complexity, its potential for interesting results and the given time constraints. It is expected that the experiment will be a process of trial and error, but all relevant results will be presented at the end of this chapter.

6.1.1. Experimental Aims

Experimental aims are defined in the list below. The aim is first and foremost to investigate whether a clustering of clinical patient data can yield interesting results, like the identification of latent subgroups of patients, clinical profiles, or patient situation similarity in CAMHS. List item 1 and 2 summarises the main objectives. Secondly, it must be assessed whether it is even feasible; clinical data like this, i.e. manually entered records with the additional challenge of a code system renewal in the middle of the data coverage period, may prove to be challenging. Furthermore, should the process be successful and yield valuable findings, additional aims related to gender differences, patterns of rejected patients and referral impact have also been included.

1. **Identify separated groups of patients that have similar situation and trajectories.**

2. **Assess the feasibility and usability of clustering as a tool for identifying latent subgroups of patients based on clinical data.**

3. Identify gender-related patterns of interest, by investigating whether boys and girls have different or similar patterns related to referring instance and referral reason that lead into the treatment or diagnose of hyperkinetic disorders.

4. Assess whether clustering is able to describe the rejected part of the cohort, and use the clusters as assistance for identifying patient situations that are more prone to rejection.

5. Identify the importance and impact the different stages of the first referral period have for the outcome of the clusters.

These aims are to be explored during the experiment, and will be revisited and assessed in section 7.2.2.

6.1.2. Experimental Steps

There are several steps in performing a clustering experiment. For this research project, these can roughly be described as:

1. **Data selection**. The dataset to be used is extracted from the database. In many other real-word examples, the dataset is already extracted, which will require some more data preparation in the next step. If one has the option of retrieving the dataset from the database itself, this will enable the user to be more restrictive in which rows and columns that are to be extracted.

2. **Data cleaning and preprocessing**. When clustering, it is desirable to have as few null or invalid values as possible. We essentially want to apply clustering to complete data (Boluki et al., 2020). Different data imputation techniques are enforced to handle missing codes, invalid codes, or null values. Rows that are overall deficient are removed, as well as irrelevant columns or columns that highly contribute to null values. Categorical value names are shortened (may also be done in the former step, dependent of extraction method). Continuous features (numerical columns) are standardised in order to ensure that one feature is not more important than the other. It is common for features to have different units of measurement, so it is good practice to standardise the data to have a standard deviation of one and a mean of zero (Kassambara, 2021). This will make the variables comparable, independent of scale.

3. **Exploratory Data Analysis**. This step is not mandatory in a clustering process, but will give initial insight into the dataset we are working with. It is a valuable analysis that can summarise the main characteristics of the data.

4. **Clustering dataset**. In order to find the optimal number of clusters, a modeling technique for identifying this is employed to find a sufficient number of k clusters. A clustering algorithm of choice is used to cluster the dataset that has been preprocessed, with the number of optimal clusters k as input. This process may require some trial and error in an attempt to find the number of clusters that is suitable for the type of data used in the experiment.

5. **Analysing results**. Results are presented and interpreted. The quality of the clustering output is an iterative and exploratory process. Results can be verified against expectations, and be improved by running new iterations of the previous steps.

6. **Result evaluation and discussion**. Clustering results are evaluated by looking at separation, meaningfullness, and by employing evaluation techniques like a SHAP-plot. Experimental results are consulted and discussed in collaboration with professionals. Findings are discussed and key takeaways from the research are defined.

Any of these steps may repeatedly be done in order to find a sufficient data selection, number of clusters, or collection of clusters. As mentioned in the last step, the ambition is to discuss possible findings with CAMHS professionals or other resourceful professionals. This is done in order to better evaluate the meaningfulness and usefulness of the research, and to validate the findings of the experiment as both an experiment evaluation and as validation of research discoveries.

All of these steps are thoroughly described throughout the next sections, as we look into the details of the experiment.

6.1.3. Experimental Time Frame

This section presents the time frame of the experiment. It is first and foremost concerned with the implementation of the experiment, and not on the thesis itself.

Experiments are prone to changes, pitfalls, rejection of solutions, change of methodology, and the researcher being overambitious, just to name a few. Due to this, experimental plans may need to be flexible and account for enough time for delays and other unexpected changes, but also drive the project forward by having time-related goals. In this experiment, these are defined by the latest week each stage must be implemented and completed.

There are five main intermediate goals to this experiment. These are: Definition of experimental scope and ambition, preprocessing and clustering of the dataset, result analysis, clinical validation, and evaluation and discussion of results. In order to facilitate the implementation of these stages, and mitigate delays, enough time must be set aside.

Latest week	**Activity**
8	Definition of experimental scope and ambition
10	Data selection
12	Preprocessing and clustering of dataset
16	Experimental result analysis
18	Clinical validation of experimental findings
22	Submit Master's thesis

Table 6.1.: Time frame of experiment.

6.2. Experimental Setup

This section presents the complete experimental setup and implementation, including tools, data selection as well as data cleaning and preprocessing. For reproducibility purposes and future work on the area, this is done in very much detail. Important assumptions and remarks are also included, to better facilitate further work and aid the identification of potential errors. This also applies to any subsequent sections in the experiment.

6.2.1. Tools

The tools used for the experimental part of this project are all services provided by or facilitated by HUNT Cloud, which have all been discussed in section 5.2. These are summarised below.

- **MobaXterm**: Digital laboratory access and file management.
- **DBeaver**: Database management and data selection.
- **X2Go**: Rapid access to DBeaver and SPSS software.
- **HUNT Cloud Workbench**: Access to the Jupyter Notebook, setup of environment, data preprocessing, cleaning, EDA-analysis, clustering and result interpretation. All Python libraries and packages are listed in the Jupyter notebook, which can be found in its entirety in appendix D.
- **IBM SPSS Software version 27**: Statistical analysis of the clustered dataset.

6.2.2. Data Selection

In this step, the data of interest is selected from the database. Having access to the database and being able to extract exactly the kind of data that is interesting, put me in a unique position of doing a lot of the data selection and preprocessing at the earliest stage. This meant both being able to precisely select the data I wanted, but also making as many reasonable limitations and requirements as I found fit. This ensured that the amount of cleaning and preprocessing could be reduced. An underlying challenge throughout this project is the poor quality of data, as will be thoroughly elaborated on in 7. This requires a careful selection and combination of patient records in order to ensure a minimum number of rows to work with.

To keep the complexity of the clusters to a reasonable level, the data covers the first

referral period for each patient, regardless if they have more referral periods. Every referral period is identifiable by a case number. This case number is used to identify any associated stays, journal entries and diagnoses. Furthermore, the data covers the patient's situation prior to the referral (care situation, relation to care takers etc.), the referring time, reason, type and instance, as well as the assessment of the referral upon arrival at the CAMHS clinic. Furthermore, the episodes related to each referral period are included, as well as the different types of contact related to each episode. This means that we have information of the patient's personal situation at home, and insight into the time from the first referral period to the end of the first referral period.

When selecting features for clustering analysis, there are some things to account for with regards to clustering and data quality. We want to avoid columns with a lot of unique values, and columns that have missing values. Due to the nature of clinical data, these columns are common. Because of this, the process of selecting data was by far one of the most time consuming steps. As these datasets are manually entered records for each patient, case, stay, diagnose and journal, they are prone to error, inconsistencies within and between different practises, mistyping and missing values. Essentially, the data quality is not optimal. To mitigate this at the earliest stage, the columns were restricted to only include the codes that are valid and have an existing mapping. Furthermore, it is desirable to only include columns that add value to the clustering, i.e. they are meaningful and relevant. Less meaningful columns may advantageously be removed. For a more thorough explanation of the choices made for this experiment, see section 7.4.4.

To join the different tables, *sak.nr* (i.e. case number for each referral period) is used as join criteria, as associated referral periods and journal entries can be identified by this value. This also ensures that only the rows related to the specific case number for each patient is extracted.

As already mentioned, new referral reasons replaced the old ones in 2009/2010. Even though some codes from the former coding system are present in the column for referral reason, the inclusion of the patients belonging to this time period do not interfere with the mapping of the rest of the codes in the other columns. It was made sure that any column included in the data selection, had the same code mapping prior to and after the new coding system was introduced. However, this tactic does not apply to all columns, so in the case of reproducing similar results or for future work, it is urged to be aware of this.

The PostgreSQL-query in its entirety can be seen below.

PostgreSQL-query for data selection:

```
select
        pasient.kjonn as "gender",
        date_part('year', age(sak.henvdato, pasient.fdt)) as "age",
        pasient.omsorg1 as "care",
        pasient.foreldre as "custody",
        pasient.morrelasj as "mrelation",
        pasient.farrelasj as "frelation",
        sak.instanskode as "refinstance",
        sak.henvgrunnb1 as "refreason",
        sak.tattimot as "assessment",
        sak.icd1,
        nrStays,
        sak.avslkode as "closingcode",
        sak.etterkode as "aftercode"
from
        sak
left join pasient on
        sak.pasient = pasient.nr
inner join (
        select
                opphold.sak,
                count(distinct(opphold.nr)) as nrStays
        from
                opphold
        group by
                opphold.sak) as opphold on
        sak.nr = opphold.sak
right join (
        select
                pasient,
                min(henvdato) as "henvdato"
        from
                sak
        group by
                pasient) as oldestCase on
        (sak.henvdato = oldestCase.henvdato
                and sak.pasient = oldestCase.pasient)
where
        ((sak.henvgrunnb1 = '4'
                or sak.henvgrunnb1 = '3'
                or sak.henvgrunnb1 = '29'
                or sak.henvgrunnb1 = '30'
```

```
                or sak.henvgrunnb2 = '4'
                or sak.henvgrunnb2 = '3'
                or sak.henvgrunnb2 = '29'
                or sak.henvgrunnb2 = '30'
                or sak.henvgrunnb3 = '4'
                or sak.henvgrunnb3 = '3'
                or sak.henvgrunnb3 = '29'
                or sak.henvgrunnb3 = '30'
                )
        or (sak.icd1 = 'F901'
                or sak.icd1 = 'F900'
                or sak.icd1 = 'F908'
                or sak.icd1 = 'F909'))
        and (date_part('year', age(sak.henvdato, pasient.fdt)) > -1
                and date_part('year', age(sak.henvdato, pasient.fdt)) < 19)
        and (pasient.morrelasj > 0
                and pasient.morrelasj < 10)
        and (pasient.farrelasj > 0
                and pasient.farrelasj < 10)
        and (pasient.omsorg1 > 0
                and pasient.omsorg1 < 10)
        and (pasient.foreldre > 0
                and pasient.foreldre < 5)
        and (sak.instanskode > 10
                and sak.instanskode < 79
                and sak.instanskode != 30)
        and (sak.henvgrunnb1 > 0
                and sak.henvgrunnb1 < 40)
        and (sak.tattimot > 0
                and sak.tattimot < 5)
        and (sak.avslkode > 0
                and sak.avslkode < 10)
        and (sak.etterkode > 0
                and sak.etterkode < 6)
        and (pasient.kjonn > 0
                and pasient.kjonn <3)
group by
        sak.pasient,
        sak.nr,
        pasient.kjonn,
        "age",
        pasient.omsorg1,
        pasient.foreldre,
```

```
        pasient.morrelasj,
        pasient.farrelasj,
        sak.henvdato,
        sak.henvgrunnb1,
        sak.instanskode,
        sak.tattimot,
        sak.icd1,
        nrStays,
        sak.avslkode,
        sak.etterkode
order by
        sak.pasient
```

Pay notice to the first selection criteria in the *where*-clause. To identify and select patients of interest, we want to extract patients that have either, or both:

1. Been referred with referral reasons (either firstly, secondly or thirdly) related to hyperkinetic disorders and behavioral difficulties, i.e. 3, 4, 29 or 30

2. Have a diagnose in the F90-group on axis 1

The reasoning for this can be found in section 2.1.3. This also includes patients that have relevant referral reasons, but (in this context) an unrelated diagnose, or no diagnose at all. Many patients do not have nor will ever receive a diagnose, but that is just one of the features to help us locate the cohort of interest. However, it is important to keep in mind that any diagnose that some patients have in this dataset, is relevant for the first referral period only. They may be given their first or another diagnose in a later referral period, which this dataset does not account for. Also, every diagnose any patient may have, is given after admission to the CAMHS clinic and not by the referring instance.

Furthermore, note that some characteristics, like second and third reason for referral, have been included as a part of the data selection, but not as columns for patient clustering. This is because these columns contain information about the different combinations of referral reasons for patients that have been referred for reasons relevant to hyperkinetic disorders, but the choice was made to not include them as columns for clustering because too many patients do not have a second or third reason for referral. For reference, see section 5.1.1. As elaborated on in chapter 4, many patients with relation to hyperkinetic disorders also suffer a high degree of comorbidity (Young et al., 2020). For this experiment, that essentially means that even though the most common referral reasons related to hyperkinetic disorders are not listed first, it may be listed second or third after symptoms that are common comorbidities. That way, all the patients of interest are included in the cohort, without having to handle null or invalid values.

Figure 6.1 illustrates how we define a referral period, and which variables it entails for

this experiment.

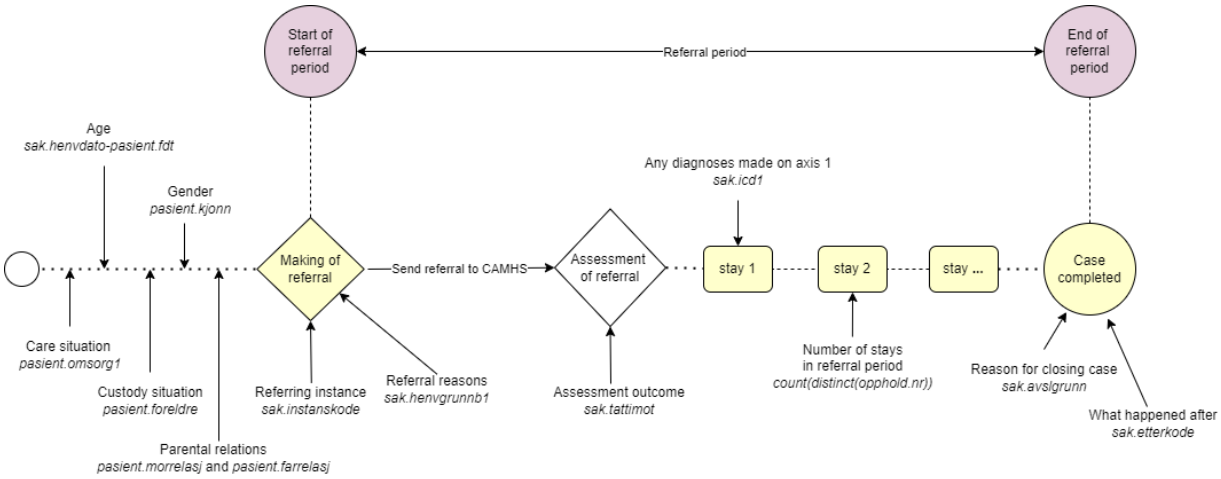


Figure 6.1.: Flow chart and timeline for referral period.

Table 6.2 describes the content of each column, as well as any associated database field. To keep the list short, only a description of each column is provided. The complete code list mappings can be found in appendix A.

6.2.3. Data Cleaning and Preprocessing

This section covers the process of data preparation. Due to the majority of data cleaning being done by careful selection of features in section 6.2.2, this step is mainly considered with the mapping of categories to their textual representation, and data imputation techniques for any missing values. Note that if there were any missing mapping after the preprocessing stage, this was detected during the actual clustering process, and enabled the opportunity to go back and improve any missing mapping. Lastly, the numerical values are standardised.

After selecting the dataset and translating the column names, it was extracted as a .csv-file. Having both numerical values for variables like age, and then categorical values like gender and custody represented by numbers, this would usually require some transformation for applying the similarity measure to both the numerical and categorical values. However, having chosen K-prototype as clustering method, this algorithm does not require any preprocessing to categorical data. This is beneficial because methods like one-hot-encoding (Scikit-learn, 2022a), which transforms categorical data to binary categories, would expand the dimensionality of the dataset in an adverse way.

Column	Field	Description
Gender	pasient.kjonn	Gender of each patient.
Age	sak.henvdato and pasient.fdt	Age of patient, calculated from birth date and date the first referral was made by referring actor.
Care situation	pasient.omsorg1	What kind of care situation the patient has at the time of referral.
Custody	pasient.foreldre	Which parent, if either, that has custody of the child.
Relation mother	pasient.morrelasj	The patient's relation to mother figure.
Relation father	pasient.farrelasj	The patient's relation to father figure.
Referring instance	sak.instanskode	Actor or instance making the referral.
Referral reason child	sak.henvgrunnb1	Reasons for making the referral, related to the child itself.
Assessment code	sak.tattimot	Outcome of assessment, whether the referral was accepted for admission or rejected.
ICD1	sak.icd1	If filled out, any diagnose the patient may have received during one of their stays in their first referral period.
Reason for closing	sak.avslgrunn	The reason for closing a case.
After code	sak.etterkode	What happened to the case after it was closed.
Number of stays	count(opphold.nr)	The number of stays each patient had during their first referral period.

Table 6.2.: Dataset column description.

Code Map

The mapping of clinical codes are based on the BUPdata to NPR (i.e. Norwegian Patient Register) code mappings, which are described in internal system documentation by

Hiadata AS, later Visma Unique. This was last updated 24.03.2010.

Since our categorical data are represented as numbers, we need to map every categorical code to its corresponding textual representation. This will essentially make every count of something (e.g. *age* or *number of stays*) to be represented as numbers, and every occurrence of a categorical code represented by text (e.g. *suspicion of ADHD* or *accepted*). This process also requires numerical columns to be standardised, which is covered in section 6.2.3.

In most cases, there is a direct map between the codes for both NPR and BUPdata. This meant there would be no incorrect mapping even though the dataset contains records prior to the replacement of the old code system. This was a trade-off process when selecting columns.

One example is that by including the old codes for referral reason, we would get a larger cohort of patients to work with. In exchange, some of the columns that had different mapping prior to and post the new code system, would have to be excluded. The trade-off essentially was between old referral reasons and *sak.saktype*, case type. Looking into the records with a registered case type, it seemed to be interchangeably which was used when, even for apparently similar referrals. Excluding case type would yield 4,274 rows. Excluding old referral reasons would yield 2,606 rows. Needless to say, case type was excluded.

Some codes were merged. This was the case with the assessment codes, *sak.tattimot*. According to the meeting between CAMHS and IDDEAS 18.11.2021, CAMHS does not have the option of rejection due to capacity; every rejection must be professionally justified. If the code for *rejection due to capacity*, it is rather a case of *rejection due to professional reasons*. Thus, these two have been merged in the mapping process. This leaves three codes; Accepted, rejection, and assessment so far.

Additionally, some columns were also merged. After mapping *pasient.morrelasj* and *pasient.farrelasj* to their respective textual values, these were merged to one column representing the patient's combination relationship to its parents or care takers. This reduced the columns from 13 to 12.

The remaining part of the mapping process was concerned with the map of old to new referral codes. In the same meeting between CAMHS and IDDEAS 18.11.2021, it was stated that there is no direct mapping between the old codes and the new ones. However, many of these share very similar definitions, and can be grouped in a meaningful way. The purpose is to assemble the referral reasons that are quite similar, so that patients that are clustered do not appear as different when they in fact have the same referral reasons. What is most important is that the data analyst knows the composition of every group. Thus, for codes that do not have any reasonable map, these may very well be stand-alone categories or included in the *other reasons*-category.

For old codes that were very similar to new codes, these were directly mapped, e.g. *eating problem* and *suspicion of eating disorder*. For codes that were not similar to any new codes and had a frequency of less than or equal to 10 rows, were put in the *other reasons*-category. There were 6 old referral reasons that were put in the *other reasons*-category, with a total count of 29 rows.

Lastly, codes that were not similar to any new codes and had a frequency of more than 10 rows, were kept as stand-alone categories. The complete overview of these categorical mappings can be seen in table 6.3 and 6.4. A complete overview of all the mappings between codes and categorical values can be found in Appendix A.

Most columns have been translated to English for readability purposes. However, referring instance names and referral reasons have been kept in Norwegian due to interpretability purposes in the upcoming clinical validation. A translation of both referring instances and referral reasons can also be found in appendix A.

Note that this process was a matter one should preferably consult with a local CAMHS clinic, but due to unavailability of resources on their end and time constraints on my own, this was done without a verification by CAMHS. However, it was consulted with and approved by other nearby sources in the IDDEAS group.

New code	Map	Old code	Map
1	Alvorlig bekymring for barn under 6 år		
2	Mistanke om gjennomgripende utviklingsforstyrrelse (autimse)	21	Autistiske trekk
3	Mistanke om trasslidelse/adferdsforstyrrelse	29	Atferdsvansker
4	Mistanke om hyperkinetisk forstyrrelse (ADHD)	30	Hyperaktiv/konsentrasjonsvansker
5	Mistanke om Tourette syndrom		
6	Skolevegring		
7	Mistanke om angstlidelse	25	Angst/fobi
8	Mistanke om tvangstanker / tvangshandlinger	26	Tvangstrekk
9	Mistanke om spiseforstyrrelse	36	Spiseproblem
10	Mistanke om depresjon	27	Tristhet/depresjon/sorg
11	Mistanke om bipolar lidelse		
12	Vedvarende og alvorlig selvskading		

Table 6.3.: Part 1: Mapping of old to new codes for referral reason (*sak.henvgrunnb1*)

New code	**Map**	**Old code**	**Map**
13	Mistanke om psykose	22	Psykotiske trekk
14	Alvorlige psykiske reaksjoner etter traumer, kriser eller katastrofer		
15	Alvorlige psykiske symptomer sekundært til somatisk sykdom		
16	Annet	38	Annet
		31	Rusmiddelmisbruk
		35	Syn/hørselsproblem
		34	Språk/talevansker
		32	Asosial/kriminalitet
		37	Andre somatiske symptomer
20	Ikke fylt ut av henviser	39	Ingen
		23	Suicidalfare
		24	Hemmet atferd
		28	Skolefravær
		33	Lærevansker

Table 6.4.: Part 2: Mapping of old to new codes for referral reason (*sak.henvgrunnb1*)

Data Imputation

Due to the careful and strict selection of data in the first step of the experiment, there were few records with zero or invalid values. However, *relation to mother* and *relation to father* have similar code map, i.e. one can theoretically map the relation *biological father* to a record in the column *relation to mother*. The imputation technique used in this case, is an intuitive one rather than traditional imputation methods like mean, median or mode (Secherla, 2021) because there is theoretical basis for believing that *biological father* was supposed to be coded as *biological mother*. Using statistical methods would in fact not be as reasonable in this case, but were of course considered. Furthermore, it must be seen in comparison to the other parental relation. If *relation to mother* is coded as *biological mother*, and *relation to father* is also coded as *biological mother*, it is reasonable to believe it should in fact be *biological father*. Another example is if *relation to mother* is coded as *adoptive mother*, and *relation to father* is also coded as *adoptive mother*. It is more reasonable to believe that both parents are adoptive parents than one of them being biological.

Numerical Standardisation

As mentioned in section 6.2.3, numerical columns need to be normalised, because they have different scales. Even though this is a step in the preprocessing stage, note that this was actually done post EDA to ensure standardised data was not included in the analysis.

Standardisation can be done by using appropriate techniques, such as min-max normalisation, Z-score normalisation, or similar techniques (Aprillant, 2021). Standardisation prevents variables with larger scales from dominating how clusters are defined. This essentially means that all numerical features will contribute evenly.

For this experiment, Z-score normalisation, often referred to as standardisation will be applied. Z-score transforms the data into a distribution of values where the mean is 0 and has a standard deviation of 1. To achieve this, the StandardScaler-class from the sklearn-module is used and applied to all numerical values (Scikit-learn, 2022b).

6.3. Exploratory Data Analysis

This section presents the exploratory data analysis (EDA), a valuable analysis usually conducted prior to clustering a dataset.

The purpose of an EDA is to analyse and investigate datasets and summarise their main characteristics. This often also involves employing data visualisation methods to build comprehension and derive insight (IBM, 2020b). The aim is to find interesting points that can be useful for capturing the phenomenon in the data, discover patterns, spot anomalies or outliers, or check assumptions. It is suitable for both detecting errors, and to better understand the data.

The EDA and findings will be presented prior to the clustering experiment. For multivariate data, we will be using bar plots and scatter plots for graphical representation of the relationships in the data. Some plots have been made with vertical bars to ensure enough room for longer column names. The reader is encouraged to keep in mind the selection criteria for the dataset when looking into these visual representations, as described in section 6.2.2. Even though there are many interesting variables to compare, specifically gender and assessment outcome have been used due to the their low dimensionality. Age, which is one of two continuous variables, has frequently been used as the numerical reference value.

6.3.1. Pre-Analysis Statistics

Before moving on to an EDA, we look into a statistical description of the dataset. Upon inspection of the categorical and numerical values, one can find the number of unique values and a statistic summary. This will also help identify anomalies.

Figure 6.2a presents the number of unique values for each column, and confirms that there are no more redundant categorical codes. Figure 6.2b presents a numerical summary prior to standardisation of the continuous variables. As can be seen, the youngest patient in the cohort is 1 years old, while the oldest is 18. However, the mean age is 9.6. Even though one patient has as many as 24 stays in connection to their first referral period, the average patient has 1.4 stays. Notice that every patient has at least one stay, regardless of assessment outcome.

```
gender          2
care            9
custody         4
relation       10
refinstance    24
refreason      21
assessment      3
icd1           56
closingcode     9
aftercode       5
dtype: int64
```

(a) Number of unique categorical values.

	age	nrstays
count	4201.000	4201.000
mean	9.555	1.365
std	3.446	1.107
min	1.000	1.000
25%	7.000	1.000
50%	9.000	1.000
75%	12.000	1.000
max	18.000	24.000

(b) Statistical summary for age and number of stays.

Figure 6.2.: Count of unique values and numerical statistical summary of the dataset.

Lastly, a count of every unique value for every column was conducted, which through multiple iterations unveiled several outlier patients, e.g. patients outside the age limitations. Additionally, every diagnostic code with frequency less than three were excluded from the dataset in order to limit the number of unique categorical values, in accordance with clustering guidelines. This process reduced the number of unique diagnostic codes from 102 to 56, and the dataset from 4,274 to our current dataset of 4,201 records.

Rejected Cohort

Before moving on to the EDA of the selected cohort, a brief statistical summary of the rejected part of the selected cohort is presented. This will give some initial indications of variables that are commonly observed with rejected referrals.

In the rejected cohort, there is a total of n=452 rejected patients (10.8% of entire cohort: n=4,201). Median age is 9 with a variance of 12.1. 68.4% are boys (n=309) and 31.6% are girls (n=143). In 86.7% of the cases, both mother and father have custody of the child, followed by mother having custody alone (9.3%). 58.2% live with both parents (n=263), then secondly 16.2% live with one parent, then thirdly 12.6% commutes between both parents.

The general physician (GP) is the most common referring instance at 56.4%, with n=255 of n=452 total rejected patients. The GP is followed by the Help Service for Children and Young People at 17.0%. The most common primary referral reason is *suspicion of defiance/conduct disorder* at 54.4% (n=246). This is followed by *suspicion*

of ADHD (30.0%, n=136), and with a large drop, *learning difficulties* at 2.7% (n=12).

6.3.2. EDA

This section covers the EDA, as well as any related comments to the visualisations. Due to the great potential an EDA has in unveiling useful insights and relevant discoveries, it has been given considerable time and effort in the thesis.

Bar Plots

This section presents the selected cohort with bar plots in order to build comprehension of the patients in our dataset. Some of these plots are seen in comparison with variables like gender and assessment outcome in order to detect initial findings of interest.

Age: Figure 6.3 displays the age distribution in the cohort. It can be noticed how the age distribution for girls is more skewed to the right in comparison with boys.

Figure 6.4a and 6.4b compares the age distribution for the patients in our dataset, and patients in the overall total cohort regardless of referral reasons and diagnose, in their first referral period. Notice how the patients in our data selection have their first referral period at an earlier age than the patients in the entire cohort. The majority of patients are between 7 and 11 years old when they have their first referral period. Compared to figure 6.4b which displays the age distribution of the entire cohort, the majority of the patients in our data selection are younger: as much as 5-8 years. The query for the selection of the entire cohort can be found in appendix B.

Note in figure 6.4b that the count scales are different, as the y-axis of the right plot is about five times bigger in count. The figures should be used as curve distribution reference only. Also note that there were no children in the age of 0 years in our dataset, but a considerable amount in the overall cohort. The selection criteria are nevertheless the same in both queries: patients in the age of 0-18.

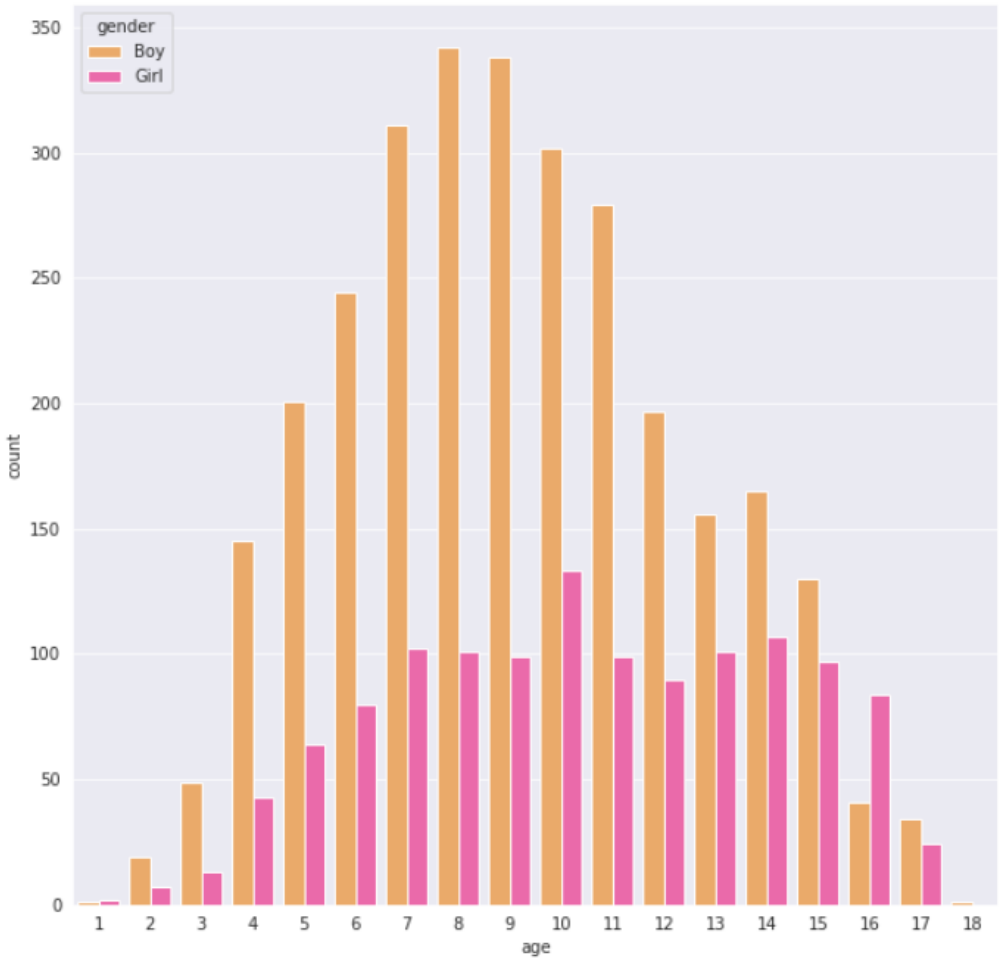
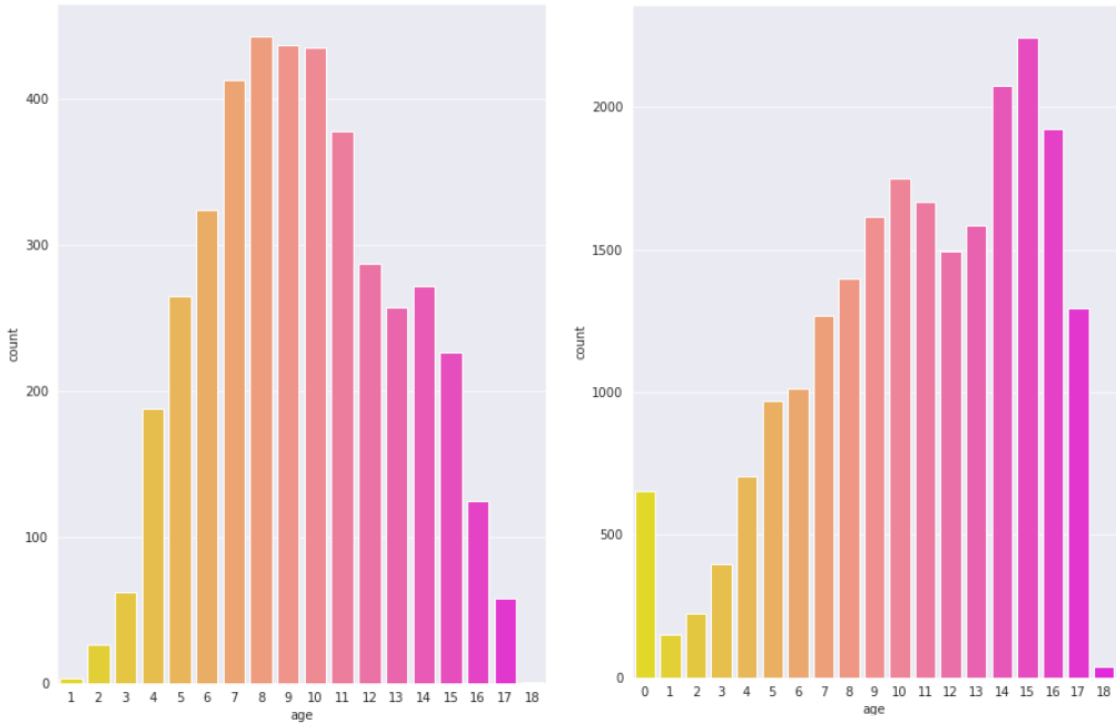


Figure 6.3.: Age distribution for the dataset.

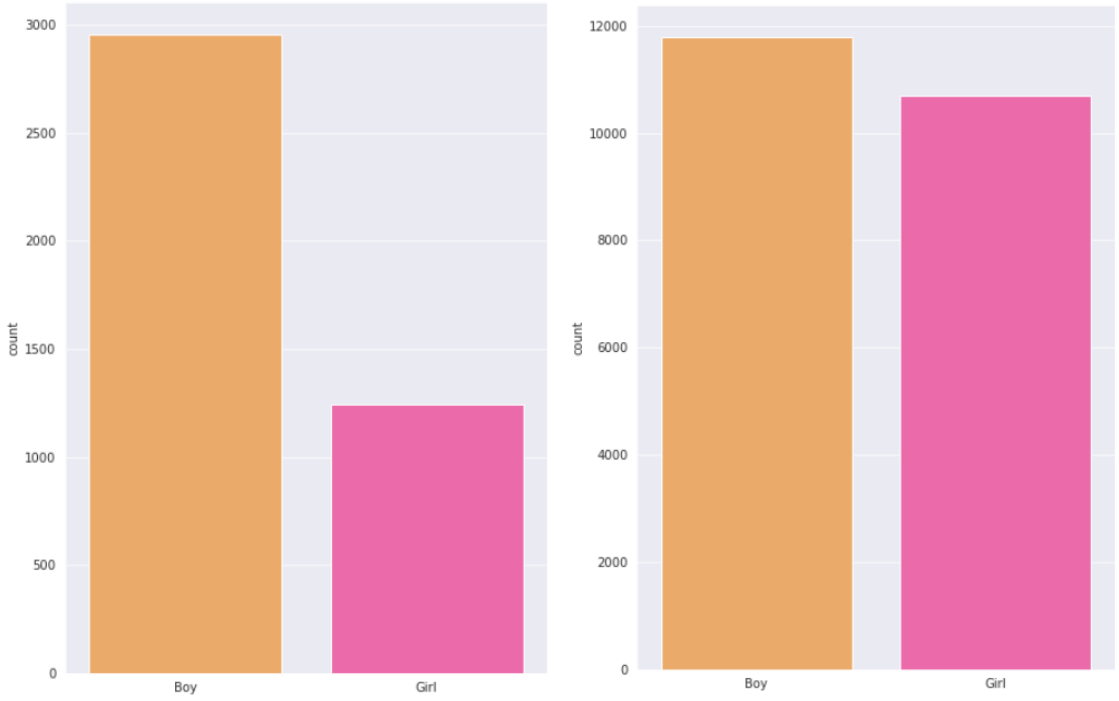


(a) Age distribution in our data selection.

(b) Age distribution in the entire cohort.

Figure 6.4.: Comparison of age distribution in the data selection and in the entire cohort.

Gender: The same comparison has been made for gender. Figure 6.5a and 6.5b compares the frequency of each gender for the patients in our data selection, and for all patients in the overall total cohort. Note that the count scales are different, and the y-axis of the right plot is about four times bigger in count. As can be seen in figure 6.5a, there are less than half the amount of girls than boys. However, in the entire cohort, the number of boys and girls is more even. As we move on, it is suggested to keep in mind that boys outnumber girls about 2.4 times while we evaluate variables related to gender.



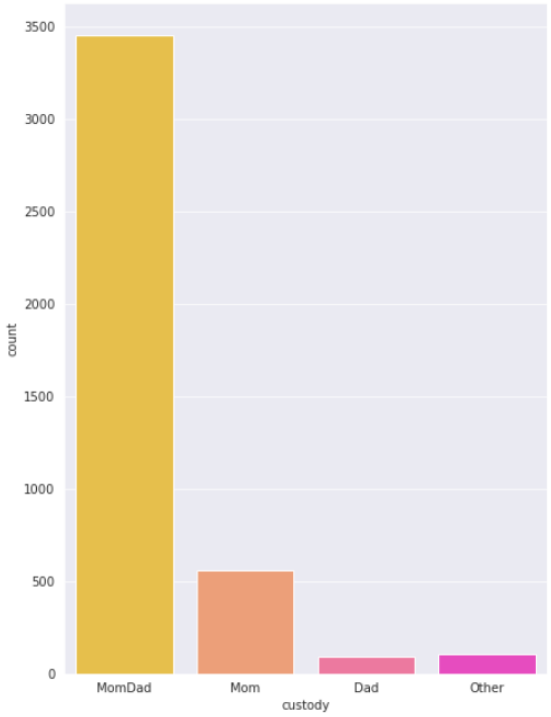
(a) Frequency of gender in our data selection.

(b) Frequency of gender in the entire cohort.

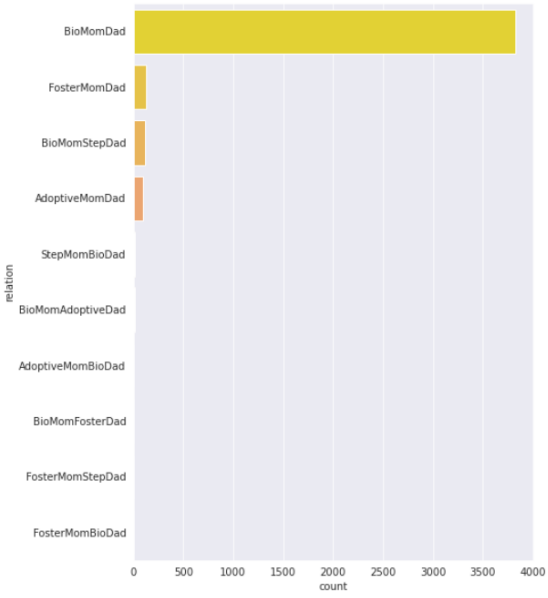
Figure 6.5.: Comparison of gender frequency in the data selection and in the entire cohort.

Custody, relation, care and assessment: Figure 6.6 depicts custody situation, relation combination, care situation and assessment outcome in the dataset.

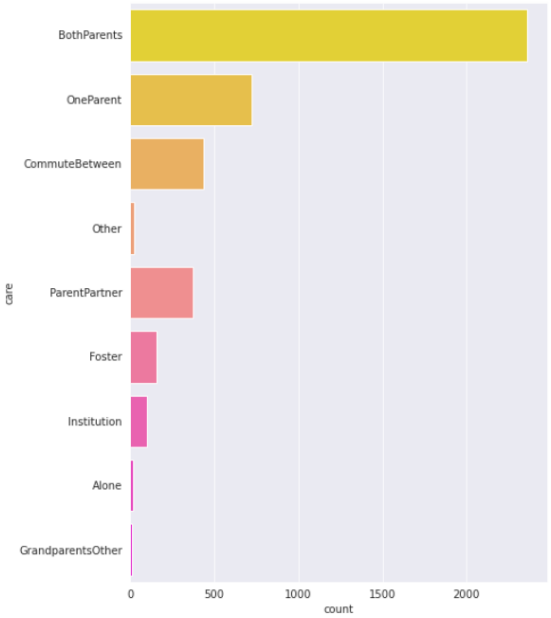
The most common custody situation is both mom and dad having custody of the child, followed by mom having custody alone. The step down to dad having the custody, is rather distinct. In most cases, the child is recorded as having both biological mom and dad, and this is by far the most dominant combination. It is also most common that the child is living with both parents, then with only one parent, followed by either commuting between parents or living with one parent and their parent. Foster care and living in an institution is less common. Furthermore, a large majority of referrals are accepted, and the number of rejected referrals are approximately 1/7 the amount of accepted ones.



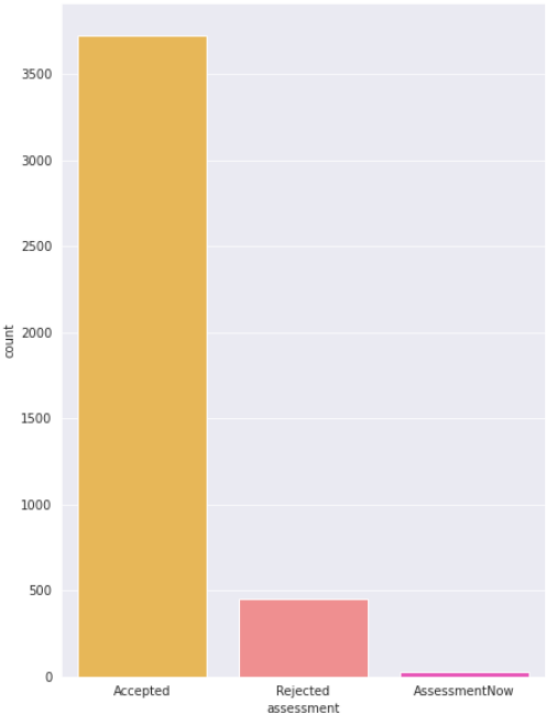
(a) Frequency of custody situation.



(b) Frequency of relational combinations.



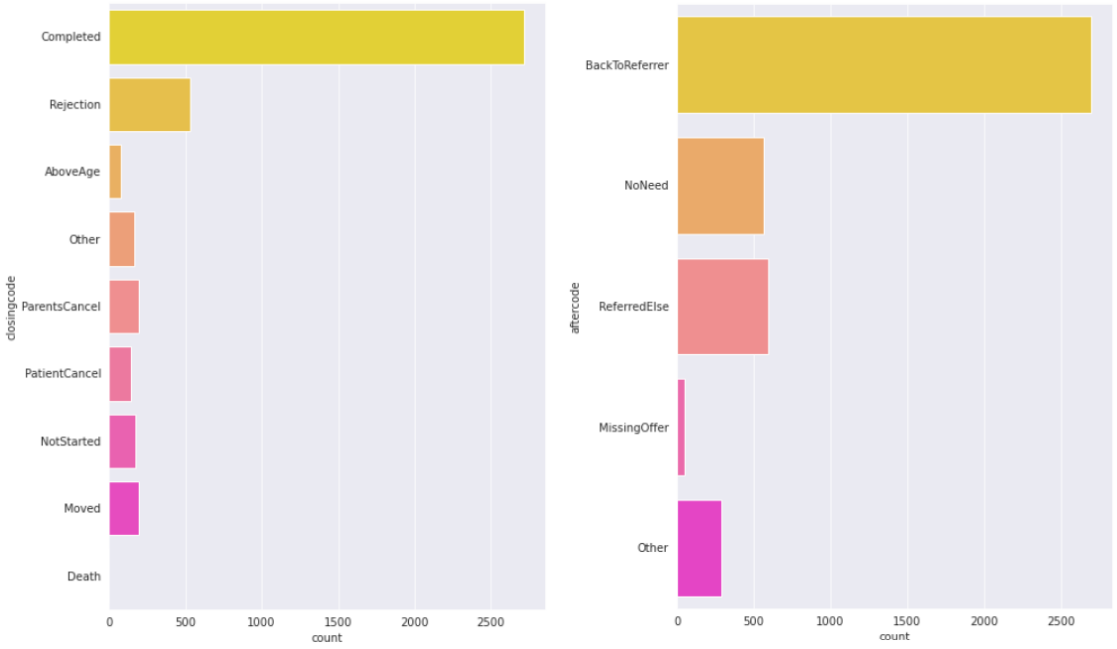
(c) Frequency of care situation.



(d) Frequency of assessment outcome.

Figure 6.6.: Count of each custody situation, relation combination, care situation and assessment outcome in the dataset.

Closing code and after code: Figure 6.7 depicts the frequency of closing codes and after codes, respectively. Note how *completed* is the most common closing code, and *back to referrer* is the most common after code.



(a) Frequency of closing code.

(b) Frequency of after code.

Figure 6.7.: Frequency of closing code and after code in the dataset.

Any diagnose on axis 1, ICD-1 : Diagnoses made on axis 1 can be identified in the *sak.icd1*-field in the database. As can be seen in figure 6.8, the top five most frequent registrations for axis 1 are:

1. F900: Disturbance of activity or attention
2. Blank: Empty field/NULL-value. Usually indicates no diagnose, either because patient was rejected, or was accepted but did not receive a diagnose
3. Z032: Observation for suspected mental and behavioural disorders
4. Missing: Code 1999. Insufficient information for making diagnose on axis 1
5. F901: Attention Deficit Hyperactivity Disorder

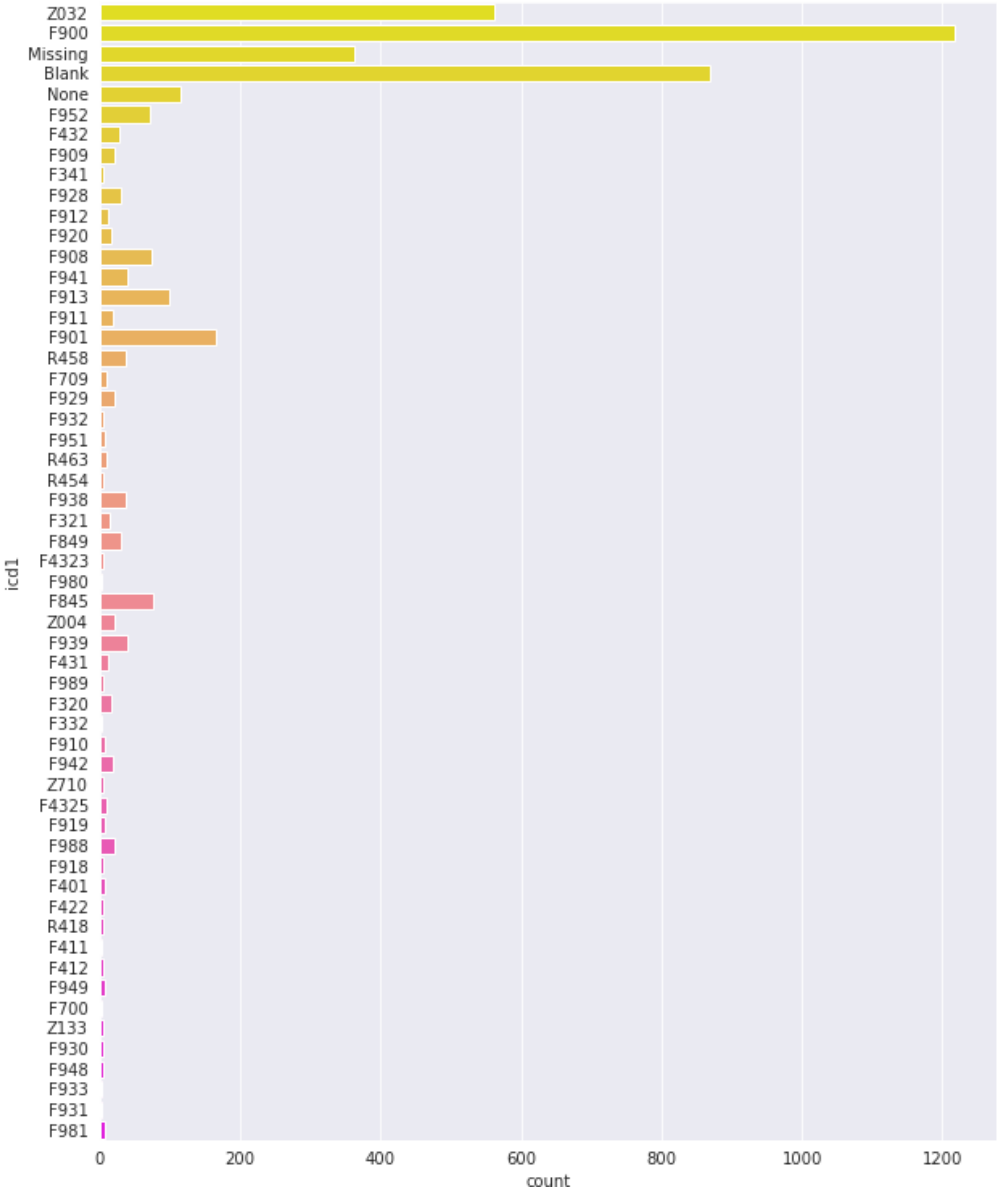


Figure 6.8.: Frequency of registrations on axis 1.

Referring instance: As can be seen in figure 6.9, the top five most frequent referring instances are:

1. Lege: General physician (GP)
2. Pedagogisk-psykiatrisk tjeneste: Educational Psychological Service
3. Hjelpetjenesten for Barn og Unge: The Help Service for Children and Young People
4. Helsestasjon: Health Station
5. Barnevern, kommunen: Child Welfare Services, municipality level

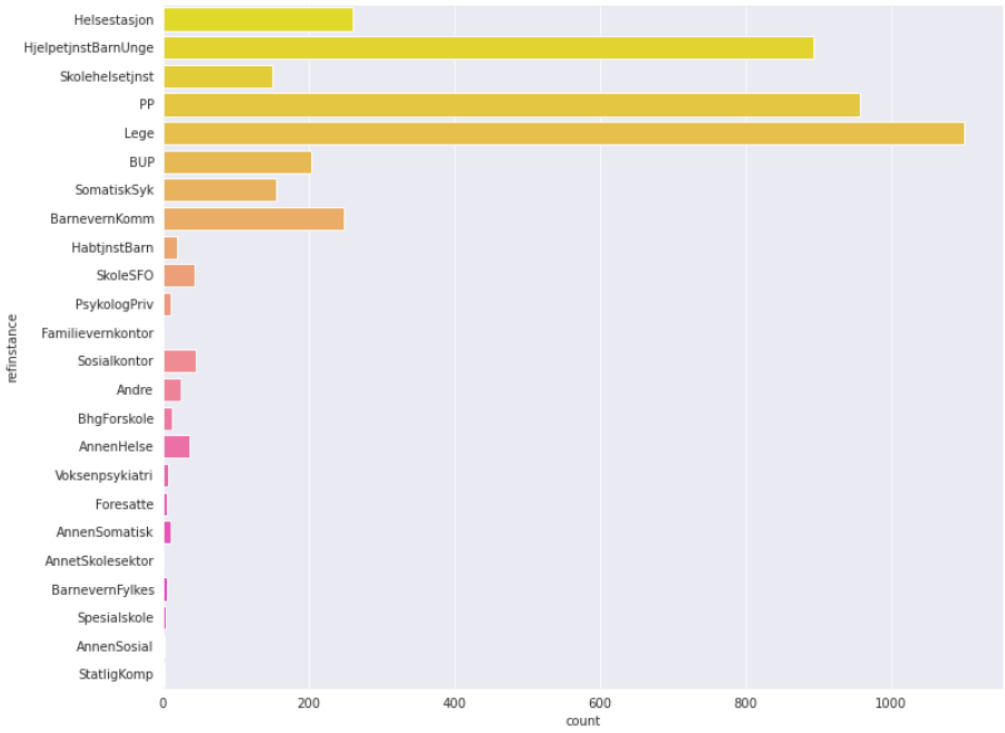


Figure 6.9.: Frequency of referring instance.

First referral reason: As can be seen in figure 6.10, the top five most frequent first referral reasons are:

1. Suspicion of hyperkinetic disorder (ADHD)
2. Suspicion of defiance/conduct disorder
3. Suspicion of depression
4. Other reasons
5. Suspicion of anxiety

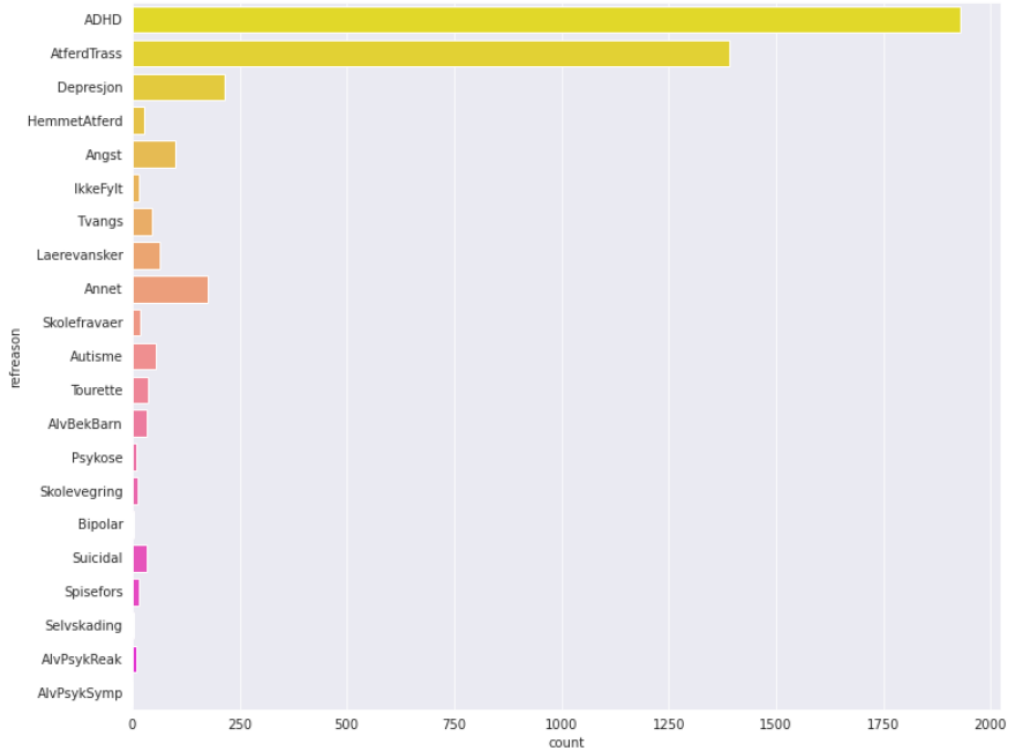


Figure 6.10.: Frequency of first referral reason.

Variables compared to gender: Figure 6.11, 6.12 and 6.13 illustrates how gender behaves with three other categorical values, namely referring instance, referral reason, and axis 1 recordings. It does not seem like some referral instances more commonly refer either boys or girls. Regarding referral reasons, these also behave as can be expected, but notice the high amount of girls being referred on suspicion of depression, in comparison to the gender balance in our dataset. Even though boys outnumber girls about 2.4 times, almost as many boys and girls have depression as primary referral reason. It is also among the top five most frequently recorded referral reasons. Additionally, referral reasons like *suspicion of Autism* and *suspicion of Tourette's syndrome* are considerably more frequently recorded for boys. Regarding diagnoses, these also behave as expected. However, in the F90-group, boys outnumber girls to a larger degree for F901 and F909, and more evenly for F900 and F908.

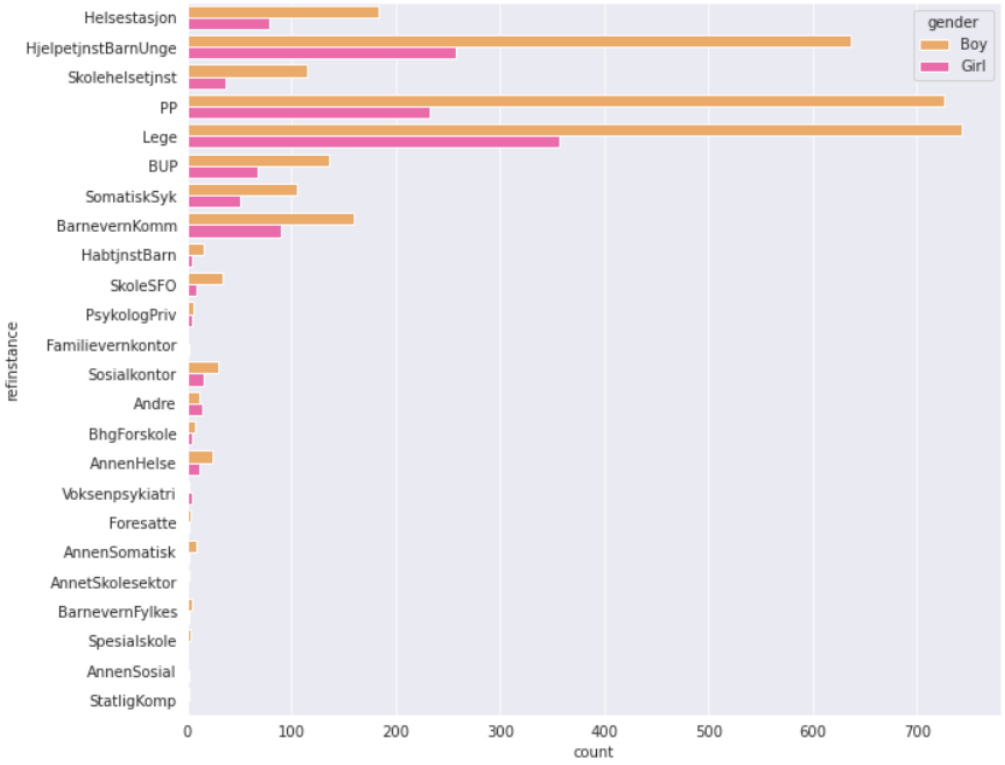


Figure 6.11.: Frequency of referral instance for each gender.

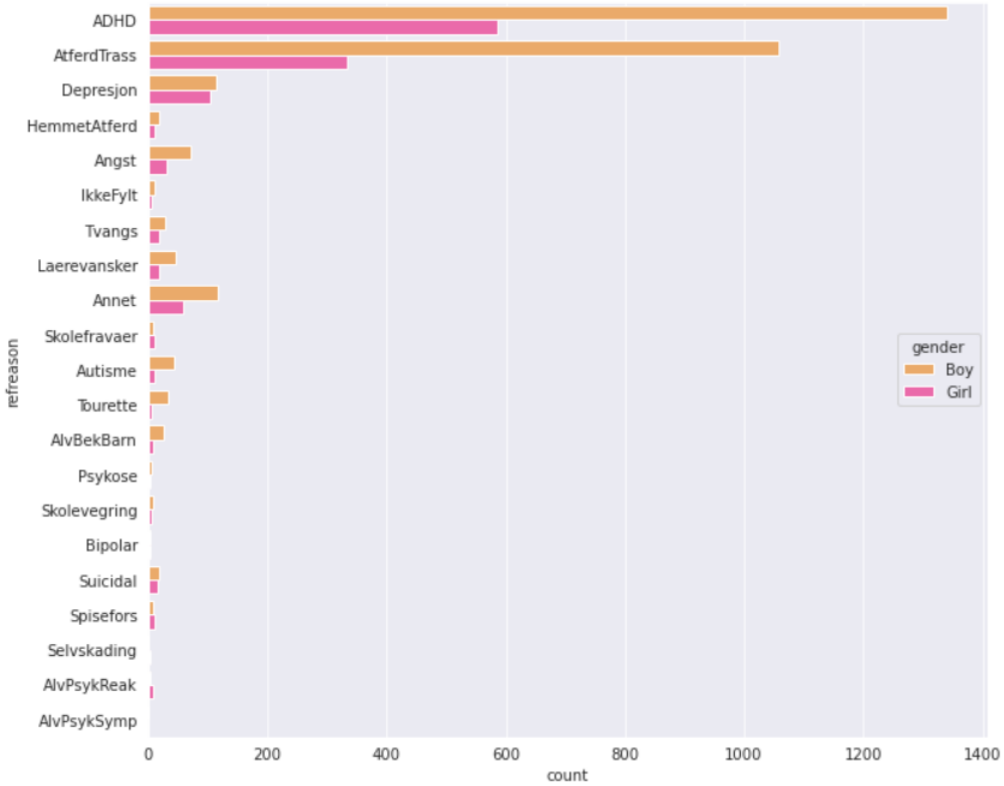


Figure 6.12.: Frequency of first referral reason for each gender.

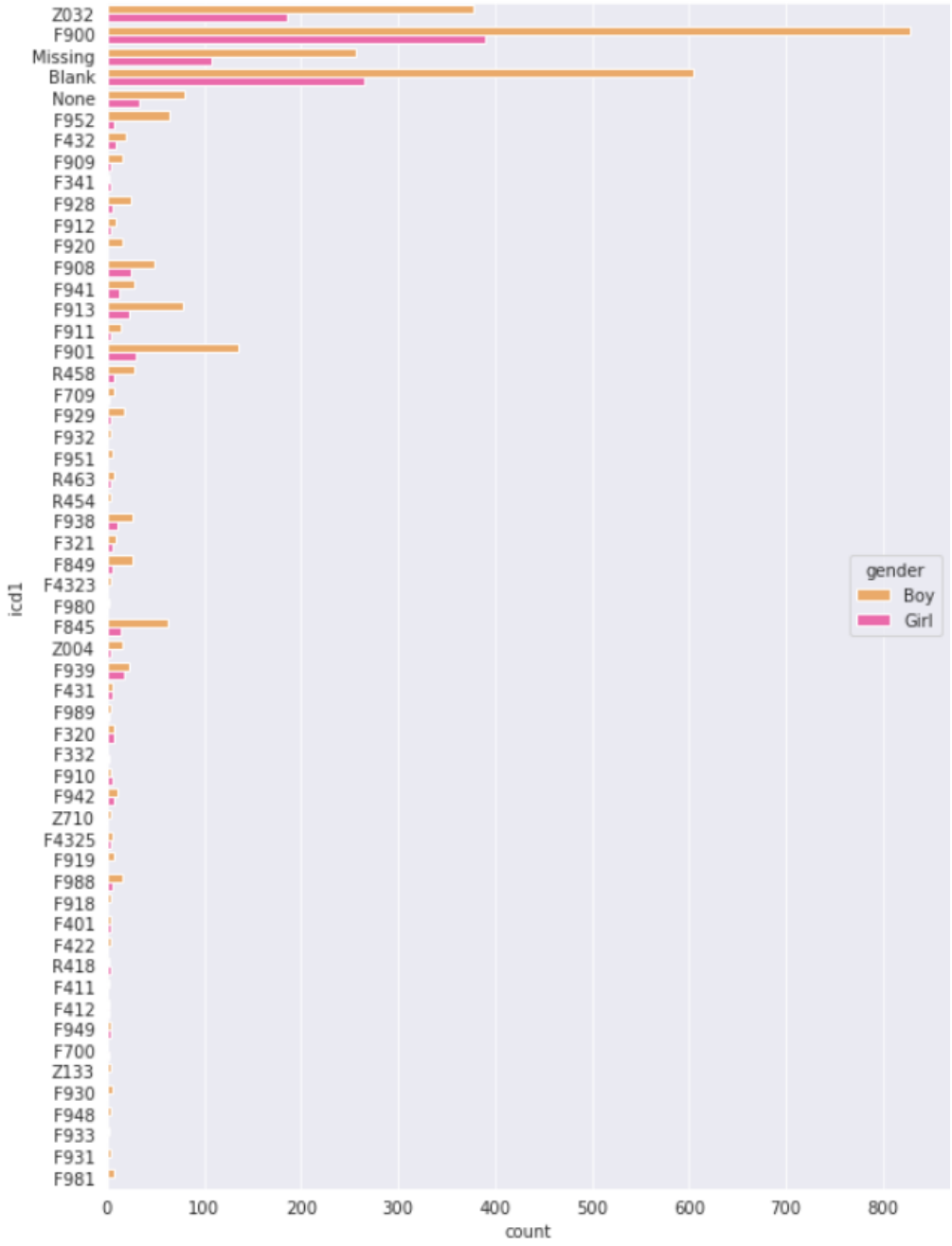
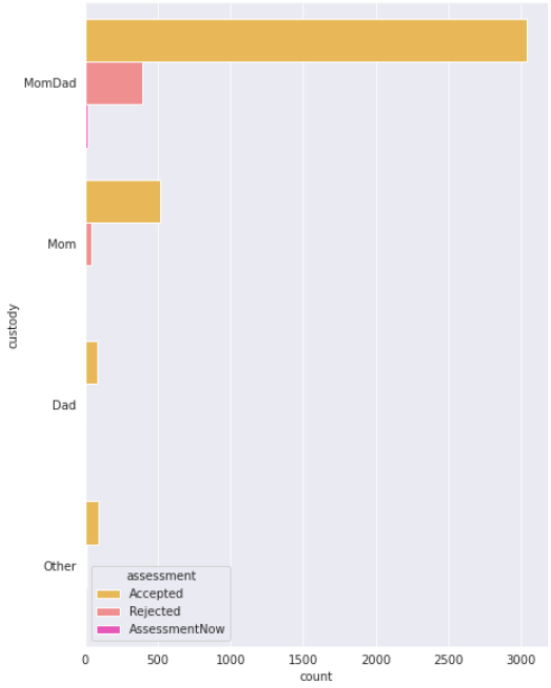
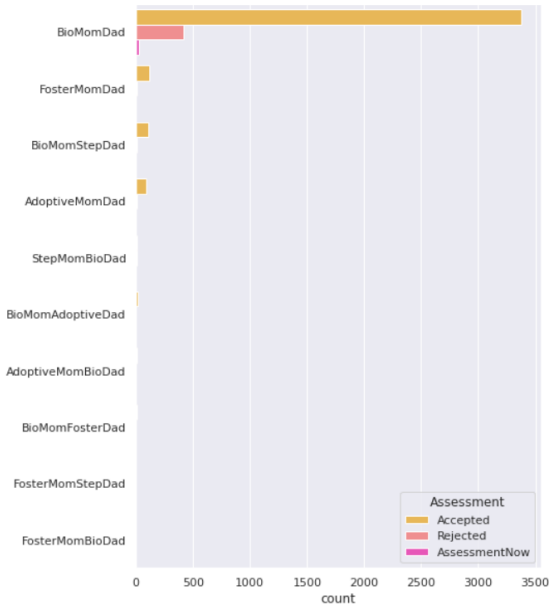


Figure 6.13.: Frequency of registrations on axis 1 for each gender.

Variables compared to assessment outcomes: Figure 6.14 depicts how assessment outcome varies with custody and relation.



(a) Custody in comparison to assessment.



(b) Relation in comparison to assessment.

Figure 6.14.: Comparison of custody situation and relation with assessment outcome.

In figure 6.14a, the combination of both mom and dad having the custody, is most prone to rejection. This is to some degree visible in figure 6.14b as well, as it can be noticed that the majority of rejections happen in combination with the child having a biological mom and dad. However, this is also the combination with significantly more referrals, and is the dominant combination of the ten relation combinations. Also notice how the other combinations have almost no rejections in comparison.

Figure 6.15 and 6.16 illustrates how assessment outcome behaves in connection with referring instance and referral reason.

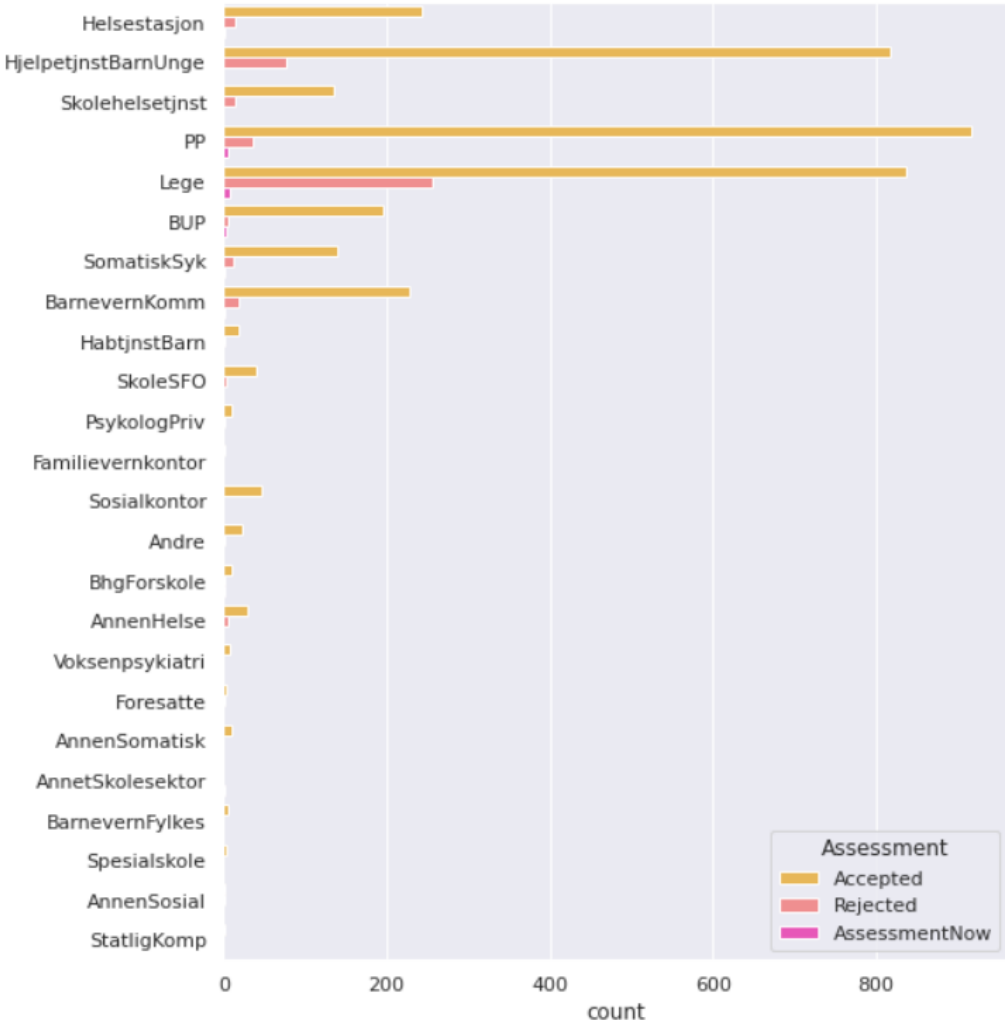


Figure 6.15.: Frequency of assessment outcome for each referring instance.

In figure 6.15, it becomes clear that even though there are almost as many referrals from the GP, the Educational Psychological Service and the Help Service for Children and Young People, significantly more of the referrals from a GP receive rejections. The Help Service for Children and Young People are the second-most prone to rejections. As elaborated on in section 2.2.2, it is often a collective service for other services like e.g. Child Welfare Services and School Health Services. Figure 6.16 indicates that *suspicion of ADHD* and *suspicion of defiance/conduct disorder* are most prone to rejection, but these are also dominantly the most recorded referral reasons.

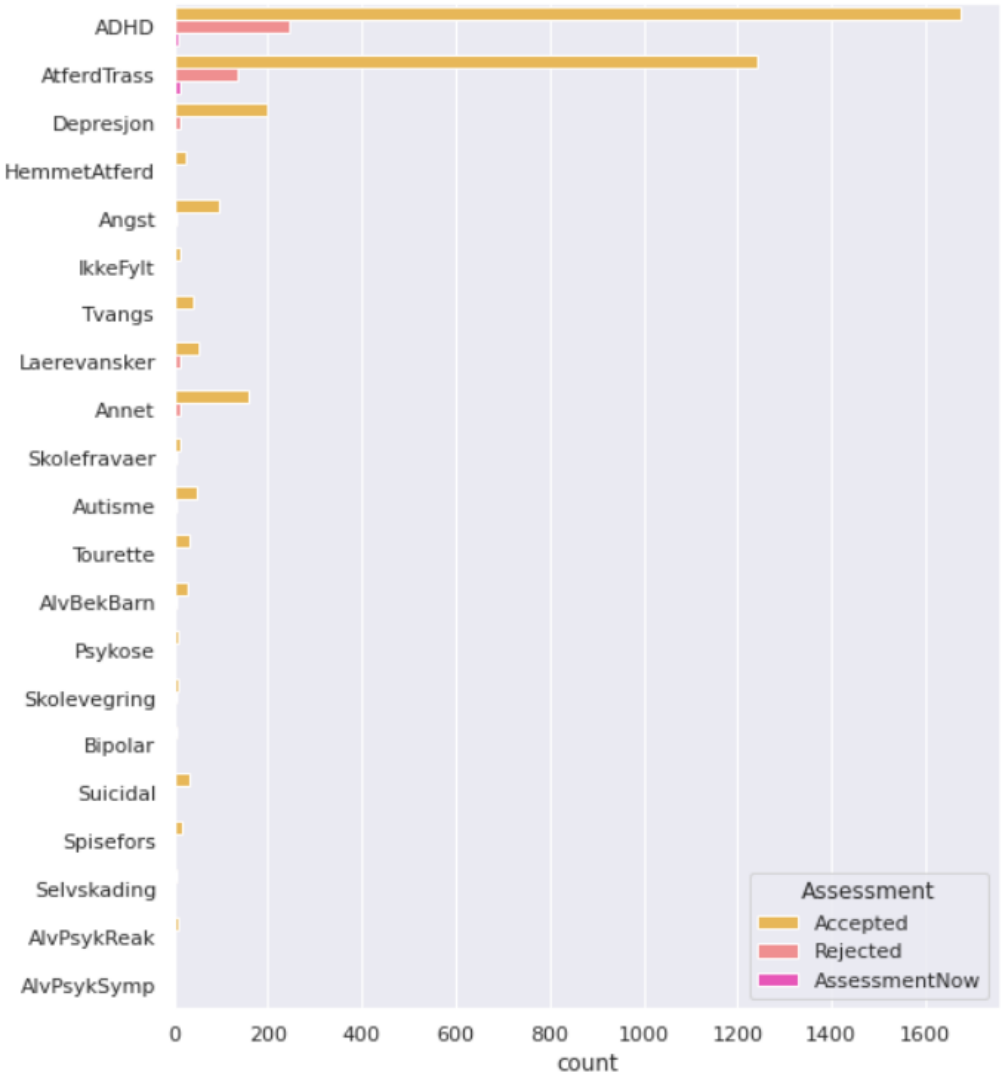


Figure 6.16.: Primary referral reason in combination with assessment outcome.

Scatter Plots

Scatter plots are useful for illustrating the datasets with regards to more dimensions, and use dots to represent values. They are useful in observing relationships between variables, both numerical and categorical.

Figure 6.17 illustrates referral reason, age and assessment outcome for each gender. Note that some referral reason have a later occurrence for girls than boys, like *learning difficulties*. However, *suspicion of ADHD* and *suspicion of defiance/conduct disorder* are actually quite similar and occur simultaneously across different age groups. *Absence from school* occurs rather late for both genders, even after the age one starts school. *Suspicion of Autism* and *suspicion of Tourette's syndrome* are by far more frequently recorded for boys across all age groups. *Suspicion of eating disorder* and *suicide risk* are slightly more frequently recorded for girls, and with a later onset than boys.

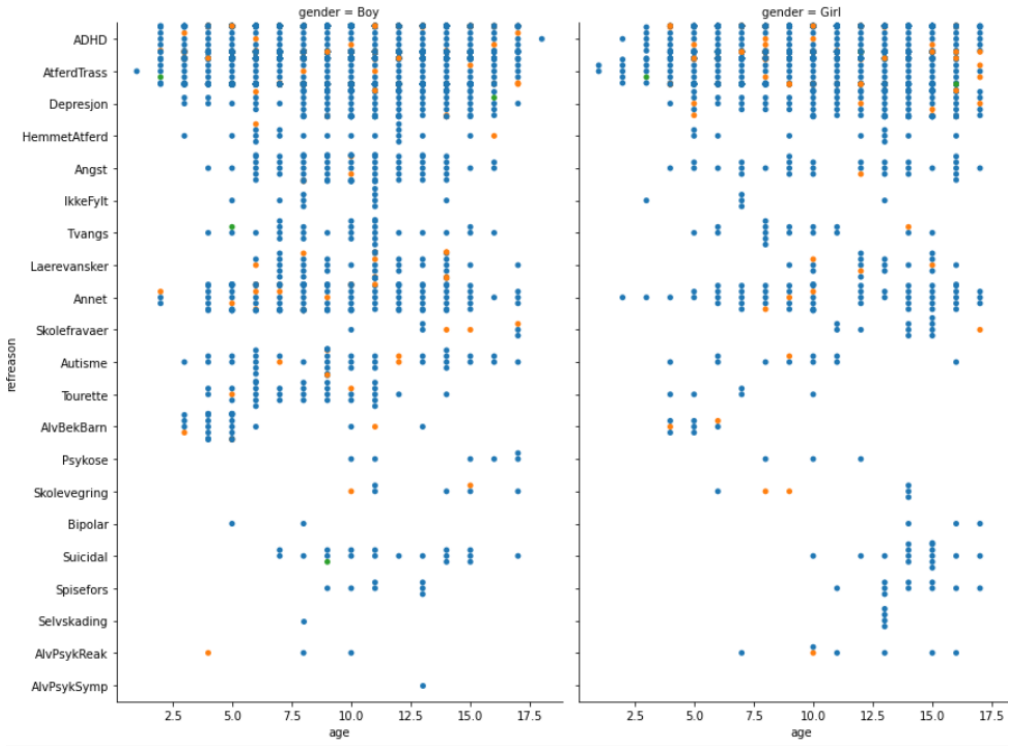


Figure 6.17.: Scatter plot illustrating referral reason, age and assessment outcome for each gender. Blue dots are accepted, yellow are rejected, and green are temporarily assessed.

With regards to care situation, children of both genders that also live in institution have their first referral at an older age than the other care situation groups. This can be seen in figure 6.18. For children that live in foster care, boys tend to be referred at an earlier age than girls, as boys' distribution is slightly skewed to the left, and girls' distribution to the right.

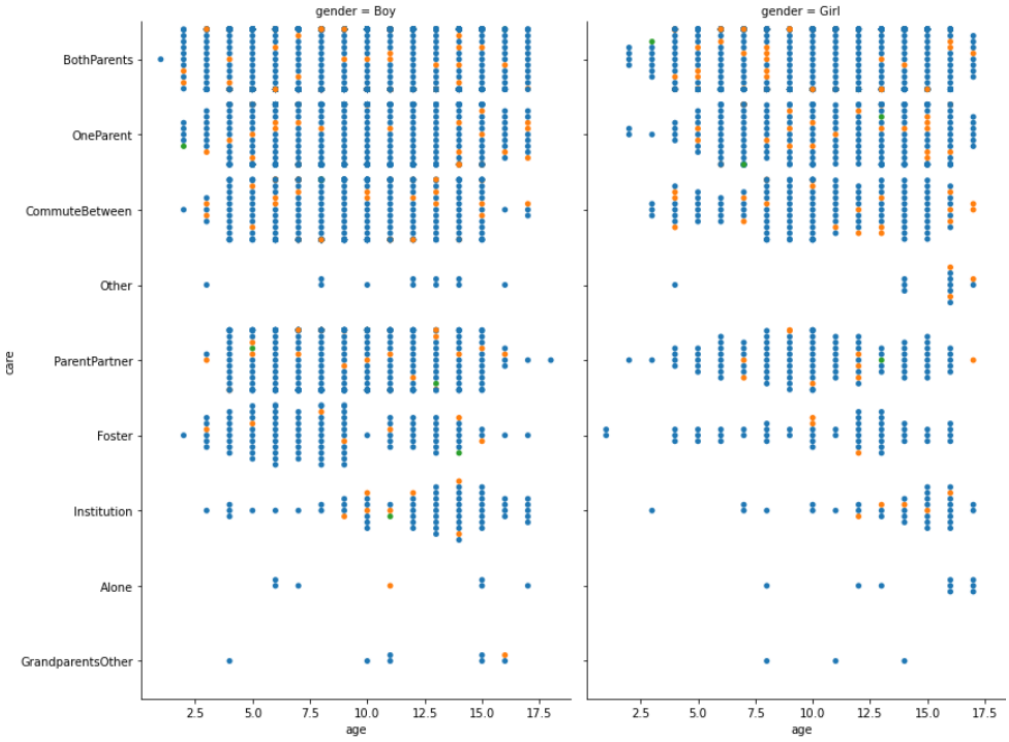


Figure 6.18.: Scatter plot illustrating care situation, age and assessment outcome for each gender. Blue dots are accepted, yellow are rejected, and green are temporarily assessed.

Figure 6.20 shows diagnose codes in relation to gender, age and assessment outcome. Regarding F900, boys have slightly earlier onset than girls, but after 2.5 years there are little to no difference. As can be expected, the rejected cases are in relation to *missing, blank* and *none*, which confirms that diagnoses are given after admission.

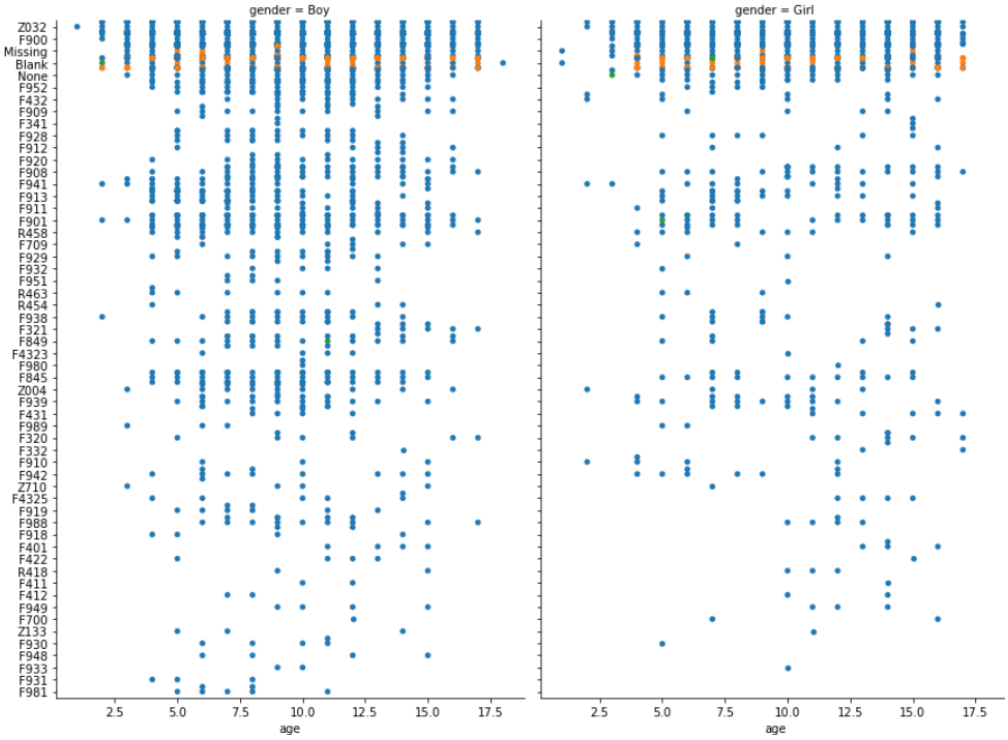


Figure 6.19.: Scatter plot illustrating registrations on axis 1, age and assessment outcome for each gender. Blue dots are accepted, yellow are rejected, and green are temporarily assessed.

Lastly, looking into referring instance, there are some minor differences. It is much more common for the School Health Service to refer boys than girls, and as can be predicted, referrals from this service do not occur until school age. The same applies to CAMHS clinics, the Social Office, as well as the school and after school program (SFO), which are referred to as one and the same instance. Otherwise there are little difference, e.g. from Child Welfare Services or GP.

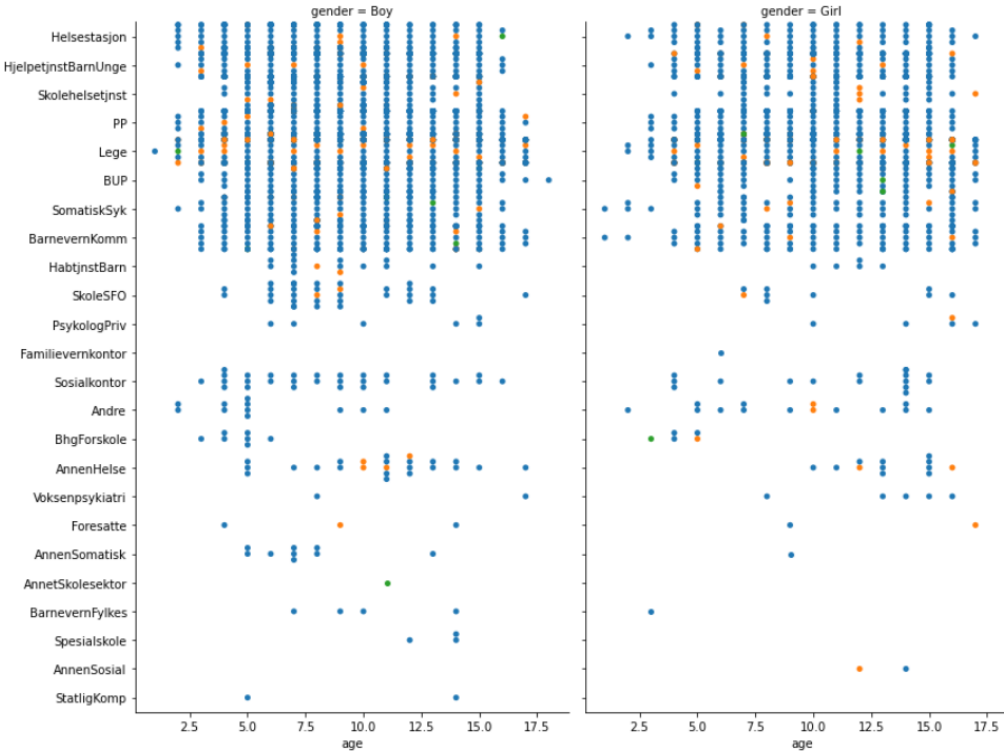


Figure 6.20.: Scatter plot illustrating referring instance, age and assessment outcome for each gender. Blue dots are accepted, yellow are rejected, and green are temporarily assessed.

Referring instance can also be seen in combination with referral reason, as figure 6.21 presents. From the figure it is evident that the Educational Psychological Service, the GP, the Help Service for Children and Young People and the Health Station have the largest amount and broadest range of referral reasons.

As can be seen in figure 6.15, one instance in particular has a large number of rejections, namely the GP. Thus it may be interesting to also look into which referral reasons that are most common for each referring instance, as figure 6.21 shows. It is noticeable that the GP has more referrals with primary referral reason *suicide risk*, *suspicion of eating disorder*, *not filled by referring instance*, *persistent and severe self-harm* or *severe psychological reactions after trauma, crises or disasters*. However, looking into the yellow

dots, the rejected cases, these are evenly distributed across the different referral reasons. In comparison to other institutions that also have rejections, the GP may have slightly more rejections for referral reasons *suspicion of ADHD*, *suspicion of depression*, *other reasons*, *suspicion of anxiety*, *serious concern for children under 6 years*, and *school refusal*. The GP is also the referring actor with widest range of referral reasons.

Even though the Educational Psychological Service has the second most referrals after the GP, they have slightly less wide spread range of referral reasons, and their rejections are mostly centered around *suspicion of depression*, *suspicion of anxiety*, *learning difficulties* and *school refusal*.

Both Educational Psychological Service and GP have the highest number of referrals with unfilled primary referral reason, but these are not the cases with rejections.

The School Health Service have a considerable amount of rejections for *learning difficulties*. Somatic hospitals have a noticeable number of rejections for *suspicion of ADHD*.

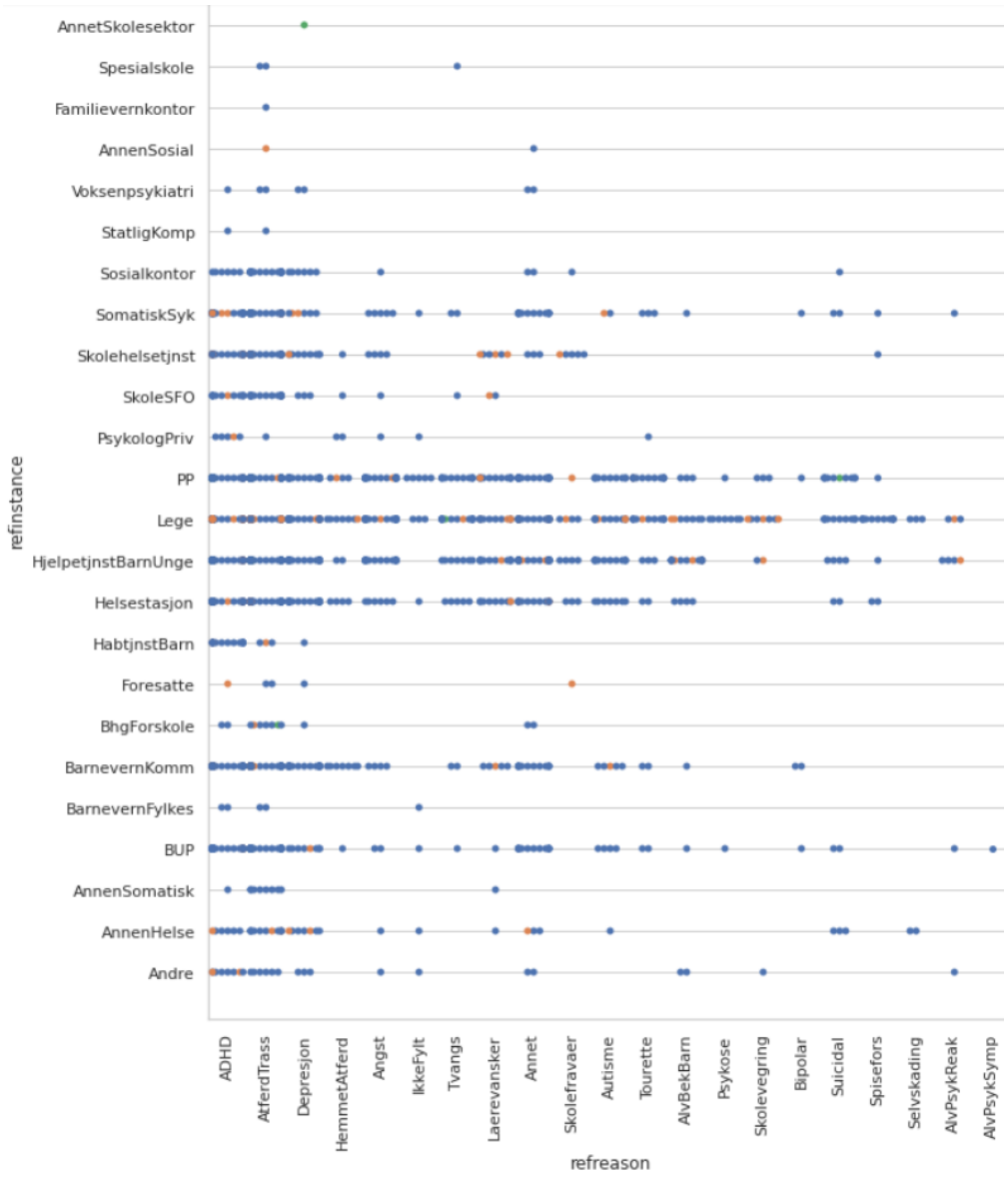


Figure 6.21.: Scatter plot illustrating referring instance, referral reason and assessment outcome. Blue dots are accepted, yellow are rejected.

Key Takeaways from EDA

To summarise, the key takeaways from EDA results are as follows.

- Compared to the entire cohort where the gender distribution was rather even, the patients in this data selection have considerable fewer girls than boys.
- The majority of girls have their first referral period later than boys, an average of two years later.
- By far the most rejections happen in combination with the child having a biological mother and father. This is also the combination with most referrals.
- *Suspicion of ADHD* and *suspicion of defiance/conduct disorder* are quite similar in occurrence in relation to both age and frequency for boys and girls, but other referral reason like *learning difficulties* have later occurrence for girls than boys.
- *Suspicion of Autism* and *suspicion of Tourette's syndrome* are more frequently recorded for boys across all age groups. *suspicion of eating disorder* and *suicide risk* are more frequently recorded for girls than boys, but boys have an earlier recorded onset.
- Children of both genders that also live in institution have their first referral at an older age than the other care situation groups. For children that live in foster care, boys tend to be referred at an earlier age than girls.
- The most frequently primary referral reasons are *suspicion of ADHD*, *suspicion of defiance/conduct disorder*, *suspicion of depression*, *other reasons* and *suspicion of anxiety*.
- The referring instances with largest number of referrals are the GP, the Educational Psychological Service and the Help Service for Children and Young People.
- Referrals from a GP has the largest number of rejections, even though the Educational Psychology Service and the Help Service for Children and Young People have almost the same number of referrals. The GP also has the broadest range of referral reasons.
- The rejected cases from a GP are evenly distributed across the different referral reasons. However, the GP has slightly more rejected referrals for referral reason *suspicion of ADHD*, *suspicion of depression*, *other reasons*, *suspicion of anxiety*, *serious concern for children under 6 years*, and *school refusal*.
- Given the number of boys to girls in the dataset, there is a considerable large amount of girls being referred on the *suspicion of depression* as first reason - almost

as many girls as boys. *Suspicion of depression* is also the third-most recorded referral reason.

- There are much less patients being referred that have the relation bio dad-step mom, than bio mom-step dad.
- Even though almost half as many girls as boys are referred with primary referral reason *Suspicion of ADHD*, only about a quarter of girls compared to the of the number of boys are diagnosed with F901. However, F900 more strongly represent the ratio of boys and girls that are referred on suspicion of ADHD.

Initial thoughts and hypotheses:

- Based on the EDA results, it is reasonable to expect that the dataset primarily has minor situation differences separating the patients prior to referral. There may be no major separation lines between the patients at this point in the temporal data coverage, and it can be somewhat difficult to separate patients into distinct groups that do not have great commonalities.
- To separate patients, it may be necessary to choose a larger number of clusters k in order to bring out the small nuances.
- The data seems to offer more distinguished differences in patient situation from the time of making the referral and onward. There may be greater potential of useful insight in the data concerning referring instance, referral reason and referral assessment, than patient situation before making the referral.
- Regarding referring instances, the GP, with the largest amount of rejections, is an actor to pay attention to.
- Regarding referral reasons, the high frequency of *other reasons* may be something to pay attention to. The large count of such an unspecified reason is noticeable.

6.4. Determining Optimal Number of Clusters

This section is concerned with the experimental execution, which includes determining the optimal number of clusters by the elbow method, and using this number as an input to the cluster algorithm. The clustering results are presented in the next section, 6.5.

Regardless of which unsupervised algorithm that is used, an important step is to determine the optimal number of clusters k. There are several methods to do this, and the *Elbow method* is one of the most popular methods to determine this optimal value of k (Aprillant, 2021). An elbow plot will show at which number of k clusters the cost begins to linearly decrease. For the K-prototypes algorithm, cost is defined as the sum distance of all points to their cluster centroids (Huang, 1997). Since K-prototype combines both numerical and categorical variables, it also provides the cost function for calculating a combined cost and similarity measure on both types of attributes.

The calculated elbow plot can be seen in figure 6.22 below.

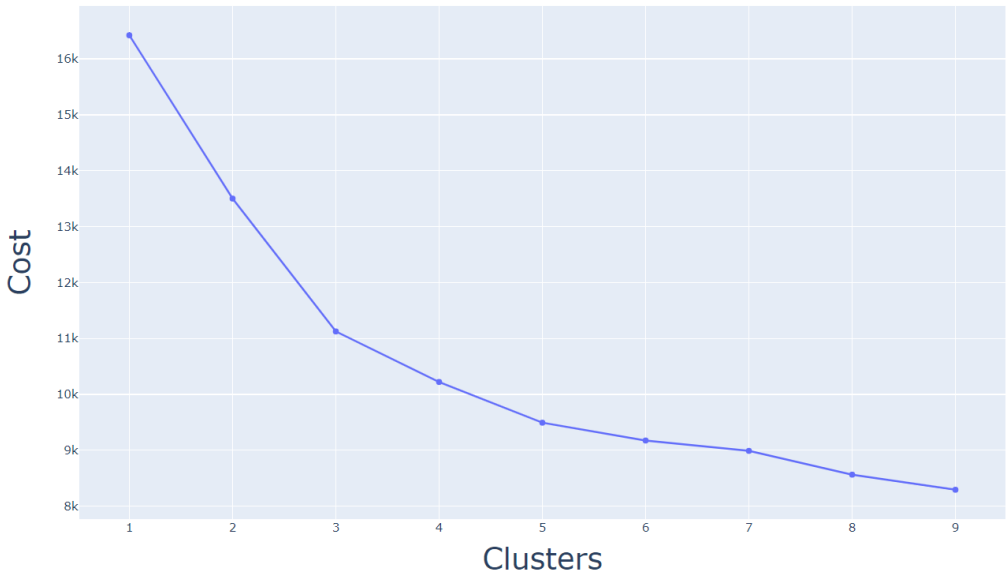


Figure 6.22.: Calculating the optimal number of clusters using the Elbow method.

What one essentially wants to look for is a change of slope from steep to shallow (an elbow) to determine the optimal number of clusters (Matt, 2019). According to the elbow plot of cost function in figure 6.22, choosing the number of clusters k=3, k=5 or k=6 can be a reasonable choice. This method will choose the optimal number of clusters for the cluster analysis with K-prototype, and any number beyond this will not necessarily give a meaningful separation of clusters.

By trial and error, k=3, k=4, k=5, k=6 and k=7 was chosen to find the most meaningful

clusters. Even though the most defined elbow can be found at k=3, larger values of k were included based on the reflections after EDA in the previous section. The resulting cluster centroids are presented in the next section.

In addition to choosing the number of clusters, we have to determine how to select the starting points for the clusters. With K-prototype, there are two methods to initialise the clusters; *Huang* and *Cao*. By selecting *Huang*, the model will select the first k-distinct objects from the dataset as initial k-modes before equally assigning the most frequent categories to the initial k-modes. By selecting *Cao*, it selects prototypes for each data object based on the density of the data point and the dissimilarity value (Zazueta, 2020). In this experiment, *Huang* was the model of choice, simply because it is the most frequently used method in similar research.

Clustering with k=3: With k=3 we get one relatively small cluster and two larger ones.

Cluster	**Count**
0	1,534
1	167
2	2,500

Table 6.5.: Count for k=3 clusters.

Clustering with k=4: With k=4, we get three larger clusters and one small. Notice how the smallest cluster from k=3 is kept in k=4. However, we now have three rather than two large clusters.

Clustering with k=5: With k=5, the smaller variations start to show, and we get two rather small clusters, and three larger clusters.

Clustering with k=6: With k=6, minor nuances appear in their own cluster, and we get one small cluster, three mid-sized and two larger clusters. The smallest is n=31 and the largest is n=1,223.

Ideally, one would choose the most defined elbow point of the elbow plot, which would be k=3. However, since k=6 is the only number of clusters k that includes a girl as a cluster centroid, we move on with k=6 as our chosen number of clusters. Having the majority of the girls in one cluster, enables us to better analyse gender differences. The six clusters are presented in section 6.5.

Cluster	Count
0	1,183
1	167
2	1,214
3	1,637

Table 6.6.: Count for k=4 clusters.

Cluster	Count
0	1,588
1	1,184
2	97
3	952
4	380

Table 6.7.: Count for k=5 clusters.

Cluster	**Count**
0	621
1	1,067
2	1,223
3	961
4	31
5	298

Table 6.8.: Count for k=6 clusters.

6.5. Experimental Results

This section presents the results from the clustering experiment. First, the cluster centroids are described, before moving on to a statistical summary of each cluster in order to be better acquainted with the patient subgroups.

To make the result presentation comprehensible and interpretable, the focus is on the key information for each cluster. Furthermore, information relevant to the experimental aims, as well as the research questions, have been prioritised. An evaluation, as well as a discussion of the results, are left for the next chapter.

6.5.1. Results

Initial data retrieval yielded a total of 4,274 patients. After the data preprocessing stage, a final number of 4,201 patients were kept. Six clusters were identified by using demographical data as well as clinical registrations from their first referral period. The cluster centroids are summarised in table 6.9.

Age and the number of stays were inversely transformed after processing and the mean value was calculated. Following table 6.9, descriptive statistics of each cluster will be presented. Note that only the most noticeable values are drawn out and highlighted, in order to clarify the differences.

Descriptive statistics of the labeled dataset were generated using SPSS version 27 (IBM, 2019). Means and range were reported for numerical values. Since the continuous data are not standardised, medians were also reported. Counts and percentages were reported for categorical data. Due to the varying size of the clusters, counts were emphasised more than percentages when comparing clusters. Missing data were not reported as any missing data was handled during the preprocessing stage.

General Characteristics

Demographics and numericals: The size of the six clusters were cluster 1: n=621, 2: n=1,067, 3: n=1,223, 4: n=961, 5: n=31 and 6: n=298. Cluster 2 has the lowest median age of 6. Cluster 4 has the highest median age of 14. Cluster 1 has the largest amount of girls with 74.72% girls, and boys constitute the majority in the rest of the clusters. Cluster 5 has the highest median of stays (n=8).

Family situation: With regards to custody situation, there were minor differences, but cluster 5 has the highest proportion of patients with mom as the only one with the custody (29.0%), and no incidents in this cluster with dad having the custody. However,

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Age	7.721	5.925	10.138	14.185	11.774	8.815
NrStays	1.127	1.127	1.112	1.101	9.677	3.738
Gender	Girl	Boy	Boy	Boy	Boy	Boy
Care	BothParents	BothParents	BothParents	BothParents	OneParent	BothParents
Custody	MomDad	MomDad	MomDad	MomDad	MomDad	MomDad
Relation	BioMomDad	BioMomDad	BioMomDad	BioMomDad	BioMomDad	BioMomDad
Ref-Instance	PP	Lege	PP	Lege	Lege	Hjelpe-tjnstBarn-Unge
Ref-Reason	ADHD	AtferdTrass	ADHD	ADHD	Atferdstrass	Atferdtrass
Assess-ment	Accepted	Accepted	Accepted	Accepted	Accepted	Accepted
ICD1	F900	Blank	F900	Blank	F900	F900
Closing-Code	Completed	Completed	Completed	Completed	Completed	Completed
After-Code	BackTo-Referrer	BackTo-Referrer	BackTo-Referrer	BackTo-Referrer	Referred-Else	BackTo-Referrer

Table 6.9.: Cluster centroids of cluster 1-6 with k=6.

across all clusters, cases with mom having custody are 4-6 times more frequent. Regarding care situation, there are very small differences in the number of patients living with both parents, varying from 50.2%-61.4%, except for cluster 5 with the lowest percentage of 29.0% living with both parents. For relational combinations, cluster 1-4 were the only clusters with patients having a step mom and bio dad; these percentages varied from 0.2% to 0.6%. All clusters had patients having biological mother and step father; these percentages varied from 1.8% to 9.7%. I.e. patients that are referred, more often have a biological mother and a step father than the other way around. Ranges across clusters for having a biological mother and father varied from 87.1% to 93.0%, making it the dominant combination.

I.e. if patients that are referred do not have a biological mother and father (the most common combination), then the mother more often has custody alone than dad having custody, and the patients more often have a biological mother and a step father than the other way around.

Cluster 2 has the highest percentage of patients in foster care (5.1%). Cluster 4 has the largest percentage of referrals from a GP (37.1%) across the clusters, and cluster 2 second largest percentage of referrals from a GP (35.1%).

Referral situation: Cluster 2 has the largest amount of patients being referred by a GP.

Across all clusters, if the frequency of *suspicion of ADHD* and *suspicion of defiance/-conduct disorder* were high, then the count of *suspicion of depression*, *suspicion of anxiety* and the count of *other reasons* were also high. *Suspicion of ADHD* and *suspicion of defiance/conduct disorder* were frequently recorded in cluster 1, 2, 3, 4 and 6, but common for all these clusters is the fact that *other reasons* were more frequently reported than *suspicion of anxiety*, and sometimes more often than *suspicion of depression*.

Cluster 1 and cluster 2 had no occurrence of *suspicion of eating disorder*.

Assessment situation: Cluster 2 and 4 had the highest percentage of rejected referrals at 17.2% and 15.6%, respectively. Cluster 5 only has accepted referrals, but is also of size n=31. Cluster 2 has the highest amount of rejections at n=183 and is also the cluster with most blank diagnoses, i.e. no given diagnose at n=339. Cluster 4 has the second most amount of ICD1 = blank at n=284. Upon ending a referral period, most are coded with closing code = completed, and after code = back to referrer. For closing code, cluster 2 has the highest number of parents cancelling (n=70), and cluster 4 the highest number of patient being above age (n=57) and the patient cancelling (90). For after code, cluster 5 is the only cluster with more patients being sent elsewhere instead of back to referrer.

Detailed cluster descriptions are given below.

Cluster 1:

There are n=621 patients in cluster 1. It is characterised by being the cluster with largest proportion of girls (74.72%). The patients are in the age 3-10.

Cluster 1 has the highest percentage of patients commuting between parents as their care situation (12.1%) and the highest percentage of patients living with a parent and their partner (9.7%).

Most patients (33.5%, n=208) in this cluster are referred from the Educational Psy-

chological Service. Only cluster 3, with twice as many patients, have more referrals from this service.

68.44% were referred on suspicion of ADHD.

Cluster 1 has the highest percent of patients being referred on *suspicion of ADHD* (68.4%), n=425.

Cluster 1 is also clearly the cluster with the least percent being referred for *suspicion of defiance/conduct disorder* (14.5%), and apart from cluster 5, the cluster with least amount (n=90).

8.4% of patients in cluster 1 are rejected, n=52. Cluster 1 has the highest percentage of ICD-1=F900 (43.6%).

Cluster 2:

Cluster 2 is characterised by being the youngest subgroup, the largest percent and amount of rejected patients, the most referrals from a GP, and has by far the largest amount of patients being referred for *suspicion of defiance/conduct disorder*.

There are n=1,067 patients in cluster 2. This cluster has the lowest mean age of 5.93 years and a median of 6. The patients are in the age 1-10. Its patients are primarily patients without a diagnose. There are 87.35% boys (n=932), which makes it the second largest group of boys.

Cluster 2 has the highest percentage of both patients living in foster care (5.1%) and amount of patients with foster mother and foster father (n=49, 4.6%). It has the largest amount of patients having a biological mother and a step father (n=37), and the highest number of patients living with one parent (n=220).

Most patients (35.1%) in this cluster are referred from a GP, and also has the highest number of referrals from a GP across the clusters (n=374). It is also the cluster with highest percentage of referrals from School Health Services (4.7%).

Cluster 2 has the highest percent of patients being referred for *suspicion of defiance/-conduct disorder* (59.8%), and by far the largest amount of patient with this referral reason (n=638), which is 2.4 times more than second largest frequency. Disregarding cluster 5, it is the cluster with the lowest percent of referrals with *suspicion of ADHD* (24.8%).

Cluster 2 also has both the highest rejection rate (17.2%) and number of rejected patients (n=183). It has the highest percentage for ICD-1=blank, i.e. no diagnose (31.8%) with n=339 cases, but also highest percentage of ICD-1=Z032 (15.7%). It has

the highest number of ICD-1=F901 across all clusters (n=46). It also has the highest amount of ICD-1=Missing (i.e. there is not enough information for coding on axis 1) at n=123.

It has the highest number of parents cancelling as reason for ending referral period (n=70).

Cluster 3:

Cluster 3 is characterised by being the cluster with most boys, highest number of *suspicion of ADHD* as referral reason (n=719) and largest amount of ICD-1=F900 (n=476) as diagnose.

There are n=1,223 patients in cluster 3, which makes it the largest cluster. The patients are in the age 8-13. Cluster 3 has the highest percent and amount of boys of all the clusters (88.1%, n=1078).

Cluster 3 has the lowest percent of both mom and dad having custody; 84.4%. Cluster 3 has the largest amount of patients with a step mother and biological father (n=7).

Most patients (34.7%) in this cluster are referred from the Psychological Educational Service, and also has the most referrals from this service (n=424). It is also the cluster that has the least percentage of referrals from a GP, at 15.9%.

After cluster 5, it is the cluster with second highest percentage of *suspicion of anxiety* as referral reason (3.1%). It has the highest number of *suspicion of ADHD* as referral reason (n=719). The strongest first referral reasons for this cluster with regards to frequency are *suspicion of ADHD* (n=719) and *suspicion of defiance/conduct disorder* (n=266).

Cluster 3 has the largest amount of ICD-1=F900 at n=476. Across the clusters, it has the largest number of ICD-1=None (n=52) (i.e. a patient was assessed and was found to meet no diagnose criteria for axis 1).

Cluster 4:

Cluster 4 is characterised by having the oldest subgroup of patients, and the most even gender balance.

There are n=961 patients in cluster 4, and this cluster has the highest mean age of 14.2 years and a median of 14. The patients are in the age 11-18. Cluster 4 has the most even gender balance; 58.1% are boys, 41.9% girls.

Cluster 4 has by far highest number of patients living in institution as their care situation (n=62).

Most patients (37.1%, n=357) in this cluster are referred from a GP, as well as being the cluster with highest percent of referrals from this referring instance. It is also the cluster with highest percentage of referrals from the Municipal Child Welfare Services (8.8%).

The strongest first referral reasons for this cluster with regards to frequency are *suspicion of ADHD* (n=416) and *suspicion of defiance/conduct disorder* (n=271).

Disregarding cluster 5 due to the size being small, cluster 4 has the highest percentage of patients being referred on *suspicion of depression* (9.3%), as well as the highest number for *suspicion of depression* (n=89). It has the highest number of *suicide risk* at n=14, and the highest number of *suspicion of eating disorder* (n=10). Additionally, it has the highest number of referral reason being *learning difficulties* (n=24), *absence from school* (n=13) and *school refusal* (n=6) across the clusters.

Cluster 4 has the second largest percentage of rejected referrals (15.6%) and number of rejections (n=150). It also has the largest amount of the patient cancelling as reason for ending referral period, n=90, in addition to the largest number of patient being above age limit as reason for ending referral period (n=57).

Cluster 5:

There are n=31 patients in cluster 5, which makes it the smallest cluster. The patients are in the age 4-15. Cluster 5 is characterised by having the highest mean number of stays; 9.7 with a median of 8. This cluster also has the highest maximum number of stays = 24, and a minimum number of stays = 7. This cluster almost has the same gender balance as cluster 4; 61.3% boys and 38.7% girls.

Cluster 5 has the highest percent of the mother having custody alone; 29%. Furthermore, no patients in cluster 5 are in a situation where the father has custody alone.

Regarding care situation, cluster 5 has the lowest proportion of the patient living with both parents (29.0%). It also has the highest rate for living in an institution (12.9%) and the highest rate of living with one parent (35.5%), which agrees with the high number of mothers having sole custody.

Most patients (29.0%) in this cluster are referred from a GP. All 31 patients in cluster 5 were accepted. The majority of patients were referred elsewhere after ending the referral period (n=11, 35.5%). It is the only cluster with more patients being sent elsewhere, than back to referrer.

Cluster 6:

There are n=298 patients in cluster 6. The patients are in the age 2-17. Cluster 6 has the second highest mean number of stays; 3.7, and a median of 3. There is a minimum of 3 stays and a maximum of 6 stays. The remaining clusters (1-4) have a mean value of 1.1 stays.

Cluster 6 has the largest percentage of patients with adoptive mother and adoptive father (3.0%).

Most patients (35.6%) in this cluster are referred from the Help Service for Children and Young People, as well as being the cluster with highest percentage of referrals from this referring instance. It is also the cluster with highest percentage of referrals from CAMHS (10.4%). After cluster 5, it is the cluster with the second highest percent of *suicide risk* as referral reason (2.3%).

After cluster 5 with an acception percentage of 100.0%, cluster 6 has 99.3% accepted referrals.

7. Evaluation

This chapter is concerned with evaluation of the experiment. It looks into the experimental implementation, model performance, results and aims, as well as the numerous challenges and limitations of the experiment. Additionally, the section presents the clinical evaluation that has been conducted. The clinical evaluation is significant for the understanding and interpretation of experimental results, and a major part of the analysis of electronic health records in CAMHS. Any revelations and understandings made along the way are also included with the purpose of better facilitating future work.

The aim is to evaluate clustering techniques and experimental results from the previous chapter, by investigating how relevant and meaningful the clusters are, and if the results were as expected. This is presented in section 7.1 and 7.2. Following this, the clinical evaluation of which the results have been evaluated and discussed in collaboration with a selection of both research- and clinical professionals, is presented in 7.3. Lastly, the experimental process itself is evaluated in section 7.4, which also includes a thorough description of experimental limitations. The discussion of results is left for chapter 8.

7.1. Model Evaluation

By using the Elbow method, the optimal number of clusters was theoretically estimated to be k=3, but as can be seen in figure 6.22, k=5, k=6 and k=7 were also possibly good candidates. In an effort of trial and error to find the most suitable number of clusters, k=3, k=4, k=5, k=6 and k=7 were experimented with. At k=6, we started seeing more nuances in the cluster sets, e.g. the first cluster with a girl as centroid appeared, as well as more variation in referring instance and referral reason.

However, even though clusters should preferably be of the same cluster size, k=6 yielded a cluster of n=31 patients. EDA prior to clustering revealed that most patients have very similar situations, especially regarding family-, care- and custody situation. Thus, the nuances in the dataset appear in smaller quantities because a minor number of patients deviate from the most common patient situations. Due to this, it was decided to move on with k=6 to also highlight the small nuances. The smallest cluster did in fact prove to capture the smaller nuances in the dataset. However, it was necessary to account for having a very small cluster when analysing cluster results. Having one much smaller

cluster, percentages could not be used as comparative measure to the other clusters to the same degree, thus the focus was more on quantities rather than percentages when comparing the clusters.

To evaluate how strongly the features affect each cluster, a model interpretation framework called SHAP is used. SHAP is useful for explaining the output of a machine learning models and is accessible as a Python library (O'Sullivan, 2021). Figure 7.1 illustrates the importance of different features for each cluster. Note that the clusters are zero-indexed in this figure, but are referred to as cluster 1, cluster 2, etc.

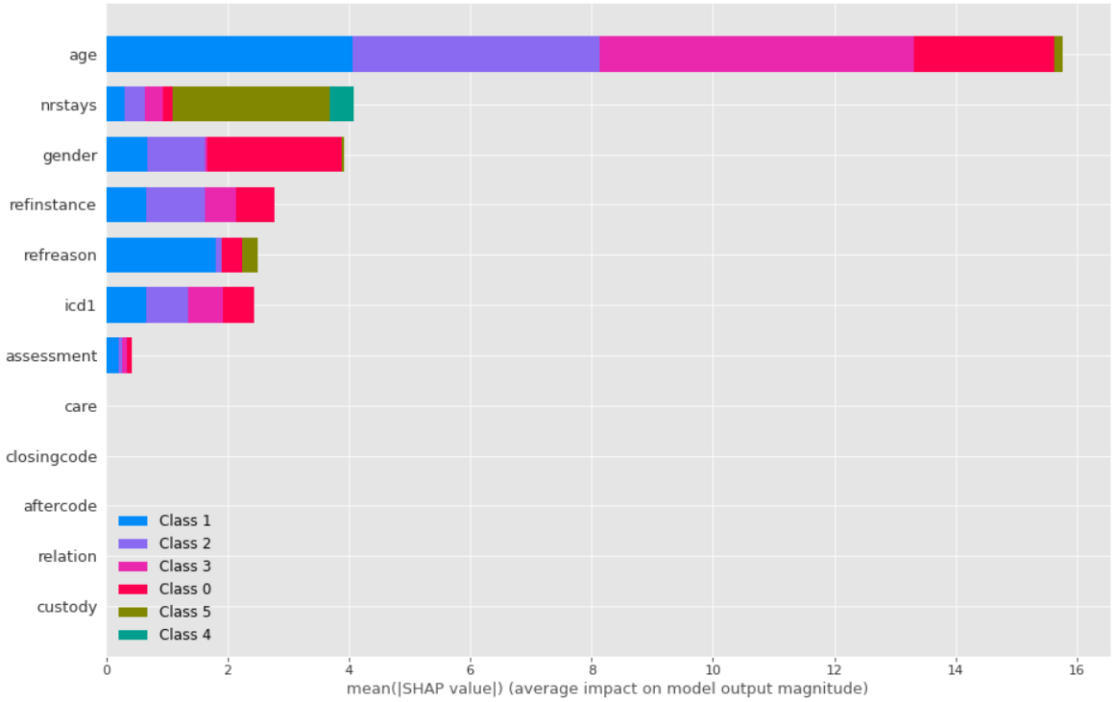


Figure 7.1.: Principal experiment: SHAP summary plot of feature importance for each cluster.

Features like age, number of stays, gender, referring instance, referral reason and diagnose on axis 1 strongly affected the separation of the clusters. Furthermore, some clusters are more dominated by certain features; Cluster 6 is largely dominated by the number of stays for each patient in the cluster. Gender has largely affected cluster 1. Referral reason very much affected cluster 2. This has already been confirmed when looking into the general characteristics of the clusters in section 6.5.1, and the SHAP plot evaluation method provides a useful visual representation of the cluster characteristics.

On the other hand, as has already been seen in the cluster centroids, care situation, closing code, after code, relation and custody do not tend to vary between the clusters, and this is also confirmed by the SHAP plot. They have little to no impact on the clusters. It was hoped for these variables to be more distinguishable, in order to properly assess a patient's situation prior to being referred to CAMHS. This is most likely due to these variables having very little variation; most of the patients have the same parental relation, custody, and care situation. As briefly mentioned above, this became apparent during the EDA in section 6.3, as shown in the different bar plots. Assessment outcome also has limited impact on the clusters, but most patients are also accepted. However, assessment outcome mostly affects cluster 1 and 2, one of which has the largest amount of rejected patients. This essentially means the cluster process was able to accumulate most rejections in one of the clusters.

F1 score is a another evaluation tool and a common metric in machine learning that can be used to assess the quality of a model. We assess the quality by using a cross-validation model, which produces an F1 score. An F1 score close to 1 is desirable. The cross-validation F1 score for our K-Prototypes clusters is 0.99024, which means that the clusters that were produced are meaningful and distinguishable. Despite figure 7.1 showing some features were less essential in the context of determining the clusters, the clusters were nonetheless distinguishable.

7.2. Result Evaluation

This section briefly evaluates the output from clustering with the K-prototype algorithm, by assessing the results and comparing them with the experimental aims. The discussion of the results itself will be presented in section 8.

7.2.1. General Evaluation

Firstly, the number of both columns and rows for such an experiment is rather low. This was, as discussed in section 7.4.4, due to having to prioritise which columns to include. Regardless of the actual clustering results, this will to some degree affect the output

of the experiment. By having to opt out so many rows (i.e. patients) due to null- and invalid values, the subgroup of the cohort is already limited in extent. This may be considered a research weakness, as the data basis is on the smaller side.

It is important to mention that in the process of removing insufficient and poorly coded patient records, one is essentially also removing records that have low data quality, e.g. poorly written referrals. This consequently means that when having to remove a large number of rows, it may also be an indication of high frequency of poor referrals and/or record keeping in CAMHS. As some records are excluded due to poor coding after CAMHS received the referral, it can not be denied that insufficient coding is a challenge in several areas of clinical psychiatric practice. Nevertheless, mitigation strategies were applied to ensure an acceptable number of patient records, and as we shall witness, the results came out strong and provided useful findings.

The clusters were able to identify outlier patients, as the smallest cluster assembled several of these. Cluster 5, with a size of n=31, collected a lot of the outlier patients, which is useful in the context of clinical clustering. Outlier patients are common, regardless of the group size, and isolating these is useful for identifying their characteristics. In this case, with cluster 5 having the highest mean number of stays at 9.7, isolating outliers like this in one group may give a more precise clustering of the other patient subgroups. However, since cluster 5 is so small, it is difficult to compare the patients with the other clusters because of the percentages not accounting for the size of the cluster. This was mitigated by mostly comparing patients across clusters by frequencies, and not by percentages.

One of the more unexpected findings was that the results weakly distinguished patient situations before referrals were made, but were able to clearly distinguish patient situations after referrals were made. Even though the ambition was to better describe patient situation prior to a referral, the results were stronger than expected for the other variables. Clustering managed to evenly distribute the age groups and number of stays into different clusters. It also captured the majority of girls, which is the minority gender group in this cohort, in cluster 1 with a percentage of 74.7%. This gives valuable insight into gender related differences. Referring instance and referral reason also have a significant impact on the clusters. The largest number of rejections, 17.3% was assembled in cluster 2, as well as most referrals from a GP (35.1%). This cluster also has the largest percent of patients being referred for *suspicion of defiance/conduct disorder*. Findings like these are especially useful in the upcoming discussion on rejection rates, and referring instances and reasons that are most prone to rejections. These are just some of the results that confirm the value of clustering clinical data, and will be further discussed in light of the clinical evaluation in chapter 8.

7.2.2. Evaluation of Experimental Aims

This section investigates whether the experimental aims have been met. The aims of the experiment were defined in section 6.1.1.

The two first aims are concerned with identifying patient subgroups, and assessing the feasibility and usability of clustering as a tool for identifying these subgroups.

Regarding the first aim, this has to some degree been accomplished. The clusters were partly able to distinguish between certain variables like age, number of stays, gender, referring instance and referral reasons. Variables related to the patient's situation prior to referral, were less distinguishable.

In short, cluster 1 is dominated by gender, age and referring instance, cluster 2 by age and referral reason, cluster 3 by age and referring instance, cluster 4 by age and registrations on axis 1, cluster 5 by number of stays and referral reason, and cluster 6 by number of stays. Those are variables to pay extra attention to when moving on to the discussion.

The second aim is entangled in the first, as it describes the *ability* the clusters have to identify subgroups. Clustering is a process which can yield interesting and useful results, as proven in this experiment. Prominent outlier patients were identified in their own cluster, and other clusters had mostly strong characteristics. According to the the F1-score in the model evaluation in section 7.1, the clusters are also distinguishable.

Clusters were indeed able to identify subgroups of patients, but maybe even more importantly to identify variables and relationships one should pay more attention to. The high rejection rate of the GP, the weak correlation between suspicion of ADHD and gender, or the high frequency of suspicion of conduct/defiance disorder in rejected referrals were all identified through clustering. These are elaborated on in the next chapter.

However, an important addition to this is that in the identification of patient subgroups, the clusters largely benefit from being used in combination with an iterative EDA. The use of bar plots and scatter plots are useful for the initial identification of phenomena, and when clusters have further detected any patterns of interest, new plots can be made to scope in on important findings in the clusters.

The third aim, which is related to gender differences, was accomplished, but in an unexpected manner. While it was hoped that boys and girls had at least somewhat distinguishable referral reasons, these were less separable than expected. Expectations were in fact contradicted; the cluster with the most girls also had the highest percent of referrals reasoned with *suspicion of ADHD*, the highest percent of ICD-1 = F900, and the least percent of *suspicion of defiance/conduct disorder*. It was expected to see more

suspicion of anxiety and *suspicion of depression* in the cluster with the largest amount of girls, but this was rather seen most frequently in the cluster with the most even gender balance. *Suspicion of anxiety* was most frequently seen in the cluster with most boys.

Regarding the fourth aim, the rejected cohort was not primarily assembled in one cluster, but in two. Cluster 3 and 4 stood out with 17.2% and 15.6% rejection rate. These were the clusters of which one cluster had the most boys, and the other cluster had the most even gender balance. It would be desirable to assemble more of the rejected patients in one cluster to better describe their situation, but these clusters both had most referrals from a GP, and were able to capture the fact that the GP is prone to rejections. Furthermore, *suspicion of defiance/conduct disorder* was also captured by the cluster with most rejections. This is one of the most common, but also one of the most rejected referral reasons, and contributes to the description of patient situations that are prone to rejection.

Regarding the fifth aim, the different stages of every patient's first referral period, the clustering gave more insight into the later part of the referral period. The results given by the cluster give less insight into patient situation prior to referral, and more insight into the differences and nuances in patient situation post referral. This was rather unexpected. It was hoped that clustering would yield more insight into family and care situation, alas it did not. On the other hand, important phenomena were captured especially around the time of making the referral, which are discussed in the next chapter.

As a result of this, the clusters do not strictly separate between patient situations like family relationships and care situations, but are able to identify some key referral situations, gender phenomena and interesting referral reason patterns, to mention some.

7.3. Clinical Evaluation

The aim of the clinical validation is to present, discuss and interpret the findings with a group of professionals, and to ensure professional anchoring of the research conducted in this Master's thesis. Yet again, it is significant to emphasise the importance of understanding all the information we analyse in the context of clinical practice in Norway, and the clinical evaluation is an important contribution to that understanding. The results from the clinical validation meeting is also a part of the experimental evaluation, as well as the discussion. If professionals can support the findings, this may strengthen the results. This section presents details regarding the clinical evaluation. In order not to repeat myself, specific details regarding the discussion and inputs given during the meeting has been left for chapter 8.

Initially, the ambition was for CAMHS professionals from St. Olavs Hospital to be present during the presentation to ensure the strongest clinical representation. Unfortu-

nately, CAMHS representatives were not able to attend due to being occupied with the implementation of Helseplattformen, a system responsible for the introduction of a new joint electronic health record solution in central Norway (Helseplattformen AS, 2019). However, an opportunity later arised to discuss the findings with a psychologist specialist from CAMHS, of which the outcome and inputs will be included in the upcoming discussion in chapter 8. Both the clinical evaluation and the meeting with CAMHS are essential to the interpretation of research results, and the research project itself.

The results were presented to the panel the 2nd of May, 2022, to a group consisting of members of the IDDEAS project. This group includes psychiatrists, computer scientists, researchers, developers and other professionals from *Regionalt kunnskapssenter for barn og unge* (RKBU), *Helse Midt-Norge IKT* (HEMIT), Department of Computer Science at NTNU, VIVIT AS, and international collaborators from USA and Germany. These are further referenced as *the group*. The aim was to present the results objectively and without my own interpretation in order to facilitate an uninfluenced discussion of the findings.

In the first part of the meeting, the group was presented with the specific data selection, including any data requirements and selection criteria, in order to understand the data basis for the research. Following this, the exploratory data analysis was presented. Any bar plots or scatter plots illustrating referral period details alone or in combination with other variables were objectively presented, before highlighting any noticeable or interesting phenomena in the data. Lastly, the clustering process was presented to the group, by firstly describing the cluster centroids and feature importance, before key findings were presented.

Previously published literature, as well as national and international clinical practice today, was used by the group as reference for comparing results. It was confirmed by the group that the findings were all consistent with previously published data. The age and gender distribution in our cohort is consistent with the age and gender distribution in published literature, with girls being referred later than boys. Family and care situation prior to referral were also as expected. Some results, like the quantity of rejections of referrals from a GP, were found to be odd, but not unexpected. Referral reasons and their prevalence were also consistent, even though some referral reasons like *suspicion of anxiety* usually is not more commonly seen in boys than girls. The prevalence of *suicide risk* in young boys also caught the group's attention, which may be interesting to look into at a later time. Any diagnoses were also as expected, however the discussion highlighted the need to investigate the prevalence of blank entries and diagnose code Z032 further. It also emphasised the need to look into why accepted patients are not given a diagnose, and what might impact this.

It was frequently repeated that the ratios of our data seem to be more accurate than older literature, and more consistent with the most recent literature. Numerous studies claim that boys outnumber girls 4:1, but the more recent studies claim the reality is more

2:1. Our data consistently presents the ratio as 2:1.

The discussion was very useful for confirming the relevance and validity of the results, to clarify any concepts not entirely comprehended on my hand, to highlight some phenomena that are quite important, and especially to identify the great potential of future work.

During the meeting, an important, potential mapping bias on my side was also identified. The frequency of blank entries is the second most common value for ICD-1, i.e. a diagnosis on axis 1. However, *blank* in this dataset indicates a patient that was either rejected, or a patient that was accepted but was not found to have a diagnose. These codes were merged during preprocessing because they both indicated no diagnose, but in retrospective, they do differentiate between two separate groups of patients. This essentially means I might have created a confounder myself.

As remarked by the group, these patients need to be analysed separately and together to see what difference it makes. Just because they are registered with *blank* on axis 1, does not mean they do not have a disorder. Our aim is to separate out what the circumstances are that make them being recorded as *blank*, *missing*, or remain Z032. We need to understand what makes these patients fall into either of those categories as opposed to getting a diagnosis. Thus it crucial to keep this in mind when looking into the diagnose codes, and avoid this misstep in future research. This is also discussed in the next chapter.

To summarise, the group could conclude that the work confirms previous knowledge on the area and also shed light on the clinical reality in CAMHS. The ratio between boys and girls in this dataset is more consistent with the most recent literature, and is therefore more precise than older studies on the area. Other data were also highly consistent with what literature tells us and what can be seen in practice. Furthermore, the study unveils and highlights several areas that are necessary to look further into, and the work provides a solid foundation for future research. It is highly relevant and useful to specifically look into how rejections of referrals from the GP can be mitigated, why the use of Z032 is so widespread in use, and to investigate what actually happens to the patients that are accepted, but is either not given a diagnose or remains at Z032.

Lastly, it is important to emphasise what the evaluation of these results tell us. Across all results, they were found to be consistent with international publications and to what is seen in practice in CAMHS clinics in Norway. This proves the research feasibility clinical data has to produce accurate and consistent results. As long as proper strategies and techniques are applied to mitigate pitfalls like incorrect coding and human error, the data can produce accurate, relevant and useful results.

7.4. Process Evaluation

This section aims to evaluate the process of the work with this Master's thesis. The evaluation emphasises choices, compliance with time frame, prioritising of solutions, and choices made regarding future work. We also discuss the process of investigating netDx as a clustering tool for building patient similarity networks, and the decision to move away from that solution. Lastly, a thorough evaluation of experimental limitations are discussed.

7.4.1. Discard of netDx

netDx, a patient classifier for building patient similarity networks (Pai et al., 2019), was initially the algorithm of choice for clustering. This was chosen due to solicitation from supervisor in order to explore its utility in the context of measuring patient similarity on the basis of clinical data. netDx was discarded as late as March 2022, and due to the impact on the experimental process and time frame, it has been given space for discussion. The aim is to provide reasoning for discarding it, give future researchers insight into the use of netDx in similar contexts, and aid the decision of whether it is suitable for their purpose.

While evaluating netDx as my research tool, this led to several interesting findings and conclusions. It was discarded after a long period of trial and error. There were several reasons for this:

1. The documentation was to some extent useful, but not nearly as extensive as more well known and popular clustering algorithms. Thus it was not always easy to find the answer or support one needed in order to install and run netDx.

2. A lot of packages, libraries and dependencies for the netDx-package made the installation and setup process cumbersome. Several reiterations of installing the programming language R, the packages and the dependencies were necessary to figure out which version of every component that was compatible with the other components.

3. In order to run analysis on the datasets, the dataset had to be transformed to specific objects that could be loaded, as opposed to directly loading a .csv-file like traditional clustering algorithms.

4. netDx had proven great success in many medical areas, e.g. for the classification of breast cancer types (Pai et al., 2019) or colonial cancer subtypes (Oslo University Hospital, 2018), but first and foremost when there was a classification task to be done. In fact, netDx is an algorithm best suited for classification rather than

clustering. Given the amount of work for setting up the environment, loading data, and using the algorithm, other more traditional clustering algorithms could easier and faster manage the task of clustering the dataset. This is also the case for my own project, as I rather late discovered that simpler, more manageable clustering techniques would be just as good, or even better, for my task.

5. netDx was developed for the purpose of building a patient classifier from patient data (Pai, 2021). Given this specific context, it would be reasonable to assume that most users would be in the medical and/or data science field. However, given its specific target group, one would also expect it to be much easier to install and implement. Due to this, I believe that the somewhat challenging use of netDx could make it unsuitable for professionals strictly in the medical field, especially without the assisted competence of a data scientist.

In retrospective, I would have done some things differently. These are summarised in the list below.

1. As a part of my fall project report, I also elaborated on the theoretical aspects of the netDx algorithm. However, I should have sooner started on the practical aspects, i.e. the setup and installation process. This could have made me faster come to the realisation that this process was too cumbersome and more advanced than necessary for my task.

2. As I was setting the limit for asking for help a bit too high, this reduced the progression of the work. When first reaching out for help, this lead to faster progress and more findings regarding the usability of netDx. Had this been anticipated at an earlier stage, more time could have been put into additional clustering experiments.

Some positive takeaways from this process are:

1. Being able realise when to move away from a solution and look in new directions, is a part of any plan. For this project, it was the right choice to make. As a result of this, I was able to change the course of direction with reasonable time left for carrying out the clustering experiment.

2. Considering the significant amount of time spent in the HUNT Cloud Workbench attempting to manage netDx, this process yielded valuable insight and competence in the use of the HUNT Cloud services. This was useful knowledge, like the use of cloud applications, data loading, working in a remote environment, and how to handle sensitive data. This meant less setup time and faster embark on the actual clustering task.

7.4.2. Evaluation of Time Frame

This section aims to briefly evaluate the experimental time frame, which was presented in chapter 6, table 6.1.

First and foremost, the implementation of the experiment was by no means as linear as the timeline suggests. This is due to a number of factors.

At the start of the semester, a significant amount of time was spent trying to set up the environment in HUNT Cloud, of which the components have been described in section 5.2. Working in a cloud environment, and additionally with sensitive data, was a new experience, which required a substantial amount of time. This was both due to not paying enough attention to this in the fall semester of 2021, and because it was not clear at the time which tools I needed. Furthermore, time was spent trying to make the netDx algorithm work, which eventually was discarded before moving on to the K-prototype clustering method mid-March 2022. For future work, it is advised to map, get acquainted with, set up and experiment with any tools at the earliest time.

Another impactful delay was due to the lack of a well-defined and clear goal of the research. Having investigated netDx for a whole semester, I still did not have a clear picture of its implementation in my project. It took me some time to realise that this was because it was not suited for my research at all, as described in the former section, 7.4.1. Another literature study on similar studies was required to clarify the ambition of the research, and with the additional insight in the dataset made in January and February, the project was redefined in March.

Regardless, even though the experimental scope was redefined around week 12 rather than week 8, the result analysis was only delayed by one week. The clinical validation was postponed twice, but eventually within the goal week of 18, and the project aligned with the experimental time frame from this point on.

7.4.3. Additional Sub-Experiments

Upon completion of the principal experiment, some sub-experiments were in fact conducted. The aim of these were to (1) explore the predictive ability of assessment outcome, and to (2) isolate referral situation. However, along the way it was realised that the results of these largely overlapped with the main experiment; the findings were already present in the first experiment. Therefore a choice was made to not include these, as they did not contribute with anything new, but overlapped with existing results. Furthermore, even without the inclusion of subsequent experiments, the report was already quite long. It was thus prioritised to focus on a good analysis and discussion of existing results, and to thoroughly describe the extensive potential of future work to facilitate future research.

7.4.4. Experimental Limitations

There are many challenges to working with clinical data, and to this research in general. As discussed in section 6.2, clustering analysis favors data with few unique codes and missing entries. In this case, it is a constant trade-off when deciding what data to include, and what data to exclude. We want to include as much information as we can, but not on the condition that majority of the information is faulty or missing, which will negatively affect the clustering analysis. This section elaborates on the experimental challenges and limitations, in order to provide reasoning for the choices made, and assist future research in working with similar clinical data.

Cohort Limitations

This section covers which limitations in the cohort that restricted the data selection. This also includes what the initial ambition was to investigate, and what kind of downscaling had to be made. These limitations are mainly due to poor data quality like mistyping, human errors and missing entries, as well as duplicates of what should have been unique codes.

Going into this experiment, the aim was to investigate whether patients can be characterised and grouped by their situation from the time of making the referral up to the different treatment and contact types at CAMHS. The scope was to investigate the first referral period of every patient relevant to the research. In retrospective, the research was able to describe in detail the process from making a referral up until assessment of the referral. However, in the context of data coverage, the ability to assess situations after the end of a referral period was limited, specifically concerning during a patient's stay at CAMHS.

Some of the data that could not be prioritised, were relevant for capturing what happens in these associated stays at the CAMHS clinic. Section 2.2.1 describes some of the terms related to these stays, like activity and contact type. It would be useful to look into which CAMHS unit that was responsible for the care of a patient, which activities and contact types that were a part of a stay, and what professional personnel that was involved. Since the number of stays/episodes for each referral period was a fairly easy task to retrieve, specifically contact type and activity type would be interesting to include. However, I was simply not able to identify the associated database fields or receive the correct answers in time to include these in the experiment. To ensure the research results were valid and reliable, the data I used had to be safe, approved and correctly mapped between numerical value and actual clinical term. This meant a more solid basis, at the cost of shortening the patient trajectory scope.

There were also some limitations to the investigation of situations prior to referral

assessment at CAMHS. Even though interesting features like number of days from referral to assessment do not necessarily have a lot of unique values, there were major drawbacks of including it as a feature. This was one of the features I was quite interested in, in order to investigate the progress time for assessment in our cohort and whether these were in appliance to recommended progress time. This aspect was discussed in section 2.2.3. Progress time can be calculated by subtracting the referral date from the assessment date, but in many cases these were mistyped and we would get a negative difference. This would require a restriction of only including a positive difference, which would reduce our dataset with almost 40%. Note that even though many have valid dates for when treatment began and when the referral period was over, it would not make sense to include these because we are interested in both those who are accepted and those who are not. Those who are not accepted will not receive treatment or care, and their closing date would either just by the same, or very close to, the assessment date.

Furthermore, note that even though *ansatt.fagkode* (i.e. the professional code for any employee registered in a case, stay or journal entry) is not included as a column in the PostgreSQL-query of section 6.2, it is still included as a restriction. This is due to the fact that for every employee ID, there were numerous associated duplicates of profession and professional code. Furthermore, for every professional code, there were tens of associated professions, which made it an unsuitable identifier for those who have worked on a case, and it was left out as a column. However, having several duplicates, without restricting the number of employees, the number of stays associated with each patient would also be distorted, which is the reason for including restrictions of such non-present columns.

Another interesting column, and yet with too many missing values, is the reason for referral related to the child's environment (field: *sak.henvgrunnm1*). When including it in our statement, as much as 65% of the records did not have a registration for this field. If this field was to be included, being a categorical column, the patients with the value *not filled in by referrer* would seem more alike, even though this is just an incident of bad referral quality. To keep the consistency and quality of the data selection, this was thus left out.

The same is also applicable to both the second and third reason for referral related to the child itself *(field: sak.henvgrunnb2 and sak.henvgrunnb3)*, as these columns had too few registrations.

Other interesting columns worth mentioning that were mainly excluded due to deficient data, are *sak.henvgrunnm2, sak.henvgrunnm3, sak.hjemmel, sak.barnevern1, pasient.etniskmor, pasient.etniskfar* and *pasient.hjemmesprk*.

Lastly, another reason for not including some columns is due to missing map for the most frequently used codes. An example is *sak.hjemmel*, which most frequently is coded as 11, of which there is no mapping (i.e. the code map list only spans from 1-8). In most cases, one could assume this is the equivalent to *not filled by referring actor*, but one

should refrain from making such unqualified assumptions.

Time and Resources

This section briefly describes the limitations regarding time and resources in the project, and how this was mitigated when having to make choices. The research, being highly interdisciplinary, is dependent on other human resources for its implementation. This was mainly due to own academic background being in computer science, and not medicine. These resources were not always available at the most convenient time.

As new discoveries and needs were unveiled throughout the project, there were also emerging needs for answers or input. As some aspects of the project were changed rather late, e.g. the use of K-prototype in place of netDx, this lead to some delay, which meant less time to figure out any emerging obstacles further along the way. Mostly, I was able to reach the right person in time, and could support my decisions on their feedback. However, if it was not possible to reach the right resource in time, this forced making own assumptions. An example of this is in the case of which referral reasons to group together, or when having to exclude any contact- or activity types that are used post assessment at CAMHS, because there was not enough time to await clarifications or reliable answers on which database fields these were associated with. If neither help or time was available, choices were made, but these were carefully documented.

A valuable lesson with regards to this, is first and foremost to initiate the work as early as possible, including getting one's hands dirty at the earliest time. This essentially means that simultaneously as one begins to build domain knowledge in a foreign field, one should also get into the practical aspects as soon as possible. This will give more time to handle upcoming obstacles and challenges, as well as sudden change of plans. Additionally, being able to move away from a solution can often be better than spending more time having no progress, like discarding netDx as the algorithm of choice. Secondly, it important to reach out and ask for help from the very beginning. Usually there are resourceful people that can, and are happy to assist with arising challenges, which can free up time to focus on engaging the research further. Leaning on domain professionals and resourceful people is crucial in a project like this.

Errors Detected

This section briefly touches on errors detected during evaluation of the results. One error was detected, and the research limitation it makes is therefore assessed. The aim is to prevent the same error in the future.

There is one weakness to the mapping between diagnose code and description that is

important to document. Numerous rows had either 000, 999, 1999, 1000 or simply no entry at all (null). According to Helsedirektoratet (2022), 1999 equals to *Insufficient information to code on axis 1*, and 1000 equals to *No condition proven on axis 1*. The latter is used for accepted patients that were not found to have a diagnose at the time. A blank field usually indicates a rejection, and is also present in the records of accepted patients. However, entries with a single value of zero were initially mistakenly merged with *blank*.

Upon further inspection in the database, the rows that have 0 as value, were actually coded as 000, but the code had been shortened during conversion to *.csv*. Furthermore, all 33 patients with code 000 were accepted patients. That means it is reasonable to believe that 000 is supposed to be 1000, and 999 is supposed to be 1999, as neither 000 or 999 are valid codes in the list of CAMHS codes from 2021 Helsedirektoratet (2022). This reasoning was also confirmed during the meeting with psychologist specialist Jostein Arntzen, 11.05.2022. Codes that are recorded as either 999 or 000 are most likely meant to be 1999 and 1000, respectively, for axis 1. For any subsequent axes, it is also reasonable to presume that similar combinations of 999 and 000 are 2999 and 2000 for axis 2, 3999 and 3000 for axis 3, and so on.

The entire dataset was re-extracted, re-mapped and re-processed to chart the extent of the mistake. Thus, 000 and 1000 were mapped to 1000: *No condition proven on axis 1*, and 999 and 1999 were mapped to 1999: *Insufficient information to code on axis 1*. A total of 33 misplaced patients were moved from *blank* to code 1000.

After re-analysing the dataset, the top most frequent ICD-1-registrations are the same as before, and in the same order, as illustrated in figure C.2 in appendix C. Code 1000: *None made* now includes 33 more patients, moved from *blank*. The error was fortunately small in extent, and it is reasonable to decide that the clusters do not need to be redone. However, this emphasises the importance of being thorough when mapping and merging codes.

7.4.5. Evaluation Summary

This section aims to briefly summarise chapter 7: Evaluation, including the evaluation of the model, the results and the process itself.

Initial fine tuning of the number of clusters led to five reasonably even clusters and one cluster of size n=31. The clusters are dominated by patients with rather similar background, but the smallest cluster managed to capture the most prominent outliers. However, the exceptionally small size of the fifth cluster needs to be accounted for in further analysis. It is evident that some features like age, number of stays, and gender are more dominant, while features like custody, care and relation have very little impact. This is naturally also reflected in the clusters; They are mostly separated by the features

that were found to have a larger impact, and the less impactful features are evenly distributed throughout the clusters.

The results are to some degree affected by having to reduce the number of rows and columns to ensure a selection of sufficient data quality, which reduced the number of patient records included in the experiment. The ambition was to better describe different patient situations prior to the referral, but the results are largely dominated by many patients with similar background, and the nuances are not as apparent as hoped for. However, results came out stronger for other variables related to the referral process and the assessment at CAMHS, as well as gender differences in the cohort. This means the results provide a good basis for analysis of the time period from making the referral to the assessment of the referral.

With regards to the experimental aims, the clusters were found to not make strict separations between patient situations like family relationships and care situations, but were able to identify some key referral situations, gender phenomena and interesting referral reason patterns. It was also emphasised that for this specific analysis, it is beneficial to apply clustering in combination with an iterative EDA to best investigate any findings and discoveries. These reflections are important when moving on to the discussion in the next chapter.

In the clinical evaluation of the results, which took place 02.05.22, the results were found to be consistent with recently published literature. The clinical evaluation also confirmed the relevance and validity of the research. Furthermore, it was found that the findings confirm practical knowledge in the area of clinical practice and also sheds light on the clinical reality of CAMHS in Norway. This process yielded valuable clarifications, insight and input which will largely impact the discussion and interpretation of results in the upcoming discussion. In addition to the later meeting with CAMHS 11.05.22, they provide significant value to the comprehension of results in the context of Norwegian clinical practice.

There were also several experimental limitations to this project, primarily concerned with the data quality of the cohort, and secondly with regards to time and resources available. Insufficient coding and journal keeping by both referring instances and CAMHS clinics has been identified as an indisputable challenge in several areas of clinical psychiatric practice. A thorough presentation of the experimental limitations was provided to argue for any choices made in this project, and may assist any future researchers in the selection and prioritisation of patient records. Limitations with regards to time and resources can be improved by taking advantage of the lessons learned, and by early focusing on the practical aspects of the research and the utilisation of available human resources.

8. Discussion

This chapter presents the discussion of findings and discoveries in the research on patient characteristics and latent subgroups. The discussion has been organised around certain topics to clarify the findings. An attempt has been made to discuss the key discoveries from this research, both from the EDA and clustering process. Inputs and possible explanations from the discussion during the clinical evaluation and from the meeting with psychologist specialist Jostein Arntzen have been integrated in the discussion that follows. Any opinions or additions made by the clinical panel is referred to as made by *the group*. The professional discussion and clarifications are important to the comprehension of findings, and are therefore emphasised as a major part of this chapter. Finally, we summarise the discussion while also revisiting the research questions formulated in section 1.

As a reminder, the reader is encouraged to keep in mind that the results throughout this discussion are from an analysis on a dataset of which patients must meet the following criteria:

- Either have been referred with referral reason (database fields *sak.henvgrunnb1, sak.henvgrunnb2* or *sak.henvgrunnb3*) equal to *suspicion of ADHD* (including hyperactivity/concentration difficulties) or *suspicion of defiance/conduct disorder* (including behavioral difficulties), **OR** have a diagnose in the F90-group (F900, F901, F908 or F909).

- Only the very first referral period of every unique patient. Any subsequent referral periods are not relevant in this experiment, and the referral period is considered historyless.

In the experiment, unsupervised clustering was used to probe the latent subgroups of 4,201 patients with relevance to hyperkinetic disorders, either by referral reason or diagnose. Based on age, gender, and 10 variables connected to their first referral period to a CAMHS clinic in Norway, six clusters were generated.

One of the clusters aggregated both the highest number of rejections (17.2%) and the largest number of referrals from a GP (35.1%). Another cluster managed to assemble the largest amount of girls at 74.2%. The smallest cluster collected the patients with highest mean number of stays at 9.7. One cluster had both the most equal gender distribution and the highest mean age, and this cluster had the highest number of common comorbidities

like *suspicion of depression* (n=89), *suicidal risk* (n=14) and *suspicion of eating disorder* (n=10), in addition to several others.

8.1. Methodology

Before discussing the clustering results, we briefly elaborate on the methodology of the experiment, in order to clarify important prerequisites prior to further discussion.

Even though the elbow method recommended most commonly k=3 clusters, the clusters had centroids that were quite similar. EDA showed that most patients do have similar trajectories. The cohort is mostly gathered around the same values. This essentially means that one would need to raise the number of clusters to find outliers or the smaller subgroups of patients. Consequently, this would also mean that the size of the clusters would not be equal, but when they are not, the smaller clusters may be of interest. These can bring out the small nuances in the patient cohort.

Poor coding and the impact this has, is an important finding itself. This issue was also highlighted by Surén et al. (2018). In their study on diagnosis of hyperkinetic disorders among children in Norway, they emphasised the poor documentation of specifically diagnoses in medical records.

Bad coding quality in the clinics affects the data analysis and the quality of its outcome. This has affected the research of this thesis, as has already been elaborated on throughout the report. To maximise the potential of future research on clinical data, it must be ensured that coding is sufficient and according to clinical coding guidelines at the earliest time. The definition and use of local codes can be challenging in the context of analysing and comparing trajectories, and should be limited if possible. Additionally, registrations in the system must be validated to a greater extent than today, to ensure that e.g. mistyping and human error can be reduced as much as possible. For example, professionals should not belong to more than one professional code, and each code should only have one professional title. Closing date of a referral period should not be later than the date it is recorded. Parental relations like mom and dad should not be possible to register on the opposite parent; e.g. mom should not be a valid code for dad, and dad should not be a valid code for mom. Assessment date and start date must be a date after referral date. By increasing code quality, reducing the number of local variants, and increase validation of incoming registrations, this will increase the potential for good data analysis, and amplify the yield of the data. This is very important in order to ensure research in the field of psychiatry in the future.

It is important to highlight that some of the challenges presented here and in section 7.4.4, may be mitigated and even avoided by an upcoming implementation of ICD-11, as briefly mentioned in section 2.1. To summarise the section, ICD-11 eliminates the need

for local variants and enables simplified coding. Correct use requires less training, and coding takes less time (World Health Organization, 2022b).

Furthermore, one of the clusters will always have the largest or least percent of some variable; but it does not necessarily mean that the differences regarding the frequencies between the clusters are large or significant. This can for example be seen for variables like care situation and relational combinations, which are on average rather low across all clusters. Percentages for these variables do not affect the separation of clusters, and their impact is low. This has been kept in mind throughout this discussion. Furthermore, since the cluster sizes are so very different, the emphasis has been on comparison of quantities rather than rates to mitigate this and ensure percentages are not used to highlight an incorrect phenomenon; it is quite a different matter comparing cluster 5 with n=31 to cluster 3 with n=1223. For most cases, cluster 5 has been disregarded if cluster 5 has the highest rate of a given variable.

In the discussion that follows, the highest amount of something refers to the largest quantity n, and the highest percent refers to the highest rate in comparison to the size of the cluster itself.

8.2. Overall Cohort

Compared to the entire patient cohort, without the criteria of relevance to hyperkinetic disorders, the mean age is much younger in our dataset, as seen in figure 6.4. In other words, patients with relevance to hyperkinetic disorders have their first referral period at an earlier time and age than overall patients. The apex of our cohort is around 8 years old, as most of these patients are between 7-10 years old. In the entire cohort, the apex is around 15 years.

During clinical evaluation, the group added that it is actually quite interesting that patients are referred earlier than the overall cohort. Other difficulties and syndromes, for example Autism spectrum disorders or developmental syndromes, usually come into the attention of CAMHS quite early. In Norway we also have something called the Rehabilitation Service, which according to the group actually takes neurodevelopmental disorders more eagerly than CAMHS. Thus it is not obvious that patients with relevance to hyperkinetic disorders are referred earlier. But if there is a situation with both Autism and ADHD, these patients should usually be referred to CAMHS, which may explain the earlier referrals.

Furthermore, another aspect which was highlighted by the group contributes to explain why children with relation to hyperkinetic disorders are referred earlier than other patients referred to CAMHS. Kids who have behavior problems, which are more likely to be boys than girls, are more likely to receive other diagnoses as well because they are more likely

to be evaluated in the first place. For kids that do not have behavior problems, they are more likely to come to these services at a later time because they are not really causing any problem for anybody else. It is just a problem for themselves.

This is also something that can explain why girls often are referred later than boys for many different referral reasons; their behavioral problems are less expressive than boys' behavioral problems, and their symptoms more intrinsic, like depression. When boys are picked up earlier due to explicit behavior, they are also more likely to receive other diagnoses earlier.

8.3. Family and Care Situation

Most patients live with both their mother and father, in addition to both parents being biological. The majority of rejections are of patients with this parental relationship. If both parents do not have custody of the child, it is more common for the mother to have custody alone.

However, regardless of which parent, institution or parental figure that has custody of the child, it was frequently seen that *Biological mother and father* was recorded for parental relation nonetheless. Thus it may be more reasonable to firstly look at what kind of custody situation a child has, in order to look for situational parameters of impact and better indicators of the situation at home.

In figure 6.14a, it may appear that in situations when the mother has custody alone, there is a higher rejection rate. This may be true, but also notice the difference between the number of mothers having custody, and the number of fathers having custody. If patients that are referred do not have a biological mother and father (which is the most common combination), then the mother more often has custody alone than dad having custody, and the patients more often have a biological mother and a step father than the other way around. This phenomenon might be important to remark, but is also consistent with Norwegian custody situations, according to the group. On a national level, it is more common for mother to have custody, than the father to have custody, and our data may simply reflect this situation.

According to figure 6.14, it looks like the more composed a family is (i.e. both parents have custody of the child), the more likely it is to suffer rejection. According to the meeting between CAMHS and IDDEAS 18.11.2021, if CAMHS can assess that a situation can be handled at home or in the child's environment, this may influence a rejection of the case. If both parents have parental rights, they are in a better position to handle the situation. Parents with single custody may not have the the same prerequisite for dealing with such a situation. Another reason that can explain this is simply because both parents having custody is the most dominant combination, and thus it is also likely

to receive more rejections.

8.4. Referring Instance

The Educational Psychological Service has the most accepted referrals. GPs have the most rejected referrals, even though GPs have more referrals than the Educational Psychological Service in total.

Most patients (33.5%, n=208) in cluster 1 are referred from the Educational Psychological Service. Only cluster 3, with twice as many patients, has more referrals from this service.

Regarding referring instance, it was pointed out during the clinical evaluation that it is a bit odd that the GPs have so many rejections, but this is not a surprising result. It was neither unexpected to see that the Educational Psychological Service has the most accepted referrals among the referring instances, even though the GPs have more referrals in general. The group informed that this service consists of more psychology professionals, and are more aware of CAMHS, the needs and the routines. More importantly, the Educational Psychological Services may have less referrals, but their referrals carry more value. They are more accepted as a referring instance, and their referral quality is consistently high. The Educational Psychological Service has more time and more expertise to prepare the referrals.

Furthermore, it would be very interesting to investigate whether the referral quality of the GPs has improved the last five to ten years. According to the group, by going back 10, or even 15 years, GPs were known to often just sign a referral, without having seen the child, in order for the referral to be sent to a CAMHS clinic. This would potentially unveil whether the quality of referrals made by the GPs has been improved over time. An illustration of the development of referrals from GP's can be seen in figure C.1 in appendix C, but is essentially left for future work due to time constraints.

8.5. Referral Reason

The top five most frequent first referral reasons across our cohort are *suspicion of ADHD*, *suspicion of defiance/conduct disorder*, *suspicion of depression*, *other reasons*, and *suspicion of anxiety*, respectively. Note that *other reasons* is its own category, translated from the Norwegian referral reason *Annet*. See appendix A for mapping details.

When there is a high frequency of both *suspicion of ADHD* and *suspicion of defi-*

ance/conduct disorder, the count for firstly *suspicion of depression*, *other reasons*, and *suspicion of anxiety* are usually also high across the clusters.

As examined in the mapping of old referral reasons to new ones (A), only a total of 29 rows in the dataset that were coded with the old referral reasons were joined with the new referral reason code *other reasons*. This means that the old referral reasons that were joined, have very limited impact on the frequency of *other reasons*. Furthermore, *other reasons* must have very frequently been used since the implementation of the new referral reasons, in order to be the fourth most used primary referral reason, as seen in figure 6.10.

In other words, it is clear that *other reasons* is frequently used in situations where common comorbidities or symptoms relevant to hyperkinetic disorders are not suitable as referral reason for the situation, which may indicate that the existing referral reasons do not have sufficient coverage. Due to this, initial hypothesis was that referral reason *other reasons* is frequently recorded because the other referral reasons were not sufficient enough to describe a patient's situation and symptoms.

This phenomenon was discussed during the clinical evaluation, and the group was not surprised by its prevalence. Contrary to own belief, it was not unexpected for the group to see that referral reason *other reasons* was so frequently recorded. GP's and school staff most likely do not have the sophistication to distinguish between referral reasons and diagnoses. They know something is wrong, but they do not have enough information, rating scales, or the sufficient diagnostics. So the referring instance uses *other reasons* as referral reason, and then puts down the details of the child's condition. Their priority is not to distinguish between the disorders when they do not exactly meet the criteria for a condition, and they may not even have the means to; they just want to make sure the child is referred to an instance with competence on the issue.

Furthermore, *other reasons* as a referral reason is not really seen in correlation with a large number of rejections; *Other reasons* as a referral reason actually has the same number of rejections as for example learning difficulties, which for the record has a lot less total referrals. So even though all the referral reasons are quite narrow, the tendency to use *other reasons* instead and as safeguarding, has little impact on the referral outcome. It is thus an area that do not require as much attention as initially believed.

Regarding the prevalence and onset of other reasons like *learning difficulties*, *suspicion of eating disorder*, *suspicion of Autism* and *suspicion of Tourette's syndrome*, it is all consistent with the literature. However, *suicidal risk*, which is slightly more common in boys, and has later occurrence in girls, was an interesting result, according to the group. Suicidal behavior is rather rare in pre-pubertal children, not uncommon, but less common than in pubertal children. The total number of cases between boys and girls may though not be different, but it appears earlier in boys. So a question that arose is why is the occurrence earlier for boys, and are other behaviors, like accidents etc. deemed

as suicidal behavior? It is necessary to look into why children in this age are referred primarily on the basis of a reason such as suicidal behavior.

However, there is another aspect to this, as stated by the group. The system learns very quickly that if someone is said to be suicidal, they are much more likely to be seen and not rejected. This also applies psychosis or other severe diagnoses. So the system may be shaping the referral reason, even though that may not be the clinical finding. If someone refers a patient to a clinic because they are suicidal or have psychosis, this will get them in much more quickly than if they have ADHD or something that can wait. Understanding the system and how it works is important for understanding how referrals are made.

It is also interesting how children are referred with Autism at the age of 15-17. According to the group, it is really late to be referred on suspicion of a disease such as Autism, but not uncommonly high. So these children that are over the age of 7-10 years, are probably all very high functioning verbal kids. In standard literature, boys outnumber girls 4-5:1, however, in our data, which especially one of the psychologists believe to be better because they more accurately describe the ratio, boys outnumber girls 2.5:1.[1]

The same is with *suspicion of Tourette's syndrome*; Boys outnumber girls with the same ratio. However, most patients are referred on suspicion of Tourette's syndrome before the age of 12. As informed by the group, probably 50% of children with Tourette's syndrome also have ADHD, which would partially explain its prevalence. Hyperkinetic disorders are known to have a high degree of comorbidity, and is also a known comorbidity to other disorders. The frequency of primary referral reasons like *suspicion of depression*, *suspicion of anxiety*, *suspicion of Autism*, *learning difficulties* and other referral reasons for patients that are later diagnosed with a hyperkinetic disorder, confirms this known clinical fact that hyperkinetic disorders have a high degree of comorbidity. Some of these comorbidities are also confirmed by the data regarding diagnoses on axis 1; F952: Tourette's syndrome and F845: Asperger's syndrome are as prevalent as e.g. F908, Attention-deficit hyperactivity disorder, other type, or F913, Oppositional defiant disorder.

Moving on to the prevalence of *suspicion of ADHD*, it was highlighted by myself that *suspicion of ADHD* occurs earlier in boys, but once they occur in both genders, they have an even prevalence. It was informed by the the group that the prevalence are indeed even when you control for behavior problems; however, once one adds in behavior problems they are different, because girls do not as often have behavior problems. So there are nuances to *suspicion of ADHD* that the referral reason do not display. Girls are more likely to have the inattentive type, i.e. they do not pay attention, but they are not as impulsive and hyperactive as boys are. This may also explain the ratio between boys and girls for F900, compared to F901, F908 and F909, which have a higher ratio of boys to girls.

[1] Note that any reference to ratios between boys and girls in standard or published literature in this discussion is based on statements by the psychologist in the evaluation group.

8.6. Gender

Regarding gender, girls in our cohort seem to have their first referral period later than boys. Boys have an apex at 7 and girls an apex at 10, as seen in figure 6.5a.

Several referral reasons are recorded at a later age than boys, e.g. *suspicion of ADHD*. During clinical evaluation, it was informed that boys are evaluated for e.g. ADHD at an early age, and when they later get learning disabilities, they are already picked up by the system. Girls are evaluated at a later time, and learning disabilities may not be found until later. This will impact the time of which boys and girls have their first referral period.

As discussed in chapter 4, girls show a modified set of behaviours, symptoms and comorbidities compared to boys, and this also makes them less likely to be identified and referred for assessment. This is very evident in our analysis. Compared to the number of girls in the entire cohort, the number of girls in our specific experiment are 1/3 of the amount of boys, as opposed to an almost even number in the entire cohort.

Cluster 1 and 4, each with 74.2% and 41.9% girls respectively, have best captured the girls in the cohort, and may better describe situations specific to girls. Both clusters have most referrals with *suspicion of ADHD* as primary reason.

Suspicion of ADHD being the most common primary referral reason in cluster 1, is a rather surprising result. Cluster 1 has the highest percent of girls, the highest percent of *suspicion of ADHD*, and the highest percent of patients with diagnose F900, 43.6% and n=271. Only cluster 3 has more patients with ICD-1 = F900, n=476 (38.9%). Cluster 3 with most boys also has the highest frequency of *suspicion of ADHD* as primary referral reason (n=719).

Cluster 1 is by far the cluster with the least percent (14.5%) being referred for *suspicion of defiance/conduct disorder* (n=90). Cluster 3 has the second-lowest percent (21.2%) being referred for *suspicion of defiance/conduct disorder* after cluster 1 (n=638).

This result may indicate that gender does not necessarily correlate as much with the suspicion of ADHD or defiance/conduct disorder as one might initially think. The low separation of *suspicion of ADHD* and *suspicion of defiance/conduct disorder* in both the group with most girls and the group with most boys, may have little to do with the actual gender. Because the referring instance do not what condition or situation they have with a patient, they use referral reasons interchangeably. They refer on the basis of a set of behaviors, not on the basis of a diagnosis, as confirmed by the clinical evaluation group.

The clinical evaluation also confirmed that an even frequency of *suspicion of depression* in girls and boys is as expected. However, it was more uncommon for boys to have

more referrals on *suspicion of anxiety*. This may actually be due to hyperactivity being confused with anxiety in boys, which may influence the number of referrals with *suspicion of anxiety* as primary referral reason.

Surén (2018) highlighted the increased prevalence of mental illnesses in girls in Norway. Our study does not investigate the development of prevalence, but supports the high count of girls with symptoms of especially depression and eating disorder. The frequency of these referral reasons, keeping in mind that boys outnumber girls 2.4 times, are higher than other diseases like *suspicion of Autism* and *suspicion of Tourette's syndrome*.

Furthermore, when comparing the gender differences for diagnoses, it is clear how a high amount of girls seem to receive F900 and Z032, but the rest of the F90-group are more frequently recorded for boys (i.e. F901, F908 and F909). In F900 and F908, boys outnumber girls 2:1, while F901 and F909 are closer to 4:1. According to the clinical evaluation group, this was nothing unexpected, however, the ratios for F900 and F908 are somewhat different than expected. Usually, it is stated in previous literature that boys outnumber girls 4:1, but across the categories many of the diagnoses in our cohort are more like 2:1. Thus, having data of 2:1 may be more accurate, because the most recent reports state the ratio really is 2:1. This ratio is often due to girls being underdiagnosed because they are less hyperactive and more inattentive, according to the group. Hence, the high frequency of girls receiving Z032 as a temporary diagnosis may indicate that it was more difficult to determine a diagnosis for the girls, as their symptoms were more intrinsic.

In other words, more girls in our cohort are being diagnosed with F900 and F908 than would be expected according to previous literature, but this may be a more realistic ratio according to the most recent studies on the area. F908 and F901 behave as expected, but the ratio of boys to girls is notable regarding F901. Thus, the results are not inconsistent with the overall literature, but may be more realistic, because they are very consistent with the most recent literature stating that the ratio is closer to 2:1 due to underdiagnosis of girls.

8.7. Registrations on Axis 1

The most frequent recordings on axis 1 are F900, *blank* (indicating rejected patients or accepted patients that have not been given a diagnose), Z032, 1999 (insufficient information to code on axis 1) and F901, respectively. The clusters centroids are dominated by F900 and *blank*. The reader is encouraged to visit appendix A and table A.12 to refresh the memory on ICD-1 codes that do not have a direct mapping.

While the frequency of F900 is just as can be expected within this group of patients, the frequency of Z032 is initially not obvious. During the clinical evaluation, some thoughts

regarding this diagnose were confirmed. Z032: *Observation on suspicion of mental illness or behavioral disorder* is actually labeled as *A pure additional code, and should never be used as primary diagnose code*, and takes a percentage of 13.4% of all diagnoses in the cohort. For patients that are diagnosed, it is the second most used diagnose in our cohort.

Upon further inspection, for all patients diagnosed with Z032, which are only accepted patients, the diagnose was made anywhere between 1997 and 2018. Just above 1200 patients have F900, and about 600 patients have Z032 on axis 1. The high frequency of this diagnose required further inspection.

The group's interpretation of its prevalence is that Z032 is an observation diagnosis which has been chosen because CAMHS clinics at a time were required to make a diagnosis after a maximum of 5 patient contacts. If this was not done, a lower score was given on a quality indicator, and there was a considerable amount of pressure from both management and the authorities to do this. When one was still unsure of what the condition might be, Z032 probably became a diagnosis that was easy to resort to. This may explain the motivation behind the use of Z032, even though the prevalence is remarkably high and something to look further into.

This phenomenon was further investigated and was the topic of discussion during the meeting with psychologist specialist Jostein Arntzen, 11.05.2022.

According to Arntzen, the use of Z032 as diagnostic code was recently changed. Z032 was for a long time used as a temporary and observational diagnosis while assessing a patient. In some cases it was used as early as directly after the first point of contact, and when there was no current basis for an F-diagnose. When little assessment had been done, but symptoms and function level indicated that further assessment was necessary, Z032 was used as a tentative diagnosis when a conclusion regarding a final diagnose was not yet made. It was also registered because there was a need to better assess the presence of comorbidities and differential disorders. In other cases, it was used even for just a few days of bed post. If no diagnose was found, Z032 would oftentimes remain the diagnosis, instead of being exchanged by e.g. code 1000: *No diagnose proven on axis 1.*

Today, new guidelines for Z032 ensure that it is only used after a full assessment of the patient has been carried out, and at the end of the referral period. It is expected that the prevalence decreased due to this, according to Arntzen.

In the investigation of this phenomenon, short versions of ICD-10 were provided by CAMHS at St. Olavs Hospital, in which the diagnostic guidelines are described. The following is stated on page 6 in the 2007-version:

At the very beginning of our contact with the child/adolescent, from the admission interview/first meeting, we must code Z03.2 (observation) if there is no current R-code to use, or a diagnosis of the specialist health service that we can use until we have had a

separate diagnostic assessment (Indredavik and Gårdvik, 2007).

However, in the 2016-version, the following is stated on page 4:

Medical observation and assessment in case of suspected diseases and conditions Z03.2: Observation in case of suspicion of mental disorders and behavioral disorders. The code should only be used after completion of the investigation, when the suspicion is ruled out and it is concluded that no further investigation or treatment is needed (Indredavik and Gårdvik, 2016).

In other words, a change in the diagnostic guidelines took place before there was a change of journal system, which essentially means that one will find different uses of Z032 in the patient records created in BUPdata, of which our data comes from. BUPdata was discontinued around 2019 (Koochakpour et al., 2022). This is an important observation that is important to keep in mind when analysing patients with Z032 as diagnostic code.

It is also reasonable to assume that some patients with Z032 actually are patients without a proven diagnose during their first referral period, and should have received code 1000, as suggested by Arntzen. Since our data only covers referral periods up until 2018, we only have approximately 2016-2018 to use as assessment basis for investigating whether the prevalence of Z032 went down after the new guidelines. Newer data may also give answers to whether the prevalence has decreased and if other diagnostic codes are used instead.

Both the reflections made by the group and by Arntzen indicate the motivation and the practical use, respectively, of Z032 in CAMHS clinics. It should be looked into whether the prevalence went down after 2016, which is a matter of future work.

Regarding the occurrence of code 1999: *Insufficient information to code on axis 1*, Arntzen states this has usually been recorded when the patient has either moved, interrupted their stay, withdrawn from treatment, a conflict has arisen, or the patient cancelled their cooperation with CAMHS due to other reasons.

The prevalence of 1999 is stable across the clusters, varying from 3.4-11.5%, except for cluster 5 with no incidents. Code 1999 is frequently used for patients across all custody situations; thus its prevalence is not characterised by a specific custody situation.

When looking into what happens with patients with diagnose code 1999 and Z032 after their referral period, these patients have the same common frequencies as other patients registered with the top five most common diagnostic codes. However, both 1999 and Z032 more often have *Missing offer* registered with regards to after code, even though F900 have more patients. However, since the use of Z032 changed around 2016, note that this may lead to inconsistencies in other variables for these patients.

Another unexpected occurrence is the high number of patients with 1999 and *Rejection*

as the reason for closing the case, which is second most recorded for these patients after patients with *blank* as diagnose code. This is not as consistent with the reasoning for the occurrence of 1999, and should be investigated in future work.

As emphasised during the clinical evaluation, not receiving a diagnose, does not mean that the patient do not have a disorder. It is necessary to understand why the patients have either *blank*, 1000, 1999, or remain at Z032. Are there characteristics about those patients that make them fall into those categories as opposed to getting a diagnosis? There may be high rates of comorbidity, divorced or separated parents, or the presence of multiple confounders, mediators or moderators that play a role. This is important to understand, and in order to do so, these patients may need to be analysed separately and together in order to see what difference it makes.

To conclude, considering both the prevalence of Z032 and the three different categories accepted patients without a diagnose fall into, it may be necessary to analyse more variables connected to the patients that do not receive a diagnose. It is useful to do this both in isolation and in comparison to the other patients. Due to Z032 being a purely observational code, this means it is of interest to further compare these with accepted patients that were not found to have a diagnose, both prior and post guideline change around 2016, in order to properly compare the use of Z032, 1000, 1999 and patients with no entry at all. This, as well as other arisen questions, are left for future work.

8.8. Rejected Cohort

Before moving on to a summary of the discussion, we briefly present the rejected part of the cohort in order to highlight its characteristics.

Cluster 2 and 4 have the highest percentage of rejected referrals, at 17.2%, n=183 and 15.6%, n=150, respectively.

The rejected patients were not distinctively assembled in one cluster, but were distributed in cluster 2 and 4. Even though the clusters were not able to collect these patients in one clusters, the characteristics of these two clusters have yielded useful findings.

Looking into referral reasons for cluster 2, with most rejections across the clusters, it has by far the most referrals with *suspicion of defiance/conduct disorder*. It also has the the lowest percent of referrals with *suspicion of ADHD* (24.8%). I.e. it may be reasonable to assume that a large amount of rejections is not necessarily in connection with a high number of patients referred on the suspicion of ADHD. As in cluster 3, the latter may be seen in connection with many boys, but not with rejection. However, note that cluster 4 actually has the highest amount of patients with *suspicion of ADHD*, but also the most even gender balance. Since only cluster 1, with the highest amount of girls,

have a higher percent with *suspicion of ADHD*, it may be reasonable to believe that a high rejection rate of *suspicion of ADHD* is seen when there also is a high amount of girls. This may be necessary to further evaluate in future work.

Another finding of interest is that cluster 2 also has the lowest mean age at 5.93. This may be due to patients that are referred at a later age, are assessed to less likely get the follow-up they need at home and in the community. Younger patients in their first referral period may be assessed to more likely receive sufficient treatment by facilitating measures at home rather than continuing treatment at CAMHS. However, on the other hand, cluster 4 actually has the highest mean age at 14.2, i.e. both the youngest and the oldest subgroup of patients are prone to most rejections in their first referral period. A reason for the older subgroup may be either that children at this age may be found to just be behaving badly, as previously mentioned, or that it is rather a connection to the amount of girls rather than the prevalence of high age. Lastly, it could also be due to age in both cluster 1 and 2 having little impact in general and it is necessary to look into other variables to explain the rejection rates. This should also be further investigated, but these reflections may provide an initial starting point.

Another surprising discovery is that cluster 2, having most rejections, also have the most patients living in foster care, most patients with foster parents, most parents having a mother and a step father, and most patients living with one parent. Cluster 4, with the second largest amount of rejections (15.6%), has by far the largest amount of patients living in institution (n=62). Thus the clusters with the most children living in foster care or in institution are also in the clusters with most rejections.

Both cluster 2 and cluster 4 share the fact that most patients in these clusters are referred from a GP. This strengthens the evolved hypothesis about referrals from a GP being more prone to rejections. During the EDA in section 6.3 and more specifically figure 6.15, this was also confirmed. The GPs are by far the referring actor with most rejections, followed by the Help Service for Children and Young People, and then the Educational Psychological Service. However, all three have almost the same amount of referrals; but the GPs nevertheless bypass by a large margin.

Considering the rejected cohort, the results were also very consistent with current literature. The ratio stay the same. If there is suspicion of ADHD, which is treatable, they are likely to be accepted, and if there is suspicion of defiance/conduct disorder, they are more likely to be rejected, because it is much harder to treat and they may not have the services for them. If there is a perception that someone is just behaving badly, the patient is also rejected.

Lastly, it is relevant to add that the reason for a rejection is recorded in a text record, and is not coded, according to CAMHS professionals during the meeting between CAMHS and IDDEAS 18.11.2021. These records were not available to me at the time of this research, but may play a key role in future research for understanding the reasoning

behind rejections.

8.9. Discussion Summary

This section aims to summarise and finalise the discussion, in order to emphasise the discussion highlights and address the utility of clustering. The summary also aims to close the research questions that have been thoroughly explored throughout this thesis, as defined in section 1.2. As the research questions are more concerned with the actual results, these are addressed as a part of the discussion, while the overall goal is left for the conclusion.

Throughout this chapter, the EDA and clustering results have been thoroughly discussed by the use of existing knowledge, information granted through meetings with professionals, and in light of input and findings from the clinical evaluation. The professional input to this discussion has been significant in the aim of interpreting all information in its relevant context, and has elevated the analysis from a conventional machine learning experiment to a professionally reasoned discussion of electronic health records.

The analysis process yielded both expected and unexpected results, and most of the findings were found to be highly consistent with published literature. Most of the ratios of boys to girls across all of the variables were very accurate and realistic, consistent to the most recent studies on the area.

As much as the discussion was able to explain a lot of the phenomena we see in the data, it was even more fruitful in revealing several topics of future work. Some of these that are especially worth highlighting are firstly to continue to build comprehension of the use of Z032, which is an observational code and not final diagnosis. Secondly, to investigate the development of referrals from the GPs and how to accommodate their needs in order to improve their referral quality. Lastly, it should be investigated why accepted patients fall into different categories of no diagnose, like.

On the basis of those findings, it may be concluded that an analysis like this can be useful for identifying phenomena of interest in the context of Norwegian psychology, as asked in the first research question. The utility value of such an analysis has also been proven, confirmed by the discoveries made.

In the identification of the latter phenomena, we were able to identify a selection bias, of which the extent was measured. The error was found to be small in extent, and it was concluded that it would not impact the clustering results to a significant degree. However, it emphasised the need to further look into this patient group, and what makes accepted patients fall into different categories of not receiving any diagnose. These are the fields that are blank (indicating that no diagnose was given), the code for no diagnose

proven (1000), and the code indicating that information was insufficient in order to code anything on axis 1 (1999).

Regarding research question 2, clustering did not excessively give more information than an exploratory data analysis at this stage in the research. It is beneficial for grouping patients, but is prone to selection error, like the number of clusters k, choice of algorithm, and what data to include. These variables may affect the clustering results. There is also no initial guarantee that clustering is able to separate the data into relatively distinct groups. However, it proved to be useful for describing the groups of patients, what their clinical profiles are, and for unveiling patterns that are useful in clinical care today. The strength of clustering is the way it enables the comparison of patterns and phenomena across several patient groups and trajectory variables. Additionally, it differentiates from an EDA by being able to compare and assess variables across more than four dimensions, which was the highest dimension of comparison in the EDA (e.g. the comparison of age, gender, diagnose on axis 1 and assessment outcome in one and the same scatter plot).

As foreshadowed by the EDA, strictly separating groups of patients proved to be a challenge. The majority of patients have similar family and care situation, and gender differences were not as prominent as initially thought. Overall, the clusters bear the imprint of several instances and reasons being more frequently recorded, as can be expected when looking into the specific group of hyperkinetic diseases. However, the clusters were increasingly able to identify subgroups of patients regarding their referral situation and assessment. They rather precisely collected the rejected patients and the referring instance associated with this group, and properly separated different demographics. They also captured important characteristics related to gender differences, unusual frequencies of referral reasons and outlier patients. Despite the fact that most patients are accepted, have their case completed, and their referral is sent back to referring instance after completion, the clusters captured these important nuances that are relevant for future work.

The unclear separation between some of the parameters may also be explained by the careful selection of patients. Intentionally, patients that are similar on several variables were chosen from the overall cohort. When choosing patients with relevance to hyperkinetic disorders, these may share several commonalities like referral reason. The aim, however, was to find the minor variations within these known similarities; which the clusters were able to find.

In short, the clusters are able to give more information about the variables that separate the patients after their referral, but did not distinguish much between features like care, custody or parental relations prior to referral. Clustering is useful for analysing several variables simultaneously and unveil phenomena across groups of patients. In this context, it was largely dependent upon the EDA, and was not sufficient as an analytical tool on its own. However, an EDA in combination with clustering, may provide as a useful and beneficial tool for analysing patient trajectories on the basis of clinical data.

According to these results, it is reasonable to believe that future work has more benefit of looking into what distinguishes patients from the time of making a referral, to the assessment and treatment at CAMHS. This includes any waiting times, what happens to accepted patients, what kind of contact and activity types that take place, and which professionals that are involved during their stay. Clustering may be used in future applications, but should be used in combination with a thorough EDA to best harvest its potential.

9. Conclusion and Future Work

This chapter aims to conclude the work and research conducted in this project, and summarises the process, its findings, and any contributions to the field in light of the overall project goal. We also address the main goal of the project. Contributions to the research field have been integrated in the conclusion as we look back on the research results in this Master's thesis, and are then briefly summarised in section 9.2. Finally, the major potential of the research is put forward in section 9.3, which presents future work in the research area.

9.1. Conclusion

In this Master's thesis, electronic health records of patients referred to CAMHS on suspicion of ADHD and behavioral difficulties have been analysed, assessed and interpreted. As stated in section 1.2, the overall goal was to analyse electronic health records of patients with relation to hyperkinetic disorders in CAMHS, investigate if patient profiles and subgroups can be identified by cluster analysis, and interpret phenomena in the context of Norwegian psychiatry. To ensure compliance with this aim, an iterative EDA process was carried out, before performing a clustering experiment of which results were interpreted in collaboration with professionals. The motivation was to provide a better understanding of children and adolescents that are referred to CAMHS, facilitate more future research, and make a small contribution to the long-term goal of enhancing medical diagnosis and care in Norway. This section confirms the compliance of the research with the project goal.

Upon initiation of the project, it was quickly realised that the research just as much concerns psychiatry and the comprehension of clinical practice, as it concerns computer science, machine learning and data analysis. As both areas required skill and specific domain knowledge, a considerable amount of time was used to understand the respective domains, learn how to utilise available resources, and plan the interconnectivity of the two areas. In order to understand the trajectories of patients referred with hyperkinetic problems, effort was made to understand the processes of CAMHS, clinical practice and coding guidelines, in addition to the many changes in the clinical environments over the last 20 years.

As previously emphasised, an essential part of the research is concerned with the comprehension of all information in the relevant context of clinical psychiatry, and is not limited only to the clustering and data analysis. Due to the importance of understanding the results in the context they are in, this has been included as a considerable part of the research. By analysing results, aiming to build comprehension and engaging in dialogue with actual clinics, people and professionals that understand the data I am using, findings have been interpreted beyond an isolated clustering of available data. By ensuring that results are interpreted in collaboration with CAMHS professionals, a better understanding has been made and a more valuable return of the research has been provided.

Another important addition to this is that since little research has been done in the area or with the specific data at hand, there were few expectations by the research group to what we wanted to find, or if any discoveries could be made at all. The initial assignment of performing a clustering experiment with the netDx algorithm to identify patient subgroups proved to be too narrow. Early insight and understanding of data from the BUPdata system was very limited, due to insufficient comprehension of the potential and limitations of the data at the earlier stages of research. Consequently, the project assignment, scope and aim were expanded throughout the project to better fit the data and the progressively more specified research goal. As more interesting findings emerged, these were also further inspected to also accommodate the need for a better comprehension of patients in CAMHS, in addition to the assessment of clustering as a tool for the identification of patient subgroups and characteristics. To facilitate this, an extensive process of literature-, documentation-, and systems archaeology was also carried out in order to assemble sources and information on the domain. This was necessary to compile existing documentation and understand the work's position in the research field.

The issue of data quality has been a consistent challenge throughout the research. Insufficient coding and journal keeping by both referring instances and CAMHS clinics were identified as a considerable challenge in several areas of clinical psychiatric practice, which provided some limitations to the research. These limitations have been thoroughly described and mitigation strategies have been proposed, in order to argue for the choices made and with the aim of aiding future scientist with research on clinical data in Norway.

Another significant aspect related to this, is the regular replacement of coding practices and guidelines in CAMHS. The replacements are reflected in the dataset, which breaks with the quality and continuity of the data. An example of this is the guideline change for use of code Z032, as elaborated on in 8, or the replacement of old referral reasons with new around 2009/2010, which do not have a direct mapping. Furthermore, clinical practice and the perception of clinical situations change, even within the same CAMHS clinic where coding practice is strictly exercised. This means that patient situations that are actually similar, may be described or coded differently, as discussed in 7.4.4. This Master's thesis is one of the first to address the impact that the change in coding practices and revisions of code systems have on data quality and data continuity, and

proves that contextual comprehension is invaluable when assessing clinical data.

The experimental part of the research yielded interesting results. By the use of unsupervised machine learning, namely clustering, supported by a thorough exploratory data analysis (EDA), we were able to explore patient profiles, identify patient trajectory characteristics that are impactful for the assessment and treatment of patients, and confirm what can be seen in Norwegian clinical reality and practice.

Both the EDA and the clustering were beneficial tools in this research. EDA was useful for identifying phenomenon to further investigate in the clustering process and for building initial hypotheses. However, the EDA comes short when trying to compare more than four variables. Clustering was very beneficial in comparing patients and assessing more variables simultaneously, and was also able to identify important characteristics and patient subgroups across the clusters. This essentially contributed to the assessment of patient situation similarity in the cohort, both confirmed and refuted hypotheses prior to the analysis, and unveiled several important findings. It was concluded that for the context of clinical research in light of clinical data quality, clustering will largely benefit from not being used on its own, but in combination with iterative EDAs. This will enable the identification of patterns and subgroups, and for further inspection of any revealed phenomena.

During clinical evaluation of the results, the results were found to be consistent with literature, and highly realistic with regards to the most recent publications. The ratio between boys and girls in this dataset is more consistent with the most recent literature, and may therefore be more precise than older studies on patients and trajectories in CAMHS. Other data, like diagnoses, referral reasons, family relations, care situation and referring instances, were also highly consistent with what literature conveys and what can be seen in practice.

Furthermore, the study unveils and highlights several areas that are necessary to look further into, as the work provides input and foundation for future research. Many of the findings in this study have not been documented nationally; they are insights derived from practical reality that are compared to international studies. Thus, this thesis is also among the first to confirm and document what we see in Norwegian clinics.

From the discussion, it was concluded that it is highly relevant and useful to further look into several of the identified phenomena. Some of these are concerned with how rejections of referrals from the GP can be mitigated and their referral quality improved, the high frequency of Z032 as a diagnostic code and whether its prevalence has developed after revision of the guideline, and to investigate and build comprehension of what actually happens to patients that are accepted, but are either not given a diagnose, or remains at Z032. These findings, as well as more potential future research, are thoroughly described in section 9.3.

Lastly, this project contributes with novel analysis and documentation of data that confirms the clinical reality in Norway. Research on a national level is sparse, as discussed in chapter 4. Even though the discoveries behave as expected and is consistent with international literature and publications, they contribute to solidify what is seen in clinical practice across CAMHS clinics in Norway with regards to hyperkinetic disorders. Additionally, the feasibility of working with clinical data has been confirmed. Even though clinical data occasionally are deficient and prone to human error, it has been proven that research on this kind of data can produce results that are aligned with international results and the practical reality in our own clinics.

9.2. Contributions

Even though the former section addresses some of the contributions made, this section aims to specifically highlight these. The contributions span several domains, and concern the novel documentation of CAMHS processes and phenomena, the assessment of clinical data and its quality, clustering as tool for analysing electronic health records, and the foundation built for future work. It is important to keep in mind that these contributions are in a field of research that to some extent is unexplored in Norway.

The in-depth information archaeology process, which assembled knowledge on clinical practice and patterns, simplifies the acquisition of domain knowledge in the future. Furthermore, the feasibility of conducting research on clinical data has been assessed and proven, as results were found to be consistent with international literature. The potential of the available data has been thoroughly elaborated and assessed, which enables agile initiation of more research. Clinical data was assessed to be of slightly poor quality, but mitigation strategies were presented in order to utilise the data potential regardless.

Despite the poor data quality, several clinical phenomena related to the characterisation of patients were identified. Some of these are the prevalence and change of use of procedural codes like Z032 as a diagnostic code, and the tendency of accepted patients to fall into three different categories when not being associated with a diagnose. Furthermore, important referral situations related to rejection rates and characteristics of rejected patients were detected. Supported by professional collaboration and interpretation, aspects of Norwegian clinical practice has been confirmed and highlighted.

Additionally, novel documentation has been provided on the clinical reality in Norway, the impact of revisions and replacements of coding systems and guidelines on clinical practice, and on the challenges and limitations of record keeping in Norwegian psychiatry. The work emphasises and documents important issues to be addressed.

In terms of clustering, it is one of the first clustering experiments on clinical in Norway. Consequently, the utility and application value have been assessed, and its potential

and limitations discussed. A review of the prerequisites for a clustering algorithm for clinical data was made, which highlights the requirements for conducting experiments on electronic health record data. Additionally, an evaluation was made of clustering as tool for analysing patient profiles and subgroups. This includes the utility is has in the context of assessing electronic health records, and the clinical data restrictions clustering must account for.

Lastly, an important contribution has been the identification of future work. This project has facilitated future research by providing a theoretical foundation, assessing the potential and limitations that exist, and identified findings that should further be investigated. This essentially aids the implementation of future work, which is presented in the upcoming section.

9.3. Future Work

The potential for future work on this area is both major and extensive. This study has proven the feasibility of working with the clinical data we have at hand, as well as contributed to the identification of several trails to follow. Being one of the first experiments on clinical data, it has largely explored the possibilities and limitations that exist, which are also applicable to electronic health records from other Norwegian CAMHS clinics.

This section presents some of the issues that others may address in the future. Limitations of the research and the data have been thoroughly elaborated on, and mitigation strategies have been presented in section 7.4.4. This creates a good starting point for subsequent researchers to continue the work with clinical data.

First of all, the potential of the dataset should be addressed. This project investigates family and care situation, the referral process, and the assessment at CAMHS. As elaborated on in section 7.4.4, several columns were left out due to numerous reasons. However, by applying the presented mitigation strategies and combining columns that looks into trajectory aspects with a more narrow time scope, interesting analysis can be made. Some of the data that are encouraged as starting point for feature researchers to explore, are:

- Waiting times between referral, assessment and start of health care. Waiting times for accepted patients should also be assessed in comparison to national guidelines for progress time.
- Which professionals or teams that are involved in the assessment, care and treatment of the patients during their stays.

- What kind, as well as the frequency, of contact and activity types that take place during their stays. Use these to deeper probe the patient trajectories after admission at CAMHS.

- What kind of stays the patients have (i.e. day, 24-hour stay, outpatient).

Furthermore, this research initially focuses on the first point of contact between the patient and CAMHS, disregarding any subsequent history of contacts. In future work one should look at cases as with a history of referral periods; how cases can be clustered with their history of previous cases, and how these affect the current case. Do previously rejected patients return with same or different set of referral reasons? Which diagnoses are patients given in their different referral periods? What kind of development regarding the referral process and assessment can be seen across referral periods? Are undiagnosed patients diagnosed at a later time, and what diagnoses do they receive? These are just some of the interesting questions to be answered.

Additionally, the selection only includes the first referral reason. Second and third referral reason should also be explored, bearing in mind that the majority of referrals only have a primary referral reason. However, even though a lot of these do not have an entry in the database, one can to an even greater extent look at the distribution of causes in connection with symptoms and diagnoses.

In CAMHS, text records are used for documenting the reasons for making rejections. These should be acquired and analysed in order to better understand the rejections. As mentioned in the previous chapter, the reason for a rejection is recorded in a text record, and is not coded. These records were not available to me at the time of this research, but have great potential for future research on rejected patients. In addition to providing useful insight into the reason for rejections, machine learning techniques like natural language processing can be applied to derive the sentiment and categorical traits of the textual assessments.

Lastly in the context of data analysis, I would like to highlight some previously identified phenomena which should be further investigated. Firstly, the high rejection rate of the GPs. According to recent discussion, this is mostly due to poor referral quality. Strategies to accommodate the user needs of the GPs and improve their referral quality can be explored. Secondly is the high frequency and use of Z032 as a diagnose code. This is purely an observational code and should not be used as a primary diagnose code on axis 1. It is relevant to explain its position as second most frequently used diagnose code. A good starting point is to investigate the development of its prevalence before and after the guideline change around 2016, and the situation similarity for patients on either side of this change. Thirdly is to look into the reasoning for accepted, but undiagnosed patients to fall in to three different categories; A blank field, which indicates that no diagnose was made, the diagnose code for when no diagnose was proven (1000), and the diagnose code that indicates that information was missing in order to code on

axis 1 (1999). If we can build better comprehension of what actually makes patients receive code 1000, 1999, Z032 or a blank field, this can contribute to better assessment, treatment and decision-making long term.

Finally, I would like to address the potential of clustering. In this thesis, the choice was to focus on one clustering technique, in order to primarily assess the feasibility of clustering clinical data and see if the results are useful and meaningful. This is however not a thesis exploring the best suitable clustering method to apply to clinical data; it is a thesis investigating how clustering can be used to probe latent subgroups of patients in CAMHS. Thus the exploration of several techniques with the same dataset is a matter of future work. I encourage to explore more clustering techniques, with different similarity measures. As proven, clustering has the potential and can produce the results, thus it may be beneficial to experiment with other techniques to see if results can be improved.

A consistent goal of mine has been to compose this thesis in such a manner that any future researchers that pick up where I left, can spend less time on acquiring the knowledge required to understand CAMHS, clinical practice and the dataset, and more time on the actual research. Because of this, a considerable effort has been made to reduce the need of information archaeology, which enables using the thesis as theoretical support and a basis for future research.

There are several paths to embark on, and the ones presented in this section can provide as good starting points. When working towards the long-term goal of improving diagnosis and treatment of patients related to hyperkinetic disorders, it is significant to properly understand the data, and the context they must be interpreted in.

Bibliography

A. Aprillant. The k-prototype as clustering algorithm for mixed data type (categorical and numerical). January 2021. https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb, accessed 04.04.2022.

S. Boluki, S. Dadaneh, X. Qian, and E. Doughert. Optimal clustering with missing values. *BMC Bioinformatics*, 20(12):1–2, 2020. doi:10.1186/s12859-019-2832-3.

C. Custer. What is a foreign key?, June 2021. https://www.cockroachlabs.com/blog/what-is-a-foreign-key/, accessed 24.05.2022.

DBeaver. Universal database tool, 2022. https://dbeaver.io/, accessed 22.04.2022.

De nasjonale forskningsetiske komiteene. Regionale komiteer for medisinsk og helsefaglig forskningsetikk (rek), October 2014. https://www.forskningsetikk.no/om-oss/komiteer-og-utvalg/rek/, accessed 19.05.2022.

Direktoratet for e-helse. Multiaksial klassifikasjon i psykisk helsevern for barn og unge (phbu), May 2022. https://www.ehelse.no/kodeverk/multiaksial-klassifikasjon-i-psykisk-helsevern-for-barn-og-unge-phbu, accessed 30.01.2022.

J. Gross, J. M. Vetter, and H. H. Lai. Clustering of patients with overactive bladder syndrome. *BMC Urology*, 21(41):1–6, March 2021. doi:10.1186/s12894-021-00812-9.

A. S. Hansen, C. H. Christoffersen, G. K. Telléus, and M. B. Lauritsen. Referral patterns to outpatient child and adolescent mental health services and factors associated with referrals being rejected. a cross-sectional observational study. *BMC health services research*, 21(1):1–2, October 2021. doi:10.1186/s12913-021-07114-8.

Helsebiblioteket. (f90-f98) atferdsforstyrrelser og følelsesmessige forstyrrelser som vanligvis oppstår i barne- og ungdomsalder, 2022. https://finnkode.ehelse.no/#icd10/0/1/0/2599550, accessed 09.04.2022.

Helsedirektoratet. Henvisning til utredning og behandling i spesialisthelsetjenesten, June 2020. https://www.helsedirektoratet.no/pakkeforlop/barnevern-kartlegging-og-utredning-av-psykisk-helse-og-rus-hos-barn-og-unge/kartlegging-og-utredning-i-helsetjenesten/henvisning-til-utredning-og-behandling-i-spesialisthelsetjenesten, accessed 12.01.2022.

Helsedirektoratet. Forløpstid for utredning i psykisk helsevern barn og unge, May 2021. https://www.helsedirektoratet.no/statistikk/kvalitetsindikatorer/psykisk-helse-for-barn-og-unge/forlopstid-for-utredning-i-psykisk-helsevern-barn-og-unge-phbu, accessed 02.02.2022.

Helsedirektoratet. Multiaksial klassifikasjon i psykisk helsevern for barn og unge (phbu): Kodelister 2022, May 2022. https://www.ehelse.no/kodeverk/multiaksial-klassifikasjon-i-psykisk-helsevern-for-barn-og-unge-phbu, accessed 04.05.2022.

Helseplattformen AS. Felles pasientjournal i midt-norge, October 2019. https://helseplattformen.no/om-oss/helseplattformen-as, accessed 29.04.2022.

J. Z. Huang. Clustering large datasets with mixed numeric and categorical columns. pages 1–6, 1997. https://www.semanticscholar.org/paper/CLUSTERING-LARGE-DATA-SETS-WITH-MIXED-NUMERIC-AND-Huang/d42bb5ad2d03be6d8fefa63d25d02c0711d19728, accessed 20.03.2022.

HUNT Cloud. Hunt cloud documentation: Technical tools, May 2022a. https://docs.hdc.ntnu.no/working-in-your-lab/technical-tools/, accessed 22.04.2022.

HUNT Cloud. Hunt cloud documentation: Activities, May 2022b. https://docs.hdc.ntnu.no/about/activities/, accessed 22.04.2022.

HUNT Cloud. Faq on security: Data classification, May 2022c. https://docs.hdc.ntnu.no/faq/security/#data-classification, accessed 28.04.2022.

IBM. Ibm spss software, November 2019. https://www.ibm.com/analytics/spss-statistics-software, accessed 28.04.2022.

IBM. Machine learning, 2020a. URL https://www.ibm.com/cloud/learn/machine-learning. https://www.ibm.com/cloud/learn/machine-learning, accessed 28.04.2022.

IBM. Exploratory data analysis, 2020b. https://www.ibm.com/cloud/learn/exploratory-data-analysis, accessed 15.04.2022.

IDDEAS. The individualized digital decision assist system, 2021a. https://www.iddeas.no/, accessed 19.02.22.

IDDEAS. Iddeas: Bakteppe for prosjektet, 2021b. https://www.ntnu.no/rkbu/iddeas, accessed 03.02.2022.

M. Indredavik and K. Gårdvik. Kortversjon av icd-10 for bruk ved barne- og ungdomspsykiatrisk klinikk, st. olavs hospital hf. page 6, June 2007.

M. Indredavik and K. Gårdvik. Kortversjon av icd-10 for bruk ved barne- og ungdomspsykiatrisk klinikk, st. olavs hospital hf. page 4, December 2016.

J. Irani, P. Pise, and M. Phatak. Clustering techniques and the similarity measures used in clustering: A survey. *International Journal of Computer Applications*, 134(7):9–14, January 2016. doi:10.1371/journal.pone.0144059.

J. Irani, P. Pise, and M. Phatak. An unsupervised machine learning clustering and prediction of differential clinical phenotypes of covid-19 patients based on blood tests—a hong kong population study. *Frontiers in Medicine*, 8:1–9, February 2022. doi:10.3389/fmed.2021.764934.

A. Kassambara. Clustering distance measures: Data standardization, March 2021. https://www.datanovia.com/en/lessons/clustering-distance-measures/#data-standardization, accessed 12.04.2022.

K. Koochakpour, Ø. Nytrø, O. S. Westbye, B. Leventhal, R. A. Koposov, V. Bakken, C. Clausen, T. B. Røst, and N. Skokauskas. Success factors of an early ehr system for child and adolescent mental health: Lessons learned for future practice data-driven decision aids. Medinfo 2021: One World, One Health. Proceedings of the 18th World Congress on Medical and Health Informatics, pages 1–5. IOS Press, 2022. ISBN 978-1-64368-002-6.

O. Matt. 10 tips for choosing the optimal number of clusters, January 2019. https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92, accessed 13.04.2022.

Norsk forening for Barne- og ungdomspsykiatriske institusjoner and Hiadata AS. Kodebok for bup. pages 20–21, March 1999.

Oslo University Hospital. Patient similarity networks for precision medicine. *Big Med*, pages 1–10, 2018. https://bigmed.no/assets/Reports/patient_similarity_networks_for_precision-medicine_version_1.pdf, accessed 12.01.2021.

C. O'Sullivan. Introduction to shap with python, December 2021. https://towardsdatascience.com/introduction-to-shap-with-python-d27edc23c454, accessed 29.04.2022.

S. Pai. netdx: Tool to build a patient classifier using similarity networks, September 2021. http://www.netdx.org/, accessed 20.01.2022.

S. Pai, S. Hui, R. Isserlin, M. A. Shah, H. Kaka, and G. D. Bader. netdx: interpretable patient classification using integrated patient similarity networks. *Molecular Systems Biology*, 15(3):1–7, March 2019. doi:10.15252/msb.20188497.

M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. Costa, and F. A. Rodrigues. Clustering algorithms: A comparative approach. *PloS one*, 14 (1):4–6, January 2019. doi:10.1371/journal.pone.0210236.

P. Rossi, I. Pretelli, D. Menghini, B. D'Aiello, S. DiVara, and S. Vicari. Gender-related clinical characteristics in children and adolescents with adhd. *Journal of Clinical Medicine*, 11(385):1–7, January 2022. doi:10.3390/jcm11020385.

A. Ruberts. K-prototypes - customer clustering with mixed data types, May 2020. https://antonsruberts.github.io/kproto-audience/, accessed 16.05.2022.

Scikit-learn. Scikit-learn: One hot encoder, 2022a. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html, accessed 07.05.2022.

Scikit-learn. Scikit-learn: Standardscaler, 2022b. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html?msclkid=5e95de21cf0711ec8269dbca32916b17, accessed 07.05.2022.

S. Secherla. Different imputation methods to handle missing data, June 2021. https://towardsdatascience.com/different-imputation-methods-to-handle-missing-data-8dd5bce97583, accessed 12.04.2022.

J. Sidey-Gibbons and C. Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19(64):1–4, March 2019. ISSN 1471-2288. doi:https://doi.org/10.1186/s12874-019-0681-4.

J. Smith, R. G. Kyle, B. Daniel, and G. Hubbard. Patterns of referral and waiting times for specialist child and adolescent mental health services. *Child and adolescent mental health*, 23(1):41–49, February 2017. doi:10.1111/camh.12207.

P. Surén. Har ungdommer dårligere psykisk helse enn før? *Tidsskriftet Den Norske Legeforeningen*, 138(14):1–2, September 2018. doi:10.4045/tidsskr.18.0558.

P. Surén, A. G. Thorstensen, M. Tørstad, P. E. Emhjellen, K. Furu, G. Biele, H. Aase, C. Stoltenberg, P. Zeiner, I. J. Bakken, and T. Reichborn-Kjennerud. Diagnostikk av hyperkinetisk forstyrrelse hos barn i norge. *Tidsskriftet Den Norske Legeforeningen*, 138(20):3–13, November 2018. doi:10.4045/tidsskr.18.0418.

World Health Organization. The icd-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines: F90 hyperkinetic disorders, January 1992a. https://www.who.int/publications/i/item/9241544228, accessed 03.02.2022.

World Health Organization. F90-f98 behavioural and emotional disorders with onset usually occurring in childhood and adolescence, January 1992b. https://www.who.int/publications/i/item/9241544228, accessed 20.04.2022.

World Health Organization. Icd–11 fact sheet, 2022a. https://icd.who.int/en/docs/icd11factsheet_en.pdf, accessed 20.04.2022.

World Health Organization. International statistical classification of diseases and related health problems (icd), 2022b. https://www.who.int/standards/classifications/classification-of-diseases, accessed 20.04.2022.

C. Xu and Y. Tian. A comprehensive survey of clustering algorithms. *Ann. Data. Sci.*, 2: 165–193, May 2015. doi:10.1007/s40745-015-0040-1.

S. Young, N. Adamo, B. B. Ásgeirsdóttir, P. Branney, M. Beckett, W. Colley, S. Cubbin, Q. Deeley, E. Farrag, Hill P. Gudjonsson, G., J. Hollingdale, O. Kilic, T. Lloyd, P. Mason, E. Paliokosta, S. Perecherla, J. Sedgwick, C. Skirrow, K. Tierney, and E. Woodhouse. Females with adhd: An expert consensus statement taking a lifespan approach providing guidance for the identification and treatment of attention-deficit/hyperactivity disorder in girls and women. *BMC Psychiatry*, 20(404):5–22, August 2020. doi:10.1186/s12888-020-02707-9.

Z. Zazueta. K-prototypes clustering: for when you're clustering dynamic, real world data, October 2020. https://zachary-a-zazueta.medium.com/k-prototypes-clustering-for-when-youre-clustering-continuous-and-categorical-data-6ea42c2ab2b9, accessed 18.04.2022.

Z. Zazueta. Github: Telecom churn take two, December 2021. https://github.com/zachzazueta/telco_churn_take_two, accessed 18.04.2022.

A. Code List Mappings

This appendix includes every database column that has been used in the data selection process and their mappings. This also includes a translation of referring instances and referral reasons, as these were intentionally not translated during clustering due to interpretability purposes.

The mappings are based on the BUPdata to NPR code mappings, which are described in internal system documentation by Hiadata AS, later Visma Unique (last updated 24.03.2010), as well as an overview of local codes found in the database. In this appendix, these are presented in a temporal order in the context of the referral period. Note that as a part of data preprocessing stage, the codes were represented as their respective map (i.e. categorical, rather than numerical representation).

A.1. Gender

Code	Gender
1	Girl
2	Boy

Table A.1.: Map between code and gender.

A.2. Care situation

Code	Care situation
1	With both parents
2	Commuting between mother and father
3	Living with one parent
4	One parent and partner
5	With grandparents/other
6	In foster care
7	In institution
8	Living alone
9	Other

Table A.2.: Map between code and care situation.

A.3. Custody

Code	Custody
1	Mother and father together
2	Mother
3	Father
4	Other

Table A.3.: Map between code and custody situation.

A.4. Relation to care takers

Note that *sak.morrelasj* and *sak.farrelasj* have been joined to one single column; *relation*. I.e., if a case has code 1 for relation mother and code 6 for relation father, the value is *BioMomStepDad*. This was done for dimensionality reduction purposes.

Code	Relation
1	Biological mother
2	Biological father
3	Adoptive mother
4	Adoptive father
5	Step mother
6	Step father
7	Foster mother
8	Foster father
9	Other

Table A.4.: Map between code and parental relation.

A.5. Referring instance

Code	Referring instance	Translation
11	Pasienten	Patient
12	Foreldre/foresatte	Parents/care takers
13	Fosterhjem	Foster care
14	Andre fra nærmiljø	Others from the local environment
21	Skole/fritidsordning	School/after school program
22	Barnehage/førskole	Kindergarten/preschool
23	Pedagogisk psykologisk tjeneste	Educational psychological service
24	Spesialskole	Special school
25	Statlig kompetansesenter	State competence center

Table A.5.: Map and translation of referring instances.

Code	Referring instance	Translation
26	Annet innen skolesektor	Other in school sector
31	Lege	General physician (GP)
32	Skolehelsetjenesten	School health service
33	Helsestasjon	Health station
34	Habiliteringstjeneste barn	Habilitation service for children
35	Somatisk sykehus	Somatic hospital
36	Flyktningehelsetjeneste	Refugee health service
37	Annen somatisk helsetjeneste	Other somatic health service
41	Rusmiddelomsorg	Substance abuse care
42	Habiliteringstjeneste voksne	Habilitation service for adults
43	BUP poliklinikk/avdeling	CAMHS outpatient clinic/department
44	Voksenpsykiatri	Adult psychiatry
45	Psykolog/psykiater privat	Private psychologist/psychiatrist
46	Annen helsetjeneste	Other health services
51	Sosialkontor	Social office

Table A.6.: Map and translation of referring instances.

Code	Referring instance	Translation
52	Barnevern (kommunen)	Child welfare services (municipality)
53	Barnevern (fylkeskommunen)	Child welfare services (county)
54	Barnevernsinstitusjon	Child welfare institution
55	Flyktning/innvandrertjeneste	Refugee/immigrant service
56	Annen sosialtjeneste	Other social services
61	Hjelpetjenesten for barn/unge	Help service for children and young people
71	Familievernkontor	Family welfare office
72	Utekontakt/uteseksjon	Outdoor contact/outdoor section
74	Krisesenter	Crisis center
75	Kriminalomsorg	Norwegian correctional service
76	Politi/lensmann/rettsvesen	Police/judiciary
77	Arbeidsmarkedsetat	Labor market agency
78	Andre	Others

Table A.7.: Map and translation of referring instances.

A.6. Referral reason

Note that all codes above code 20 are the old referral reasons. When possible, old codes have been mapped to the new codes.

Code	Referral reason	Translation
1	Alvorlig bekymring for barn under 6 år	Serious concern for children under 6 years
2	Mistanke om gjennomgripende utviklingsforstyrrelse (autimse) Also includes: 21: Autistiske trekk	Suspicion of Autism
3	Mistanke om trasslidelse/adferdsforstyrrelse Also includes: 29: Atferdsvansker	Suspicion of defiance/conduct disorder
4	Mistanke om hyperkinetisk forstyrrelse (ADHD) Also includes: 30: Hyperaktiv/konsentrasjonsvansker	Suspicion of hyperkinetic disorder (ADHD)
5	Mistanke om Tourettes syndrom	Suspicion of Tourette's syndrome
6	Skolevegring	School refusal
7	Mistanke om angstlidelse Also includes: 25: Angst/Fobi	Suspicion of anxiety

Table A.8.: Map between code and description for referral reasons.

Code	Referral reason	Translation
8	Mistanke om tvangstanker-/tvangshandlinger Also includes: 26: Tvangstrekk	Suspicion of obsession
9	Mistanke om spiseforstyrrelse Also includes: 36: Spiseproblem	Suspicion of eating disorder
10	Mistanke om depresjon Also includes: 27: Tristhet/Depresjon/Sorg	Suspicion of depression
11	Mistanke om bipolar lidelse	Suspicion of bipolar disorder
12	Vedvarende og alvorlig selvskading	Persistent and severe self-harm
13	Mistanke om psykose Also includes: 22: Psykotiske trekk	Suspicion of psychosis
14	Alvorlige psykiske reaksjoner etter traumer, kriser eller katastrofer	Severe psychological reactions after trauma, crises or disasters
15	Alvorlige psykiske symptomer sekundært til somatisk sykdom	Severe mental symptoms secondary to somatic illness

Table A.9.: Map between code and description for referral reasons.

Code	Referral reason child	Translation
16	Annet Also includes: 38: Annet 31: Rusmiddelmisbruk 35: Syn/hørselsproblem 34: Språk/talevansker 32: Asosial/kriminalitet 37: Andre somatiske symptomer	Other reasons
20	Ikke fylt ut av henviser Also includes: 39: Ingen	Not filled in by referrer
24	Hemmet atferd	Inhibited behavior
23	Suicidalfare	Suicide risk
28	Skolefravær	Absence from school
33	Lærevansker	Learning difficulties

Table A.10.: Map between code and description for referral reasons.

A.7. Assessment

Code	Assessment
1	Accepted
2	Rejection due to capacity
3	Rejection due to professional reasons
4	Assessment so far

Table A.11.: Map between code and assessment outcome

A.8. ICD-1

Code	ICD-1
1000 and 000	No condition proven on axis 1
1999 and 999	Missing information to code on axis 1
NULL	Blank

Table A.12.: Map between code and registrations on axis 1. The other codes remain in their original form.

A.9. Closing code

Code	Closing code
1	Completed assignment
2	Patient cancelled
3	Parents cancelled
4	Above age
5	Moved/wrong district
6	Death
7	Rejection
8	Did not get started
9	Other

Table A.13.: Map between code and closing code.

A.10. After code

Code	After code
1	Back to referring instance
2	Referred to another instance
3	Missing offer
4	No need for follow-up
5	Other

Table A.14.: Map between code and after code.

B. Additional PostgreSQL-Queries

PostgreSQL-query for data analysis of the entire cohort: To compare age and gender between the entire cohort and the data selection used in the experiments, a separate query was made to extract all patients, regardless of referral reasons or any diagnoses. However, the criteria of just including the first referral period is kept. This selection yields 2 columns and 22,482 rows.

```
select
        pasient.kjonn as "gender",
        date_part('year', age(sak.henvdato, pasient.fdt)) as "age"
from
        sak
left join pasient on
        sak.pasient = pasient.nr
right join (
        select
                pasient,
                min(henvdato) as "henvdato"
        from
                sak
        group by
                pasient) as oldestCase on
        (sak.henvdato = oldestCase.henvdato
                and sak.pasient = oldestCase.pasient)
where
(date_part('year', age(sak.henvdato, pasient.fdt)) > -1
                and date_part('year', age(sak.henvdato, pasient.fdt)) < 19)
        and (pasient.kjonn > 0
                and pasient.kjonn <3)
group by
        sak.pasient,
        sak.nr,
        pasient.kjonn,
        "age"
```

```
order by
        sak.pasient
```

C. Additional Cluster Process Figures

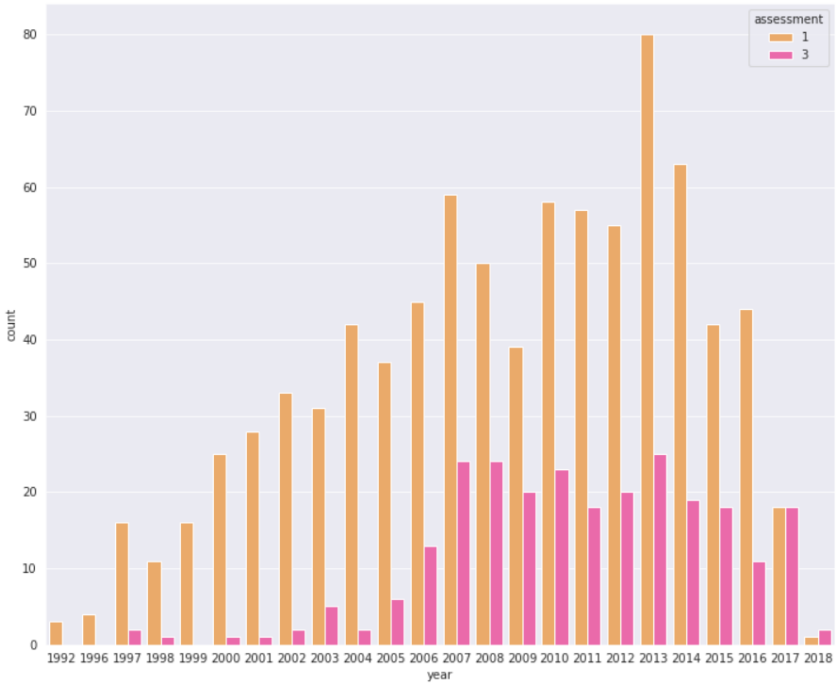


Figure C.1.: Development of the assessments of referrals from GPs from 1992-2018.

Figure C.1 illustrates the development of assessment of referrals from the GPs from 1992-2018. To emphasise rejections and admissions, the eight rows with *sak.tattimot*=4 (assessment so far) were removed. There are 857 admitted patients and 235 rejected patients referred by the GP in our cohort of patients with relevance to hyperkinetic disorders.

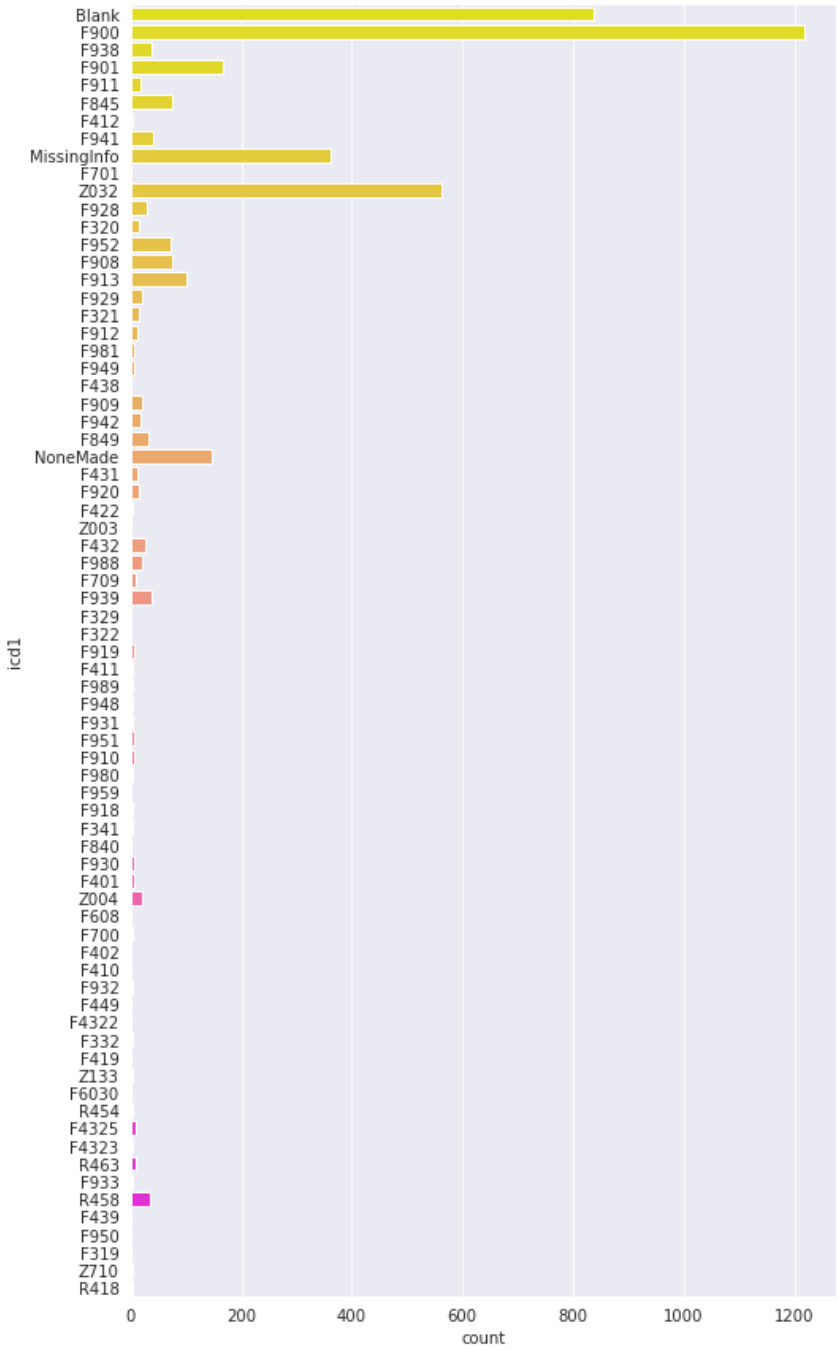


Figure C.2.: Frequency of ICD-1 after correction of mapping error.

D. Jupyter Notebook

This appendix includes the complete code from Jupyter Notebook, in which data was processed and analysed. The notebook is available in the Workbench provided by HUNT Cloud.

Sections of the Python code used in this research are influenced by a Telecom Churn-project conducted in 2021 by Zack Zazueta (Zazueta, 2021), but have greatly been adapted to fit the scope of this project.

Note that the outputs have been removed in order to emphasise the code and keep the appendix short. It is recommended to run each code block before the start of any new comment.

D.1. EDA-analysis

```
# Import libraries and packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from matplotlib import style

from kmodes.kmodes import KModes
from kmodes.kprototypes import KPrototypes
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix
from plotly import graph_objects as go

from scipy.stats import chi2_contingency
from scipy.stats import chi2
from tqdm import tqdm
from dython.model_utils import metric_graph
from dython.nominal import associations

from lightgbm import LGBMClassifier
```

```
import shapely
from sklearn.model_selection import cross_val_score

import warnings
warnings.filterwarnings("ignore")

# Load the data
df = pd.read_csv("/home/fridss/datasets/clustersets/cluster-12-4201-v1.
    csv")

# The dimension of data
print('Dimension data: {} rows and {} columns'.format(len(df), len(df.
    columns)))

# Print the first 5 rows
df.head()

# Inspect non-null objects
df.info()

# Inspect the categorical variables
df.select_dtypes('object').nunique()

# look for null values
df.isna().sum()

# Examine the data for anomalies
for i in df.columns:
    print(df[i].value_counts())

# Make bar plots as a part of the EDA
X = list(i for i in df._get_numeric_data().columns)
Y = list(i for i in df.columns) #if i not in X

for i in Y:
    plt.figure(figsize=(8,10))
    #sns.color_palette("Paired")
    sns.set_style('darkgrid')
    sns.countplot(y=i, data=df, palette='spring_r', hue='assessment')
    #hue='gender'
    sns.color_palette("Set3")
    plt.legend(loc='lower right', title='Assessment')

# Inspect the categorical variables
df.select_dtypes('object').nunique()

# Inspect the numerical variables
df.describe()

# Make scatterplots as a part of the EDA.
# The code below is for different scatter plots with different variables
plt.figure(figsize=(10,22))
sns.catplot(x="age", y="icd1", data=df,
```

```
                height=9, aspect=1, palette='hot', hue="gender")

sns.set_style('whitegrid')
ax = sns.catplot(y="refinstance", x="refreason", data=df,
                height=11, aspect=0.95, kind='swarm', hue='assessment')
locs, labels = plt.xticks()
plt.setp(labels, rotation=90)

sns.catplot(x="age", y="refreason", hue="assessment",
            col="gender", aspect=.7, height=9,
            kind="swarm", data=df)

sns.catplot(x="age", y="custody", hue="assessment",
            col="gender", aspect=.7, height=9,
            kind="swarm", data=df)

sns.catplot(x="age", y="care", hue="assessment",
            col="gender", aspect=.7, height=9,
            kind="swarm", data=df)

sns.catplot(x="age", y="icd1", hue="assessment",
            col="gender", aspect=.7, height=9,
            kind="swarm", data=df)

sns.catplot(x="age", y="refinstance", hue="assessment",
            col="gender", aspect=.7, height=9,
            kind="swarm", data=df)
```

Listing D.1: EDA analysis of selected dataset.

D.2. Minor EDA of entire cohort

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from matplotlib import style
# Import module for data visualisation
from plotnine import *
import plotnine

from kmodes.kmodes import KModes
from kmodes.kprototypes import KPrototypes
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix
from plotly import graph_objects as go

```

```
from scipy.stats import chi2_contingency
from scipy.stats import chi2
from tqdm import tqdm
from dython.model_utils import metric_graph
from dython.nominal import associations

from lightgbm import LGBMClassifier
import shapely
from sklearn.model_selection import cross_val_score

import warnings
warnings.filterwarnings("ignore")
warnings.filterwarnings('ignore', category = FutureWarning)
# Format scientific notation from Pandas
pd.set_option('display.float_format', lambda x: '%.3f' % x)

# Load the data
df = pd.read_csv("/home/fridss/datasets/clustersets/gender-age-entire-
    cohort.csv")

# The dimension of data
print('Dimension data: {} rows and {} columns'.format(len(df), len(df.
    columns)))

# Print the first 5 rows
df.head()

# Make bar plots for EDA
X = list(i for i in df._get_numeric_data().columns)
Y = list(i for i in df.columns) #if i not in X

for i in Y:
    plt.figure(figsize=(12,9))
    #sns.color_palette("Paired")
    sns.set_style('darkgrid')
    sns.countplot(x=i, data=df, palette='spring_r')
    #hue='gender'
    sns.color_palette("Set3")
```

Listing D.2: Minor EDA of entire cohort.

D.3. Clustering with K-prototype

```
import seaborn as sns
from matplotlib import style
# Import module for data visualisation
from plotnine import *
import plotnine

from kmodes.kmodes import KModes
from kmodes.kprototypes import KPrototypes
```

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix
from plotly import graph_objects as go

from scipy.stats import chi2_contingency
from scipy.stats import chi2
from tqdm import tqdm
from dython.model_utils import metric_graph
from dython.nominal import associations

from lightgbm import LGBMClassifier
import shapely
from sklearn.model_selection import cross_val_score

import warnings
warnings.filterwarnings("ignore")
warnings.filterwarnings('ignore', category = FutureWarning)
# Format scientific notation from Pandas
pd.set_option('display.float_format', lambda x: '%.3f' % x)

# Load the data
df = pd.read_csv("/home/fridss/datasets/clustersets/cluster-12-4201-v1.
    csv")

# The dimension of data
print('Dimension data: {} rows and {} columns'.format(len(df), len(df.
    columns)))

# Print the first 5 rows
df.head()

# Make dataframe
for i in df.columns:
    if df[i].dtype == 'float64':
        df[i] = df[i].astype('int64')

data_corr = df.copy()

# Preprocess numericals (standardisation)
Num_features = data_corr.select_dtypes(include='int64').columns
data_corr[Num_features] = StandardScaler().fit_transform(data_corr[
    Num_features])
data_corr.head()

# Fit and train K-prototype
style.use("ggplot")
colors = ['b', 'orange', 'g', 'r', 'c', 'm', 'y', 'k', 'Brown', '
    ForestGreen']
# List the categorical columns
cat_cols = [0, 2, 3, 4, 5, 6, 7, 8, 10, 11]
```

```
58
59 # Verbose is the degree to which output is produced
60 kproto = KPrototypes(n_clusters=6, init='Huang', verbose=2)
61 clusters = kproto.fit_predict(data_corr, categorical=cat_cols)
62
63 # Print cluster centroids of the trained model.
64 print(kproto.cluster_centroids_)
65
66 # Print training statistics
67 print(kproto.cost_)
68 print(kproto.n_iter_)
69
70 # Print count of patients in each cluster
71 print(pd.Series(clusters).value_counts())
72
73 # Setting the objects to category
74 cat_data = data_corr.copy()
75 for i in cat_data.select_dtypes(include='object'):
76     cat_data[i] = cat_data[i].astype('category')
77
78 proto_labs = kproto.labels_
79
80 # Produce cross validation score
81 clf_kp = LGBMClassifier(colsample_by_tree=0.8)
82 cv_scores_kp = cross_val_score(clf_kp, cat_data, proto_labs, scoring='
    f1_weighted')
83 print(f'CV F1 score for K-Prototypes clusters is {np.mean(cv_scores_kp)}'
    )
84
85 clf_kp.fit(cat_data, proto_labs)
86
87 # Produce SHAP plot
88 from lightgbm import LGBMClassifier
89 import shap
90 from sklearn.model_selection import cross_val_score
91
92 import warnings
93 warnings.filterwarnings("ignore")
94
95 explainer_kp = shap.TreeExplainer(clf_kp)
96 shap_values_kp = explainer_kp.shap_values(cat_data)
97
98 shap.summary_plot(shap_values_kp, cat_data, plot_type="bar", plot_size
    =(15, 10))
99
100 # Merge labeled data
101 labeled_data = pd.merge(df, pd.DataFrame(proto_labs, columns=['Cluster'])
    , left_index=True, right_index=True)
102
103 # Get statistics for numericals
104 labeled_data.groupby('Cluster').mean()
105
106 # Show labeled dataset
```

```
107 labeled_data
108
109 # Save labeled dataset to file
110 labeled_data.to_csv("labeled_data.csv")
```

Listing D.3: Clustering of dataset.

Frida Sofie Solheim

Characterising Patients Referred on Suspicion of ADHD and Behavioral Difficulties

NTNU
Kunnskap for en bedre verden

IDDEAS