Master's thesis

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

Ronja Linnéa Bævre Ek
Elin Schanke Funnemark

Clustering Hyperkinetic Child and Adolescent Patient Trajectories

Using Norwegian Electronic Health Record Data

Master's thesis in Computer Science
Supervisor: Øystein Nytrø
June 2023

NTNU
Norwegian University of
Science and Technology

Ronja Linnéa Bævre Ek
Elin Schanke Funnemark

Clustering Hyperkinetic Child and Adolescent Patient Trajectories

Using Norwegian Electronic Health Record Data

Master's thesis in Computer Science
Supervisor: Øystein Nytrø
June 2023

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

NTNU
Norwegian University of
Science and Technology

Abstract

The extensive amount of complex and heterogenous healthcare data in Electronic Health Records (EHRs) presents a challenge for analysis, encouraging data mining techniques to derive meaningful insights. Clustering, a reliable and efficient unsupervised data mining tool, is known for identifying natural groupings and hidden relationships in large data sets. Despite extensive clustering research in various domains, the application of clustering in mental health, particularly in the context of hyperkinetic disorders, remains relatively unexplored. Hyperkinetic disorders, which represent 29% of all mental health disorders among Norwegian children and adolescents, serve as the focal point of this Master's Thesis.

This thesis employs a cluster analytic approach to identify subgroups of hyperkinetic patient trajectories within Child and Adolescent Mental Health Services (CAMHS) in Norway, using EHR data collected at St. Olavs Hospital, Trondheim. The experiment consists of three sub-experiments wherein each iteration utilises the k-prototypes algorithm to cluster *Episodes of Care* (EoCs) and *Episodes of Care Bundles* (EoC Bundles) in a stepwise manner. The first cluster step identifies subgroups of the EoC data. Then the next step includes these subgroups when clustering the EoC Bundle data.

The final results distinguish patient trajectories by identifying three EoC subgroups and four EoC Bundle subgroups. These subgroups differ in patient characteristics such as age and gender, trajectory length, diagnoses given, clinical resources needed, and other trajectory aspects. In short, some trajectories are distinguished by being more "typical" hyperkinetic trajectories based on clinicians' experiences. These are often longer EoC Bundles, including planned polycyclic EoCs. Other trajectories are more varied in length and care and immediacy level.

The intermediate findings and final results are assessed through a combination of clustering validation and clinical assessment. This evaluation process facilitates the interpretation of findings and ensures their clinical relevance. Additionally, it enables ongoing improvements throughout the iterative process, as the feedback is incorporated by modifying, adding and removing features. The iterative clinical feedback and evaluation show an increase in the meaningfulness of the clustering results.

This study pioneers the clustering of hyperkinetic patient trajectories within CAMHS in Norway, contributing to the field of data mining within healthcare. The results reveal the presence of distinct subgroups within these patient trajectories, characterised by unique factors. This evidence supports the feasibility of clustering EHR data to identify clinically meaningful subgroups, opening up new avenues for future research.

Abstrakt

Store mengder kompleks og heterogen helsedata i elektroniske pasientjournaler gjør dataen utfordrende å analysere, og åpner for bruk av datautvinningsmetoder for få innsikt. Dataklynging er et pålitelig og effektivt verktøy for datautvinning, kjent for sin evne til å identifisere naturlige grupperinger og skjulte relasjoner i store datasett. Til tross for omfattende forskning på dataklynging i ulike domener, er bruken innen mental helse, spesielt i sammenheng med hyperkinetiske lidelser, relativt lite forsket på. Hyperkinetiske lidelser utgjør 29% av alle psykiske lidelser blant norske barn og ungdommer og er fokuset for denne masteroppgaven.

Denne avhandlingen dataklynger data hentet fra St. Olavs hospital i Trondheim for å identifisere undergrupper av hyperkinetiske pasientforløp innen psykisk helsevern for barn og unge. Eksperimentet består av tre deleksperimenter, der hvert deleksperiment bruker k-prototypes algoritmen til å dataklynge omsorgsepisoder og omsorgsperioder stegvis. Det første dataklyngesteget identifiserer undergrupper av omsorgsepisoder. Deretter brukes disse undergruppene av omsorgsepisoder til å dataklynge omsorgsperioder.

Sluttresultatene skiller hyperkinetiske pasientforløp ved å identifisere tre undergrupper av omsorgsepisoder og fire undergrupper av omsorgsperioder. Disse undergruppene differensieres av pasientkarakteristikker som alder og kjønn, forløpslengde, diagnoser, kliniske ressurser og andre forløpsaspekter. Kort oppsummert utpeker noen pasientforløp seg som "tradisjonelle" hyperkinetiske pasientforløp basert på klinikeres erfaringer. Dette er ofte lengre omsorgsperioder med planlagte, polikliniske omsorgsepisoder. Andre pasientforløp utpeker seg ved å ha variert omsorgsperiodelengde, behov for øyeklikkeling hjelp og omsorgsnivå.

Delresultatene og de endelige resultatene evalueres gjennom dataklyngevalidering og ved å presentere resultatene til klinikere. Denne evalueringsprosessen muliggjør tolkning av resultatene og sikrer klinisk relevans. I tillegg muliggjør evalueringen av delresultatene kontinuerlige forbedringer gjennom den iterative prosessen, da variabelendringer blir gjort ut ifra tilbakemeldingene. Iterative kliniske tilbakemeldinger og endelig evaluering av sluttresultatene viser at resultatene har blitt mer meningsfylte i løpet av prosessen.

Denne studien er en pionér innenfor dataklynging av hyperkinetiske pasientforløp innenfor psykisk helsevern for barn og unge. Resultatene avslører distinkte undergrupper av pasientforløp, kjennetegnet av unike faktorer. Dette åpner for fremtidig forsking innenfor området.

Preface

This Master's Thesis is written as the final work of the Master of Science degree in Computer Science from the *Norwegian University of Science and Technology* in Trondheim, Norway. The work was conducted in the Spring of 2023 in collaboration with the *IDDEAS* project, an interdisciplinary research group. Associate Professor Øystein Nytrø, Department of Computer Science, supervised the project.

We want to express a special thanks to our supervisor Øystein Nytrø, for his guidance throughout the process of writing this thesis. His guidance has brought many new ideas and enlarged our interest in data mining. Furthermore, we have greatly appreciated the help of the PhD Student Kaban Koochakpour for her continuous support throughout this project.

We would also like to thank the IDDEAS team and the clinicians participating in the experiment for their support and feedback. Particularly, we want to thank Odd-Sverre Westbye for continuously offering his time to evaluate the results and give detailed clinical feedback. We enjoyed these discussions, which helped us gain clinical understanding and improve our results.

Abbreviations

ACM	Association for Computing Machinery
ADHD	Attention Deficit Hyperactivity Disorder
CAMHS	Child and Adolescent Mental Health Services
CDSS	Clinical Decision Support System
CGAS	Children's Global Assessment Scale
EHR	Electronic Health Record
EMR	Electronic Medical Record
EoC	Episode of Care
EoC Bundle	Episode of Care Bundle
HUNT	Trøndelag Health Study
ICD	International Classification System of Diseases
ICD-10	The 10th version of the International Classification System of Diseases
IDDEAS	Individualised Digital DEcision Assist System
NPR	National Patient Register
NTNU	Norwegian University of Science and Technology
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PubMed	United States National Library of Medicine
REK	Regional Committees for Medical and Health Research Ethics
RKBU	Regionalt kunnskapssenter for barn og unge - psykisk helse og barnevern
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction
SHAP	SHapley Additive exPlanations
WHO	World Health Organization

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the Master's Thesis by first providing the background and motivation behind the research. Following that, the goal and research questions are defined to establish the purpose of the thesis. Subsequently, the research method is presented, outlining the approach used to address the research questions. Finally, this chapter summarises the structure of the thesis, giving the readers an overview of what to expect in the following chapters.

1.1 Background and Motivation

Today, Electronic Health Records (EHRs) offer extensive access to a large volume of healthcare data. This data is complex and heterogenous due to its reliance on medical expertise combining clinical guidelines, individual physician experiences, and patient-specific information and conditions (Evans, 2016). Consequently, the analysis of healthcare data can be challenging. With this comes a need for data mining to convert data into meaningful information. Clustering is a useful data mining tool because of its consistency, speed, and reliability in discovering natural groupings and hidden relationships in large data amounts (Berner, 2016). In recent years, clustering has been subject to wide research in multiple domains, including the healthcare sector (Negi and Chawla, 2021). Although clustering techniques have shown use within the healthcare sector, clustering within healthcare still lacks research and is not fully explored (Berner, 2016).

Hyperkinetic disorders is a group of disorders characterised by early onset, lack of persistence in activities requiring cognitive engagement, and a tendency to move from one activity to another without completing any of them. These disorders are also associated with disorganised, ill-regulated, and excessive activity. Children with hyperkinetic disorders are often reckless and impulsive making them prone to accidents and disciplinary trouble. These challenges stem from unthinking rule violations rather than deliberate defiance. Impairment of cognitive functions is common, and specific motor and language development delays are disproportionately frequent (Direktoratet for e-helse, 2022).

Nearly 4% of all 12-year-olds in Norway have hyperkinetic disorders, and the disorders constitute 29% of all mental health disorders among Norwegian children and adolescents (IDDEAS, n.d.; Young et al., 2013). Hyperkinetic disorders are some of Europe's most neglected and misunderstood psychiatric conditions. Due to the lack of public awareness and the widespread social stigma surrounding hyperkinetic disorders, very few people affected receive appropriate diagnoses and support. This lack of access to diagnoses and support often worsens the condition and may deteriorate the quality of life (Young et al., 2013).

This Master's Thesis is written in collaboration with the *IDDEAS team*. The IDDEAS team is a project group dedicated to improving patient care. Their primary objective is to provide healthcare professionals with real-time access to data-driven and evidence-based guidelines, enabling earlier and more precise decision-making. The team comprises various professionals, including developers, researchers, health IT specialists, and clinicians. Specifically, the team is focused on developing the *Individualised Digital DEcision Assist System*, the first decision support system implemented within the Child and Adolescent Mental Health Services (CAMHS) in Norway. The IDDEAS project's current focus is preventive treatment, early intervention, early diagnosis, treatment, and management of hyperkinetic disorders (IDDEAS, n.d.).

1.2 Goal and Research Questions

This Master's Thesis aims to unite the field of data mining and child and adolescent mental health by exploring the use of clustering to identify patient trajectories. The overall project goal is:

> **Goal** Analyse patient trajectories of hyperkinetic disorders in child and adolescent mental health using clustering of electronic health record data to identify subgroups.

Subgroups are divisions within a data set where data points within each subgroup exhibit distinct characteristics compared to data points in other subgroups. The identification of subgroups enables the discovery of united similarities within a subgroup while differentiating them from other subgroups.

A fundamental part of this research goal is to obtain insight into and use CAMHS EHR data. From this, the following research question is derived:

> **Research Question 1** How can hyperkinetic patient trajectories in an electronic health record be identified?

Using a CAMHS EHR data to identify subgroups, it is important to question the implications of the clustering. To evaluate this, clinicians having expertise within the mental health field should be included. Therefore, the second research question is as follows:

> **Research Question 2** How can patient trajectory clusters be made meaningful to clinicians?

This goal and the two research questions will provide a clear focus for this project, establishing a direction for the research and guiding the experiment.

1.3 Research Method

This Master's Thesis adopts a framework for understanding, executing, and evaluating a Design Science Research (DSR). DSR is a problem-solving paradigm seeking to enhance knowledge by using innovative solutions to real-world problems (Brocke et al., 2020). In alignment with this framework, the initial steps in this research method involve identifying the problem, establishing motivation, and defining the research objectives, as presented in the previous sections.

The subsequent phase entails the design and development of the project artefact. First, this phase includes building the necessary knowledge by exploring foundational concepts and methodologies. In the context of this project, a crucial aspect involves exploring and extracting data from an EHR. This necessitates reviewing clinical codes, systems, and procedures employed within CAMHS in Norway. This exploration will be presented as the Theoretical Background in Chapter 2. Thereafter, the clustering methodology will be investigated and presented in Chapter 3. This methodology will provide the framework for clustering the EHR data. To ground this clustering methodology, relevant prior work is examined and presented in Chapter 4. This will facilitate an understanding of mental health data clustering and existing research. Lastly, the data utilised in this process will be presented in Chapter 5, giving an overview of the relevant data to achieve the research goal and answer the research questions.

An experiment will be conducted to demonstrate the artefact, which is represented by the clustering process. This experiment aims to showcase the use of a clustering methodology to identify patient trajectories within CAMHS in Norway. The concrete steps for this experiment will be outlined in Section 6.1.3.

The experiment will consist of three sub-experiments, each designed to achieve predetermined experimental aims. These aims will be derived from the research questions stated to provide clear guidance for the experiment. The sub-experiments will be performed to evaluate how well the use of the artefact answers the research questions iteratively. Including clinical feedback throughout these sub-experiments may improve the clustering outcomes.

Several criteria will be considered when selecting clinicians to participate in these sub-experiments. First, including clinicians with years of experience treating patients in relation to hyperkinetic disorders within CAMHS in Norway is essential. Their extensive experiences will enrich the analysis of the cluster findings. Moreover, this may ensure the inclusion of clinicians having years of familiarity with EHR systems to ensure that their thoughts reflect real-world practices. Additionally, including clinicians who hold or have held a leadership role is desired, as their input can draw upon the experiences of other clinicians. Lastly, the clinicians should work at different clinics to offer a broader clinical perspective.

Furthermore, professionals from the IDDEAS team will participate in the experiment to obtain feedback from clinicians and domain experts familiar with this project's goal and the EHR data. The feedback from one sub-experiment will inform and enhance subsequent iterations to address the research questions better.

After completing the final iteration, a concluding evaluation will be performed as presented in Chapter 8. This evaluation will include cluster validation, result evaluation aligned with the research aims and clinical assessment. Then, a discussion will be presented in Chapter 9 regarding this research method and the results. This discussion will explore different factors influencing the clustering process and its outcomes, along with limitations and recommendations. Finally, the Master's Thesis will conclude by answering the research questions and determining if the research goal is met in Chapter 10.

1.4 Thesis Structure

The remaining of this Master's Thesis is organised in the following manner:

- Chapter 2 presents the essential background theory to familiarise the reader with relevant topics for this project.
- Chapter 3 describes the clustering methodology used during the experiment.
- Chapter 4 covers related work within CAMHS in Norway and a literature review regarding clustering of EHR data.
- Chapter 5 presents the data used in this project, the environment and the tools used when handling the data, and the agreements required to access the data.
- Chapter 6 details the experiment, including three iterations of the clustering process and the intermediate findings.
- Chapter 7 presents the final results of the last clustering iteration.
- Chapter 8 evaluates the methodology followed and the obtained results.
- Chapter 9 discusses the methodology and results in light of research, knowledge, clinical evaluation, and the project goal.
- Chapter 10 concludes the research and delivers final thoughts regarding the contributions of this project.

Chapter 2

Theoretical Background

This chapter introduces the foundational theory necessary for understanding the theoretical concepts utilised in this project. It provides the reader with essential background information that forms the basis of the project and offers an overview of the relevant fields studied prior to handling the EHR data and conducting the experiment. The chapter begins by introducing CAMHS in Norway and then explains the process for establishing clinical diagnoses. Next, important terminology is provided to ensure a shared understanding of hyperkinetic patient trajectories within CAMHS in Norway. Finally, the theoretical information is contextualised to explain patient trajectory guidelines followed.

2.1 CAMHS in Norway

CAMHS in Norway provides psychiatric assessment, counselling, treatment, and facilitation for children and adolescents aged 18 and under. CAMHS is a specialist health service organised as part of public hospitals. They serve municipal health services, schools, child protection, and general practitioners acting as gatekeepers who can refer to CAMHS (Koochakpour et al., 2022). The *Norwegian Directorate of Health* states that all children and adolescents in Norway with clear signs or symptoms of mental difficulties or disorders should be offered the help they need from the level of care required, either from the municipal health services or CAMHS (Helsedirektoratet, 2021). The predominance of referral reasons to CAMHS is due to externalising disorders, where the biggest patient group is referred to CAMHS due to hyperkinetic disorders (Breivik, 2020).

CAMHS in Norway can be considered a pioneer in comprehensive and rich coding of patient data by having the first EHR that reported individual patient treatment to the *National Patient Register* (NPR) following national requirements. CAMHS Norway has since 1984 provided standardised coding of patients' conditions, progressions, and status, including diagnoses, interventions, activities, and clinicians' notes, as well as patient demographics, family situations, and care collaborators. The extensive data collected in this EHR system, called *BUPdata*, allowed for multi-diagnosis and state-based encoding of diseases, and it can be considered the first secure patient portal in Norway. By utilising BUPdata, CAMHS in Norway has managed to support clinical work, promote quality in clinical practice, and ensure uniform quality of care across the country (Koochakpour et al., 2022).

2.2 Clinical Diagnoses

CAMHS in Norway gives clinical diagnoses to administer long-term collaborative treatments to their patients and stores these in the EHR for future reference. The diagnoses are given using the *International Classification System of Diseases* (ICD) and a multi-axial classification system (WHO, n.d.; Gårdvik, 2007). Understanding these two systems is crucial when analysing patient trajectories.

2.2.1 ICD-10

To classify mental health conditions CAMHS in Norway employs *the 10th version of the International Classification of Diseases* (ICD-10). ICD-10 is a worldwide standard for health data, clinical documentation, and statistical aggregation. The main objective of ICD-10 is to ensure that recorded data has semantic interoperability and reusability, catering to various uses beyond mere health statistics, such as decision support, resource allocation, reimbursement, and guidelines (WHO, n.d.).

As ICD-10 has been used in Norway since 1999, this project will rely on the ICD-10 guidelines for health-related issues and information. ICD-10 consists of 21 chapters, including codes for various health-related issues. To classify child and adolescent mental health disorders, relevant chapters are Chapter F (also coded as V), Chapter R (also coded as XVII), and Chapter Z (also coded as XXI). These codes are defined as follows by *The Norwegian Directorate of eHealth*:

- **Chapter F** presents a range of mental and behavioural disorders. This chapter includes disorders of psychological development but excludes symptoms, signs, and abnormal clinical and laboratory findings. Chapter F encompasses the subsequent categories of disorders:
  - **F00-F09**: Organic, including symptomatic, mental disorders.
  - **F10-F19**: Mental and behavioural disorders due to psychoactive substance use.
  - **F20-F29**: Schizophrenia, schizotypal, and delusional disorders.
  - **F30-F39**: Mood affecting disorders.
  - **F40-F48**: Neurotic, stress-related, and somatoform disorders.
  - **F50-F59**: Behavioural syndromes associated with physiological disturbances and physical factors.
  - **F60-F69**: Disorders of adult personality and behaviour.
  - **F70-F79**: Mental retardation.
  - **F80-F89**: Disorders of psychological development.
  - **F90-F98**: Behavioural and emotional disorders with onset usually occurring in childhood and adolescence.
  - **F99**: Unspecified mental disorder.

- **Chapter R** encompasses symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified. While there are R-codes covering all organs and their functions, the codes R40-R46 are especially important in CAMHS in Norway. These codes are categorised under "Cognition, perception, emotional state and behaviour". Relevant codes in this category are:
  - **R40**: Somnolence, stupor, and coma.
  - **R41**: Other symptoms and signs involving cognitive functions and awareness.
  - **R42**: Dizziness.
  - **R43**: Disturbances of smell and taste.
  - **R44**: Other symptoms and signs involving general sensations and perceptions.
  - **R45**: Symptoms and signs involving emotional state.
  - **R46**: Symptoms and signs involving appearance and behaviour.
- **Chapter Z** includes codes for special purposes. Relevant Z-codes for CAMHS in Norway include:
  - **Z00-Z13**: Persons encountering health services for examinations.
  - **Z55-Z65**: Persons with potential health hazards related to socioeconomic and psychosocial circumstances.
  - **Z70-Z76**: Persons encountering health services in other circumstances.
  - **Z80-Z99**: Persons with potential health hazards related to family and personal history and certain conditions influencing health status.

(Direktoratet for e-helse, 2022)

2.2.2 The Multi-Axial Classification System

Since 2008, CAMHS in Norway has utilised a multi-axial classification system developed by the *World Health Organization* (WHO) when registering and reporting patient diagnoses. This system allows for a comprehensive classification of the different aspects of mental health conditions, which are often complex and require a more extensive approach than other clinical conditions (Direktoratet for e-helse, 2023). The system includes six axes for coding diseases, with ICD-10 codes used on axes 1 to 5.

To better understand the patient trajectories within CAMHS in Norway, it is necessary to understand the diagnoses given on each of the six axes. Referring to the *Retningslinjer for koding Multiaksial klassifikasjon i psykisk helsevern for barn og unge* the six axes are described below (Direktoratet for e-helse, 2023).

Axis 1: Clinical Psychiatric Syndrome

The first axis of the multi-axial classification system includes all necessary diagnoses to provide a complete picture of the patient's condition. The primary diagnosis, representing the patient's main condition, should be stated first. It is mandatory to include at least one code in the first axis. The codes in this axis may include F-, R- and Z-codes.

Axis 2: Specific Disorders of Psychological Development

The second axis should contain diagnoses that exhibit the following characteristics:

- Onset during childhood.
- Deficient of delayed development of functions related to the central nervous system's biological maturation.
- A stable course that does not involve relapses or remissions.

These characteristics frequently entail that the patient experiences difficulties with their language skills, visuospatial skills, and motor coordination. To assign a diagnosis to Axis 2, it is required to conduct a standardised psychological test, perform a comprehensive medical examination, or undertake another reliable medical assessment. All codings on the second axis are F-codes.

Axis 3: Intellectual Level

The codes on the third axis should indicate potential developmental disabilities and their severity. A separate classification on this axis should also describe the extent of behavioural problems. Only one code may be assigned on Axis 3, and it must result from a standardised psychological test and comprehensive evaluation of social maturity and adoption. All pertinent codes on the third axis are from the F-chapter.

Axis 4: Co-existent Medical Conditions

The fourth axis includes codes corresponding to diseases, injuries, or causes of death / suicide attempts. Only conditions pertinent to the treatment process should be coded on this axis. This implies that the conditions coded are considered and addressed during the treatment process or impact the examinations or treatment. Only physicians are authorised to assign codes on Axis 4. The fourth axis comprises code from chapters A - T (excluding chapter F).

Axis 5: Associated Abnormal Psycho-Social Situations

Axis 5 includes codes referring to the patient's family relationships, other close connections, and the environment during the last six months of the patient's life. It is possible to assign multiple codes on this axis, and the codes may describe aspects such as the patient being raised in an institution, family illness, or experiences of abuse. On Axis 5, all codes are numbered from 1-10, corresponding to Z-codes.

Axis 6: Global Assessment of Disability

The sixth axis employs the Children's Global Assessment Scale (CGAS). The CGAS ranges from 1 to 100 to characterise the patient's level of disability, varying from "excellent function in all areas" to "requires constant supervision". The CGAS score is determined at the outset and may be updated throughout a patient's treatment. The code is assigned based on the most reduced level of disability observed during the previous month.

2.3 Terminology

Establishing a common understanding of key terminology is crucial to analyse hyperkinetic patient trajectories in CAMHS in Norway effectively. Table 2.1 outlines official terminology presented in *Bupdata brukerhåndbok* (2009), *Volven* (2001), and other relevant terms for this project.

Term	Definition
Patient Trajectories	The assembling, scheduling, monitoring, and coordinating of all steps necessary to complete the work of patient care. The term trajectory refers not only to the pathophysiological process of a patient's disease state but also to the total organisation of work done throughout all interactions and the impact of the patient care processes (Direktoratet for e-helse, 2023).
Episodes of Care Bundle (EoC Bundle)	A time period for contacts and admissions at healthcare institutions for a condition's assessment, treatment, and rehabilitation. An EoC Bundle comprises one or more Episodes of Care (Direktoratet for e-helse, 2001).
Episode of Care (EoC)	A continuous period of time during which a patient receives care at one healthcare institution for one condition. An EoC have a determined care and immediacy level and may involve multiple contacts (Direktoratet for e-helse, 2001).
Contact	An uninterrupted interaction between a patient and health personnel where the patient receives healthcare at one healthcare institution for one health issue. All contacts in CAMHS in Norway are categorised as one of the following contact types: • Therapy: Measures to cure, combat, alleviate, and prevent discomfort, diseases, injury or disability based on science and knowledge. • Examination: Conversations and examinations to map a patient's illness, situation, and need for treatment. • Indirect contact: Work or activity related to the healthcare provided to a patient without the patient's participation. • Planning: Work where only healthcare professionals are present. Time spent planning the patient's treatment / future contacts with the patient. • No-show: A planned contact that is not performed since the patient did not show (Direktoratet for e-helse, 2001).
Diagnosis	A diagnosis refers to the multi-axial classification system based on ICD-10 codes, elaborated in Section 2.2.2. Within CAMHS in Norway, clinicians have a general rule to register as many diagnoses as needed to get a clear view of a patient's health situation (Direktoratet for e-helse, 2023). A diagnosis may be the patient's main diagnosis or not.

Continues on next page

Term	Definition
Care Level	Organise the different levels of treatment given to the patient. The care level may be one of the following three: • Polyclinic: EoC at a healthcare institution that provides assessment, treatment or rehabilitation without the patient staying overnight or participating in activities other than consultations. • Outpatient: EoC where the examination or treatment is more extensive than on the polyclinic care level but where the patient does not stay overnight. • Inpatient: EoC where the patient stays overnight at the healthcare institution (Direktoratet for e-helse, 2001).
Immedicay Level	Indicates how immediate the EoC is. The immediacy level may be one of the following five: • Acute: No waiting time before treatment. • Non-acute: Treatment within 6 hours. • 6-24 hour wait. • Planned: Treatment is planned in advance. • Return from another hospital: The patient returns after treatment in another hospital. (Direktoratet for e-helse, 2001)

Table 2.1: Terminology regarding patient trajectories within CAMHS in Norway.

2.4 Hyperkinetic Disorders Patient Trajectories

After presenting an overview of CAMHS in Norway, encompassing a description of the organisation, clinical diagnoses offering, and important terminology, patient trajectories for patients with hyperkinetic disorders within CAMHS in Norway can be introduced.

Once a patient is referred or admitted acutely to CAMHS in Norway, an EoC Bundle is started. The two referral reasons in relation to hyperkinetic disorders are suspicion of defiance/behavioural disorder and suspicion of hyperkinetic disorders. Upon an EoC Bundle start for the patient referred, *the Norwegian Directorate of Health* has established a guideline for the patient trajectory (Helsedirektoratet, 2022). This guideline begins with the initial assessment of hyperkinetic disorders. The diagnosis should be determined through the following procedures:

1. Assessment and documentation of the patient's psychosocial, developmental, somatic, and psychiatric history and status, and the patient's strengths and interests. This assessment should include the following:
    - Use of diagnostic criteria stating that at least six symptoms of inattention, three symptoms of hyperactivity, and one symptom of impulsivity must be present. These symptoms must also have been persistent for at least six months.
    - Conversation with the patient's guardians.
    - Conversation with the patient.
    - Assessment of the patient's developmental history.
    - Assessment of symptoms and function in different areas.
    - Assessment of concurring difficulties.
    - Assessment performed by a doctor.
2. A physical and neurological examination.
3. Potential additional examinations such as laboratory testing.

(Helsedirektoratet, 2022)

From these procedures, the patient may get one of the following diagnoses in the hyperkinetic disorders category (ICD-10 chapter F90) on the first axis of the multi-axial classification system:

- F900 - Disturbance of Activity and Attention
- F901 - Attention Deficit Hyperactivity Disorder (ADHD)
- F908 - Other Hyperkinetic Disorders
- F909 - Hyperkinetic Disorder, Unspecified

Additionally, the guideline emphasises the significance of considering other potential disorders during the performance of these procedures, seeing that other conditions frequently accompany hyperkinetic disorders. These diagnoses are based on the principles outlined in Section 2.2.2. When a diagnosis is given, it is determined whether it is the patient's primary diagnosis.

After the initial assessment, the treatment process for patients with hyperkinetic disorders begins. CAMHS in Norway follows guidelines by *the Norwegian Directorate of Health*, including treatment steps and principles, to ensure high-quality service. The guideline aims to establish proper priorities in the service, address interaction challenges, and ensure comprehensive patient processes. For all steps in the treatment plan, the following principles should be followed:

- All treatment measures should be determined by an assessment considering the individual patient.
- All treatment measures should have a plan of action.
- Different healthcare institutions should collaborate to complete the treatment.
- When the symptoms of hyperkinetic disorders are severe, multiple measures should be considered simultaneously.

The following are the treatment steps recommended to include when treating the patient with hyperkinetic disorders:

1. Explaining diagnoses, symptoms and the treatment plan to the patient, the guardians, and the school.
2. Parent and child training programs.
3. Prescription of drugs. This includes the following steps:
   (a) A four-week trial period.
   (b) Evaluation of symptom improvement to decide the continuation of medication.
   (c) Consideration of other treatment options if side effects are present and/or the patient, is missing improvement.
4. Social skills training should be performed if the patient is experiencing difficulties interacting with others.
5. Cognitive behavioural therapy.
6. Coaching: Providing patients with guidance, motivation, and training to support their personal development and enable them to make informed decisions.
7. Computer-based training program to improve concentration and working memory.
8. Neurofeedback: Method for creating changes in the brain's activity to find relations between observed brain waves and behaviour.
9. Nutritional interventions.
10. Facilitation and special education measures in kindergartens and schools.

All steps in the abovementioned guidelines include different types of patient contacts. The diagnostic process normally includes mostly examinations, while the treatment process may consist of all five contact types. The parent and child training program is an example of a therapy session. Creating a plan of action is a typical planning contact, and if the patient does not show up to a scheduled contact, this is an example of a no-show contact. If, for any of the contacts, the patient is not present, but the healthcare professional work with, for instance, child protection services speaking on behalf of the patient, this is an example of indirect contact.

The diagnostic and treatment steps recommended may be separated into different EoCs within one EoC Bundle, depending on where the patient is treated and the care and immediacy level. Examples of separate EoCs within an EoC Bundle are:

- If the patient is transferred to a new healthcare institution between the diagnostic steps and the treatment steps, a new EoC is started.
- If the level of immediacy changes during the treatment process, for instance, from acute to non-acute, a new EoC is started.
- If the patient at a polyclinic needs to stay at a healthcare institution for 24 hours, a new EoC is started.

Once treatment is completed, the patient or guardians cancel the process, the patient gets above age, moves, dies, or the healthcare professionals determine that the patient should be rejected, the ongoing EoC and EoC Bundle are closed. If the patient is referred to CAMHS in Norway again later in life, a new EoC Bundle is started.

Chapter 3

Clustering Methodology

This chapter presents the clustering methodology implemented in this project, aiming to provide the reader with the technical aspects employed in the experiment. This chapter first presents general information on data mining, outlining the overall data analytic approach adopted for this project. Subsequently, the clustering approach and relevant use cases are presented, detailing project considerations and the specific clustering algorithm utilised in the experiment.

3.1 Data Mining

Data mining refers to the process of identifying valid, novel, and easily interpretable patterns within a data set (Fayyad et al., 1996). The main goal of data mining is to convert data into meaningful information. Given the significant volume of data generated in the healthcare sector, data mining is highly suitable for providing decision support. There are two main categories of data mining: *supervised* and *unsupervised* machine learning (Berner, 2016). Supervised machine learning enables algorithms to learn from labelled cases and generalise knowledge to predict new cases (Berry et al., 2020). Unsupervised machine learning leaves algorithms to discover patterns and hidden structures from data without predefined outcomes (Peiffer-Smadja et al., 2020).

For this project the unsupervised machine learning approach *clustering* is utilised to identify hyperkinetic patient trajectory subgroups in CAMHS in Norway.

3.2 Clustering

The unsupervised machine learning technique clustering is used to group unlabeled data into clusters containing data points "similar" to each other and "dissimilar" from data points in other clusters (Ahmad and Khan, 2019). The clustering technique is helpful in data mining because of its consistency, speed, and reliability (Huang, 1997b). Clustering organises objects into groups whose members are similar according to, most often, some proximity criteria defined by introducing distances. When utilising clustering, the aim is to derive a description that succinctly characterises the elements of a cluster (Ahmad and Dey, 2007).

Cluster analysis is widely used across various fields, serving as a vital tool for numerous applications (Halkidi et al., 2001b). The following presents applications of clustering relevant to this project:

- **Data compression:** One of the key advantages of cluster analysis is its ability to compress information contained in the data by partitioning it into multiple clusters. Instead of processing the entire data set as an entity, clustering enables using the identified clusters to represent the data (Halkidi et al., 2001a).
- **Natural classification:** Clustering offers an evaluation of the data's similarity degree based on natural groupings. It can be considered a statistical classification technique that quantitatively compares multiple characteristics to determine whether individuals in a population fall into different groups. As a result, clustering can be a valuable descriptive tool when one wants to understand the general characteristics of high-dimensional data but does not have pre-specified models or hypotheses (Jain, 2010).

When dealing with complex and heterogeneous data, combining the two mentioned clustering applications in a stepwise manner can be beneficial. This implies first using clustering to compress information before including this information in a second clustering that aims to characterise the data by constructing and identifying higher-level subgroups. These clustering applications, utilised stepwise, can discover underlying patterns or structures in unlabeled data sets, making it an ideal solution for patient trajectory segmentation.

3.2.1 Clustering Considerations

While clustering analysis is a fundamental technique in data mining, it is important to consider certain factors when applying this method. Following is a presentation of clustering considerations pertinent to this project that should be considered when selecting a clustering algorithm and carrying out the clustering process.

Large Data Sets

A fundamental data mining problem is efficiently partitioning large data sets into homogeneous clusters (Huang, 1997b). Clustering algorithms can be broadly categorised as either hierarchical or partitional, depending on how they partition a data set into clusters. Hierarchical clustering algorithms create clusters in succession based on previously formed clusters, while partitional clustering algorithms determine all clusters simultaneously (Soni Madhulatha, 2012). Comparing the two categorisations, partitional clustering algorithms are generally more computationally efficient than hierarchical clustering algorithms. Partitional clustering methods have a linear time complexity with respect to the number of data points, whereas hierarchical clustering has a time complexity of $O(n^3)$ (Soni Madhulatha, 2012). Given this computational advantage, partitional clustering methods may be a more favourable choice for clustering large high-dimensional data in data mining (Huang, 1997b).

Mixed Data

Choosing the appropriate clustering algorithm heavily relies on the type of data that needs to be clustered. When dealing with large data sets containing both numerical and categorical data, partitioning the data into homogeneous clusters becomes particularly difficult (Ahmad and Dey, 2007). Such data sets can be referred to as mixed data and are common when using real-world data sets (Ahmad and Khan, 2019). Clustering mixed data requires computing the similarity between different types of data. Distance-based similarity measures compute the similarity between numerical data, but computing the similarity between categorical data is more complex (Ahmad and Khan, 2019). Since categorical data is inherently unordered, the distance between the features cannot be directly computed. Therefore, a distance measure capable of adequately capturing similarities within the data set is needed to cluster mixed data. Additionally, the distance measure must be compatible with an efficient clustering algorithm to produce effective clustering results.

To overcome the challenge of clustering mixed data, the following three strategies may be considered:

1. Convert categorical values to numeric integer values and then apply numeric distance measures to compute the similarity between object pairs. This approach has limitations, such as accurately assigning the appropriate numeric value to categorical variables (Ahmad and Dey, 2007). Furthermore, this method may not yield meaningful results when dealing with categorical domains that are not ordered (Huang, 1997a).

2. Discretise numeric attributes before applying a categorical clustering algorithm. This discretisation process tends to lead to loss of information, resulting in misleading outcomes (Ahmad and Dey, 2007). Moreover, choosing a specific discretisation is not trivial; for some variables, there is no obvious way of splitting a range of numerical variables (van de Velden et al., 2019).

3. Consider numeric and categorical attributes separately by having a cost function that computes the similarity between two elements in terms of two distance values - one for numeric attributes and the other for categorical attributes (Ahmad and Dey, 2007).

High Dimensionality Data

Reducing the dimensionality of data prior to clustering is recommended, as high dimensionality data can cause computational inefficiency for clustering algorithms, and the presence of irrelevant features can hinder the identification of relevant underlying structures in the data (Boutsidis et al., 2015). Dimensionality reduction is usually done by utilising feature selection and/or feature extraction (Mladenić, 2006). While most clustering techniques assume that all features are equally important, in reality, different features may have varying effects on the desired clustering result. Irrelevant features can potentially blur the clusters, whereas essential features play a crucial role in creating them (Dash and Liu, 2000). To address this issue, feature selection can be used to reduce the data dimensionality by selecting only the relevant features for the clustering. Feature extraction constructs new features to be used instead of the original features. The construction of new features combines original features based on domain-specific calculations or statistical methods (Mladenić, 2006).

3.2.2 Clustering Algorithm

The k-prototypes algorithm was developed, alongside the k-modes algorithm, by Zahexue Huang. The two algorithms are extensions of the k-means paradigm, specifically designed to handle categorical and mixed attributes (He et al., 2005). By combining the k-modes and k-means algorithms, the k-prototypes algorithm can effectively operate on mixed data, and it does so by defining a dissimilarity measure that takes into account both numerical and categorical attributes (Huang, 1998). This makes the k-prototypes algorithm highly desirable for data mining and for this project (Huang, 1998).

Since k-means and k-modes lay the foundation for k-prototypes, these two algorithms will first be introduced before delving further into the details of the k-prototypes algorithm.

k-means

The k-means algorithm is the most widely recognised and utilised clustering technique (Sinaga and Yang, 2020). The algorithm is well known for its efficiency in clustering large data sets (Huang, 1998). The clustering method falls under the category of partitional clustering methods. The k-means algorithm divides a given data set into a predetermined number of clusters. The central concept behind k-means is to establish k centroids, each representing a cluster. k-means has the dual objective of making each cluster as compact and distinct from the others as possible (Ahmad and Dey, 2007). Hence, the objective function J that aims to minimise the within groups sum of squared errors can be expressed as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \tag{3.1}$$

where $||x_i^{(j)} - c_j||^2$ is the distance measure between a single data point $x_i^{(j)}$ and a specific cluster centre c_j (Saxena et al., 2017).

The procedure of the k-means algorithm consists of the following steps:

1. **Initial means selection:** Once the number of clusters has been determined, the k-means algorithm selects k distinct points from the data set to serve as the initial centroids for each cluster. The means of the centroids are initially calculated and then updated during each iteration of the algorithm before the clusters are finalised.
2. **Initial allocation:** Each point in the data set is assigned to the centroid whose mean is nearest. Once a point is assigned to a centroid, the mean of that centroid is adjusted to reflect the addition of the new data point. Thus at each stage, the k-means represent the mean of each centroid.
3. **Re-allocation:** Step 2 is repeated until the centroids no longer move. This separates the objects into groups from which the within groups sum of squared errors is minimised.

(MacQueen, 1967; Saxena et al., 2017)

The k-means algorithm has the following important properties:

- It is efficient in processing large data sets.
- It often terminates at a local optimum.
- It works only on numeric values.
- The clusters have convex shapes.

(Huang, 1998)

k-modes

The k-modes algorithm overcomes k-means' limitation in handling categorical data by implementing the following modifications:

- A simple matching similarity measure to handle categorical objects is introduced. This measure can be defined as the total matches between the corresponding attributes of two objects. The smaller the number of mismatches, the more similar the two objects are.
- The means are replaced with modes. Like the k-means algorithm, the k-modes algorithm assigns objects to the cluster with the nearest mode according to the dissimilarity measure. After each allocation, the mode of each cluster is updated accordingly.
- A frequency-based method is used to update modes in the clustering process to minimise the clustering cost function.

With these modifications, the k-modes algorithm enables the clustering of categorical data similar to how k-means clusters numerical data (Huang, 1998). Additionally, k-modes enables easier interpretations of the clustering results as the modes provide characteristic descriptions of the resulting clusters (Huang, 1997a).

k-prototypes

The clustering process of the k-prototypes algorithm is similar to k-means since it utilises a distance metric to assess the dissimilarity between observations. However, k-prototypes incorporates the k-modes technique to update the categorical values of cluster prototypes. In essence, the k-prototypes uses the mean values for numerical features and mode values for categorical features.

The dissimilarity between two mixed-type objects X and Y can be measured by the following:

$$d(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j) \tag{3.2}$$

with $\gamma > 0$ and where

$$\delta(x_j, y_j) = \begin{cases} 0 & \text{if } x_j = y_j, \\ 1 & \text{if } x_j \neq y_j \end{cases} \tag{3.3}$$

The first term in Equation 3.2 calculates the squared Euclidean distance for the numerical attributes. The second term measures the dissimilarity of the categorical attributes by looking at the number of mismatches between objects and cluster prototypes. The weight parameter γ ensures equal treatment of both attributes and prevents any favouring of either (Huang, 1998).

The procedure of the k-prototypes algorithm is built upon the following three steps:

1. **Initial prototypes selection:** k-prototypes selects k initial prototypes from a data set, one for each cluster.
2. **Initial allocation:** The algorithm allocates each object in the data set to the cluster with the nearest prototype, determined by equation 3.2. The prototype of a cluster is updated after each allocation. The following three situations can occur when the algorithm determines the nearest prototype:
   (a) A data point is assigned to a cluster if the cluster's prototype matches the data point's categorical and numerical values.
   (b) A data point can be assigned to a cluster if the numerical distance between the cluster's prototype and the data point is long, provided that the data point's categorical value matches the prototype.
   (c) A point can be assigned to a cluster if the cluster's prototype has a different categorical value, provided that the distance between the point and the cluster prototype is small enough in the numerical space.
3. **Re-allocation:** Once all objects have been allocated to a cluster, the similarity of each object is tested against the current prototypes. If an object's nearest prototype belongs to a different cluster than its current one, the object is reassigned to that cluster, and the prototypes of both clusters are updated. This step is repeated until no object changes clusters during a full cycle of the data set.

(Huang, 1997b)

The k-prototypes algorithm has the following important properties:

- It is efficient in processing large data sets.
- Like the k-means algorithm, k-prototypes produces local optimal solutions, which are affected by the selection of the initial cluster prototypes.
- It can cluster a mixture of both numerical and categorical variables.

(Huang, 1997b)

Limitations with k-prototypes

k-prototypes is a commonly used algorithm for clustering mixed data (Aschenbruck and Szepannek, 2020). However, the following limitations should be considered when utilising this clustering algorithm:

- **Potential information loss when measuring dissimilarity of categorical attributes:** k-prototypes incorporates the k-modes technique to measure dissimilarity of categorical attributes. However, this technique does not consider conceptual relations between the categorical values to simplify the dissimilarity measure. Consequently, two closely related categorical values are treated equally dissimilar as two categorical values in entirely different domains (Huang, 1997a).
- **Potential bias towards either numerical or categorical data:** Selecting a suitable weight to ensure equal treatment of numerical and categorical attributes is a challenging aspect of k-prototypes. One approach to tackle this challenge is utilising the average standard deviation of the numerical attributes as a reference for determining the weight. However, this method lacks sufficient research to be considered a generally applicable rule. Alternatively, weight assignment can be based on domain knowledge. Due to the absence of definitive guidelines for determining the weight, there is a risk of bias towards one of the data types (Huang, 1997a).

Chapter 4

Related work

This chapter offers an overview of previous research conducted within this project's field of interest, aiming to provide valuable insight, inspiration, and a basis for comparison. The chapter begins by briefly introducing research conducted by the IDDEAS team, highlighting its contributions and relevance to this project. Subsequently, a review of international papers is presented, focusing on studies that have employed similar clustering methods using EHRs or Electronic Medical Records (EMRs). Including related work is crucial to highlight the existing research gap and position this project within the field. Additionally, it aids in making informed methodological choices throughout the study.

4.1 Related Work by the IDDEAS Team

Related work done within Norwegian borders and by the IDDEAS team is interesting to investigate. The IDDEAS team is, as introduced in Section 1.1, a project dedicated to developing the first decision support system within CAMHS in Norway. Their focus is preventive treatment, early intervention, early diagnosis, treatment, and management of hyperkinetic disorders. This section shortly presents key findings from their published papers to contextualise the IDDEAS team's current phase. Then, a master thesis written in collaboration with IDDEAS being the first to cluster Norwegian EHR data is presented to give a foundation for this project. By examining this work, this project can draw upon its strengths and learn from its shortcomings.

The previous studies conducted by the IDDEAS team have focused on different aspects of implementing a *Clinical Decision Support System* (CDSS) and its potential benefits (IDDEAS, n.d.). Their research has revealed that many individuals support sharing EHRs for both research and clinical care purposes, demonstrating their awareness and endorsement of this practice (Bakken et al., 2022). Furthermore, their research has shown that implementing a CDSS can enhance the effectiveness and efficiency of healthcare delivery, leading to improved quality of care and clinical outcomes within CAMHS in Norway (Clausen et al., 2020). Lastly, the research emphasises that for the CDSS to be effective, it should integrate existing heterogeneous, geographically distinct, current, and historical patient-specific and population-specific data to generate new information and models for clinical decision support at the individual patient level. This facilitation should leverage already existing informatics frameworks (Raballo et al., 2020).

The first research done to utilise EHR data to generate new information regarding patients with relation to hyperkinetic disorders within CAMHS in Norway is the master thesis written by Frida Solheim (Solheim, 2022). Her work identified characteristics and a latent subgroup of patients, and natural patterns and phenomena were uncovered. Specifically, Solheim focused on the first period of patients' EoC Bundles to identify patients' situations more prone to rejection. Using k-prototypes, she captured important aspects of the referral process, identified patient profiles related to gender and rejection rates, and unanticipated referral and diagnostic phenomena. The methodology used included one clustering experiment, using data related to patient features, care situations, main diagnosis, and information regarding the end of an EoC Bundle. For every patient in her experiment, she considered their first referral period within CAMHS in Norway.

Solheim identifies subgroups of patients more prone to rejection. However, her work does not continue after this initial patient assessment. In her work, she states multiple areas for future work and details how the thesis lays the foundation for further analysis.

4.2 Literature Review

A structured literature review is conducted to gain sufficient knowledge within the field of cluster analysis applied to patient data derived from EHRs or EMRs. The method used for the literature review is based on *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA).

4.2.1 PRISMA Screening Process

A template from PRISMA is used to perform a systematic literature review. The PRISMA flow diagram visually summarises the screening process. It initially records the number of articles found and then makes the selection process transparent by reporting on decisions made at the various stages of the systematic review (Page et al., 2021). The number of articles is recorded at the different stages. When excluding articles at the full-text stage, including the reasons for exclusion is essential.

The screening process is presented in Figure 4.1. The *United States National Library of Medicine* (PubMed) and the *Association for Computing Machinery* (ACM) are used for the review. The following queries are used to identify records from the libraries:

ACM query:

```
([Title: patient] )
OR [Title: patients]
AND ([Abstract: clustering]
    OR [Abstract: cluster])
```

PubMed query:

```
([Title/Abstract: electronic medical records]
OR [Title/Abstract: electronic health records]
OR [Title/Abstract: EMR]
OR [Title/Abstract: EHR])
AND ([k-means]
    OR [k-modes]
    OR [k-prototypes])
```

The following are the reasons for exclusion found relevant when screening the papers:

1. Not relevant to the research questions and outcomes.
2. Wrong population/setting/intervention.

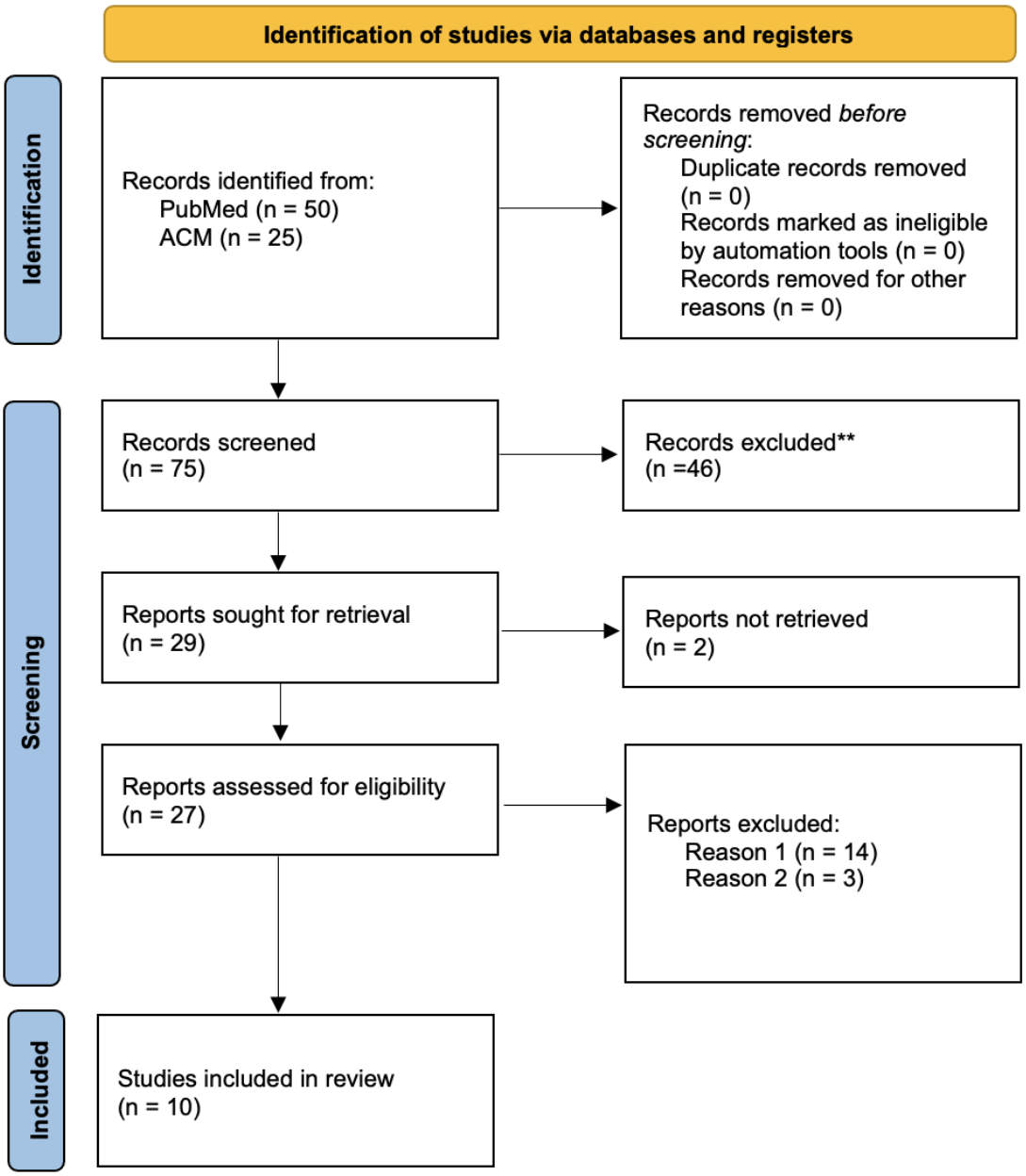


Figure 4.1: Flow diagram of the PRISMA screening process

After completing the screening process, 10 articles are left to review. These will be further presented in the following section.

4.2.2 Papers Reviewed

From the screening process conducted, 10 papers are left to review. This section presents a synopsis of the findings before a report concerning each paper's aim, data, methodology, evaluation, and conclusion is given in Table 4.1. The presented findings focus on the procedural aspects of the related work to inspire and guide the methodology of this project.

The research aim to identify and characterise patient trajectory subgroups (Table 4.1, ID 2, 3, 5, 6, 7, 9, and 10) or disease subgroups (ID 1, 4, and 8). All the papers utilise k-means-based partitional clustering algorithms, either exclusively or combined with other clustering algorithms, to identify these subgroups. While most papers rely on k-means, suitable for clustering numerical data, ID 6 uses k-prototypes to cluster mixed data directly. The differentiation among these research is how they preprocess the data before conducting the clustering analysis. Except for ID 10, the data used in the different research is not solely numerical, as indicated in the "Data" column in Table 4.1.

It is important to mention that some of the research papers (ID 1, 4, and 9) compare k-means clustering with other clustering algorithms. In contrast, ID 6 combines k-prototypes with another clustering algorithm, and ID 7 applies k-means twice on the same data. Since Table 4.1 specifically focuses on the k-means-based partitional clustering algorithm, the methodologies and evaluation measures highlighted are directly relevant to this particular clustering approach.

Moreover, the experimental findings of the reviewed research papers are briefly summarised in the "Conclusion" column of Table 4.1. However, it is important to emphasise that the primary emphasis of this structured literature review is on the procedural aspects of the related work rather than solely focusing on the final results.

Nr.	Title	Year	Author	Aim	Data	Methodology	Evaluation	Conclusions
1	Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning	2021	Alexander, Alexander, Barkhof, Denaxas	Examine the clinical heterogeneity of Alzheimer's disease patients using EHR to identify and characterise disease subgroups using multiple clustering methods.	Anonymised patient EHR from the Clinical Practice Research Datalink Numerical and categorical data	One-hot encoding Multiple correspondence analysis (MCA) Elbow plots, Silhouette coefficient, Bayesian information criterion k-means, Kernel k-means, Affinity propagation, Latent class analysis	Silhouette coefficient Jaccard coefficient Cluster comparison	Each clustering approach produced substantially different clusters. k-means performed the best.
2	Multimorbidity patterns with K-means nonhierarchical cluster analysis	2018	Violán, Roso-Llorach, Foguet-Boreu, Guisado-Clavero, Pons-Vigués, Pujol-Ribera, Valderas	Ascertain multimorbidity patterns using a non-hierarchical cluster analysis in adult primary patients with multimorbidity attended in primary care centres in Catalonia	Information System for the Development of Research in Primary Care Numerical and categorical data	MCA Calinski-Harabasz index value k-means	Jaccard coefficient	Non-hierarchical cluster analysis identified multimorbidity patterns consistent with clinical practice, identifying phenotypic subgroups of patients.

Continues on next page

Nr.	Title	Year	Author	Aim	Data	Methodology	Evaluation	Conclusions
3	Clinical and temporal characterization of COVID-19 subgroups using patient vector embeddings of electronic health records	2022	Ta, Zucker, Chiu, Fang, Natarajan, Weng	Identify and characterise clinical subgroups of hospitalised Coronavirus Disease 2019 (COVID-19) patients.	Patient EHR from Columbia University Irving Medical Center Medical coding sequences	Paragraph Vector embedding models Elbow method k-means	Chi-square test Mann-Whitney U test	20 subgroups of hospitalised COVID-19 patients, labelled by increasing severity, were categorised by their demographics, conditions, outcomes, and severity.
4	Identifying clinically important COPD subtypes using data-driven approaches in primary care population based electronic health records	2019	Pikoula, Quint, Nissen, Hemingway, Smeeth, Denaxas	Sought to discover, describe and validate chronic obstructive pulmonary disease (COPD) subtypes using cluster analysis from EHR data.	CALIBER resource EHR Numerical and categorical data	Paragraph Vector models MCA Calinski-Harabasz index value k-means and hierarchical clustering	Jaccard coefficient Silhouette coefficient	COPD patients can be sub-classified into groups with differing risk factors, comorbidities, and prognosis. The identified clusters confirm previous clustering studies and draw attention to anxiety and depression as important drivers of the disease in young, female patients.

Continues on next page

Nr.	Title	Year	Author	Aim	Data	Methodology	Evaluation	Conclusions
5	Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis	2018	Guisado-Clavero, Roso-Llorach, López-Jimenez, Pons-Vigués, Foguet-Boreu, Muñoz, Violán	Identify multimorbidity patterns and their variability over a 6-year period in patients older than 65 who attended primary health care.	EHR from Catalan Health Institute's information system Numerical and categorical data	Chi-square test Mann-Whitney test MCA Calinski-Harabaz index value k-means with random initialisation	Jaccard coefficient	Identified six multimorbidity patterns per group; one non-specific and five with a specific pattern related to an organic system. These data are useful to improve the clinical management of each specific subgroup of patients showing a particular multimorbidity pattern.
6	A two-step approach for mining patient treatment pathways in administrative healthcare databases	2018	Najjar, Reinharz, Girouardc, Gagnéa	Propose a methodology allowing the construction of patient treatment pathways from administrative relational databases to cluster them in homogenous clusters and analyse and describe them.	Administrative health care databases of the RAMQ and MSSS Numerical, categorical, and multivalued data	Tree structure categorisation First k-prototypes, then hierarchical clustering Hidden Markov models	Not specified	Designing, building, and clustering treatment patient pathways that allow the differentiation between present patterns in the data, even if patients have the same chronic disease.

Continues on next page

Nr.	Title	Year	Author	Aim	Data	Methodology	Evaluation	Conclusions
7	Weighting Primary Care Patient Panel Size: A Novel Electronic Health Record-Derived Measure Using Machine Learning	2016	Rajkomar, Yim, Grumbach, Parekh	Characterise the utilisation patterns of primary care patients and create weighted panel sizes for providers based on work required to care for patients with different patterns.	EHR from Epic, Madison, WI, USA Numerical and categorical data	k-means performed twice using Hartigan-Wong with random initialisation	Domain expert evaluation Adjusted R-squared criterion Akaike information criterion	Individual patients' health care utilisation may be useful for classifying patients by primary care work effort and predicting future primary care usage.
8	Learning Clinical Workflows to Identify Subgroups of Heart Failure Patients	2017	Yan, Chen, Li, Liebovitz, Malin	Introduce a method to identify subgroups of heart failure through a similarity analysis of event sequences documented in the clinical setting.	EMR from Northwestern Memorial Hospital Event sequences	k-means	Domain expert evaluation	Identified 8 subgroups of heart feailure, each associated with a canonical workflow inferred through an inductive mining algorithm. Each subgroup was further confirmed to be affiliated with specific comorbidities, such as hyperthyroidism and hypothyroidism.

Continues on next page

Nr.	Title	Year	Author	Aim	Data	Methodology	Evaluation	Conclusions
9	Effective Patient Similarity Computation for Clinical Decision Support Using Time Series and Static Data	2020	Masud, Hayawi, Mathew, Dirir, Cheratta	Presents a technique for computing patient similarity using time series data effectively combined with static data.	Multi-parameter Intelligent Monitoring in Intensive Care database, MIT Numerical, categorical, and times series data	Dynamic time warping, Minhash, DelMinhash ANF, Neighborhood population k-means and Spectral clustering	F1 score	Effectively combined different types of clinical data and developed an efficient unsupervised framework for computing patient similarity for CDSS.
10	Data-Driven Patient Segmentation Using K-Means Clustering: The Case of Hip Fracture Care in Ireland	2017	Elbattah, Molloy	Embraces a mere data-driven approach for segmenting patients with application to hip fracture care in Ireland.	Irish Hip Fracture Database Numerical data	Min-max normalisation Principal component analysis k-means with random initialisation	Not specified	Explored correlations between patient characteristics, care-related factors, and patient outcomes. The findings can benefit healthcare executives in Ireland to develop patient-centred care strategies.

Table 4.1: Literature review papers

Chapter 5

Data

This chapter provides an overview of the data used in this project, offering readers a comprehensive understanding of the available data before delving into specific aspects of it during the experimental phase. Firstly, the data set to be employed is presented, detailing the cohort of interest for this project. Then, the environment and tools to be employed for accessing, exploring, and analysing the data are presented, aiming to enhance comprehension of the software and facilitate future replication and expansion of the research. Finally, the specific authorisations and agreements required for this project are outlined to inform readers of important considerations when handling sensitive data.

5.1 St. Olav's Data

The IDDEAS project utilises parts of BUPdata, presented in section 2.1, collected and stored by St. Olav's University Hospital in Trondheim. The database consists of interdisciplinary patient medical records collected by *Norsk forening for barn og unges psykiske helse* (IDDEAS, n.d.). The entries in the database are composed of therapeutic and diagnostics steps that together make up patient trajectories. From this point, this database is referred to as *St. Olavs data*.

St. Olavs data comprises information on 22 643 patients referred to St. Olavs from 1982 to 2018. Data was collected from 30 938 EoC Bundles involving 41 411 EoCs during this period. The data includes records of 1 840 000 contacts and 222 165 registered diagnoses.

In this project, patients with relevance to hyperkinetic disorders are included. Of the 22 643 patients in the data, 3 856 are diagnosed on Axis 1 from the F90-group. Additionally, considering the patients referred for referral reasons related to behavioural issues or hyperkinetic disorders, the cohort of interest increased to 9 562 patients. The final cohort of interest includes 8 754 patients, as ongoing EoC Bundles when the data was selected for the data set are excluded. These patients were referred to St. Olavs Hospital between 1985 and 2018 and have had a total of 11 128 EoC Bundles, comprising 15 026 EoCs.

5.2 Environment and Tools

The IDDEAS project employs *Helseundersøkelsen i Trøndelag* (HUNT) Cloud to ensure a secure digital environment. HUNT Cloud is a cloud service provided by the HUNT Research Centre at the Norwegian University of Science and Technology (NTNU) to elevate the collection, accessibility, and exploration of large-scale biomedical data. The HUNT cloud environments are specifically designed to analyse and store sensitive data (HUNT Cloud, n.d.-b).

The data stored in HUNT Cloud can be classified into two categories:

- **Sensitive data:** Data that can indirectly identify research participants, such as phenotype or genotype data.
- **Internal data:** Data that can *not* identify research participants, such as figures, summary statistics, computer code, or non-human data.

(HUNT Cloud, n.d.-a)

The IDDEAS lab in HUNT cloud contains sensitive and internal data. Nevertheless, the project's outcomes will be non-confidential through summarised visualisations or cluster descriptions. These presentations will ensure that personal details are not revealed and that sensitive information cannot be reconstructed. Before any potential public dissemination, the lab owner will thoroughly review and approve all results.

The IDDEAS project also benefits from web-based access to analytical tools through the HUNT Workbench, provided by HUNT Cloud. Among the various tools available, Jupyter Notebook and Python will be the most frequently used (HUNT Cloud, n.d.-b). HUNT Workbench also includes Conda, which will be utilised in this experiment as a package and environment manager (Conda, n.d.).

Other relevant tools to be used in the experiment are the following:

- **DBeaver:** A database management tool to be used when exploring and retrieving relevant data from the PostgreSQL database available for this project (DBeaver, 2021).
- **NumPy:** A Python library that provides a multidimensional array object and many mathematical operations to perform (NumPy, n.d.).
- **Pandas:** A useful Python library that provides a data frame object that will be used for data loading, storing, cleaning, and manipulations (Pandas, 2023).
- **Matplotlib:** A Python library to be used to plot and visualise data (Hunter et al., n.d.).
- **Seaborn:** Another Python visualisation library based on matplotlib that will be utilised to create informative statistical graphics (Waskom, 2021).

5.3 Data Approval

Since this Master's Thesis is an NTNU project, it is necessary to ensure the correct NTNU authorisations and approvals before initialising the experiment. The project is written in collaboration with the IDDEAS team. Thus, the authorisation and approvals depend on the overall IDDEAS project's agreements. This section outlines the necessary authorisations and agreements that were required before initiating this project.

As a health research project involving personal data, the IDDEAS project has to be approved by both the *Regional Committees for Medical and Health Research Ethics* (REK) and *Sikt* (previously the *Norwegian Social Science Data Services*) (NTNU, n.d.). Personal data refers to information that can identify a person, either directly or indirectly. On October 9th, 2019, REK confirmed the following: *The project falls outside the scope of the Health Research Act, cf. § 2, and can therefore be carried out without the approval of REK* (IDDEAS, n.d.). In 2020, the IDDEAS project was granted access to the St. Olavs data by the regional health authority at St. Olavs Hospital (IDDEAS, n.d.).

To participate in the IDDEAS project as master students, additional agreements were necessary. These include a non-disclosure agreement to get access to the St. Olavs data and a HUNT Cloud user agreement to be granted access to the digital lab.

Chapter 6

Experiment

This chapter outlines the conducted experiment, detailing the various steps to cluster patient trajectories. It begins by informing the readers of the experimental aims, schedule, and steps. Then the initial data preparation is presented to detail the investigation of available data and the selection process and the cleaning performed. This initial data preparation lays the foundation for conducting an exploratory data analysis, which facilitates understanding the feature distribution within the cohort and identifies relevant relationships that inform subsequent clustering decisions. Finally, this chapter describes the clustering phase, documenting the work conducted throughout three iterations, including intermediary findings and feedback obtained during the process in all three iterations.

6.1 Experiment Plan

Prior to commencing the experiment, a plan was developed to provide a systematic outline of the steps to be completed. This plan was formulated in accordance with the research goal introduced in Section 1.2, taking into account the available resources and time constraints.

6.1.1 Experimental Aims

To ensure a clear direction for the experiment and alignment with the overall project goal and research questions, experimental aims have been defined. As a reminder, the research questions are the following:

> **Research Question 1** How can hyperkinetic patient trajectories in an electronic health record be identified?
>
> **Research Question 2** How can patient trajectory clusters be made meaningful to clinicians?

The defined experimental aims are:

1. Assess the feasibility of clustering for identifying patient trajectory subgroups.
2. Identify subgroups of EoCs that have similar characteristics.
3. Identify subgroups of EoC Bundles that have similar characteristics.
4. Identify similarities in patient characteristics.
5. Identify commonalities based on key characteristics defining the EoCs and EoC Bundles.
6. Identify similarities related to trajectory actions.

These aims will be explored during the experiment and used as a benchmark when evaluating the experiment.

6.1.2 Experiment Schedule

Table 6.1 provides an overview of the project schedule, with an emphasis on the experiment. The schedule encompasses six interim goals, namely, defining the experiment scope and aim, preparing the data, clustering the data, analysing and evaluating the result, and delivering the thesis.

For all six parts presented in Table 6.1, potential challenges and changes may occur. Since it is impossible to foresee all potential obstacles, the schedule may change as the experiment progresses. Nonetheless, the objective is that this plan allows for enough time to complete all phases before the predetermined deadline.

Deadline, week	**Task**
7	Definition of experiment scope and aim
10	Data preparation
15	Clustering
20	Result analysis and evaluation
22	Thesis delivery

Table 6.1: Planned experiment schedule.

6.1.3 Experiment Steps

For this clustering experiment, the following are the steps included:

1. **Data Preparation:** Initial investigation, selection, and cleaning of the available data. The initial preparation focuses on limiting the available data to ensure a sufficient starting point for an initial exploratory analysis and the first iteration of the clustering process.
2. **Exploratory Data Analysis:** Initial investigation of the prepared data to better understand the data used in the clustering.
3. **Clustering Process:** The clustering process consists of three iterations to pursue the research aims effectively. This iterative clustering approach enables the collection of intermediate feedback and facilitates the exploration of various strategies. Each iteration will consist of the following steps:
   (a) **Data Preparation:** In the initial iteration, the data selection will be based solely on initial data investigation, selection, and cleaning. In the two subsequent iterations, the selection will comprise assessing the outcomes from the previous iteration and interim feedback. In this phase, potential changes in the selected features may occur, leading to improved clustering outcomes. This phase also includes scaling the selected features.
   (b) **Clustering:** The k-prototypes algorithm will cluster the prepared data. Together with the algorithm, an initialisation method is required to determine the initial cluster centres, and the ideal number of clusters needs to be identified. Once the centres and the number of clusters are established, k-prototypes will be applied to the data set. In each iteration, clustering using the k-prototypes algorithm will be performed twice - once for EoC level data and once for the EoC Bundle level data.
   (c) **Intermediate Cluster Findings:** The findings from the first two cluster iterations will be examined and visualised at the end of each iteration. This entails visualising the features independently and displaying the distribution of feature values across each cluster. In addition, simplified summaries of the clusters will be presented to provide clinicians with an overview of the findings.
   (d) **Intermediate feedback:** The final phase of each iteration will present the findings to clinicians to determine potential adjustments that may enhance the quality of the findings.

6.2 Data Preparation

The first step in the experiment is data preparation. This step aims to provide high-quality data relevant to clinicians in CAMHS in Norway to ensure an effective clustering outcome. The data preparation involves investigating the data set, selecting, and cleaning the data.

When utilising the available St. Olavs data, understanding its codes is necessary. The St. Olavs data table *Koder* and *NPR Kodeverk* are employed to map the codes into insightful values (Direktoratet for e-helse, 2001). To interpret the codes correctly, the mapping is done in collaboration with a university lecturer at NTNU, Odd-Sverre Westbye, and a psychologist at *BUP poliklinikk Klostergata*, Sanja Prodanovic. The complete mappings are presented in Appendix A. Initial consultation with Westbye also lay the foundation for the initial data selection.

6.2.1 Data Selection

The data selection aims to identify the most informative features related to patient trajectories. The St. Olavs data consists of 49 tables with over 3 million entries. Data selection focuses on achieving a trade-off between the quantity of data required for clustering and the data quality. This involves selecting an appropriate amount of data that ensures sufficient rows for clustering while setting reasonable requirements and limitations to ensure high-quality, informative data (Zhang et al., 2003).

Table Selection

To investigate patient trajectories, some tables from the St. Olavs data are more informative. EoC Bundles form the basis of patient trajectories, and the St. Olavs data table *Sak* contains essential details about these EoC Bundles. One EoC Bundle may include multiple EoCs. Data regarding the individual EoCs are saved in the table *Opphold*. Within one EoC, a patient is potentially given one or more diagnoses and potentially has one or more contacts. The tables *Diagnose* and *Journal* include information regarding the two. Lastly, information regarding the patients in the *Pasient* table is interesting. The selected St. Olavs data tables and the number of entries in these tables are presented in Table 6.2. The tables selected are joined based on patient ID *pasient.nr*, EoC Bundle ID *sak.nr*, and EoC ID *opphold.id*.

Table Name	**Nr. of Enteries**
Pasient	22 643
Sak	30 938
Opphold	41 411
Journal	1 840 045
Diagnose	222 165

Table 6.2: Number of entries in the selected St. Olavs data tables.

Table Entries Selection

Not all EoC Bundles in the St. Olavs data are relevant for the experiment. To extract only relevant data, the two following initial criteria are determined (referring to the meeting with Odd-Sverre Westbye 16.02.2023):

1. Selection based on relation to hyperkinetic disorders
   For this experiment, only patients with relation to hyperkinetic disorders are of interest. To limit the data to EoC Bundles and EoCs regarding such patients, table entries based on the diagnosis given to a patient at the beginning of an EoC Bundle and/or the patient's referral reason are selected. Therefore, the focus is on EoC Bundles where either one or both of the following criteria are met:

   - The patient has a diagnosis on Axis 1 from the F90-group. The specific diagnostic codes in this group are as presented in Section 2.2.1. These are coded in *sak.icd1*.
   - The patient was referred for referral reasons related to behavioural issues or hyperkinetic disorders. Patients' referral reasons are stated in the features *sak.henvgrunnb1*, *sak.henvgrunnb2*, and *sak.henvgrunnb3* with mappings given in *Koder 11*.

2. Selection based on whether an EoC Bundle is closed
   Only closed EoC Bundles should be included in the experiment to compare the EoC Bundles on a similar basis. When an EoC Bundle is closed, the EoC Bundle should have both a closing date and a closing code. However, due to missing data, an EoC Bundle is presumed closed if either of the two dates has a valid value. Data related to closing dates and codes are presented in *sak.avsdato* and *sak.avslkode*.

From these criteria, the initial selection of EoC Bundles to be used when creating the tables for the clustering process is as follows:

```
SELECT
    *
FROM
    sak
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 ='F900'
        OR sak.icd1 ='F901'
        OR sak.icd1 ='F908'
        OR sak.icd1 ='F909'))
    AND NOT (sak.avslkode = 0 AND sak.avsldato IS NULL)
;
```

Feature Selection

When selecting features from the chosen tables, the following criteria need to be met:

- **Feature documentation in *Koder*, *NPR Kodeverk*, or given by CAMHS specialists:** For a feature to be selected, it must be documented in either *Koder* or *NPR Kodebok*, or a specialist has to provide the necessary explanation of the data. This is important to ensure interpretable results.
- **Not much missing/error-prone data:** The St. Olav's data consists of clinical data created by professionals within CAMHS in Norway. The available data is messy, real-world data with missing, error-prone, and outlying values. Columns with many such missing and/or error-prone data should not be selected.

After considering these criteria, many columns are eliminated from further inspection. The remaining features are evaluated and chosen based on the experimental aims. Furthermore, this evaluation was done in collaboration with clinicians to ensure the most informative features.

6.2.2 Result of the Data Selection

After stating the initial data criteria, the remaining entries in the chosen five tables are presented in Table 6.3.

Table Name	**Nr. Entries**
Pasient	8 758
Sak	11 128
Opphold	15 026
Journal	779 776
Diagnose	83 256

Table 6.3: Number of entries in the St. Olavs data after the initial data selection criteria.

The chosen features for further data handling are presented in Table 6.4. One should note that features might change during the experiment depending on the results and feedback.

St. Olavs Data Table	Column Name	Description
Pasient	Id	Patient ID to use when joining tables.
	Fdtnr	The patient's date of birth.
	Kjonn	The patient's gender.
Sak	Id	EoC Bundle ID for joining the tables.
	Opphold	EoC ID for joining the tables.
	Igangdato	An EoC Bundle's start date.
	Icd1	ICD-10 code on Axis1 at the beginning of an EoC Bundle.
	Icd2	ICD-10 code on Axis2 at the beginning of an EoC Bundle.
	Icd3	ICD-10 code on Axis3 at the beginning of an EoC Bundle.
	Icd4	ICD-10 code on Axis4 at the beginning of an EoC Bundle.
	Icd5	ICD-10 code on Axis5 at the beginning of an EoC Bundle.
	Icd6	CGAS score on Axis6 at the beginning of an EoC Bundle.
Opphold	Id	EoC ID for joining the tables.
	Igangdato	An EoC's start date.
	Avsldato	An EoC's end date.
	Omsniva	Care level.
	Ohjelp	Immediacy level.
Journal	Opphold	Journal ID for joining the tables.
	Type1	Contact type.
	Dato1	Date of a contact.
Diagnose	Opphold	EoC ID for joining the tables.
	Diagnose	ICD-10 diagnosis.
	Akse	Number from 1-6 indicating which axis a diagnosis is given.
	Hoved	Boolean, indicating if a diagnosis is the main diagnosis.

Table 6.4: Selected St. Olavs data features.

6.2.3 Data Cleaning and Preprocessing

After the initial data selection, the next step is handling the selected features from the St. Olavs data to obtain a data set suitable for the clustering experiment. For the experiment, it is desired to have two initial feature tables: one for clustering the EoC data and the other for clustering the EoC Bundle data. The selected features from the St. Olavs data will undergo a cleaning and transformation process to obtain the two tables. The data cleaning aims to eliminate low-quality data and ensure that the remaining data corresponds to the documentation given in *Koder*, without excluding more data than necessary. The feature transformation aims to enhance the informativeness and suitability of the features for the clustering experiment. This section explains the transformation from the selected St. Olavs features into the experiment features if not identically mapped.

The data extracted from the St. Olavs data and the resulting features to be used in the experiment are presented in Table 6.5 and Table 6.6. To transform the St. Olavs data into experiment features, a thorough investigation was done using PostgreSQL and Python.

EoC Features	
St. Olavs Data Feature	**Experiment Data Feature**
opphold.igangdato opphold.avsldato journal.dato1	EoC length
opphold.omsnivå	Care level
opphold.ohjelp	Immediacy level
journal.type1	Nr. of contacts
	Nr. of therapy
	Nr. of examination
	Nr. of indirect contact
	Nr. of planning
	Nr. of no-shows
diagnose.diagnose diagnose.akse1	Nr. of unique diagnoses 1
	Nr. of unique diagnoses 2
	Nr. of unique diagnoses 3
	Nr. of unique diagnoses 4
	Nr. of unique diagnoses 5
	Nr. of unique diagnoses 6
diagnose.hoved	Nr. main diagnoses

Table 6.5: Mapping from St. Olavs data features to experiment data features in the EoC table.

EoC Bundle Features	
St.Olavs Data Feature	**Experiment Data Feature**
pasient.fdt sak.igangdato	Age at EoC Bundle start
pasient.kjonn	Gender
sak.igangdato sak.avsldato journal.dato1	EoC Bundle length
sak.icd1	Diagnosis Axis 1
sak.icd2	Diagnosis Axis 2
sak.icd3	Diagnosis Axis 3
sak.icd4	Diagnosis Axis 4
sak.icd5	Diagnosis Axis 5
sak.icd6	Diagnosis Axis 6

Table 6.6: Mapping from St. Olavs data features to experiment data features in the EoC Bundle table.

EoC Length
The St. Olavs data variables *opphold.igangdato*, *opphold.avsldato*, and *journal.dato1* are transformed into the experiment feature *EoC length*. The *EoC length* feature represents the duration of a patient's EoC in days. The St. Olavs data variables *opphold.igangdato* and *opphold.avsldato* are intended to indicate the start and end date of an EoC. However, upon examination of these variables, it was found that the start date was NULL in 2.45% of the EoCs, while the end date was NULL in 7.44% of the EoCs. Additionally, for 4.5% of the EoCs the end date was recorded before the start date. Consequently, 14.17% of the EoCs were affected by one or more errors related to the EoCs' length.

To handle these missing or error-prone EoC dates, the *journal* table is utilised. This table contains information about all patient contacts made during an EoC. Each entry should include a *journal.dato1* field specifying the date of the contact (referring to the meeting with Westbye 16.02.2023). In EoCs where the start or end date is missing, the *journal* table entries can be used to determine the start or end date of the EoC. Specifically, the maximum and minimum values of "journal.dato1" are employed to determine the duration of an EoC if either the start or end date is missing or if the start date occurs after the end date. As a result of this procedure, *EoC length* contains 1.1% NULL values and 0.6% negative lengths. These NULL and negative values are subsequently removed.

Care Level
The St. Olavs data feature *opphold.omsnivå* is transformed into the experiment feature *Care level*. The *Care level* feature indicates the type of care provided, with possible values of "Polyclinic", "Outpatient", or "Inpatient". The *opphold.omsnivå* feature contains integer values mapped to the corresponding categorical values using *NPR Kodeverk 8406*. Specifically, values 1, 2, and 3 are transformed to "Polyclinic", "Outpatient", and "Inpatient". However, 3.55% of the EoCs have *opphold.omsnivå* values that do not have a mapping in the *NPR Kodeverk*, and these values are changed to "Missing data".

Immediacy Level
The St. Olavs data features *opphold.ohjelp* is transformed into the experiment feature *Immediacy level*. *Immediacy level* details the level of urgency of a patient's EoC. This feature can take on one of five values: "Acute", "Non-acute", "6-24 hour wait", "Planned", or "Return from another hospital". Using *Koder 13*, integer values in the range 1-5 are mapped to corresponding categorical values. 0.14% of the *opphold.ohjelp* values are found outside this range and changed to "Missing data".

Contacts
The St. Olavs data feature *journal.type1* is transformed into several features regarding the number of patient contacts. Based on the mapping in *Koder 31* integer values in *journal.type1*, categorise a contact as either "Therapy", "Examinations", "Indirect contact", "Planning", or "No-show". By categorising the contacts, the features *Nr. of therapy*, *Nr. of examinations*, *Nr. of indirect contact*, *Nr. of planning*, and *Nr. of No-Show* are extracted, identifying the total number of a contact type a patient has had within an EoC. 2.35% of *journal* entries contain values outside the range of 1-5 for the *journal.type1* features and are changed to "Missing data".

Diagnoses
The St. Olavs data features *diagnose.akse*, *diagnose.diagnose*, and *diagnose.hoved* are transformed into features regarding the number of diagnoses given, the axes on which they are given, and whether they are the patient's primary diagnosis on one of the axes. The features *Nr. of unique diagnoses 1*, *Nr. of unique diagnoses 2*, *Nr. of unique diagnoses 3*, *Nr. of unique diagnoses 4*, *Nr. of unique diagnoses 5*, and *Nr. of unique diagnoses 6* may be determined by looking at distinct diagnoses on the corresponding axes. The feature *Primary axis diagnosis* may be extracted from the total number of diagnoses where *diagnose.hoved* equals "1". For all these features derived, only rows with a specific diagnosis on the St. Olavs data feature *diagnose.diagnose* are valid. Furthermore, rows in the *diagnose* table where *diagnose.akse* is not between 1-6 are excluded. From these criteria, 0.75% of the diagnoses are excluded from the data set.

Gender
The St. Olavs data feature *pasient.kjonn* is transformed into the experiment feature *Gender* by using *Koder 13*. Integer values "1" and "2" in *pasient.kjonn* are changed to "Male" and "Female", respectively. Other values are mapped to "Missing data", as this should not be possible (referring to mail from Tove Olse Aasbø 10.03.2023).

EoC Bundle Length
The St. Olavs data features *sak.igangdato*, *sak.avsldato*, and *journal.dato1* are used to derive the experiment feature *EoC Bundle length*. The feature *EoC Bundle length* states the length of an EoC Bundle in days. For entries where neither *sak.igangdato* nor *sak.avsldato* are NULL, these are used to derive the EoC Bundle length. For the 9.5% of entries where *sak.avsldato* is NULL, *journal.dato1* is used to derive *EoC Bundle length*. There are no instances in the St. Olavs data where both *sak.avsldato* and *journal.dato1* are NULL.

Age at the Start of an EoC Bundle
The St. Olavs data features *pasient.fdt* and *sak.igangdato* are used to derive the experiment feature *Age at EoC Bundle start* presenting a patient's age at the beginning of an EoC Bundle. If *sak.igangdato* is NULL, *journal.dato1* is used as the EoC Bundle's start date. *pasient.fdt* is never NULL in the St. Olavs data.

Diagnoses on Axes 1-6
The St. Olavs data features *sak.icd1*, *sak.icd2*, *sak.icd3*, *sak.icd4*, *sak.icd5*, and *sak.icd6* are used to derive the six experiment features *Diagnosis Axis 1*, *Diagnosis Axis 2*, *Diagnosis Axis 3*, *Diagnosis Axis 4*, *Diagnosis Axis 5*, and *Diagnosis Axis 6* . These features present the diagnosis on the six axes at the beginning of an EoC Bundle. The St. Olavs data features include error-prone values not documented in *the Directorate of e-health* ICD-10 documentation. Based on feedback from Westbye, these values are changed to valid ICD-10 diagnoses (referring to the meeting with Westbye, 16.02.2023). Furthermore, similar codes are grouped to get a clearer result from the clustering. This grouping is based on the *the Directorate of e-health* documentation (Helsedirektoratet, 2022). Table 6.7 presents the cleaning done on the different ICD-10 codes in the six axes.

Feature Name	Value in the final EoC table	Value in St. Olavs data
Diagnosis Axis 1	Somnolence stupor coma.	R400-R409
	Symptoms associated with cognitive functions.	R410-R419
	Dizziness.	R420-R429
	Disturbances smell and taste.	R430-R439
	Symptoms associated general sensations and perceptions.	R440-R449
	Symptoms associated with emotional state.	R450-R459
	Symptoms associated with looks.	R450-R459
	Contact for examination and investigation.	Z004, Z032, Z133, Z134
	Contact due to potential health risk socio-economic and psychosocial conditions.	Z550-Z560
	Contact for other circumstances.	Z700-Z760
	Contact due to information regarding potential health risk family/personal history.	Z800-Z990
	Organic including symptomatic psychological disorders.	F000-F099
	Mental / behavioral disorders caused by psychoactive substances.	F100-F199
	Schizophrenia / schizotypy / other mental disorders.	F200-F299
	Mood disorders.	F300-F399
	Neurotic or stress related or somatoform disorders.	F400-F499
	Behavioral syndromes associated with physiological disturbances / physical factors.	F500-F599
	Personality and behavioral disorders in adults.	F600-F699
	Intellectual disability.	F840-F849
	Hyperkinetic disorders.	F900-F909
	Other behavioural/emotional disorders usually occurring in children and adolescents.	F910-F989
	Missing information.	999
		1999
	No diagnose.	000
		1000
	None.	NULL
		Other values
Diagnosis Axis 2	Speech and language.	F80, F800, F801, F802, F803, F808, F809
	Learning disabilities.	F81, F810, F811, F812, F813, F818, F819,
	Motor skills.	F82
	Specific skills.	F83
	Other.	F82
	Unspecified.	F82
	Missing information.	999, 2999
	No diagnose.	2000, 000
	None.	NULL, other

Continues on next page

Feature Name	Value in the final EoC table	Value in St. Olavs data
Diagnosis Axis 3	Very high intelligence.	1
	High intelligence.	2
	Normal intelligence.	3
	Slightly below average intelligence.	4
	Slight intellectual disability.	5 , F7
	Moderate intellectual disability.	6
	Severe intellectual disability.	7
	Profound intellectual disability.	8
	Unspecified intelligence level.	9
	Unknown intelligence level.	999
	Missing information.	39
		99
	No diagnose.	30
	None.	NULL
Diagnosis Axis 4	Certain infectious diseases and parasitic diseases.	A-chapter, B-chapter.
	Tumors.	C-chapter, D000-D489
	Diseases of the blood and blood-forming organs and certain conditions affecting the immune system.	D500-D999
	Endocrine diseases nutritional diseases and metabolic disorders.	E-chapter
	Diseases of the nervous system.	G-chapter
	Diseases of the eye or ear.	H-chapter
	Diseases of the circulatory system.	I-chapter
	Diseases of the respiratory system.	J-chapter
	Diseases of the digestive system.	K-chapter
	Diseases of the skin and subcutaneous tissue.	L-chapter
	Diseases of the musculoskeletal system and connective tissue.	M-chapter
	Diseases of the urinary and genital organs.	N-chapter
	Pregnancy birth and maternity.	O-chapter
	Certain conditions occurring in the perinatal period.	P-chapter
	Congenital malformations deformities chromosomal abnormalities.	Q-chapter
	Symptoms / signs / abnormal clinical/laboratory findings not elsewhere classified.	R-chapter
	Factors impacting health status and contact with the health service.	Z-chapter
	Injuries / poisonings / other consequences of external causes.	S-chapter, T-chapter
	External causes of diseases/injuries/deaths.	V-chapter, W-chapter, X-chapter, Y-chapter
	None.	F-chapter, other

Continues on next page

Feature Name	Value in the final EoC table	Value in St. Olavs data
Diagnosis Axis 5	Missing information.	0.0, 99.0, 599.0
	No diagnose.	000, 500.0
	Deviant relationships.	1.0, 1.1, 1.2, 1.3, 1.4, 1.8
	Mental illness/Deviations/Disability.	2.0, 2.1, 2.2, 2.8
	Inadequate/disturbed communication.	3.0
	Deviant aspects of upbringing.	4.0, 4.1, 4.2, 4.3, 4.8
	Deviant environment.	5.0, 5.1, 5.2, 5.3, 5.8
	Emergent life changes.	6.0, 6.1, 6.2, 6.3, 6.4, 6.5, 6.8
	Social strain factors.	7.0, 7.1, 7.8
	Chronic interpersonal strain at school/work.	8.0, 8.1, 8.2, 8.8
	Straining events/conditions resulting from child disorder/condition.	9.0, 9.1, 9.2, 9.8
Diagnosis Axis 6	Excellent function.	10.0
	Good function.	9.0
	Slight disturbance.	8.0
	Difficulties in single area.	7.0
	Varied function.	6.0
	Moderate function social areas / severe disturbance one area.	5.0
	Severely impaired several areas.	4.0
	Unable to function almost all areas.	3.0
	Considerable supervision and care.	2.0
	Constant supervision.	1.0
	None.	0.0, Other

Table 6.7: Cleaning of the ICD-10 codes and CGAS scores on the six axes.

6.3 Exploratory Data Analysis

During the experimental stage, a preliminary examination of the data is conducted before proceeding with the clustering process. This initial investigation involves producing a visual representation of the data so that its structure, the chosen features, and their distribution can be understood easier. By exploring the data, complex relationships between items, trends, and anomalies might be identified, which in turn can potentially lead to more informed decisions during the subsequent clustering process. Additionally, more knowledge of the selected features may imply more informative clustering results (Chen et al., 2008).

Figure 6.1 provides an overview of the patient distribution based on their gender and age at the start of their EoC Bundle. The graph highlights a greater proportion of males compared to females, with a rough ratio of 70:30. Moreover, it is evident that the female patients consist of a significant number of older individuals compared to the male patients, who tend to be generally younger.

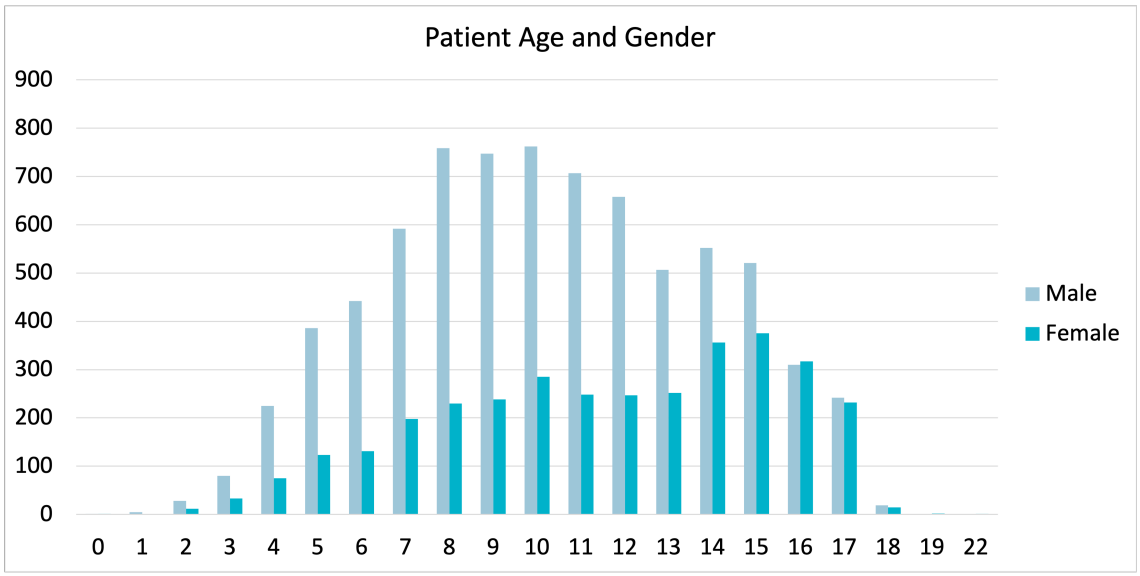


Figure 6.1: The cohort distribution of age and gender.

The patients in the cohort of interest all have an EoC Bundle related to hyperkinetic disorders. Each EoC Bundle comprises at least one EoC. These EoCs can last from a few days to multiple years. To visualise the distribution of the durations, the different EoC lengths are grouped as shown in figure 6.2.

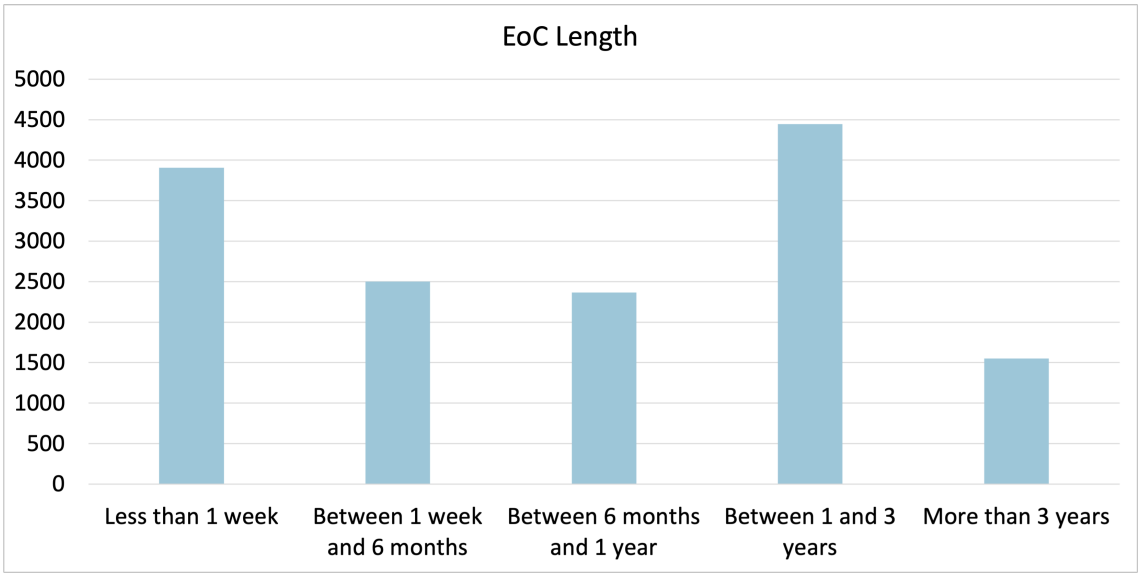


Figure 6.2: The cohort distribution of the EoC lengths.

Figure 6.3 presents the value distribution of the EoCs' care levels. As illustrated, most EoCs in the cohort are polyclinic.

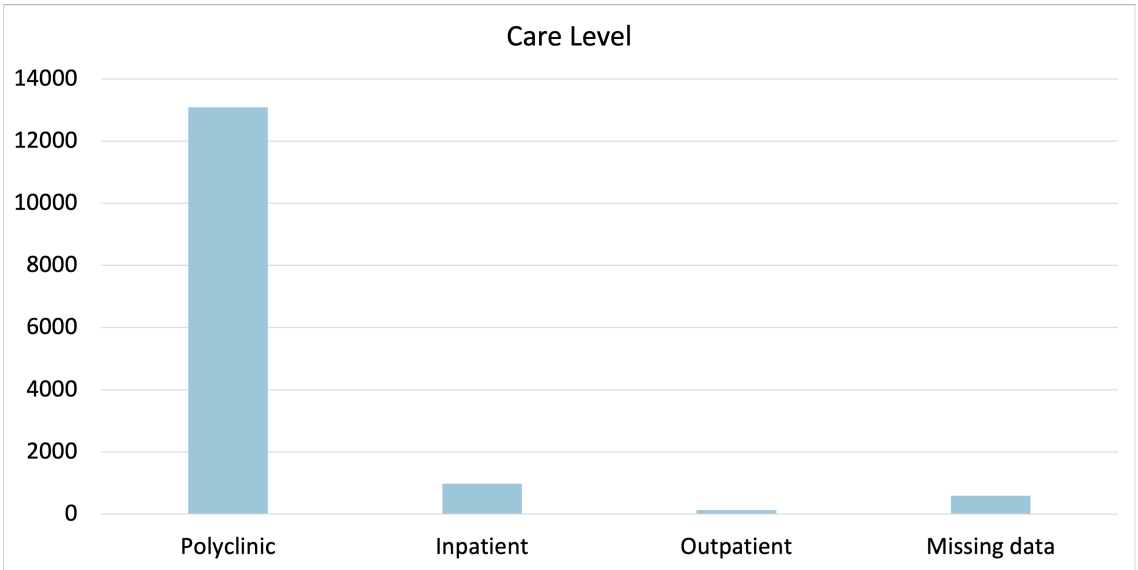


Figure 6.3: The cohort distribution of EoCs' care levels.

Most EoCs in the cohort of interest are planned, followed by acute EoCs. This finding aligns with the *Norwegian Directorate of Health's* guidelines for patient trajectories related to hyperkinetic disorders, presented in Section 2.1. The complete value distribution of immediacy levels is presented in Figure 6.4.

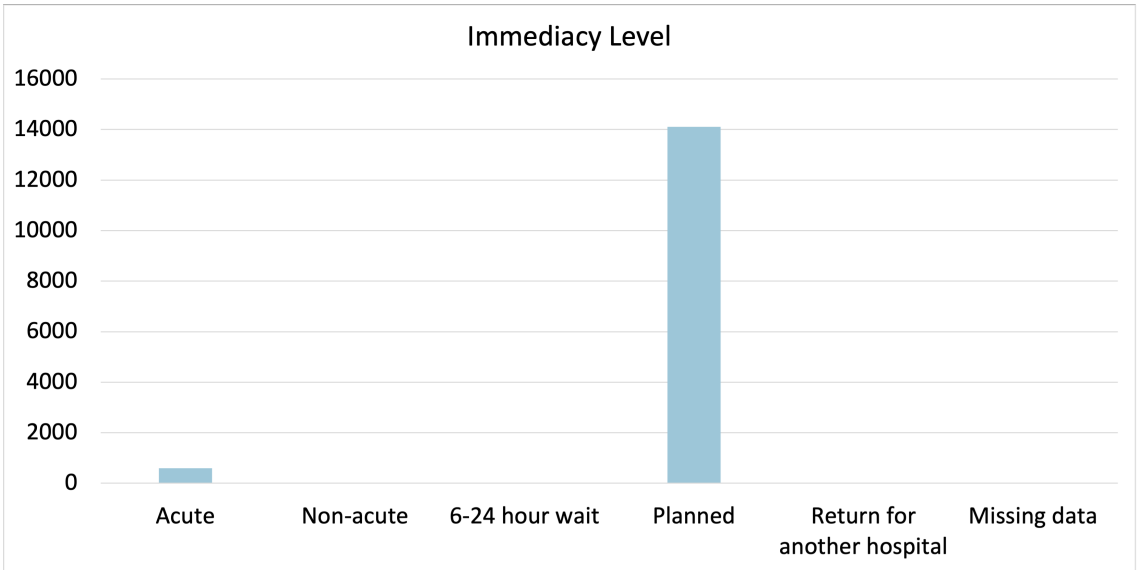


Figure 6.4: The cohort distribution of the EoCs' immediacy levels.

Figure 6.5 presents the total amounts of different contact types recorded for the cohort of interest.

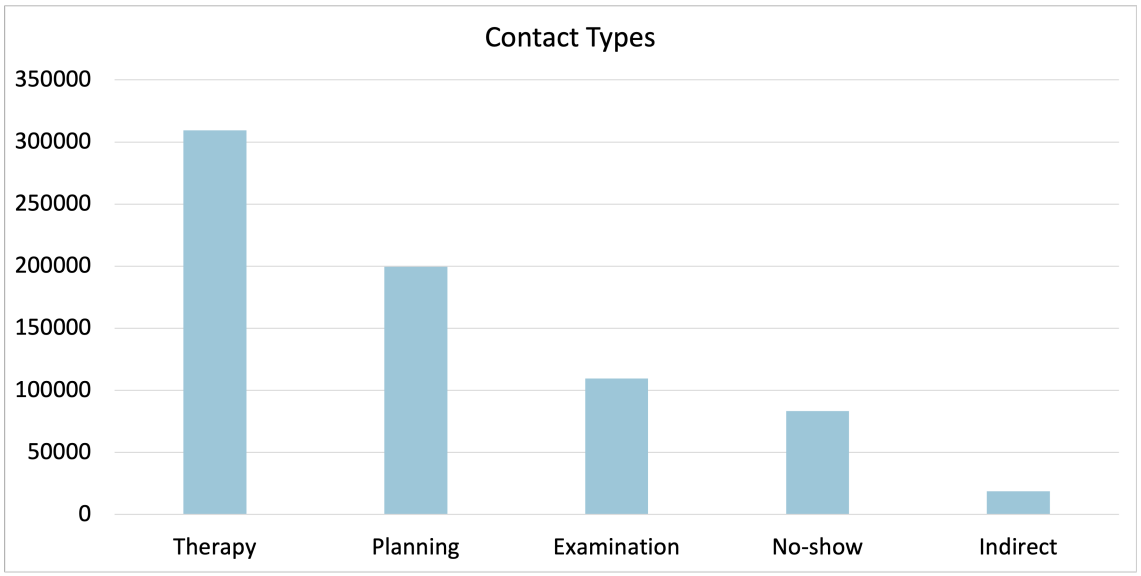


Figure 6.5: The cohort distribution of contact types.

Figure 6.6 illustrates the total count of diagnoses assigned within a patient's EoC, across all six axes, along with the count of diagnoses designated as the primary diagnosis on one of the axes. These two counts are presented together to demonstrate the relationship between all diagnoses and those given as primary diagnoses.

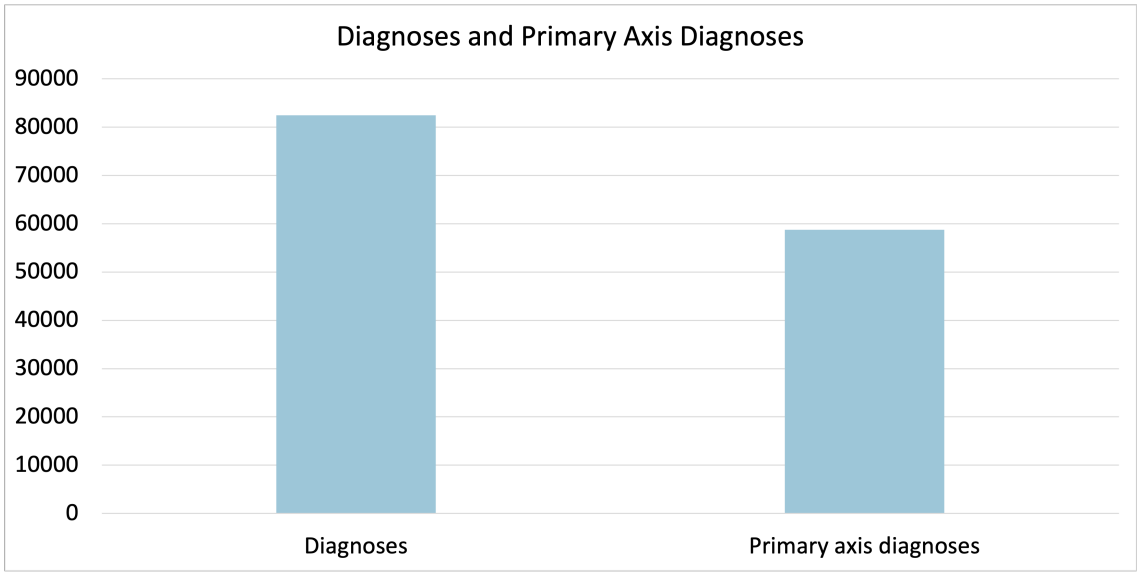


Figure 6.6: The cohort distribution of the number of diagnoses given and how many were given as a primary diagnosis.

Figure 6.7 overviews the unique number of diagnoses recorded on each of the six axes within an EoC. It offers insights into the diversity of diagnoses registered across these axes.

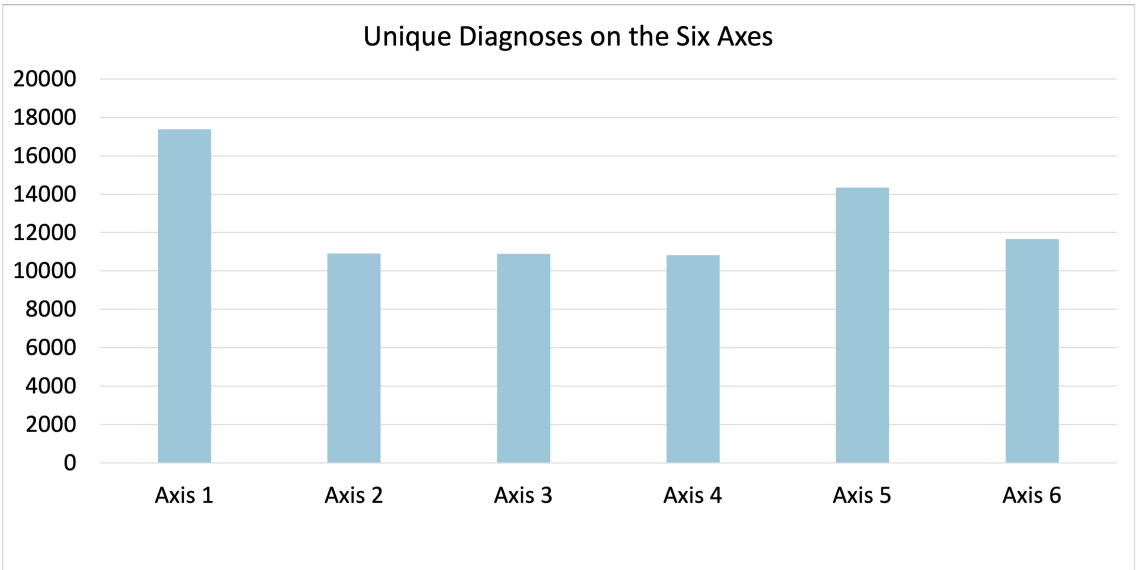


Figure 6.7: The cohort distribution of the diagnoses given on the different axes.

Moving on to visualising the features designated on the EoC Bundle level, the distribution of EoC Bundle lengths is visualised in Figure 6.8. The different lengths are again grouped to clarify the visualisation.

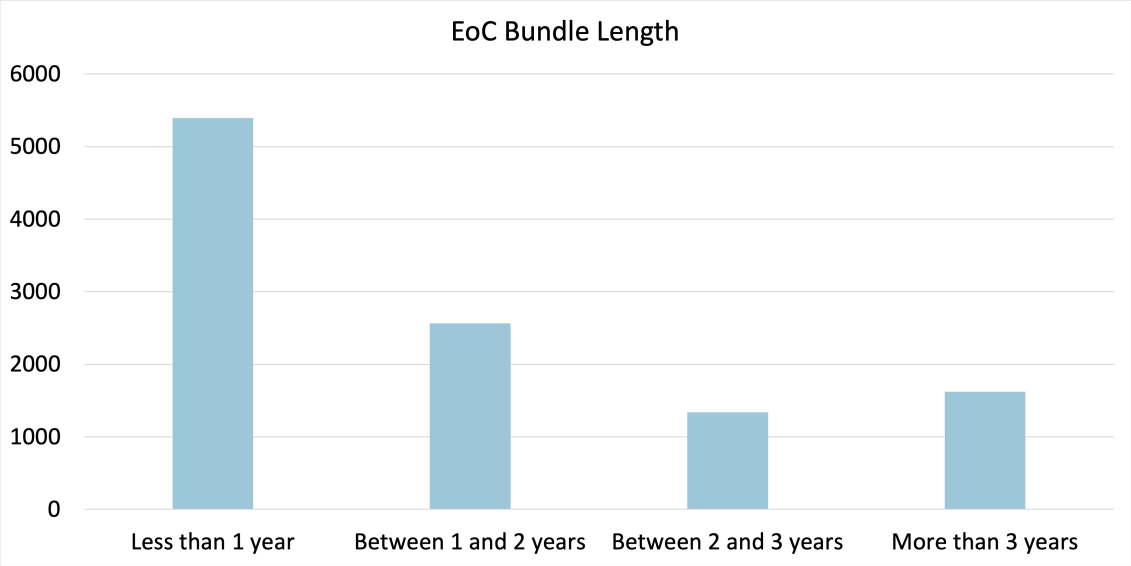


Figure 6.8: The cohort distribution of the different EoC Bundle lengths.

Figure 6.9 depicts the diagnoses given to the cohort on Axis 1 at the beginning of an EoC Bundle. The graph highlights that the most prevalent diagnosis on Axis 1 on the EoC Bundle level is hyperkinetic disorders. This observation aligns with the project's focus on patients related to hyperkinetic disorders. It is worth noting the difference between "Missing information", indicating EoC Bundles where clinicians lacked sufficient data to make a diagnosis, and "Missing data", indicating the absence of recorded diagnoses on Axis 1. The substantial amount of missing data raises concerns about the reliability and quality of the data, as it has the potential to introduce bias and obscure potential patterns within the data. The category "Other diagnoses" is created for this exploratory data analysis to collectively visualise diagnoses given in less than 100 EoC Bundles.

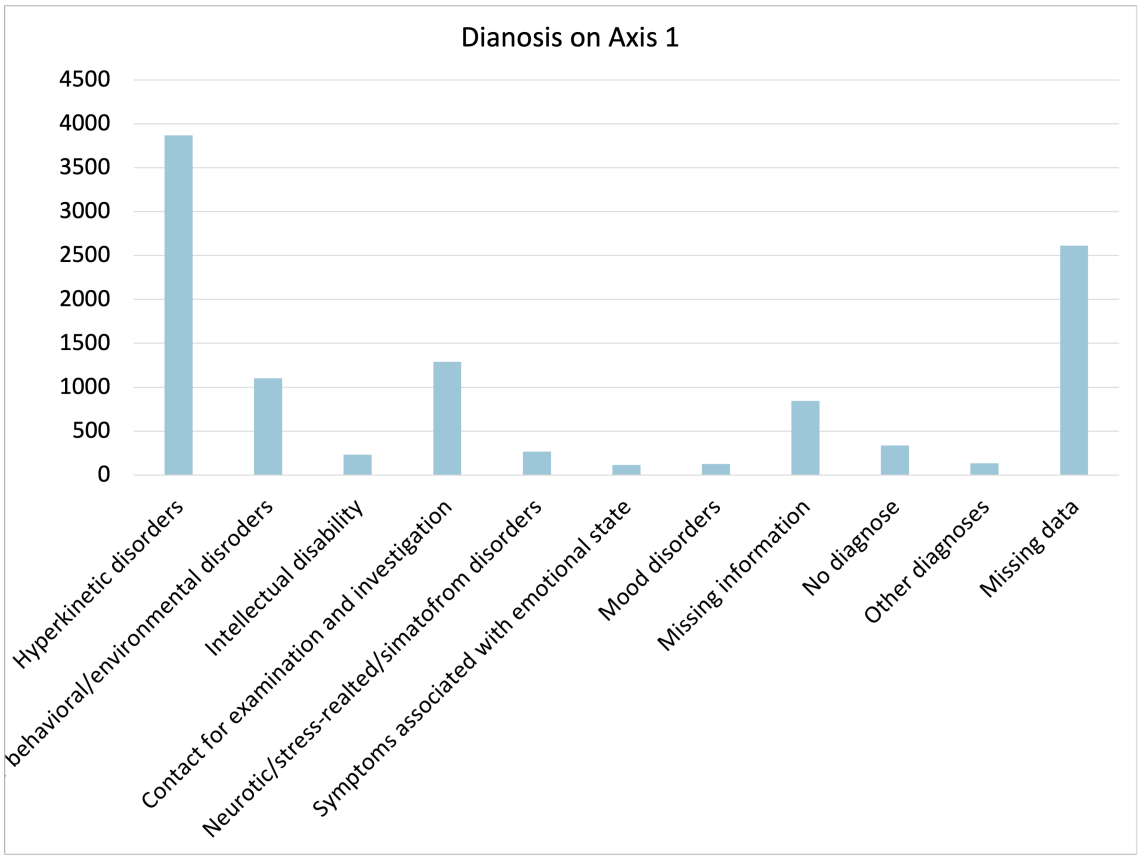


Figure 6.9: The cohort distribution of diagnoses given on Axis 1 at the beginning of an EoC Bundle.

Figure 6.10 displays the distribution of the number of EoCs within an EoC Bundle. It provides insights into how many EoCs are typically included within each EoC Bundle.

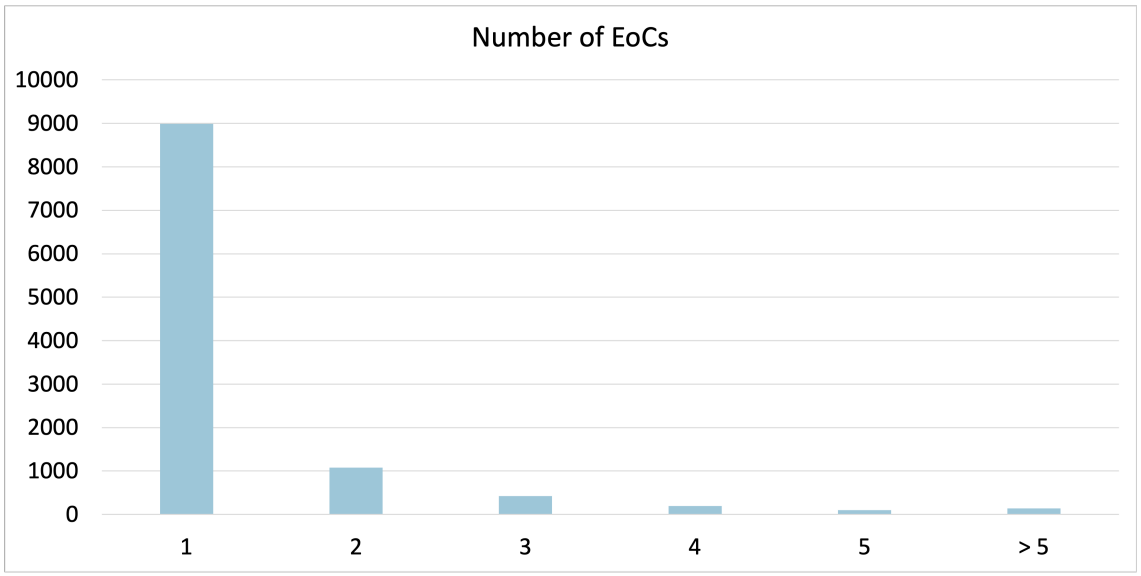


Figure 6.10: The cohort distribution of the number of EoCs within an EoC Bundle.

To assess the correlation between the features used in the first clustering iteration, the heatmap in Figure 6.11 is employed. The heatmap provides valuable insights, revealing a strong correlation between contact types and between the features related to the number of diagnoses registered on each axis. Specifically, the correlation among the features related to diagnoses suggests that EoCs with more diagnoses registered on one axis are also likely to have more diagnoses registered on the other axes.

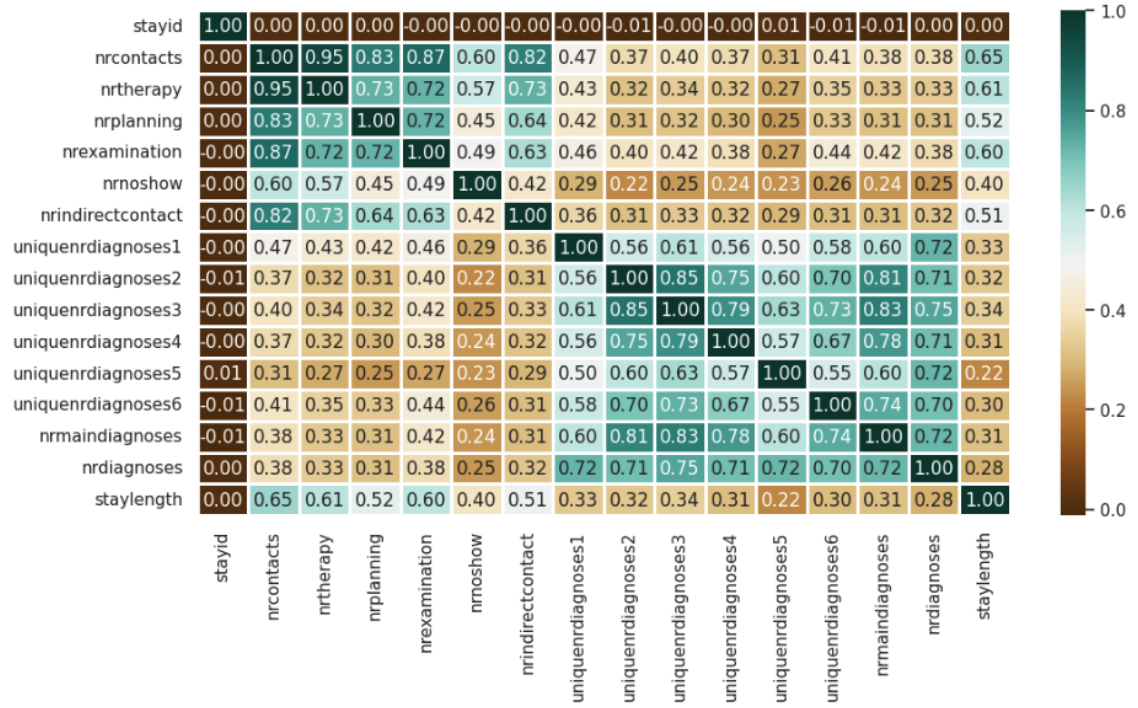


Figure 6.11: The correlation between the different features to be used in the first clustering iteration.

6.4 Clustering Process

The next step in the experiment is the clustering process. To ensure the quality and relevance of the clustering results, it is performed in three iterations, with continuous feedback from clinicians and technological evaluations provided at each stage. By incorporating feedback and evaluation, modifications can continuously improve the findings. Each iteration consists of data preparation, clustering, examination of intermediate cluster findings, and collection of intermediate feedback.

To effectively analyse the complex and heterogeneous data, a stepwise clustering approach is employed. First, the EoC level data is clustered into homogenous groups. Subsequently, the resulting clusters are used to label the identified EoC subgroups before clustering at the EoC Bundle level, which includes the categorised EoCs. This stepwise clustering methodology enables the potential identification of subgroups representing higher-level patient treatment pathways. Following the stepwise clustering, the resulting clusters are examined and visualised. Evaluation is then conducted from both a technological perspective and by clinicians who utilise their specific knowledge to assess the relevance and quality of the results.

Figure 6.12 provides an overview of the clustering process and the following sub-experiments to be performed three times. Appendix B.1.3 presents the code written for the third iteration to demonstrate the work performed in this experiment.

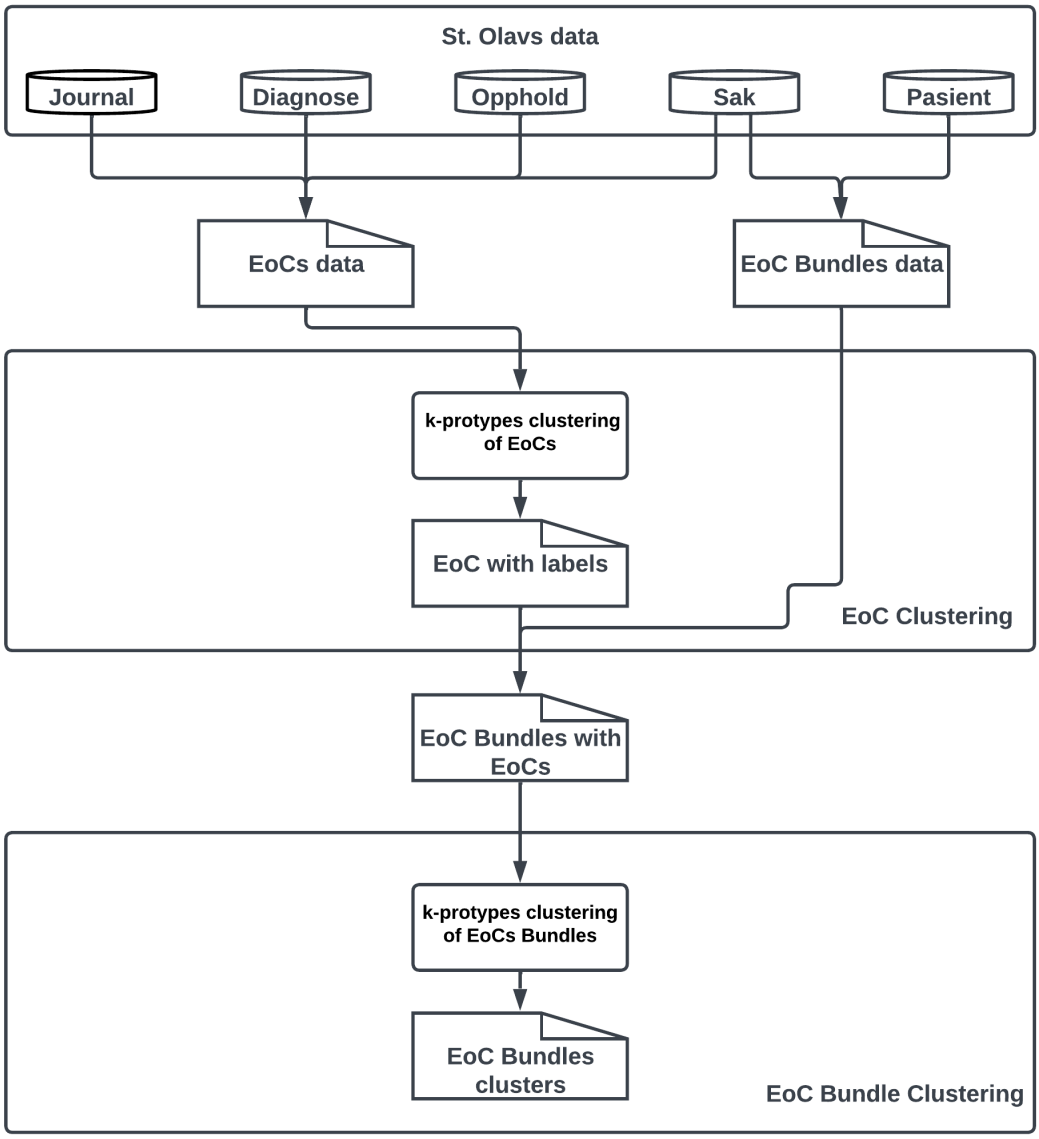


Figure 6.12: Description of the clustering process.

6.4.1 First Clustering Iteration

The aim of the first clustering iteration is to perform the first stepwise clustering based on obtained domain knowledge, data documentation, and initial consultations with clinicians. The findings obtained in this iteration are evaluated from both a technological and clinical perspective and lay the foundation for the second iteration.

Data Preparation

The features selected in the first iteration of the EoC clustering and the EoC Bundle clustering are detailed in table 6.8 and 6.9, respectively. The tables also include a short description of the features as a reminder of what each entails.

EoC Table	
Feature	**Description**
EoC length	Length of the EoC in days.
Care level	The values in this field may be: - Missing data - Day - 24-hour - Polyclinic
Immediacy level	The values in this field may be: - Missing data - Acute - Non-acute - 6-34 hour wait - Planned - Return from another hospital
Nr. of contacts	The total number of contacts a patient had during an EoC.
Nr. of therapy	Number of contacts, of the different types, a patient had during an EoC.
Nr. of examination	
Nr. of indirect contacts	
Nr. of planning	
Nr. of no-shows	
Nr. of diagnoses	The total number of diagnoses given during an EoC.
Nr. of primary axis diagnoses	The number of diagnoses given as the primary diagnosis.
Nr. of unique diagnoses 1	The number of unique diagnoses on the six axes during an EoC.
Nr. of unique diagnoses 2	
Nr. of unique diagnoses 3	
Nr. of unique diagnoses 4	
Nr. of unique diagnoses 5	
Nr. of unique diagnoses 6	

Table 6.8: Description of the first iteration's EoC features.

EoC Bundle Table	
Feature	**Description**
Age at EoC Bundle start	The patient's age at the beginning of the EoC Bundle.
Gender	The patient's gender, being of the following three: - Missing data - Female - Male
EoC Bundle length	Length of the EoC Bundle in days.
Diagnosis Axis 1	ICD-10 diagnoses declared at the beginning of an EoC Bundle, grouped based on *the Directorate of e-health* documentation as presented in Table 6.7.
Diagnosis Axis 2	
Diagnosis Axis 3	
Diagnosis Axis 4	
Diagnosis Axis 5	
Diagnosis Axis 6	CGAS score declared at the beginning of an EoC Bundle, ranging from 1-10. The numbers correspond to CGAS integrals. For instance, the value "1" corresponds to the integral 1-10.

Table 6.9: Description of the first iteration's EoC Bundle features.

The next step before clustering the EoCs and EoC Bundles is to scale the chosen features. Feature scaling is a crucial data-mining preprocessing step when clustering numerical data. Scaling the data through standardisation is necessary to control the data set's variability, as features with large sizes or great variability can strongly impact the clustering result (Johor Bahru et al., 2013). The technique *Power Transform* standardises this iteration's numerical features. Power Transform is a Python library provided by *Scikit-learn* that can be applied to data to featurewise map the data from any distribution to a more *Gaussian-like* distribution (Pedregosa et al., 2011).

Clustering

The first step of clustering is choosing an initialisation method to determine the initial cluster centres. This step is crucial as it directly impacts the final clustering outcome. The chosen initialisation method for this project is *Cao*. The Cao method chooses the initial prototypes by considering the density of each data point and the distance between them. By evaluating both density and distance, this method ensures that outliers are not chosen as the new cluster centres and that multiple cluster centres are positioned in the surrounding of one centre (Cao et al., 2009).

The next step in clustering is choosing a method to find an optimal number of clusters (k). Choosing an optimal number of clusters is a major challenge in cluster analysis as the effectivity of the clustering depends on if a reasonable k can be estimated (Ankerst et al., 1999). The Elbow method is used to find an optimal number of clusters. The Elbow method looks at the percentage of within-cluster dispersion as a function of the number of clusters (Tibshirani et al., 2001).

With the Cao initialisation and Elbow methods chosen, they are initially employed to determine the initial number of clusters for the EoC clustering. Using the Cao method with different values of k ranging from 1-10, the elbow plot in Figure 6.13 was created. This plot illustrates how increasing the number of clusters k contributes to separating the selected EoC features into meaningful clusters. The optimal value of k is determined by identifying the point of maximum curvature on the elbow plot, which is done using the Python repository called *Kneed* (Satopää et al., 2011). In Figure 6.13, the maximum curvature on the elbow plot occurs when k equals 3.

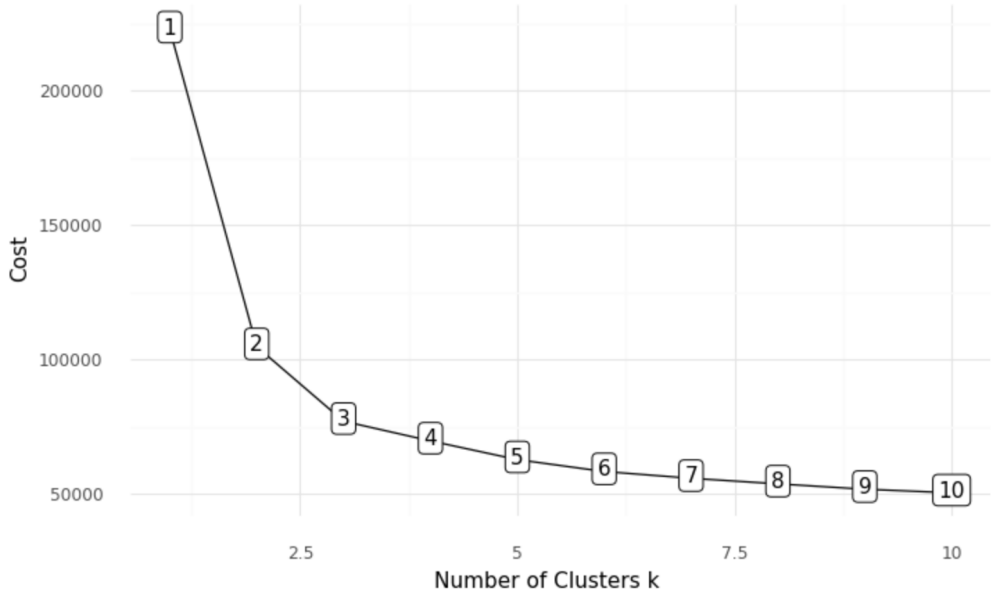


Figure 6.13: First iteration's elbow method for the EoCs.

Using the k-prototypes algorithm, the EoC data is then grouped into three clusters named *EoC Type 0*, *EoC Type 1*, and *EoC Type 2*. These clusters include 4 045, 4 574 and 6 594 EoCs.

SHapley Additive exPlanations (SHAP) is then used to highlight the impact and the importance of the different EoC features on the clustering model's decision-making process (Lundberg et al., 2017). The SHAP values illustrated in Figure 6.14 is computed by a classifier model that is fitted using the first iteration's EoC clustering findings as labels.

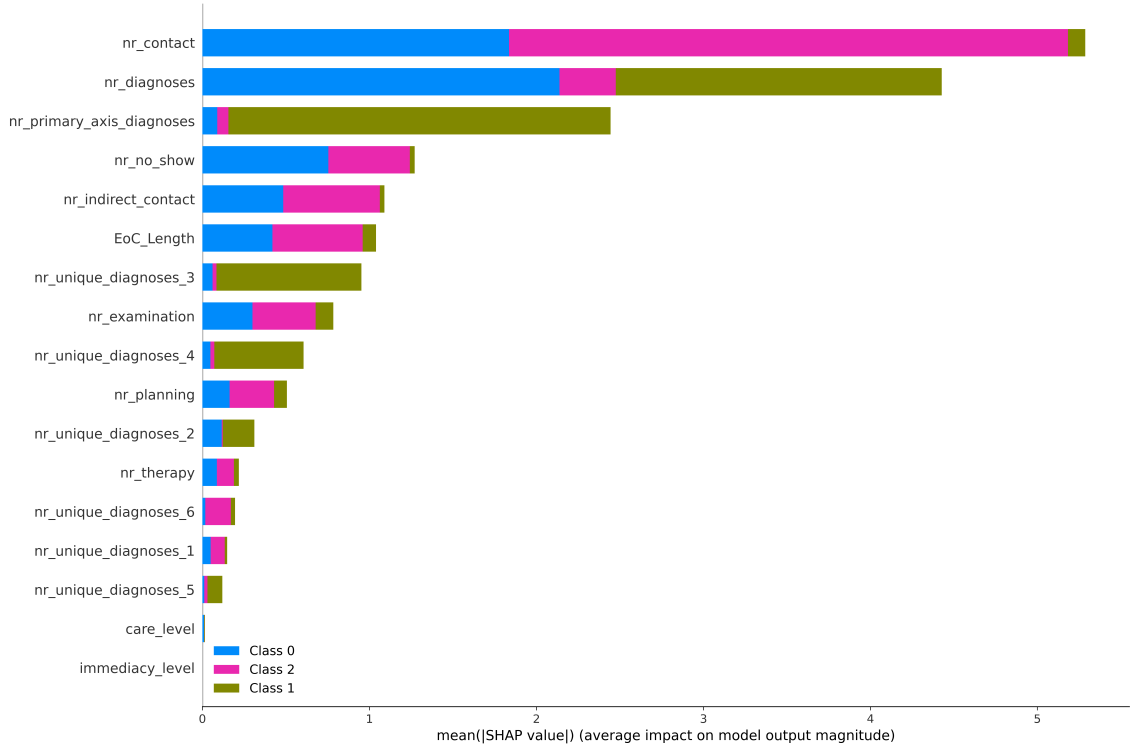


Figure 6.14: SHAP plot of the first iteration's EoC features.

The subsequent step in the clustering process involves incorporating the identified groups of EoC clusters to perform the EoC Bundle clustering. Instead of considering the entire EoC data set as an entity in the EoC Bundle clustering, the EoC data is incorporated using the three identified EoC types. To integrate the EoC types, the EoC and EoC Bundle data are merged based on their unique EoC IDs. This allows for counting the EoCs of each type within each EoC Bundle, providing valuable insights into the distribution and composition of EoCs within the EoC Bundles.

With the same initialisation method as in the EoC clustering, the optimal number of clusters for the EoC Bundle clustering is again identified using the Elbow method. The maximum curvature detected in the elbow plot visualised in Figure 6.15 is 4.

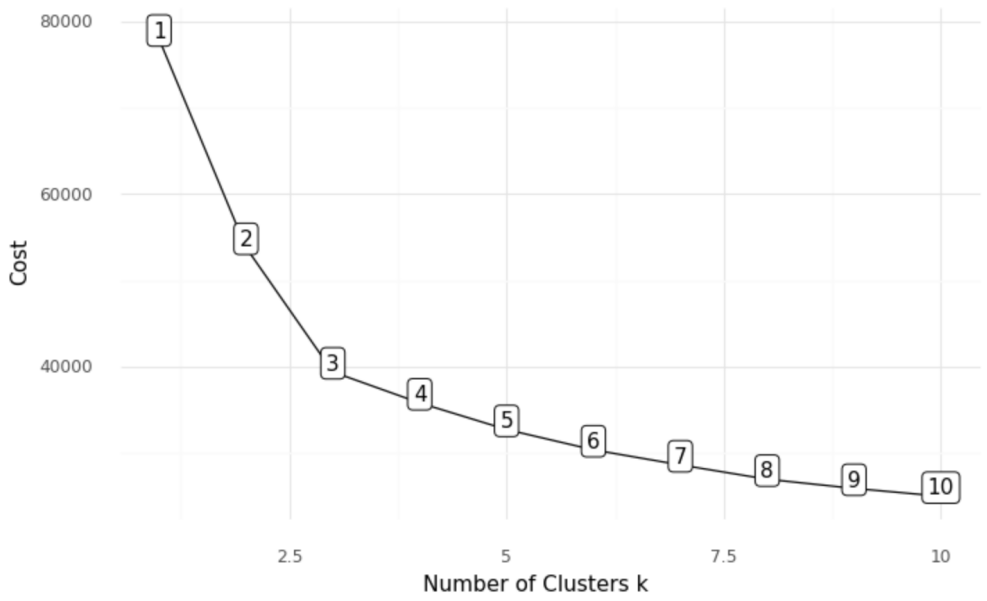


Figure 6.15: First iteration's elbow method for the EoC Bundles.

Applying the k-prototypes algorithm to the EoC Bundle data, it is grouped into three clusters: *EoC Bundle Type 0*, *EoC Bundle Type 1*, and *EoC Bundle Type 2*. These clusters comprise 5 714, 2 406 and 2 925 EoC Bundles.

A SHAP plot explaining the impact and the importance of the EoC Bundle features can be visualised in Figure 6.16.

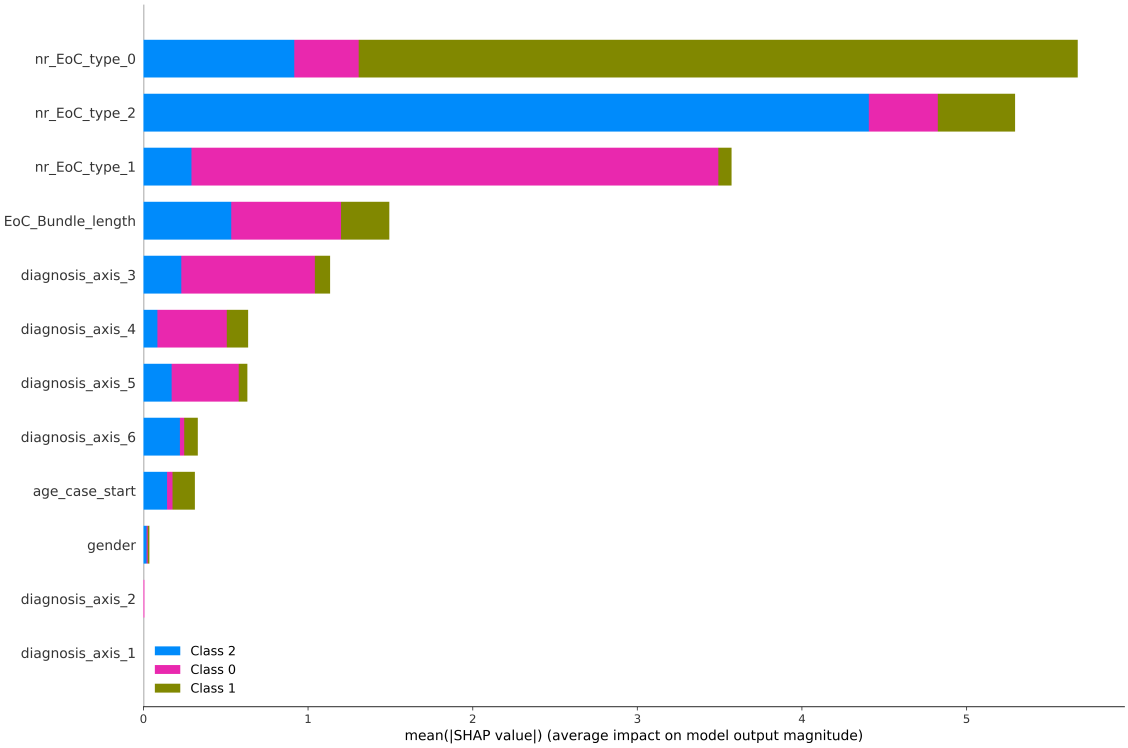


Figure 6.16: SHAP plot of the first iteration's EoC Bundle features.

Intermediate Cluster Findings

After completing the first iteration's clustering, the next step involves investigating the findings. The findings are visualised comprehensively to facilitate clinical interpretation, as visualisation is an effective communication tool (Chen et al., 2008). The following findings visualisations represent the ones presented to clinicians to discuss the findings. For the complete representation of all visualisations made for the first iteration, refer to Appendix C.1

These visualisations describe the data points distributed in the three EoC clusters and the three EoC Bundle clusters. To do so, the distribution of values of each feature included in both clustering processes is presented, first the EoC features and then the EoC Bundle features. To better illustrate the differences and similarities in the value distributions in each cluster, the percentage of the values is given, not the total. A summary of the value distribution of each EoC and EoC Bundle clusters is also presented.

EoC Cluster Findings

From the first iteration's clustering on the EoC level, the clusters identified and their size are presented in Table 6.10. Figures 6.17, 6.18, and 6.19 present the distribution of the different EoC lengths and care and immediacy level. To give an example of how to use the visualisation, Figure 6.17 shows that of the EoCs within the EoC Type 0 cluster, almost 30% are shorter than a week long.

Clusters	Nr. of Data Points
EoC Type 0	4 045
EoC Type 1	4 574
EoC Type 2	6 594

Table 6.10: First iteration's distribution of EoCs in the EoC clusters.

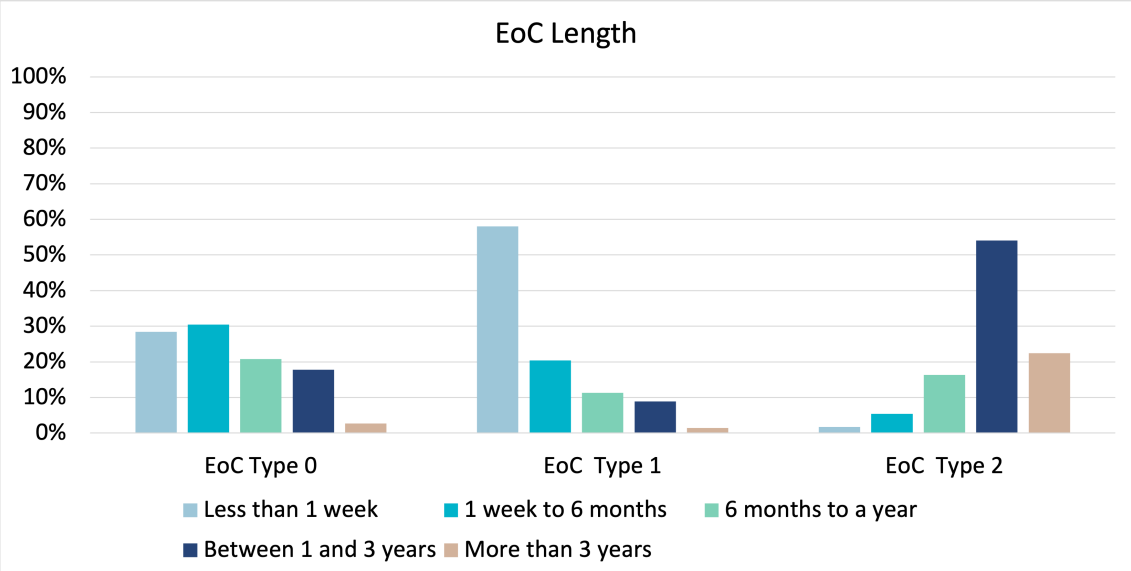


Figure 6.17: First iteration's distribution of EoC lengths.

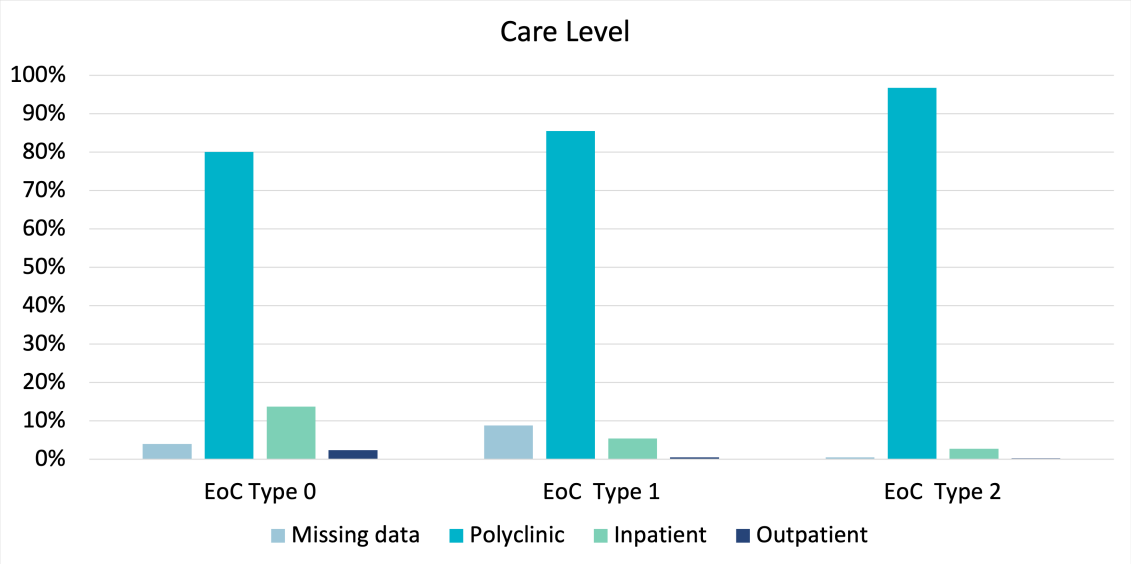


Figure 6.18: First iteration's distribution of care levels.

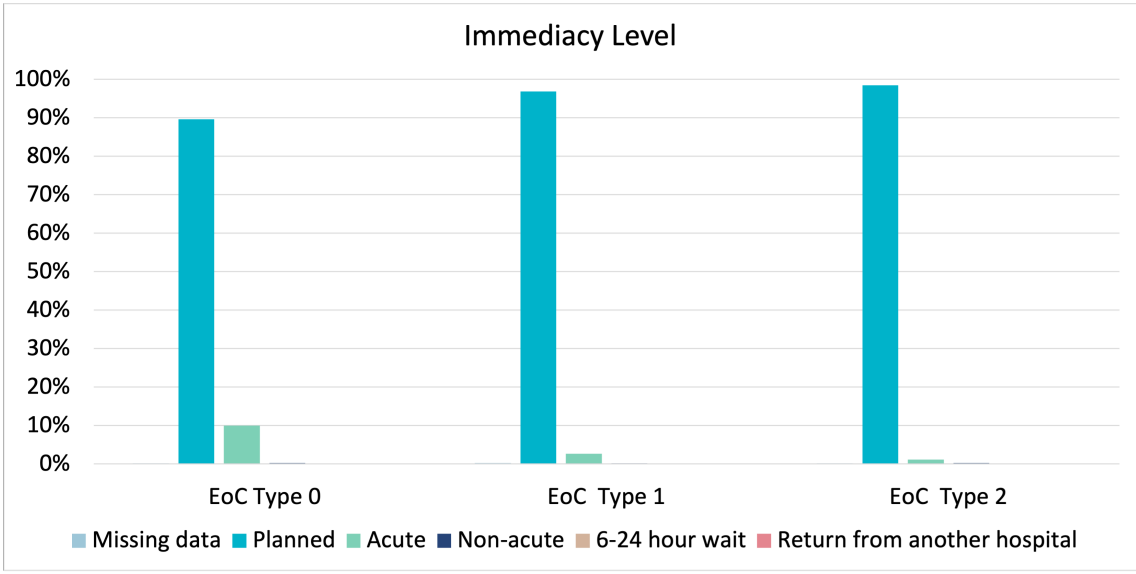


Figure 6.19: First iteration's distribution of immediacy levels.

Figure 6.20 shows the distribution of the total number of contacts within an EoC in each cluster, considering all contact types. Table 6.11 provides a more detailed breakdown of the number of contacts for each contact type, highlighting the dominant number of contacts within each cluster in bold text.

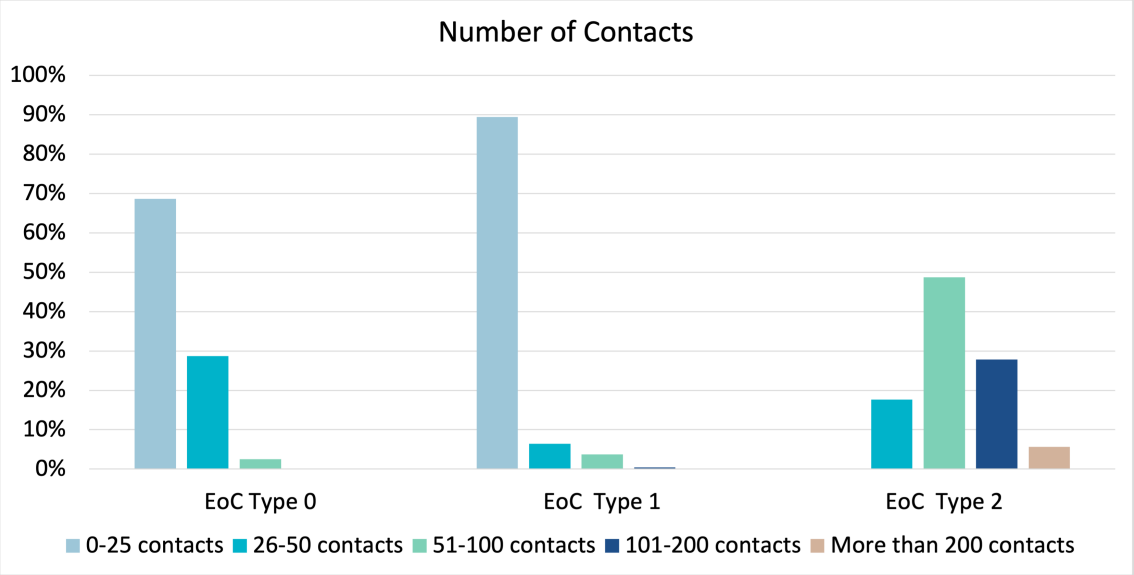


Figure 6.20: First iteration's distribution of the total number of contacts.

Contact Feature	Values	EoC Type 0	EoC Type 1	EoC Type 2
Nr. of therapy contacts	None	10%	13%	0%
	1-5 contacts	**38%**	**67%**	1%
	6-10 contacts	25%	9%	4%
	11-20 contacts	19%	6%	21%
	More than 20 contacts	8%	5%	**74%**
Nr. of planning contacts	None	9%	13%	0%
	1-5 contacts	**53%**	**74%**	2%
	6-10 contacts	22%	8%	9%
	11-20 contacts	14%	4%	29%
	More than 20 contacts	1%	0%	**60%**
Nr. of examination contacts	None	25%	**52%**	1%
	1-5 contacts	**57%**	39%	20%
	6-10 contacts	14%	6%	26%
	11-20 contacts	4%	2%	**32%**
	More than 20 contacts	0%	1%	21%
Nr. of no-show contacts	None	22%	17	1%
	1-5 contacts	**70%**	**77%**	25%
	6-10 contacts	7%	4%	**37%**
	More than 10 contacts	1%	2%	**37%**
Nr. of indirect contacts	None	**82%**	**90%**	27%
	1-5 contacts	17%	9%	**59%**
	6-10 contacts	1%	1%	11%
	11-20 contacts	0%	0%	**3%**
	More than 20 contacts	0%	1%	0%

Table 6.11: First iteration's distribution of the contact types.

During an EoC, a patient may be given several diagnoses on all six axes. These diagnoses may or may not be given as the patient's primary diagnosis on a specific axis. The distribution of the number of diagnoses in total is presented in Figure 6.21. In contrast, Figure 6.22 presents how many of these diagnoses are given as the primary diagnosis on an axis. The detailed information regarding the distribution of the number of unique diagnoses on all six axes is presented in Table 6.12.

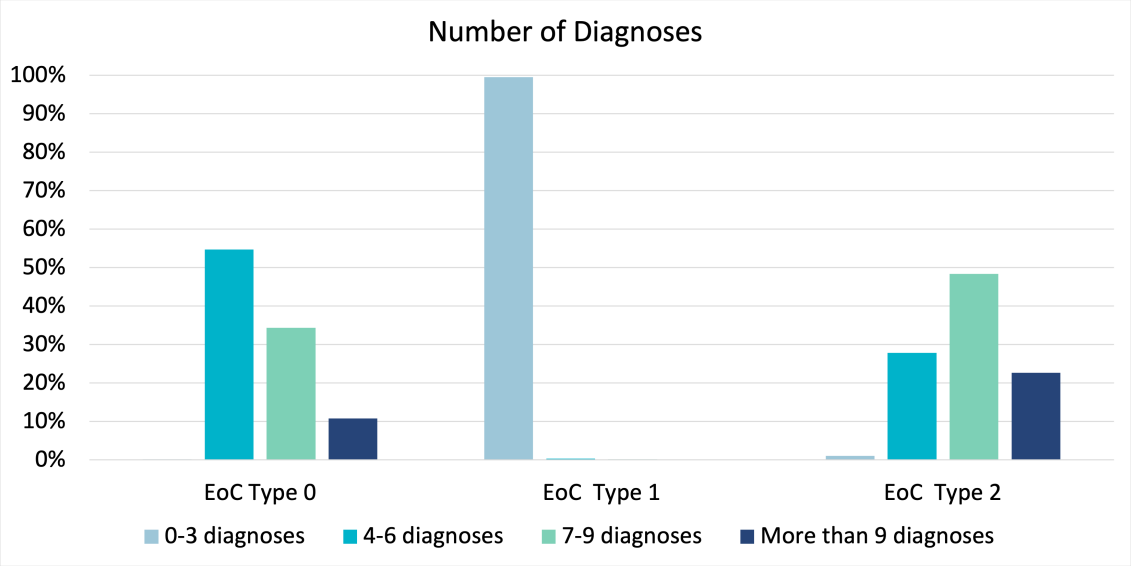


Figure 6.21: First iteration's distribution of the total number of diagnoses given.

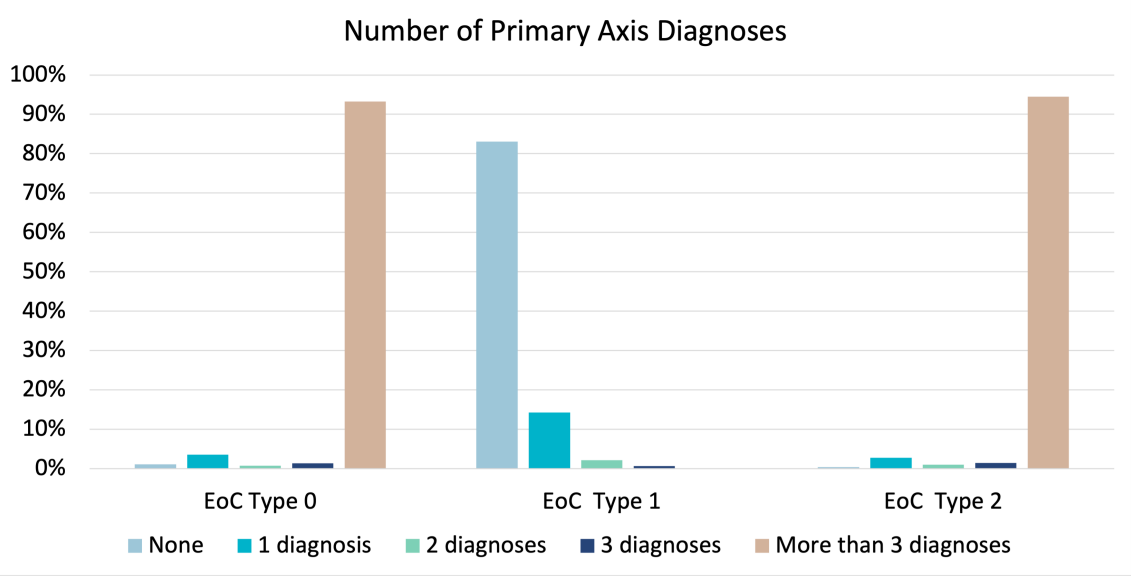


Figure 6.22: First iteration's distribution of the number of diagnoses given as a primary diagnosis on one of the six axes.

Diagnostic Feature	**Values**	**EoC Type 0**	**EoC Type 1**	**EoC Type 2**
Nr. of unique diagnoses on Axis 1	None	0%	**83%**	0%
	1 diagnosis	**72%**	15%	**45%**
	2 diangoses	20%	2%	39%
	3 diagnoses	6%	0%	12%
	More than 3 diagnoses	2%	0%	4%
Nr. of unique diagnoses on Axis 2	None	2%	**99%**	2%
	1 diagnosis	**95%**	1%	**87%**
	2 diangoses	3%	0%	10%
	3 diagnoses	0%	0%	1%
	More than 3 diagnoses	2%	0%	0%
Nr. of unique diagnoses on Axis 3	None	1%	**99%**	2%
	1 diagnosis	**96%**	1%	**88%**
	2 diangoses	3%	0%	10%
	3 diagnoses	0%	0%	0%
	More than 3 diagnoses	0%	0%	0%
Nr. of unique diagnoses on Axis 4	None	5%	**100%**	5%
	1 diagnosis	**90%**	0%	**84%**
	2 diangoses	5%	0%	10%
	3 diagnoses	0%	0%	1%
	More than 3 diagnoses	0%	0%	0%
Nr. of unique diagnoses on Axis 5	None	2%	**99%**	2%
	1 diagnosis	**76%**	1%	**68%**
	2 diangoses	12%	0%	19%
	3 diagnoses	7%	0%	7%
	More than 3 diagnoses	3%	0%	4%
Nr. of unique diagnoses on Axis 6	None	7%	**98%**	3%
	1 diagnosis	**85%**	2%	**75%**
	2 diangoses	8%	0%	20%
	3 diagnoses	0%	0%	2%
	More than 3 diagnoses	0%	0%	0%

Table 6.12: First iteration's distribution of the number of diagnoses on the different axes.

When presenting these value distributions, the numerical features' means and medians and the categorical features' modes may be interesting as a comparison. Table 6.13 was presented to clinicians to provide a more detailed description for further comparing and discussing the cluster distributions.

Feature	Meassure	EoC Type 0	EoC Type 1	EoC Type 2
EoC lenght	Mean	243	130	789
	Median	96	2	637
Care level	Mode	Polyclinic	Polyclinic	Polyclinic
Immediacy level	Mode	Planned	Planned	Planned
Nr. of contacts	Mean	19	12	96
	Median	17	6	32
Nr. of therapy contacts	Mean	8	5	41
	Median	6	2	32
Nr. of planning contacts	Mean	6	3	26
	Median	4	1	24
Nr. of examination contacts	Mean	3	2	14
	Median	2	0	11
Nr. of no-show contacts	Mean	2	2	11
	Median	2	1	8
Nr. of indirect contacts	Mean	0	0	3
	Median	0	0	2
Nr. of diagnoses	Mean	7.5	0.3	8.2
	Median	6	0	7
Nr. of primary axis diagnoses	Mean	5.6	0.2	5.7
	Median	6.0	0.0	6.0
Nr. of unique diagnoses on Axis 1	Mean	1.4	0.2	1.75
	Median	1.0	0.0	2.0
Nr. of unique diagnoses on Axis 2	Mean	1.0	0.0	1.1
	Median	1.0	0.0	1.0
Nr. of unique diagnoses on Axis 3	Mean	1.0	0.0	1.1
	Median	1.0	0.0	1.0
Nr. of unique diagnoses on Axis 4	Mean	1.0	0.0	1.1
	Median	1.0	0.0	1.0
Nr. of unique diagnoses on Axis 5	Mean	1.3	0.0	1.4
	Median	1.0	0.0	1.0
Nr. of unique diagnoses on Axis 6	Mean	1.0	0.0	1.2
	Median	1.0	0.0	1.0

Table 6.13: First iteration's EoC feature measurements.

The EoC cluster distributions may be summarised to provide an overview of each EoC type. Table 6.14 summarises the most prominent feature value distributions across the three EoC clusters. This does not give a complete picture of the clusters but can be instructive when further presenting the EoC Bundle clusters, which include the distribution of each EoC type.

EoC Type 0 4 045 EoCs	**EoC Type 1** 4 574 EoCs	**EoC Type 2** 6 594 EoCs
• Medium length. • Polyclinic. • Planned. • 25 contacts. • 1-5 contacts of each type. • 4-9 diagnoses given with more than 3 being the primary axis diagnosis. • The number of unique diagnoses is equally distributed.	• Short. • Polyclinic. • Planned. • Less than 25 contacts. • 1-5 contacts of each type. • Less than three diagnoses given and no primary axis diagnosis.	• Longer. • Polyclinic. • Planned. • 50-100 contacts. • More than ten examinations, planning and therapy contacts. • 7-9 diagnoses given with more than 3 being the primary axis diagnosis. • The number of unique diagnoses is equally distributed.

Table 6.14: First iteration's EoC clusters summary.

EoC Bundle Cluster Findings

The first iteration's clustering on the EoC Bundle level revealed three different clusters as presented in Table 6.15. The EoC Bundle features' value distribution is presented as bar charts with value percentages. Figure 6.23 presents the distribution of the EoC Bundle lengths, while Figure 6.24 and 6.24 present the distribution of the age and gender of the patients.

EoC Bundle Cluster	**Nr. Data Points**
EoC Bundle Type 0	5 714
EoC Bundle Type 1	2 406
EoC Bundle Type 2	2 925

Table 6.15: First iteration's distribution of EoCs Bundles in the EoC Bundle clusters.

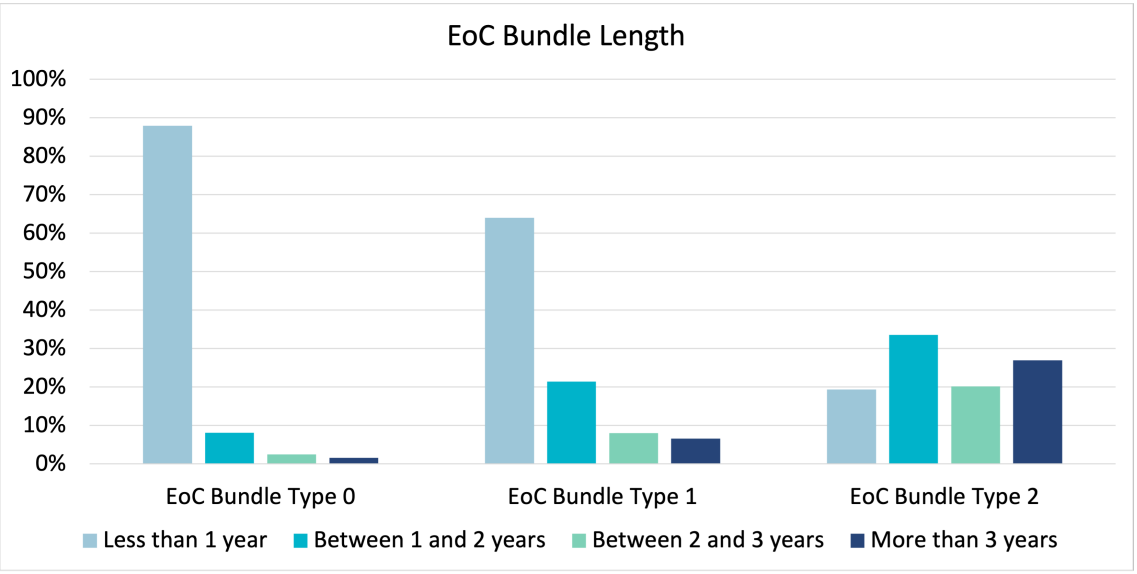


Figure 6.23: First iteration's distribution of EoC Bundle lengths.

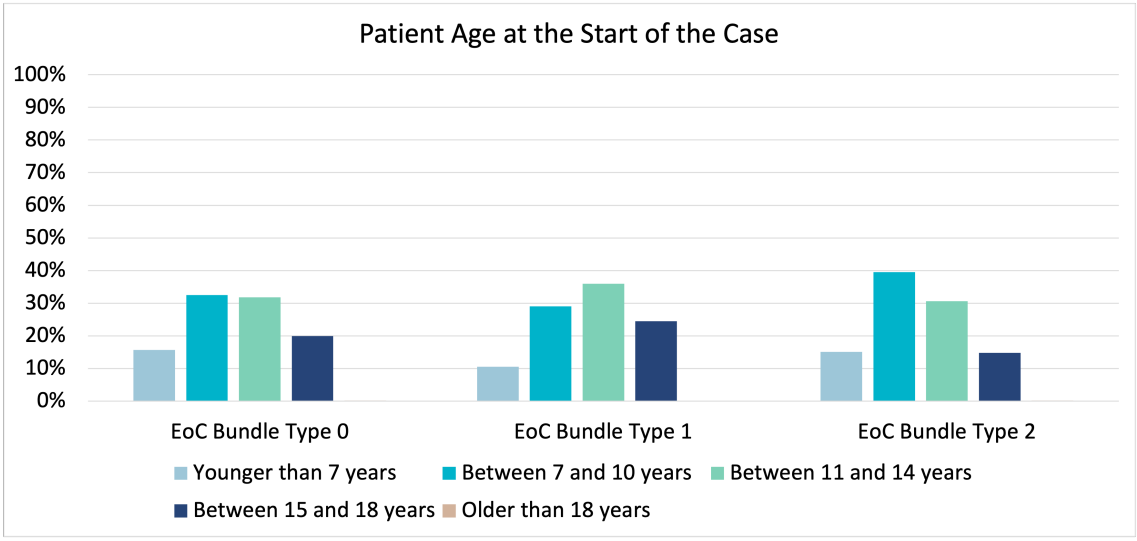


Figure 6.24: First iteration's distribution of patient's age.

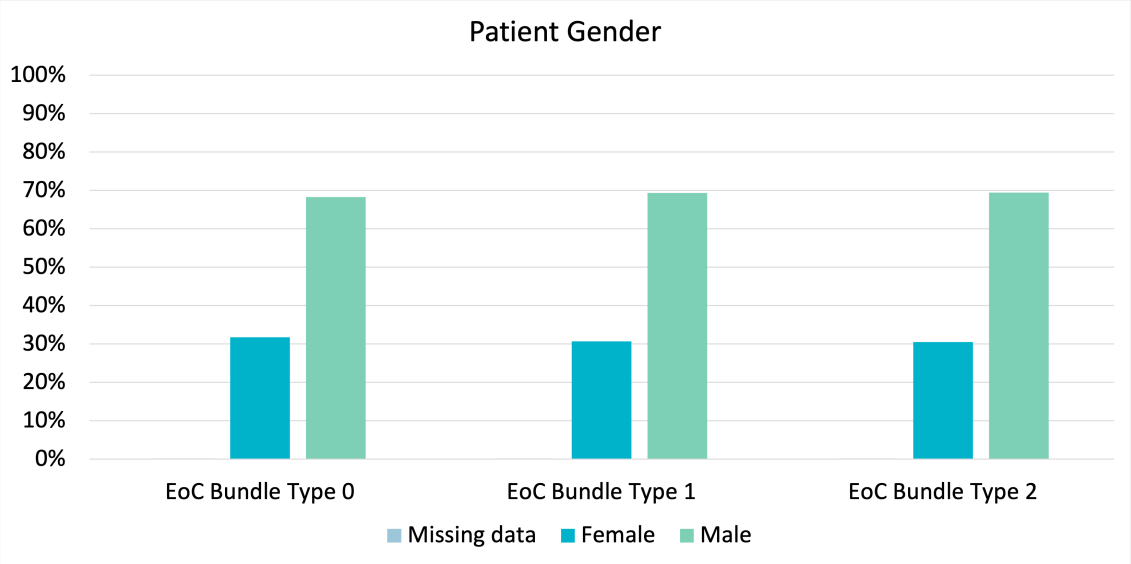


Figure 6.25: First iteration's distribution of patient's gender.

Figures 6.26, 6.27, 6.28, 6.29, 6.30, and 6.31 illustrate the distribution of the diagnoses given at the beginning of an EoC Bundle on each of the six axes. Only the most common diagnostic codes on the six axes are included in the figures to enhance interpretability.

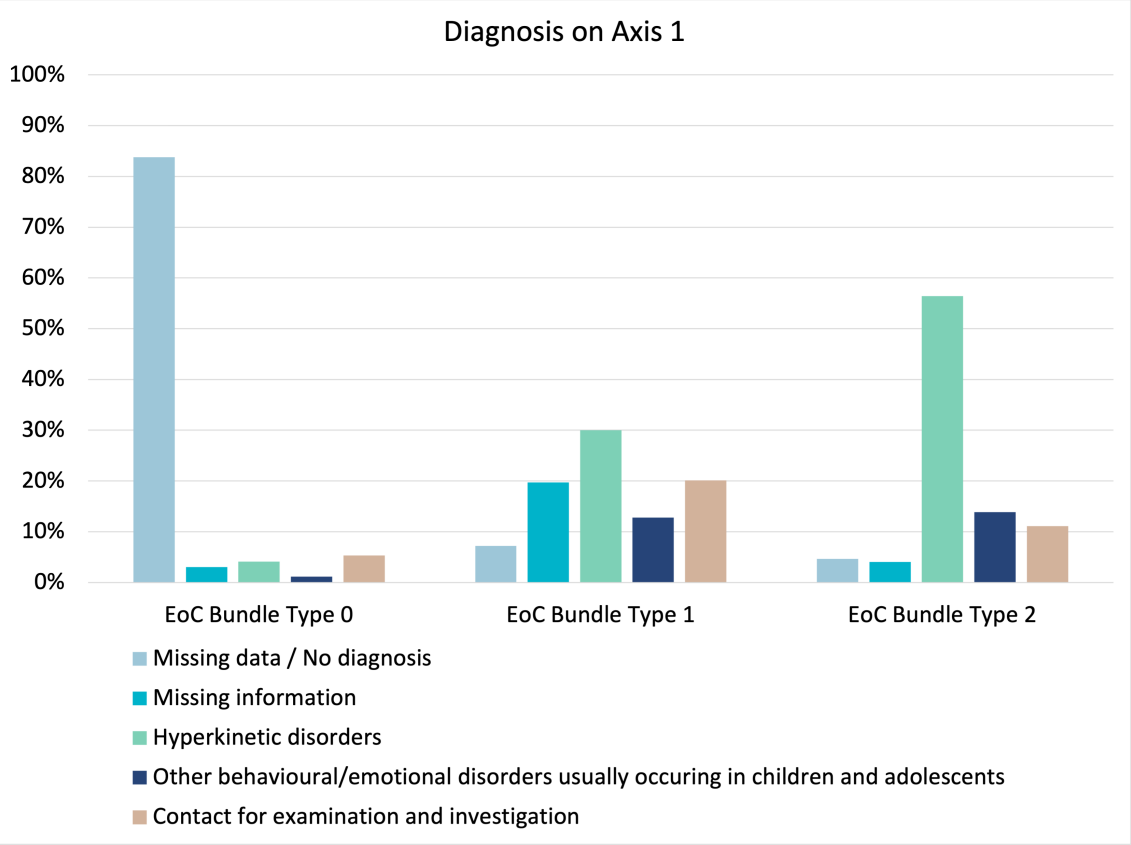


Figure 6.26: First iteration's distribution of diagnoses on Axis 1 at the beginning of an EoC Bundle.

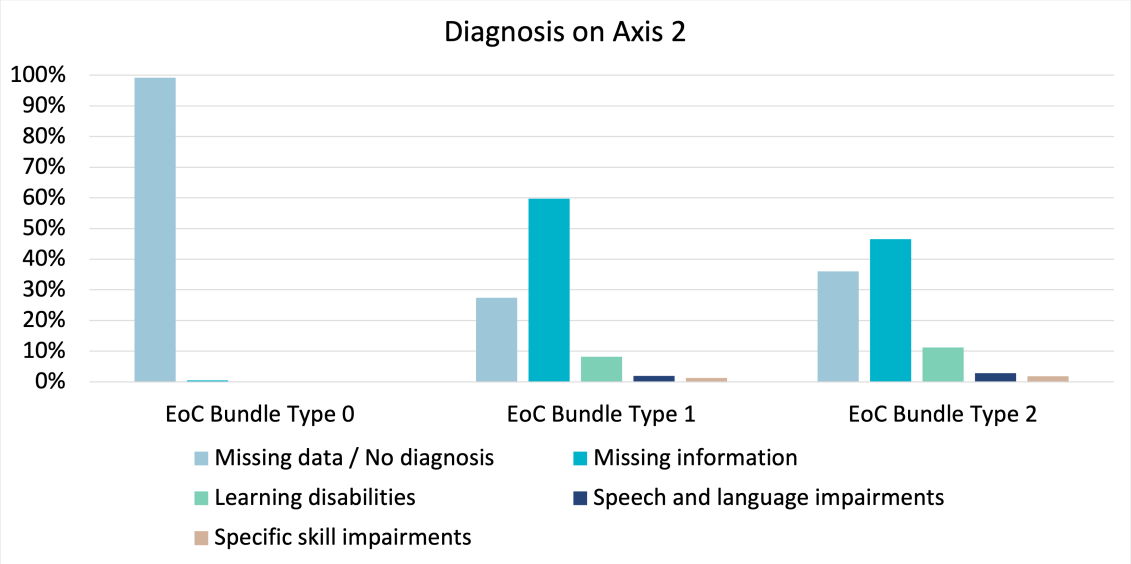


Figure 6.27: First iteration's distribution of diagnoses on Axis 2 at the beginning of an EoC Bundle.

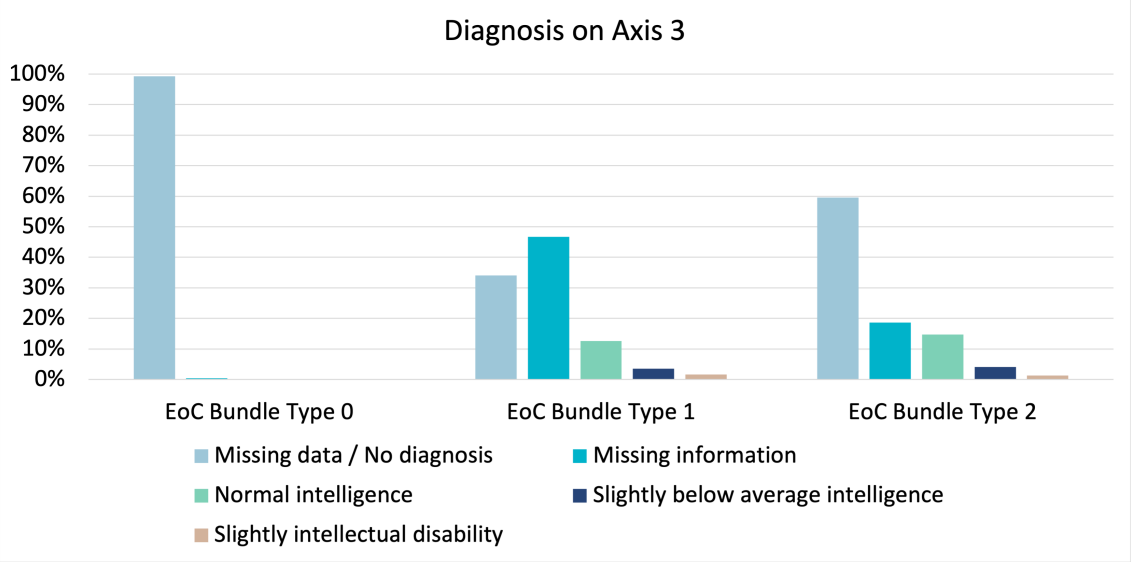


Figure 6.28: First iteration's distribution of diagnoses on Axis 3 at the beginning of an EoC Bundle.

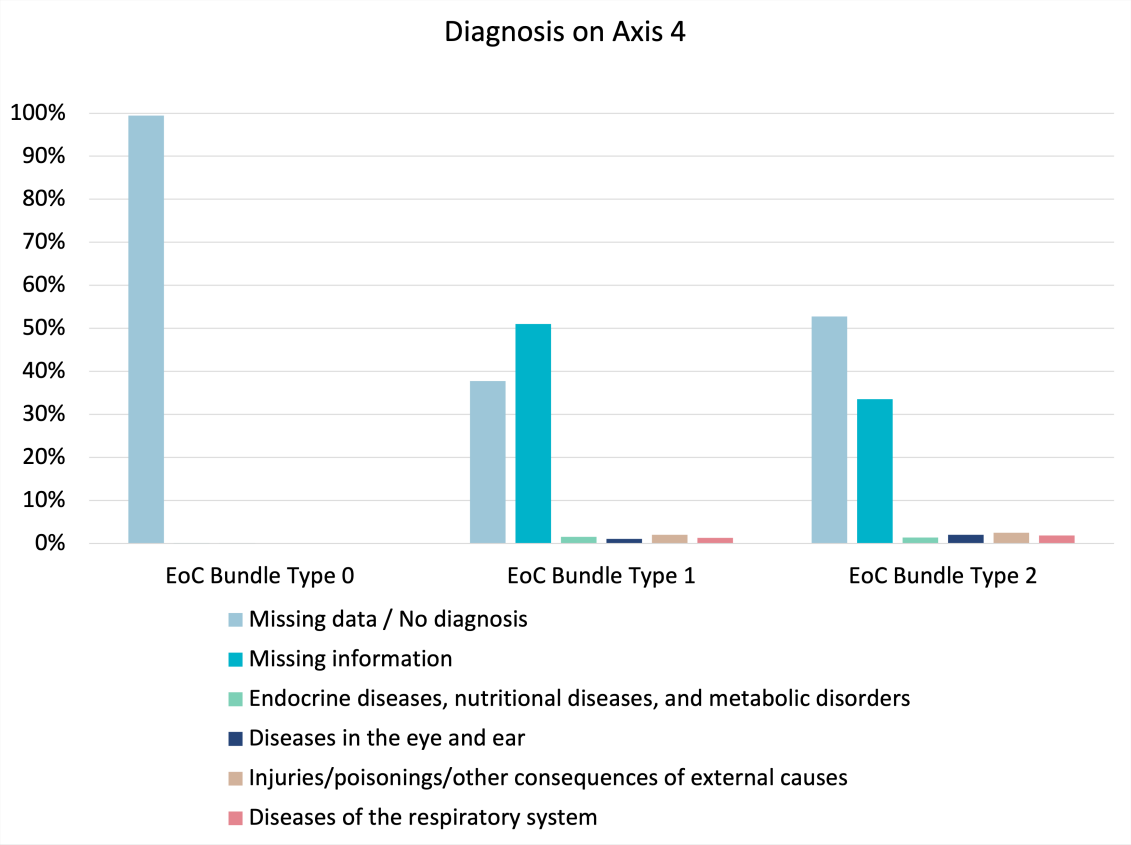


Figure 6.29: First iteration's distribution of diagnoses on Axis 4 at the beginning of an EoC Bundle.

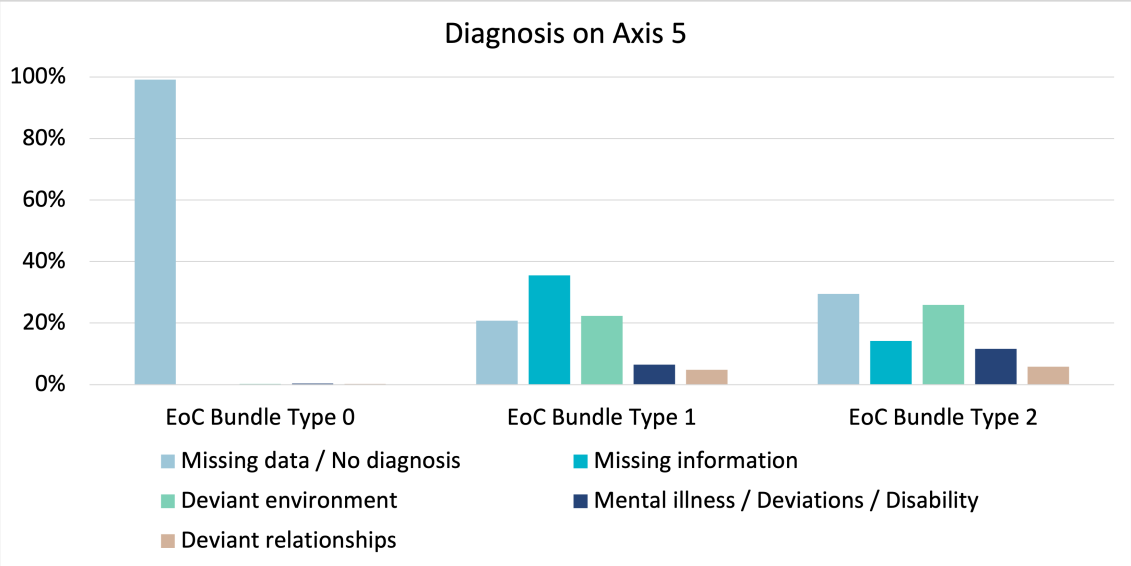


Figure 6.30: First iteration's distribution of diagnoses on Axis 5 at the beginning of an EoC Bundle.

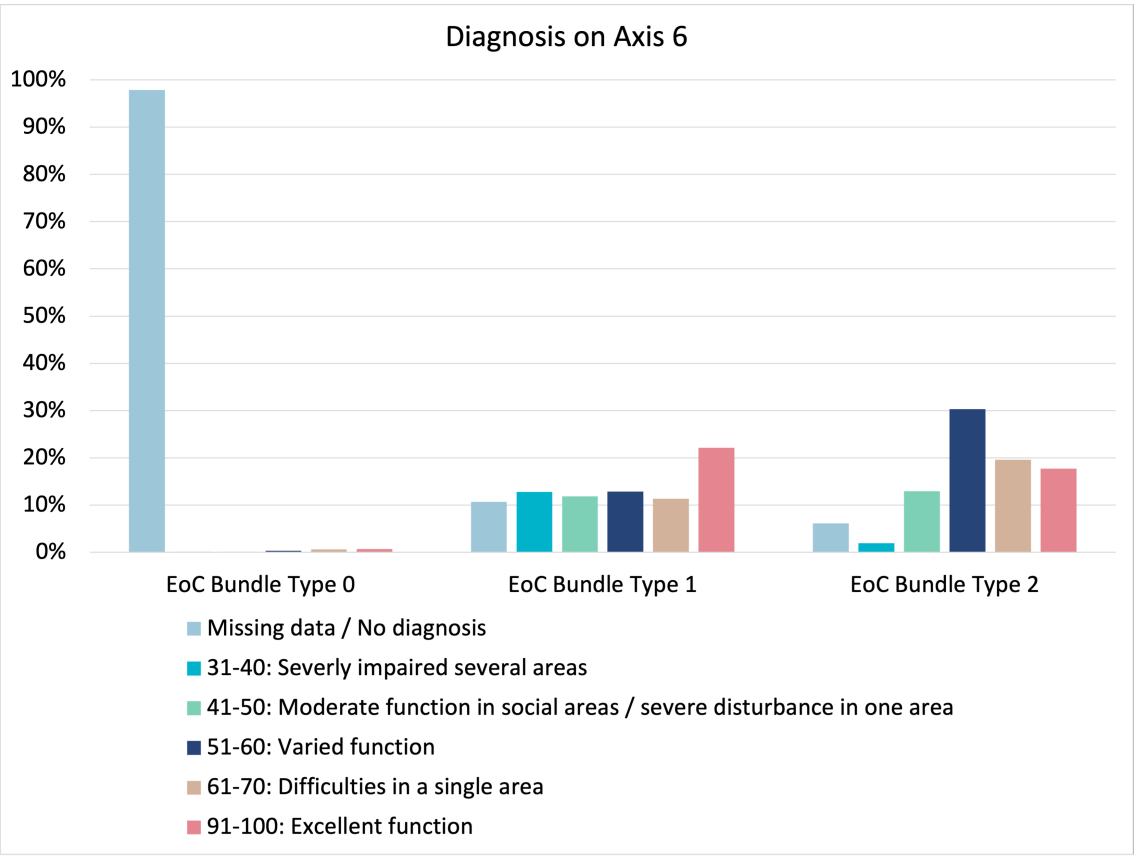


Figure 6.31: First iteration's distribution of diagnoses on Axis 6 at the beginning of an EoC Bundle.

An EoC Bundle includes one or more EoCs. The distribution of EoCs within EoC Type 0, EoC Type 1, and EoC Type 2 are presented in Figures 6.32, 6.33, and 6.34, respectively.

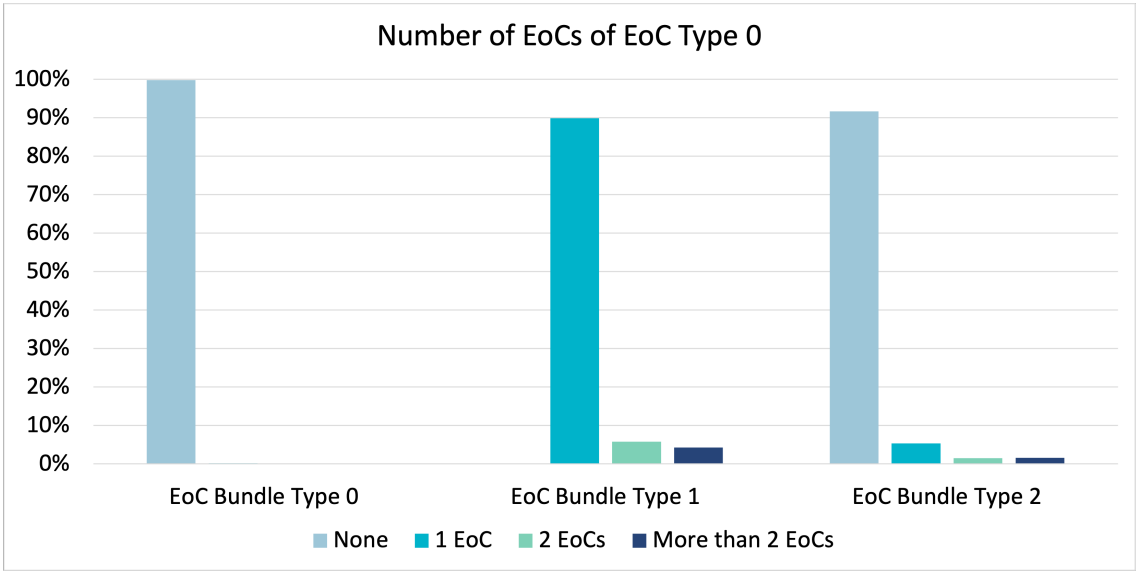


Figure 6.32: First iteration's distribution of the number of EoCs of type 0.

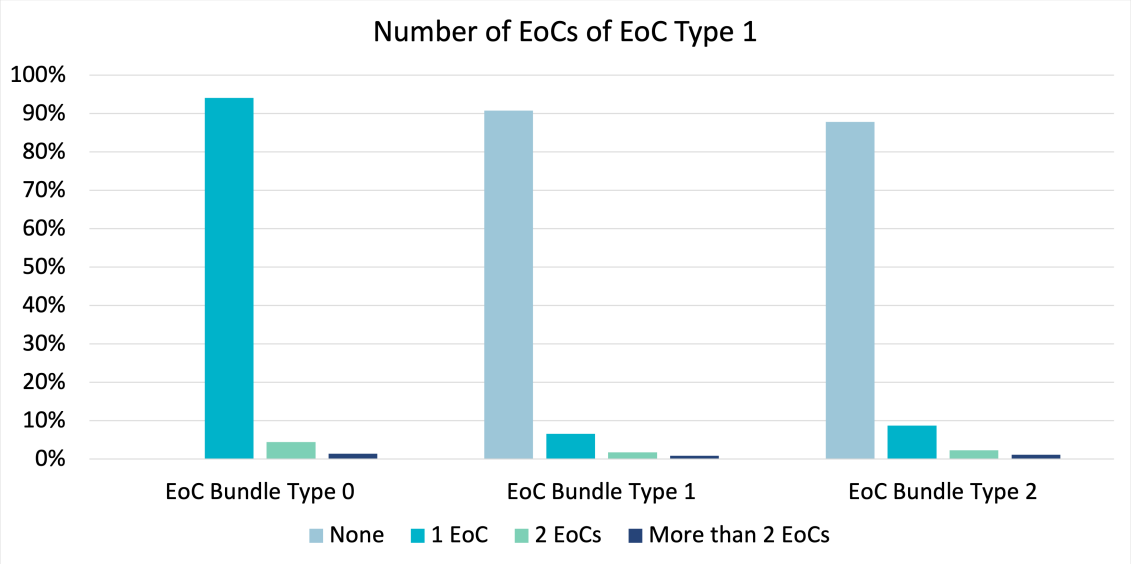


Figure 6.33: First iteration's distribution of the number of EoCs of type 1.

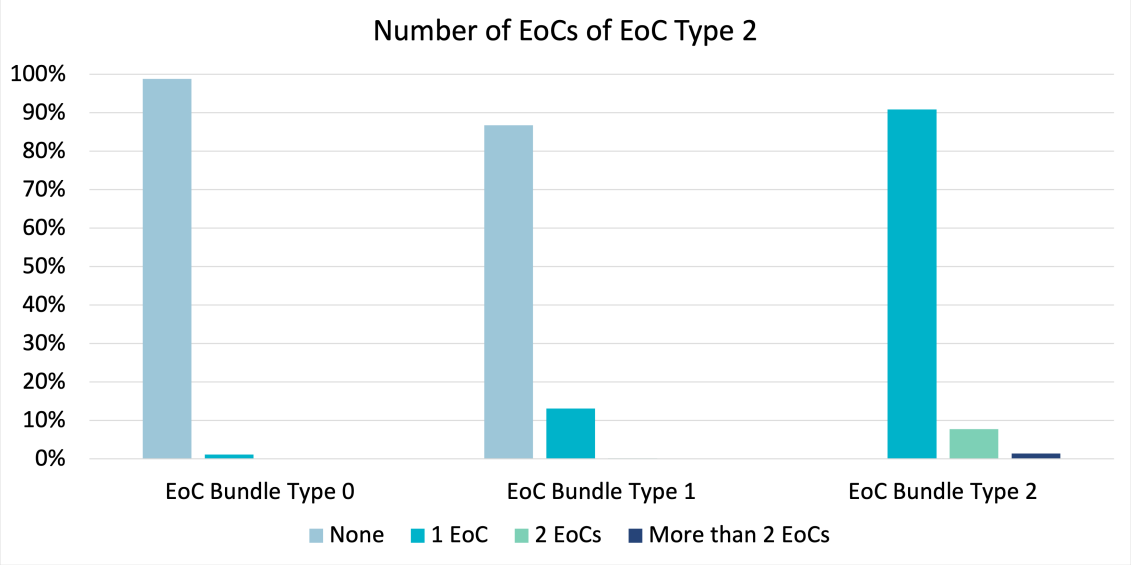


Figure 6.34: First iteration's distribution of the number of EoCs of type 2.

The numerical features' medians and means and the categorical features' modes are calculated on the EoC Bundle level and presented in table 6.16.

Feature	**Meassure**	**EoC Bundle Type 0**	**EoC Bundle Type 1**	**EoC Bundle Type 2**
EoC Bundle lenght	Mean	140	340	179
	Median	1	272	727
Age at EoC Bundle start	Mean	11	11	10
	Median	11	12	10
Patient gender	Mode	Male	Male	Male
Diagnosis on Axis 1	Mode	Missing data	Hyperkinetic disorders	Hyperkinetic disorders
Diagnosis on Axis 2	Mode	Missing data	Missing information	Missing information
Diagnosis on Axis 4	Mode	Missing data	No diagnose	Missing information
Diagnosis on Axis 4	Mode	Missing data	Missing information	No diagnose
Diagnosis on Axis 5	Mode	Missing data	Missing information	Deviant environment
Diagnosis on Axis 6	Mode	Missing data	Difficulties in single area	Varied function
Nr. EoC of type 0	Mean	0.0	1.2	0.2
	Median	0	1	0
Nr. EoC of type 1	Mean	1.1	0.14	0.18
	Median	1	0	0
Nr. EoC of type 2	Mean	0.0	0.1	1.1
	Median	0	0	1

Table 6.16: First iteration's EoC Bundle feature measurements.

Table 6.17 summarises the distribution of EoC Bundles in the three EoC Bundle clusters based on the first iteration's findings. It provides a concise overview of the findings for easy comprehension. For each cluster, the table also details the most frequently occurring EoC type as a brief reminder of what it entails.

EoC Bundle Type 0	**EoC Bundle Type 1**	**EoC Bundle Type 2**
5 714 EoC Bundles	2 406 EoC Bundles	2 925 EoC Bundles

• Short EoC Bundles.
• All ages.
• 70:30 male-to-female ratio.
• Missing data / no diagnoses on most axes.
• No EoCs of type 0.
• One EoC of type 1.
  – Short.
  – Planned.
  – Polyclinic.
  – < 25 contacts.
  – < 3 diagnoses.
• No EoCs of type 2.
	• Short - medium long.
	• All ages.
	• 70:30 male-to-female ratio.
	• Missing data / no diagnoses or missing information on axes 2-5.
	• Most EoC Bundles have a diagnosis on axes 1 and 6.
	• One or, in some EoC Bundles, multiple EoCs of type 0.
	  – Medium length.
	  – Planned.
	  – Polyclinic.
	  – 25 contacts.
	  – 4-9 diagnoses.
	• Very few have an EoC of type 1.
	• Few have an EoC of type 2.
		• Mix of all lengths, including longer.
		• All ages.
		• 70:30 male-to-female ratio.
		• Missing data / no diagnoses or missing information on axes 2-4.
		• Most EoC Bundles have a diagnosis on axes 1, 5, and 6.
		• Most EoC Bundles have no EoC of type 0.
		• Most EoC Bundles have no EoC of type 1.
		• One or, in some EoC Bundles, multiple EoCs of type 2.
		  – Longer.
		  – Planned.
		  – Polyclinic.
		  – 50-100 contacts.
		  – 7-9 diagnoses.

Table 6.17: First iteration's EoC Bundle clusters summary.

Cluster Exploration

After completing the clustering process and presenting the findings visually and interpretably, the next step is to explore these findings. This is done by examining and evaluating the work from a data perspective and collaborating with clinicians. By evaluating the findings from a data perspective, the aim is to discover potential coding mistakes or choices that impact the findings. Collaborating with clinicians is done to explore potential improvements before the second iteration.

The clustering outcomes are evaluated from a domain expert perspective through two meetings with clinicians. The first meeting was held on April 13th 2023, with Odd-Sverre Westbye. Westbye is a university lecturer at the *Faculty of Medicine and Health Sciences* at NTNU with a background as a nurse director at *RKBU Midt-Norge* within CAMHS Norway. The second meeting, held on April 14th 2023, was with Birgit Kleinau and Åsmund S. Bang. Kleinau is a senior physician at *Barne- og ungdomspsykiatrisk avdeling, Helse Nordtrøndelag*, and Bang is an innovation advisor at the same healthcare institution. During the meetings, the clustering findings were presented for review and discussion.

From the examination of the clustering findings from a data and experts' opinions perspective, the following are reflections from the first iteration:

- **Remove the total number of contacts and diagnoses** The total numbers of both diagnoses and contacts are imprecise features not giving valuable information regarding an EoC. All three experts agreed this might give an inaccurate picture of the EoCs. Representing the contacts using the number of contacts of each type gave a more accurate picture. Furthermore, Kleinau stated that giving a total of diagnoses without considering the axes might give a wrong picture of an EoC, as the six axes largely differ. It was, therefore, advised only to represent the number of diagnoses in combination with which axis the diagnoses were given on.
- **Focus on all diagnostic changes, not only unique diagnoses** All three experts pointed out that the changes back and forth between the same diagnoses should be included in the analysis. Therefore, a recommendation was to change from unique diagnoses to including all diagnostic changes.
- **Focus on axes 1 and 6 at EoC Bundle level** Both Kleinau and Westbye stated that axes one and six are particularly interesting to analyse. The first axis represents a patient's main condition, while Axis 6 states the patient's CGAS score, detailing the patient's disability level. When considering the different diagnoses given on axes 2-5 in an EoC Bundle, the most used diagnoses were "No diagnose" or "Missing information". Except for these two alternatives, the number of recurrent diagnoses for the patients was very low. Consequently, Westbye agreed that knowing only whether a diagnosis is given on one of the axes from 2-5 was sufficient information.
- **Axis 6 on the EoC Bundle level should be a numerical value** Based on *the Directorate of e-Health*, the values in Axis 6 were transformed into corresponding categorical values. However, after investigating the findings, one can see that grouping these diagnostic categories is not very insightful. By presenting them as categorical values category "91-100 Excellent function" is as different to "81-90 Good function" as it is from "1-10 Constant supervision", giving an incorrect picture of the CGAS scale.

- **Investigate correlations between features** A trend that one can see in the clustering findings is that it looks like there is a correlation between the length of the EoCs, the number of contacts patients have had, and the number of diagnoses given. According to Westbye, it is self-explanatory that longer EoCs have more contacts and diagnoses. To ensure that this trend is not covering any interesting findings, a suggestion was to change the contact types and the diagnoses given on the axes from the total amount to the frequency based on the length of the EoC. Westbye and Kleinau confirmed that they thought this was a good idea and could yield more interesting findings.
- **Adding the number of contacts before a main diagnosis is given** Adding a feature detailing the number of contacts a patient had before a diagnosis is given at a patient's main diagnosis could detail the resources used and give interesting information regarding the patient trajectories.

6.4.2 Second Clustering Iteration

The second iteration aims to utilise the insights gained from the exploration and interpretation in the first iteration to refine the clustering and improve the quality of the findings.

Data Preparation

The data preparation for the second iteration involves incorporating the feedback from the clinicians and the exploration done in the first iteration to re-evaluate the initial feature selection. A detailed explanation of the changes to the features used in the second clustering iteration follows.

- **Removing unnecessary features** All features included in the clustering process will impact the results. Therefore, it is important only to include features giving valuable insight to clinicians. From the feedback during the first iteration, the total number of contacts and diagnoses are removed from the second iteration features.
- **Including all diagnoses on the six axes** After receiving feedback that all diagnostic changes on the six axes are insightful, the number of diagnoses within an EoC is no longer limited to unique diagnoses.
- **Limiting the correlation between features** Cluster analysis involves grouping data points based on their similarity and dissimilarity to points in other clusters. However, when highly correlated variables are included, there is a risk of increasing the influence of a particular variable in the cluster formation process. Figure 6.11 illustrates the correlation between the features used in the first clustering iteration. One can see a high correlation between the total number of contacts and the specific contact types. To avoid including highly correlated features, the number of diagnoses and contacts within an EoC is divided by the length of the EoC. This might also prevent the trend of longer EoCs generally including more contacts of each type.
- **Changing the values of the features representing diagnoses on axes 2-5** After discussing with the clinicians and seeing that the feature values "Missing data" and "Missing information" dominated the diagnoses given on axes 2-5 the categorical values of these features were changed to either "Yes" or "No". "Missing data" and "Missing information" correspond to "No" while the other values correspond to "Yes".

- **Changing the feature *Diagnosis Axis 6* to a numerical feature** Seeing that numerical and categorical values are treated differently by the k-prototypes algorithm, it makes more sense to have the values on the sixth axis as a numerical variable to capture the numerical differences between the CGAS scores. The CGAS score ranges from 1-10 on Axis 6, indicating a patient's disability level. This change makes it necessary to handle the "0" values differently than in the first iteration. The "0" value is invalid, most likely resulting from missing registration on this axis (referring to the meeting with Westbye on 16.02.2023). Here a decision was made to convert the "0" value to "5", as this is the mean of this feature's values. Using a feature's mean value represents one of the most traditional data mining techniques for missing data (Theodoridis and Koutroumbas, 2008).

The number of contacts a patient has before getting a diagnosis on Axis 1 could be an interesting feature to add. Due to time limitations and many changes already decided on, this feature addition is not prioritised to include in the second iteration.

Table 6.18 and Table 6.19 summarise the features to be used in the second iteration of the EoC and EoC Bundle clustering, respectively. The green shading in the tables indicates the added or changed features from the first iteration. Red shading indicates that a feature included in the first iteration is excluded from the second iteration.

EoC Table - Second Iteration	
Feature	**Description of addition/change/removal**
EoC length	
Care level	
Immediacy level	
Nr. of contacts	Deleted not to have repeating features impacting the clustering twice.
Nr. of therapy per day Nr. of examination per day Nr. of indirect contact per day Nr. of planning per day Nr. of no-shows per day	Changed to the number of contact types per day to represent the frequency instead of the count.
Nr. of diagnoses	Not considered informative, and also deleted to not have repeating features impacting the clustering twice.
Percentage of primary axis diagnoses	The percentage of diagnoses given as the primary diagnosis on one of the six axes.
Nr. of diagnoses on Axis 1 per day Nr. of diagnoses on Axis 2 per day Nr. of diagnoses on Axis 3 per day Nr. of diagnoses on Axis 4 per day Nr. of diagnoses on Axis 5 per day Nr. of diagnoses on Axis 6 per day	Changed to consider the frequency of diagnoses given per day instead of the count.

Table 6.18: EoC table features and feature description for the second iteration

EoC Bundle Table - Second Iteration	
Feature	**Description of addition/change/removal**
Age at EoC Bundle start	
Gender	
EoC Bundle length	
Diagnosis Axis 1	
Diagnosis Axis 2 Diagnosis Axis 3 Diagnosis Axis 4 Diagnosis Axis 5	"Yes"/"No" value declaring whether the patient has a diagnosis on the corresponding axes.
Diagnosis Axis 6	Changed from categorical to numerical values.

Table 6.19: EoC Bundle table features and feature description for the second iteration.

Once the features for the second clustering iteration have been selected, they are standardised using the same approach as in the first iteration.

Clustering

Using a similar approach as the clustering performed in the previous clustering iteration, the prepared data from Section 6.19 is clustered. The Cao method is employed as the k-prototypes' initialisation method. By conducting the clustering analysis with varying values of k, ranging from 1-10, on the EoC data, the elbow plot in Figure 6.35 is generated. The plot shows that the elbow point occurs when k equals 3, indicating that three clusters are the most suitable choice for the EoC data clustering.

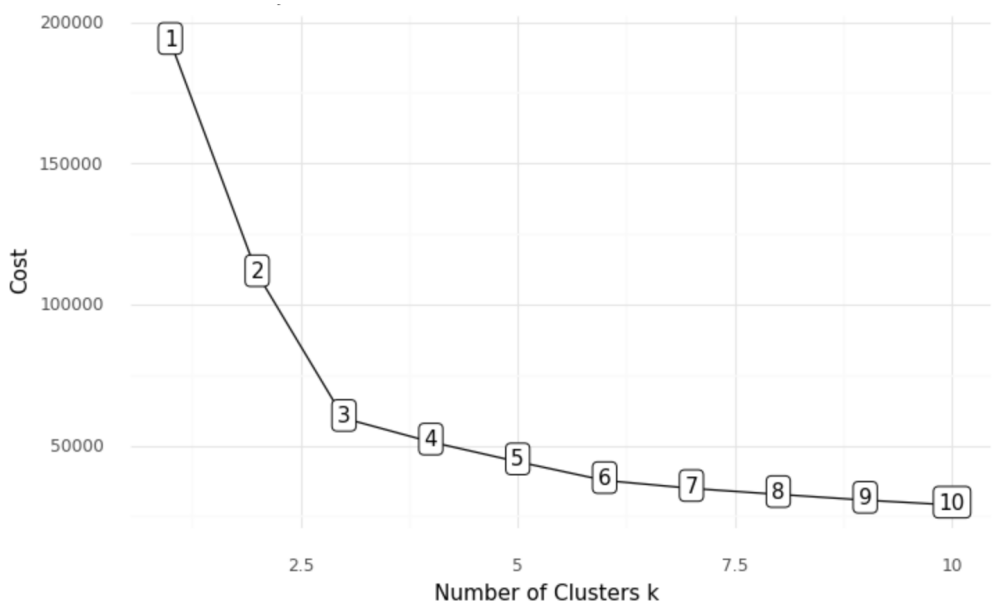


Figure 6.35: Second iteration's elbow method for the EoCs.

Using the k-prototypes algorithm, the EoC data is clustered into three groups: *EoC Type 0*, *EoC Type 1*, and *EoC Type 2*. These clusters consist of 2 768, 924 and 11 078 EoCs.

A SHAP plot explaining the impact and the importance of the EoC features is visualised in Figure 6.36.

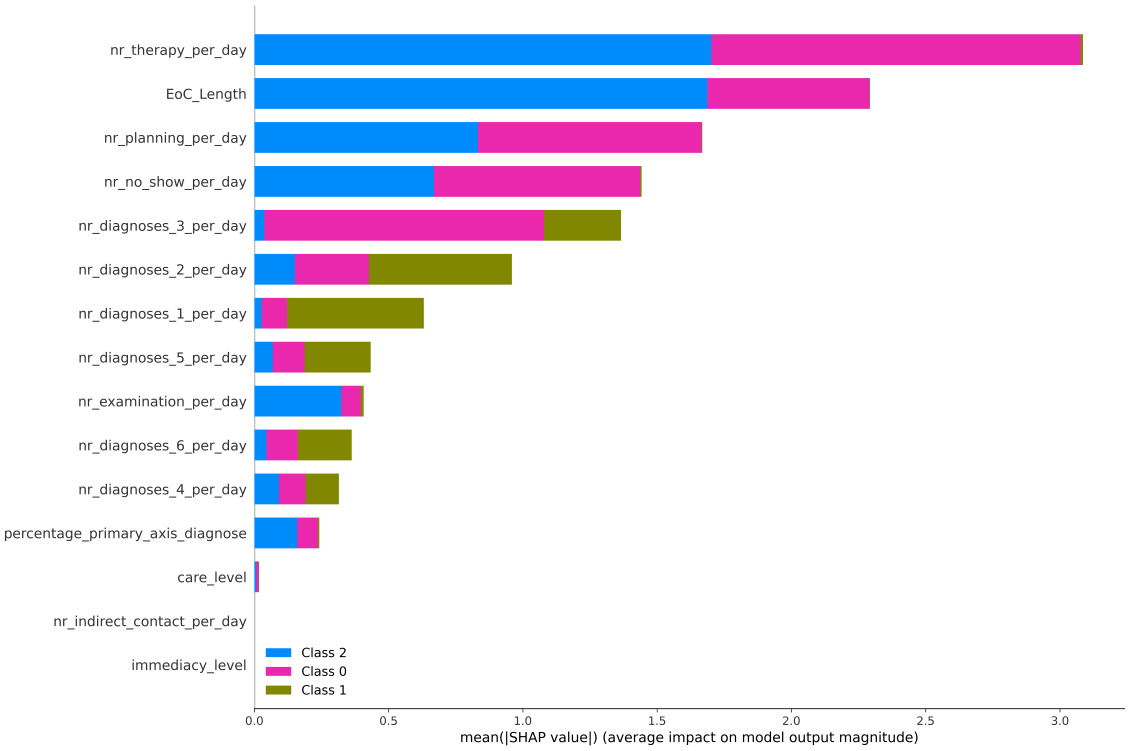


Figure 6.36: SHAP plot of the second iteration's EoC features.

The next step in the clustering process involves incorporating the identified groups of EoC clusters to perform the EoC Bundle clustering. Rather than considering the entire EoC data set as an entity in the EoC Bundle clustering, the EoC data is incorporated using the three identified EoC clusters.

Again, the optimal number of clusters for the EoC Bundle clustering is identified using the Elbow method illustrated in Figure 6.37. The elbow plot indicates that the point of maximum curvature occurs at 4 clusters.

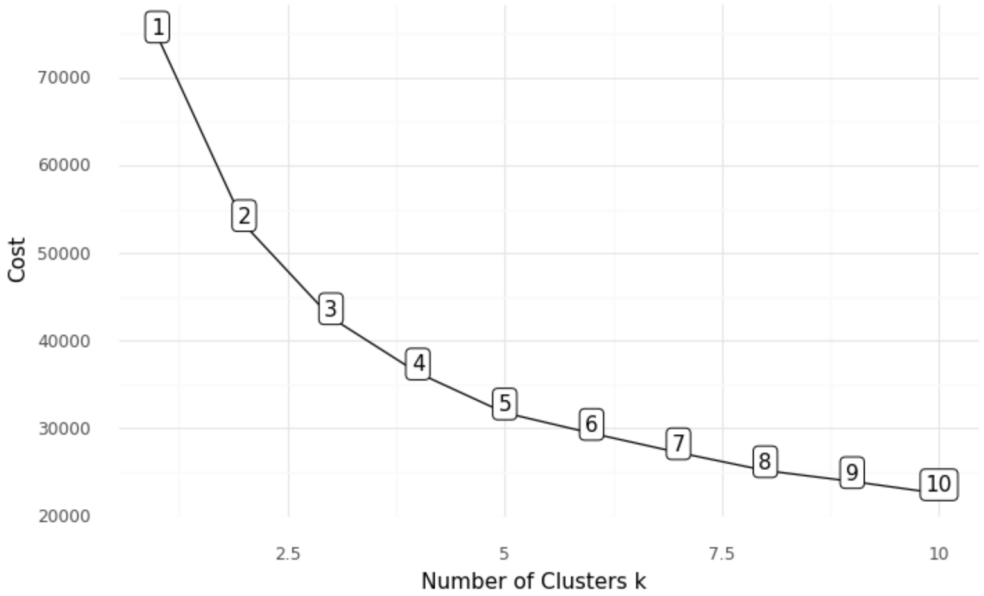


Figure 6.37: Second iteration's elbow method for the EoC Bundles.

Applying the k-prototypes algorithm to the EoC Bundle data, it is grouped into four clusters labelled: *EoC Bundle Type 0*, *EoC Bundle Type 1*, *EoC Bundle Type 2*, and *EoC Bundle Type 3*. These clusters comprise 670, 1 945, 3 919, and 4 384 EoC Bundles.

Figure 6.38 visualises the impact and the importance of the EoC Bundle features using SHAP.

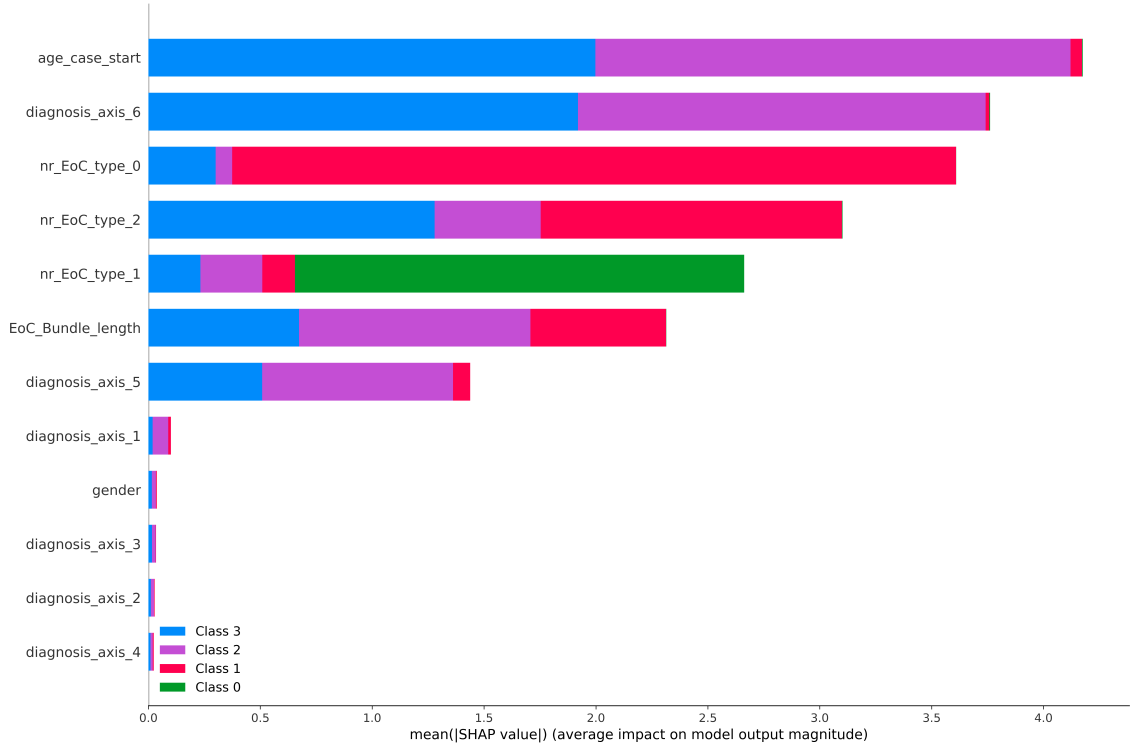


Figure 6.38: SHAP plot of the second iteration's EoC Bundle features.

Intermediate Cluster Findings

Similar to the findings from the first iteration, the clustering outcomes from the second iteration are presented. First, the feature value distribution of the EoC clusters is presented, and then the distribution within the EoC Bundle clusters is presented. The complete findings from the second iteration are presented in Appendix C.2.

EoC Cluster Findings

Clustering the data on the EoC level identified the clusters presented in Table 6.20. Figures 6.39, 6.40, and 6.41 present the distribution of the different EoC lengths and care and immediacy levels.

Clusters	Nr. Data Points
EoC Type 0	2 768
EoC Type 1	924
EoC Type 2	11 078

Table 6.20: Second iteration's distribution of EoCs in the EoC clusters.

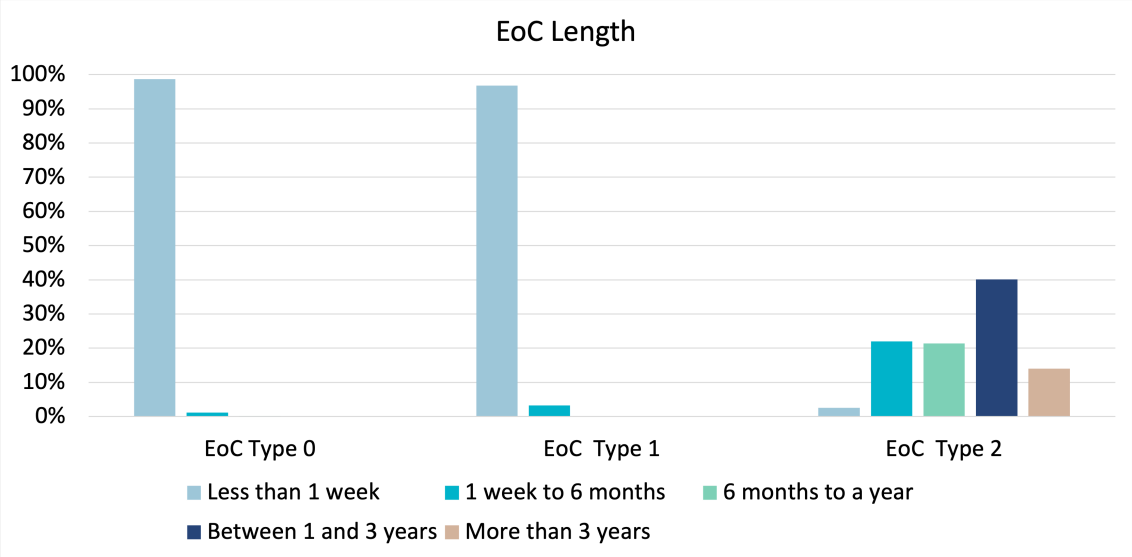


Figure 6.39: Second iteration's distribution of EoC lengths.

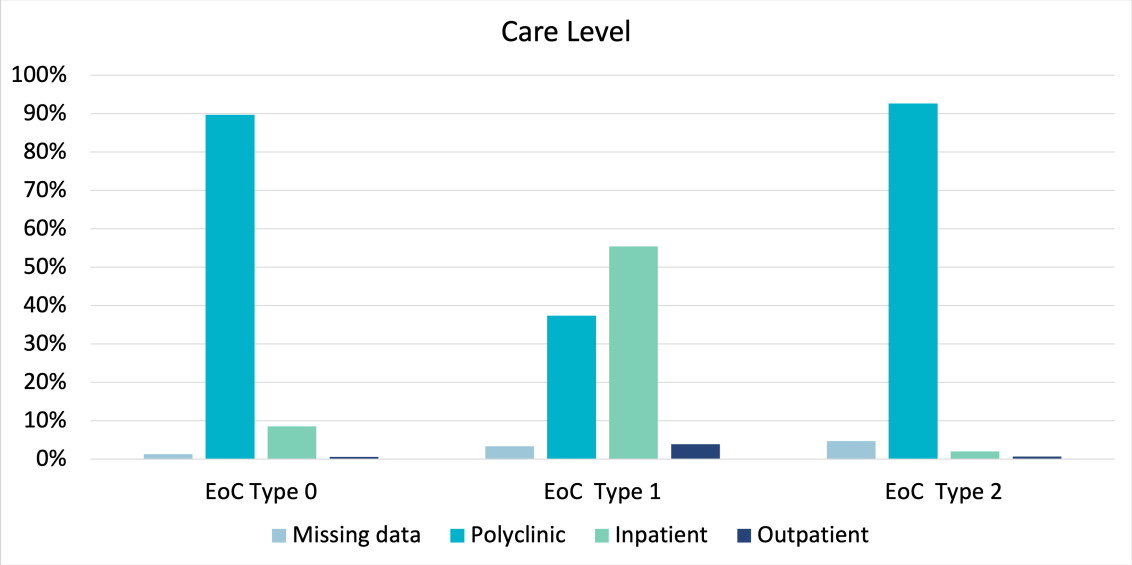


Figure 6.40: Second iteration's distribution of care levels.

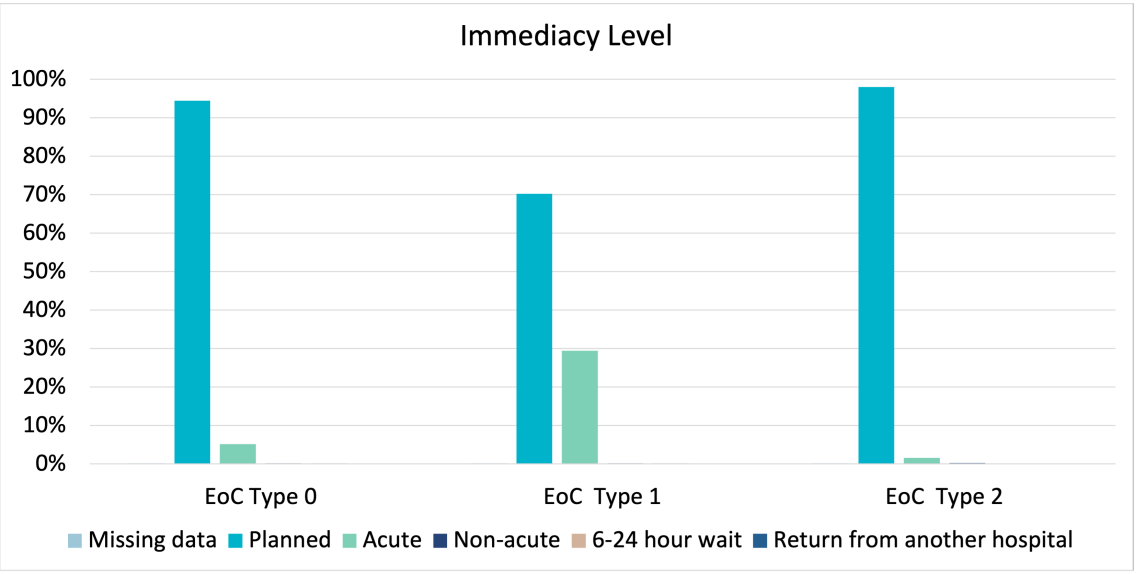


Figure 6.41: Second iteration's distribution of immediacy levels.

The frequency distribution of contacts of each contact type is presented in Table 6.21. The most common frequency for each contact type is highlighted with bold text for each cluster.

Contact Feature	Values	EoC Type 0	EoC Type 1	EoC Type 2
Frequency of therapy contacts	Never.	7%	12%	6%
	Less than once a year.	0%	0%	2%
	Between once a year and once a month.	0%	0%	34%
	Between once a month and once a week.	0%	1%	**52%**
	Between once a week and once a day.	**53%**	22%	6%
	More than once a day.	40%	**65%**	0%
Frequency of planning contacts	Never.	7%	13%	6%
	Less than once a year.	0%	0%	2%
	Between once a year and once a month.	0%	1%	**47%**
	Between once a month and once a week.	0%	2%	42%
	Between once a week and once a day.	**60%**	32%	3%
	More than once a day.	33%	**52%**	0%
Frequency of examination contacts	Never.	7%	12%	6%
	Less than once a year.	0%	0%	2%
	Between once a year and once a month.	0%	0%	34%
	Between once a month and once a week.	0%	1%	**52%**
	Between once a week and once a day.	**53%**	22%	6%
	More than once a day.	40%	**65%**	0%
Frequency of no-show contacts	Never.	11%	33%	10%
	Less than once a year.	0%	0%	5%
	Between once a year and once a month.	0%	0%	**73%**
	Between once a month and once a week.	1%	2%	11%
	Between once a week and once a day.	**59%**	**39%**	1%
	More than once a day.	29%	26%	0%
Frequency of indirect contacts	Never.	7%	12%	6%
	Less than once a year.	0%	0%	1%
	Between once a year and once a month.	0%	0%	34%
	Between once a month and once a week.	0%	1%	**52%**
	Between once a week and once a day.	**53%**	22%	6%
	More than once a day.	40%	**65%**	0%

Table 6.21: Second iteration's distribution of the frequency of the contact types.

Figure 6.42 presents the percentage distribution of how many diagnoses were given during the EoC as the primary diagnosis on one of the six axes.

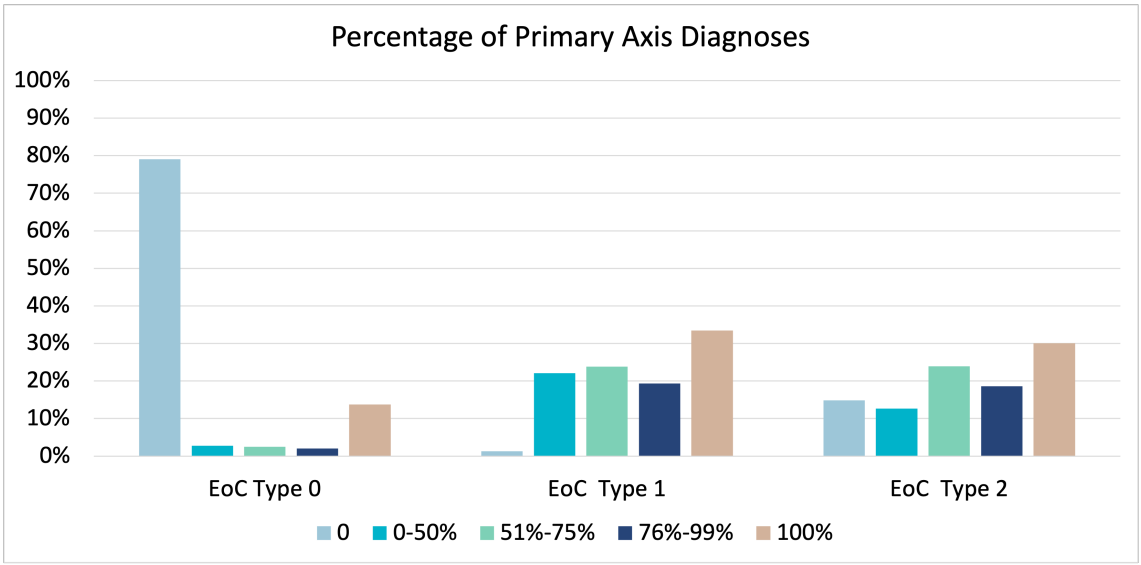


Figure 6.42: Second iteration's distribution of the number of diagnoses given as the primary diagnosis on one of the six axes.

The frequency distribution of diagnoses on the six axes is detailed in Table 6.22.

Diagnostic Feature	**Values**	**EoC Type 0**	**EoC Type 1**	**EoC Type 2**
Frequency of diagnoses on Axis 1	Never.	**95%**	0%	15%
	Less than once a year.	0%	0%	34%
	Between once a year and once a month.	1%	0%	**50%**
	Between once a month and once a week.	0%	3%	1%
	Between once a week and once a day.	3%	**79%**	0%
	More than once a day.	1%	17%	0%
Frequency of diagnoses on Axis 2	Never.	**99%**	2%	21%
	Less than once a year.	0%	0%	**43%**
	Between once a year and once a month.	1%	2%	36%
	Between once a month and once a week.	0%	2%	0%
	Between once a week and once a day.	0%	**88%**	0%
	More than once a day.	0%	6%	0%
Frequency of diagnoses on Axis 3	Never.	**99%**	2%	21%
	Less than once a year.	0%	0%	**43%**
	Between once a year and once a month.	1%	1%	36%
	Between once a month and once a week.	0%	2%	0%
	Between once a week and once a day.	0%	**90%**	0%
	More than once a day.	0%	5%	0%
Frequency of diagnoses on Axis 4	Never.	**99%**	6%	23%
	Less than once a year.	0%	0%	**42%**
	Between once a year and once a month.	1%	1%	35%
	Between once a month and once a week.	0%	3%	0%
	Between once a week and once a day.	0%	**84%**	0%
	More than once a day.	0%	6%	0%
Frequency of diagnoses on Axis 5	Never.	**99%**	2%	21%
	Less than once a year.	0%	0%	**38%**
	Between once a year and once a month.	1%	0%	40%
	Between once a month and once a week.	0%	3%	1%
	Between once a week and once a day.	0%	**77%**	0%
	More than once a day.	0%	18%	0%
Frequency of diagnoses on Axis 6	Never.	**99%**	6%	23%
	Less than once a year.	0%	0%	**39%**
	Between once a year and once a month.	1%	1%	38%
	Between once a month and once a week.	0%	2%	0%
	Between once a week and once a day.	0%	**84%**	0%
	More than once a day.	0%	7%	0%

Table 6.22: Second iteration's distribution of the frequency of diagnoses on the different axes.

Table 6.23 presents the second iteration's EoC features' calculated means, medians, and modes.

Feature	Meassure	EoC Type 0	EoC Type 1	EoC Type 2
EoC lenght	Mean	2	6	579
	Median	0	3	452
Care level	Mode	Polyclinic	Inpatient	Polyclinic
Immediacy level	Mode	Planned	Planned	Planned
Nr. of therapy contacts per day	Mean	2.24	4.36	0.05
	Median	1.0	2.33	0.04
Nr. of planning contacts per day	Mean	1.73	2.31	0.05
	Median	1.0	1.25	0.03
Nr. of examination contacts per day	Mean	0.78	2.01	0.04
	Median	0.0	1.0	0.01
Nr. of no-show contacts per day	Mean	1.30	0.96	0.03
	Median	1.0	0.5	0.01
Nr. of indirect contacts per day	Mean	2.24	4.36	0.05
	Median	1.0	2.33	0.04
Percentage of primary axis diagnoses	Mean	18%	73%	67%
	Median	0%	83%	75%
Nr. of diagnoses on Axis 1 per day	Mean	0.03	1.02	0.01
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 2 per day	Mean	0.0	0.62	0.0
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 3 per day	Mean	0.0	0.63	0.0
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 4 per day	Mean	0.0	0.64	0.0
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 5 per day	Mean	0.0	1.04	0.0
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 6 per day	Mean	0.0	0.63	0.0
	Median	0.0	0.5	0.0

Table 6.23: Second iteration's EoC feature measurements.

Table 6.24 presents an overview of the EoC clusters identified in the second iteration. This table summarises the most prominent features and simplifies the findings.

EoC Type 0	**EoC Type 1**	**EoC Type 2**
2 768 EoCs	924 EoCs	11 078 EoCs
• Shorter than a week. • Polyclinic. • Planned. • All contacts weekly to daily. • Seldom given a diagnosis on any axis.	• Shorter than a week. • Inpatient. • Planned or acute. • All contacts (except no-show) multiple times a day. • Diagnoses given on all axes weekly to daily.	• Longer than a week. • Polyclinic. • Planned. • All contacts (except no-show) yearly to monthly. • Diagnoses given on all axes (except axis 1) less than once a year.

Table 6.24: Second iteration's EoC clusters summary.

EoC Bundle Cluster Findings

The clustering of the EoC Bundle data revealed four clusters, as illustrated in Table 6.25. Figures 6.43, 6.44, and 6.45 present the distribution of the different EoC Bundle lengths and the patients' age and gender.

EoC Bundle Cluster	**Nr. Data Points**
EoC Bundle type 0	670
EoC Bundle type 1	1 945
EoC Bundle type 2	3 919
EoC Bundle type 3	4 384

Table 6.25: Second iteration's distribution of EoC Bundles in the EoC Bundle clusters.

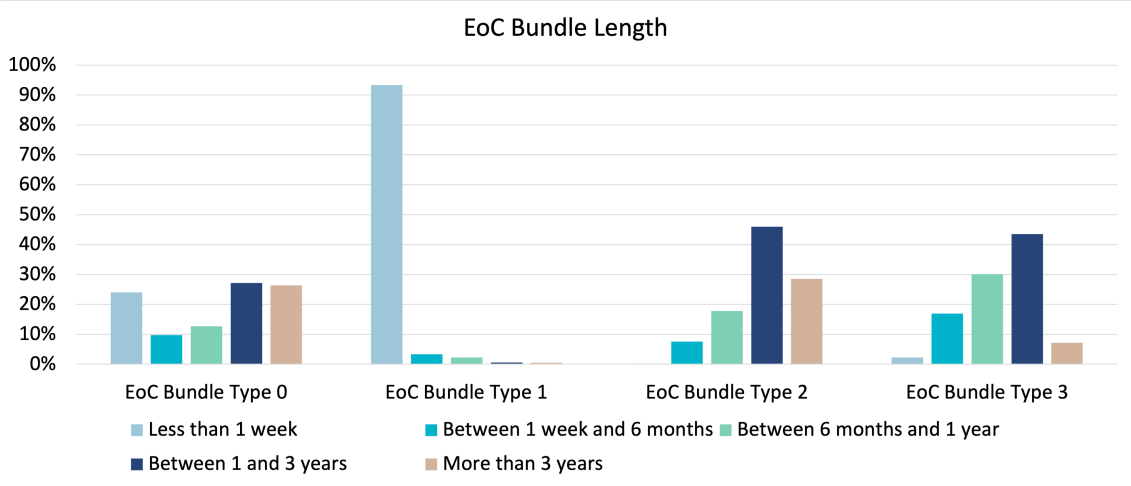


Figure 6.43: Second iteration's distribution of EoC Bundle lengths.

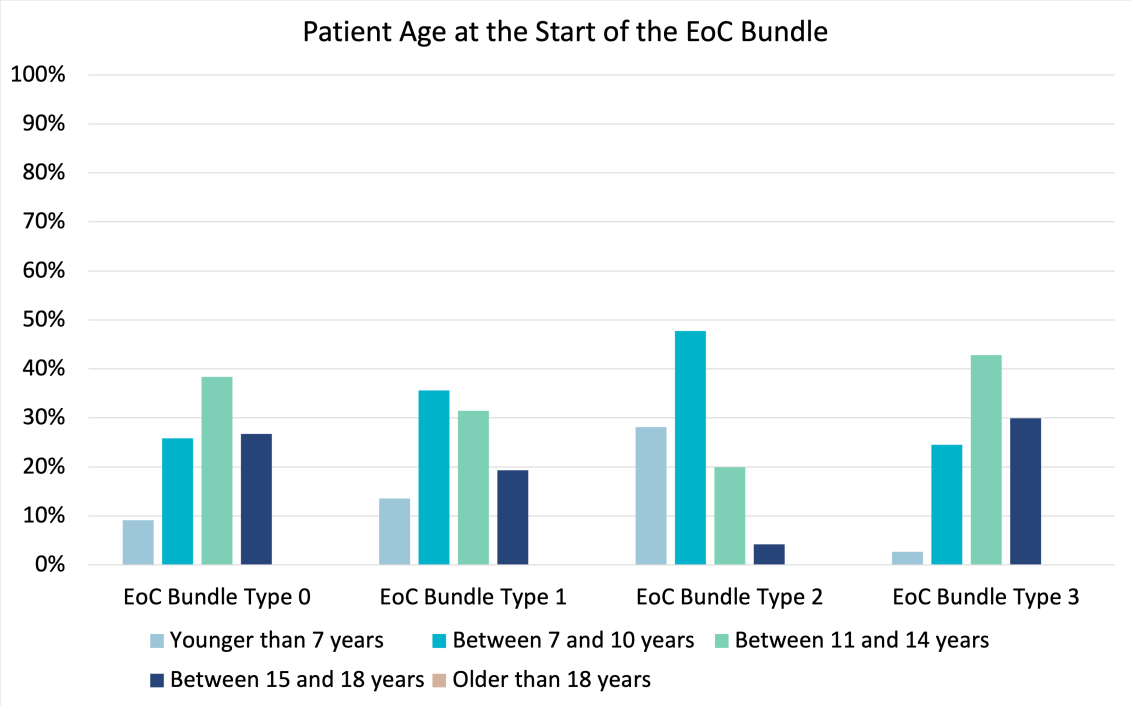


Figure 6.44: Second iteration's distribution of patients' age.

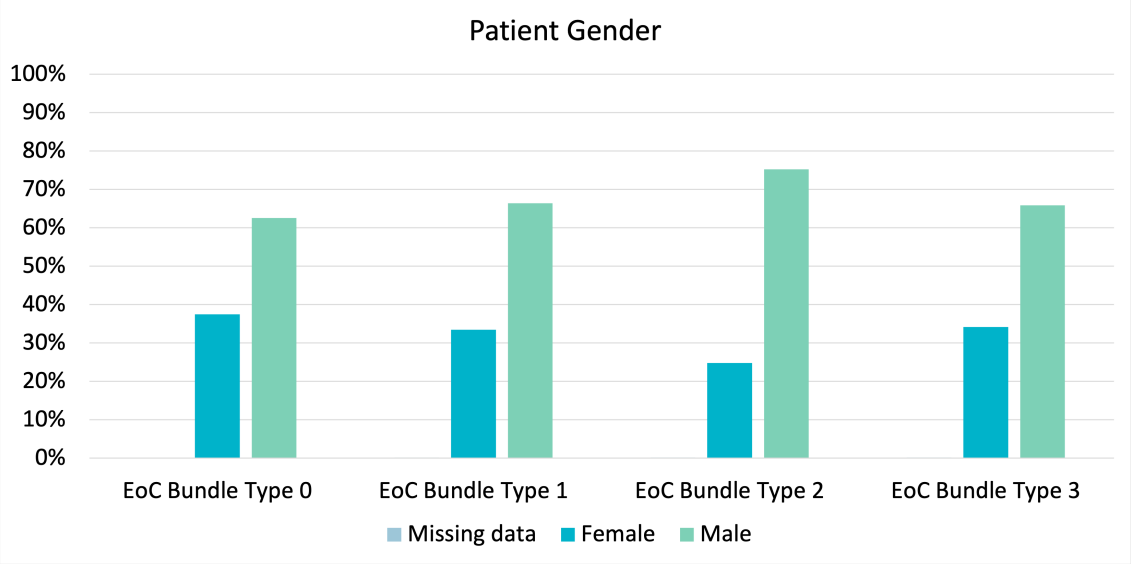


Figure 6.45: Second iteration's distribution of patients' gender.

Figure 6.46 presents the distribution of the most common diagnoses given on the first axis on the EoC Bundle level. The distribution of whether a diagnosis is given on Axis 2-5 is visualised in Figure 6.47. Then, the distribution of the most common CGAS scores is given in Figure 6.48. Note that the percentage of CGAS score equalling five is impacted by the choice to change the "0" values to "5" as described in the data preparation in Section 6.19.

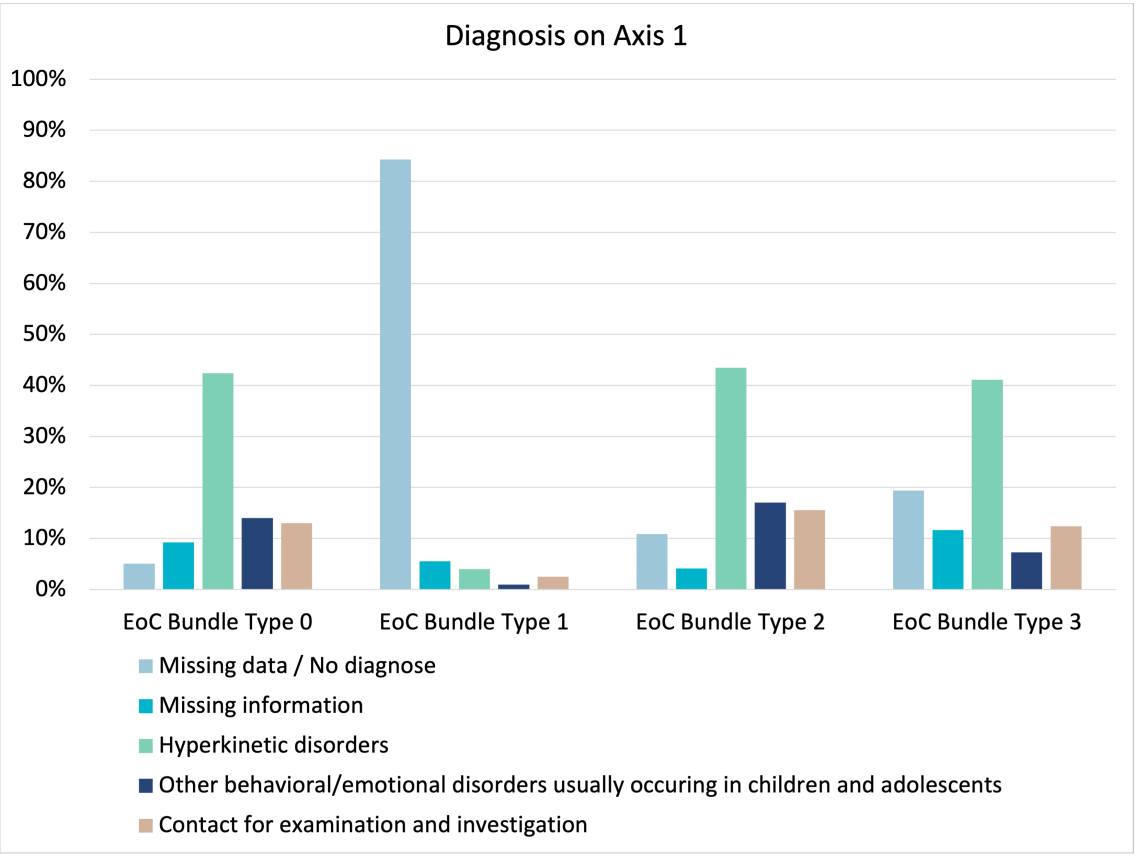


Figure 6.46: Second iteration's distribution of diagnoses on Axis 1 at the beginning of an EoC Bundle.

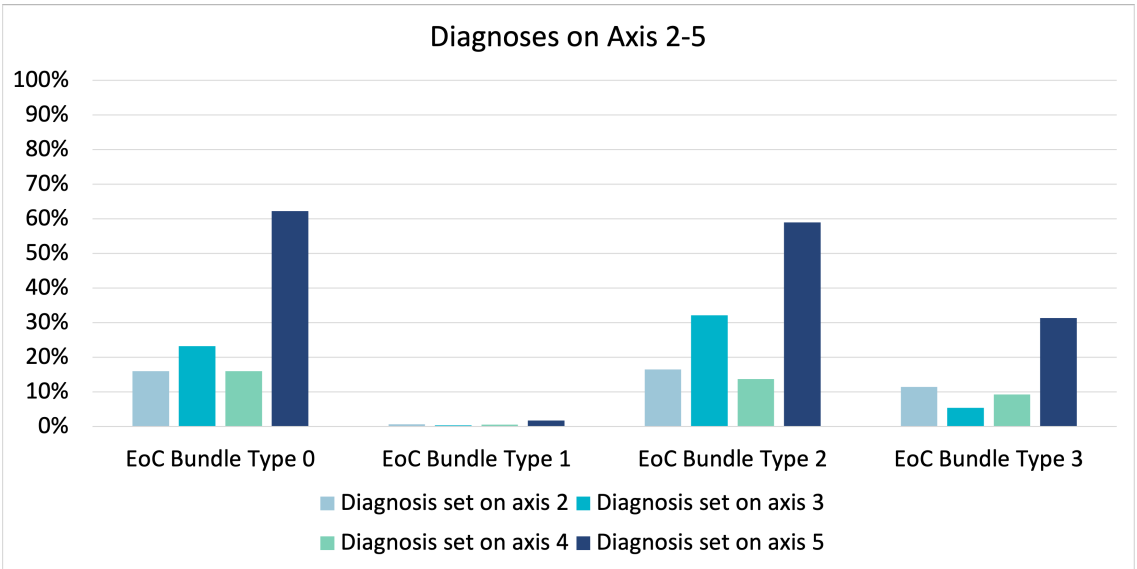


Figure 6.47: Second iteration's distribution of diagnoses on axes 2-5 at the beginning of an EoC Bundle.

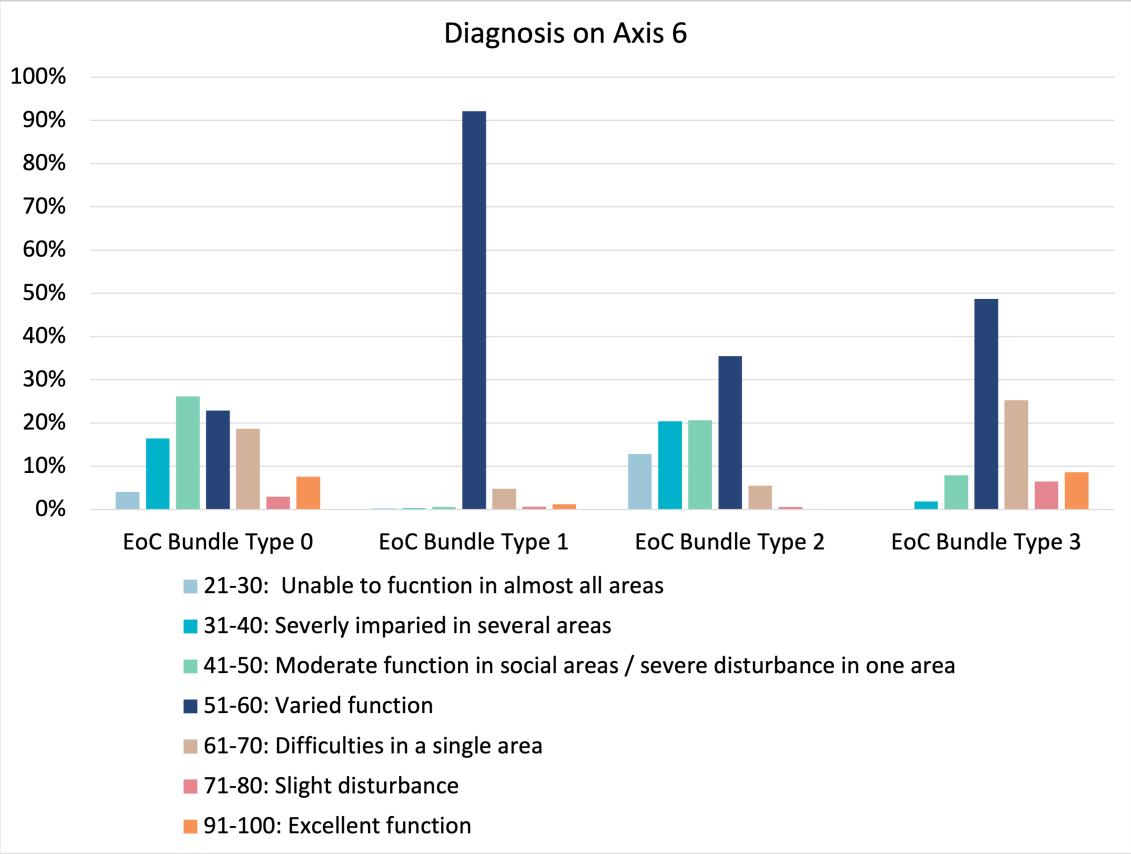


Figure 6.48: Second iteration's distribution of diagnoses on Axis 6 at the beginning of an EoC Bundle.

The distributions of the number of EoCs of the three different EoC types are presented in Figures 6.49, 6.50, and 6.51.

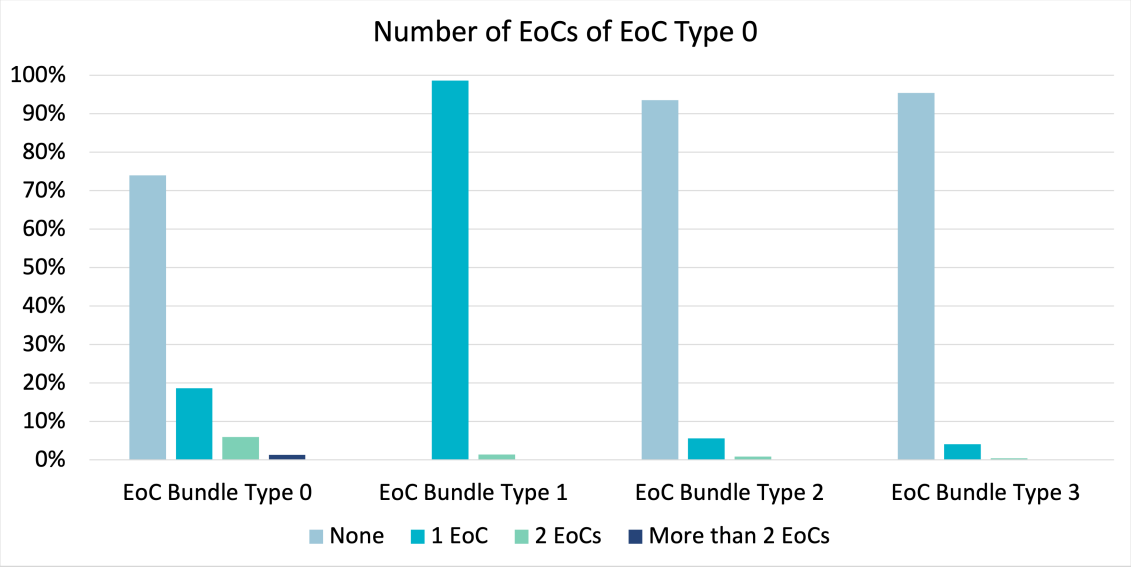


Figure 6.49: Second iteration's distribution of the number of EoCs of type 0.

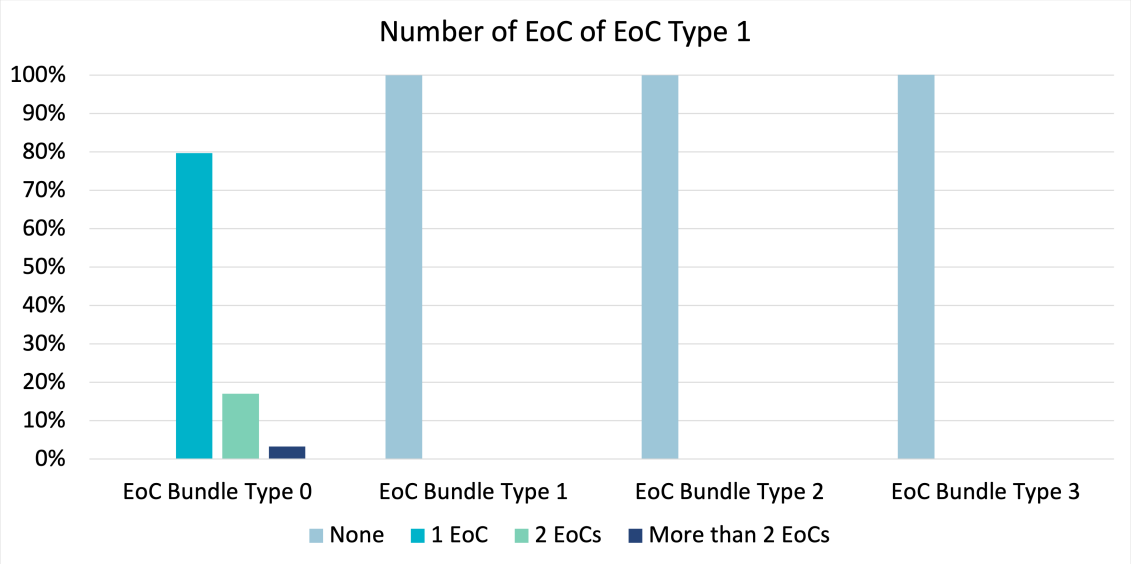


Figure 6.50: Second iteration's distribution of the number of EoCs of type 1.

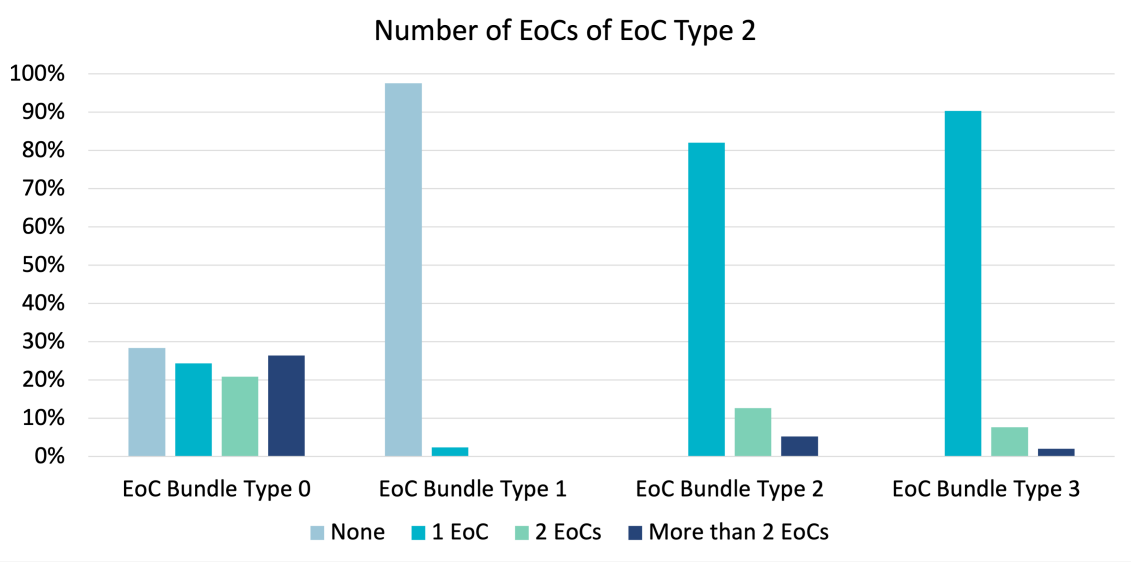


Figure 6.51: Second iteration's distribution of the number of EoCs of type 2.

For the EoC Bundle features included in the second iteration, the modes, medians, and means are presented in Table 6.26.

Feature	**Meassure**	**EoC Bundle Type 0**	**EoC Bundle Type 1**	**EoC Bundle Type 2**	**EoC Bundle Type 2**
EoC Bundle lenght	Mean	730	19	884	471
	Median	455	0	727	366
Age at EoC Bundle start	Mean	12	11	9	12
	Median	12	11	8	13
Patients' gender	Mode	Male	Male	Male	Male
Diagnosis on Axis 1	Mode	Hyperkinetic disorders	Missing data	Hyperkinetic disorders	Hyperkinetic disorders
Diagnosis on Axis 2	Mode	No	No	No	No
Diagnosis on Axis 3	Mode	No	No	No	No
Diagnosis on Axis 4	Mode	No	No	No	No
Diagnosis on Axis 5	Mode	Yes	No	Yes	No
Diagnosis on Axis 6	Mean	4.8	5.1	3.9	5.6
	Median	5	5	4	5
Nr. EoC of type 0	Mean	0.4	1.0	0.1	0.1
	Median	0	1	0	0
Nr. EoC of type 1	Mean	1.4	0.0	0.0	0.0
	Median	1	0	0	0
Nr. EoC of type 2	Mean	1.7	0.0	1.3	1.1
	Median	1	0	1	1

Table 6.26: Second iteration's EoC Bundle feature measurements.

From all the findings presented in this section, a summarisation is created. As in the first iteration, this summarisation is a simplification and starting point for feedback and discussion, not a complete picture of the clustering findings. For the second iteration, the simplification is presented in Table 6.27. Keep in mind when examining the table, the key features of the identified EoC types, noted again following the table.

EoC Bundle Type 0 670 EoC Bundles	**EoC Bundle Type 1** 1 945 EoC Bundles
• All lengths. • Mostly older than 7 years. • 65:35 male-to-female ratio. • Missing information on Axis 1 • Mostly no diagnoses on axes 2-4 • Moderate function in social areas / severe disturbance in one area • Few have an EoC of type 0. • One or sometimes more EoC of type 1. • One or often more EoC of type 2.	• Less than a week long. • Mostly between 7 and 14 years. • 70:30 male-to-female ratio. • Missing data / no diagnoses on Axis 1. • No diagnoses on axes 2-5. • Varied function • One EoC of type 0. • No EoCs of type 1. • No EoCs of type 2.
EoC Bundle Type 2 3 919 EoC Bundles	**EoC Bundle Type 3** 4 385 EoC Bundles
• Mostly longer than a year. • Mostly younger than 10 years. • 75:25 male-to-female ratio. • Hyperkinetic disorders on Axis 1. • Rarly diagnoses on axes 2-4. • Moderate function. • No EoC of type 0. • No EoC of type 1. • One or sometimes more EoC of type 2.	• Mostly longer than six months. • All ages, but very few from 7 to 10 years. • 65:35 male-to-female ratio. • Hyperkinetic disorders on Axis 1. • Rarly diagnoses on axes 2-5. • Moderate function. • No EoC of type 0. • No EoC of type 1. • One EoC of type 2.

Table 6.27: Second iteration's EoC Bundle clusters summary.

- **EoC type 0:** Shorter than a week, planned polyclinic EoCs with contacts on a weekly to daily basis and seldom diagnoses given.
- **EoC type 1:** Shorter than a week, planned inpatient EoCs with many contacts daily and diagnoses given weekly to daily.
- **EoC type 2:** Longer than a week, planned polyclinic EoCs with contacts yearly to monthly and diagnoses are given less than once yearly.

D) Cluster Exploration

The findings from the second iteration are explored similarly as in the first iteration. However, for this iteration, the findings were presented to the IDDEAS team on April 21st, 2023 and Odd-Sverre Westbye on April 28th, 2023. From the second iteration exploration, the following reflections are made:

- **Identify diagnostic patterns** Westbye described that an interesting factor when investigating patient trajectories is the changes throughout an EoC and an EoC Bundle. Investigating potential patterns in the changes made to the diagnoses on the different axes is a topic he found particularly interesting. Westbye and the IDDEAS team found information regarding axes one and six changes worth investigating. A suggestion was to include the difference in the CGAS score given on Axis 6 from the beginning to the end of an EoC Bundle.
- **Add information regarding the first main diagnosis given** The IDDEAS team and Westbye suggested adding information regarding the period before a main diagnosis is given on the first axis. This concurs with the feedback from Kleinau in the first iteration (referring to the meeting with Birgit Kleinau 14.04.23). Adding the number of days or contacts before the main diagnosis is given could inform about clinical resources spent and how difficult the diagnostic process was.
- **Focus more on the "typical" hyperkinetic disorders patients** The IDDEAS team and Westbye noted that including rejected patients might impact the findings. They mentioned that including these patients in the first two iterations might have contributed to identifying differences between the more "typical" hyperkinetic disorders trajectories and other trajectories. Specifically, they pointed out that EoC Type 2 might identify more "typical" EoCs for patients with hyperkinetic disorders. Removing rejected patients from the last iteration could yield more detailed findings regarding typical hyperkinetic disorders trajectories.
- **Add more information regarding a patient** Including more patient features could be interesting to get a more detailed view of the patients. The IDDEAS team hypothesised that life situations often impact the patient trajectories and, therefore, could be interesting to include. A suggestion given was to include patients' care situation.

6.4.3 Third Clustering Iteration

Utilising the knowledge and feedback from the two prior iterations, the third and final iteration is performed. For this iteration, the data preparation and clustering process is presented in this section, while the results are presented separately in Chapter 7. The feedback is given as a part of the final evaluation, presented in Chapter 8.

Data Preparation

Considering the findings and exploration in the second iteration, changes regarding the data set are made. Following is a presentation of all changes made:

- **Exluding rejected patients** To potentially obtain more detailed findings regarding trajectories of "typical" hyperkinetic disorders patients, the inclusion of rejected patients is evaluated. To do so, multiple features from the St. Olavs data are investigated. Firstly, the feature *sak.tattimot* is investigated to study the initial assessment of the patients. Here, it is decided to exclude the EoC Bundles and the corresponding EoCs, where the assessment is either "Rejected due to capacity" or "Rejected due to professional reasons" (using *Koder 13)*.

  Additionally, the closing code for an EoC Bundle is investigated using the St. Olavs value *sak.avslkode* (using *Koder 22*). Here the values "Rejected" or "Did not get started" are excluded. This decision was made in cooperation with psychologist Sanja Prodanovic who explained that the "Rejected" value represents patients who were rejected during their EoC Bundle, indicating that these EoC Bundles can be considered incomplete. The value "Did not get started" may indicate that a patient initially accepted a treatment offer from CAMHS but later rejected it or did not attend the appointments (referring to the meeting with Prodanovic 05.05.2023).

  Applying these criteria eliminates 2 209 distinct EoC Bundles and 2 298 corresponding EoCs.

- **Including EoC Bundles' closing codes** The feature *Cloding Code* is included on the EoC Bundle level to include the possible reasons to close an EoC Bundle that is not "Rejected" or "Did not get started" (based on *Koder 22*). This feature inclusion aims to detail how the patient trajectories ended. The distribution of the cohort's different closing codes can be visualised in Figure 6.52.

- **Including more information regarding diagnostic changes and first main diagnosis** For the third iteration, diagnostic changes on the first and sixth axes are investigated. This investigation is done with the new data set, excluding rejected patients, containing 8 919 EoC Bundles and 12 728 corresponding EoCs. Investigating diagnoses on Axis 1, one can see that only 30% of the EoCs have more than one diagnosis on the first axis, and only 0.3% of the EoCs include more than one main diagnosis. From this, one can conclude that including changes in the patient's main diagnosis almost completely would consist of missing values. Similarly, investigating the CGAS scores given on Axis 6, only 13% of the scores have been changed. Therefore, also this value is excluded from the third iteration.

  Although most EoCs have few diagnostic changes, 68% of the EoCs have at least one main diagnosis given that is neither "Missing information" nor "No diagnosis". Therefore, the feature *Nr contacts before the main diagnosis* is extracted using the date of the first main diagnosis given and counting all contacts up until this date. For the 32% of the EoCs not having a main diagnosis during an EoC, or only "Missing information" or "No diagnosis", the diagnosis on the EoC Bundle level on Axis 1 is investigated. If an EoC does not have a main diagnosis, but there is a main diagnosis on the EoC Bundle level, the *Nr contacts before the main diagnosis* feature is converted to "0", indicating that the patient started out having a main diagnosis. If the patient does not have a diagnosis on Axis 1 on the EoC Bundle level, *Nr contacts before the main diagnosis* is converted to "1 000" to indicate that no diagnosis is ever given. This number is decided to distinguish this value from values representing an actual number of contacts before the main diagnosis is given. The distribution of the number of contacts the cohort had before getting a main diagnosis on Axis 1 is visualised in Figure 6.53.

- **Including patients' care situations** Considering the feedback to include more information regarding the patients' life situations, the feature *Care Situation* was included in the third iteration. This was decided after investigating multiple possible patient features and concluding that this was the most insightful feature, including the least error-prone data. To include this feature, *Koder 7* is used. In the EoC Bundles where the St. Olavs value equals "0", the value is changed to "Missing data". The distribution of the cohort's care situations can be visualised in Figure 6.54.

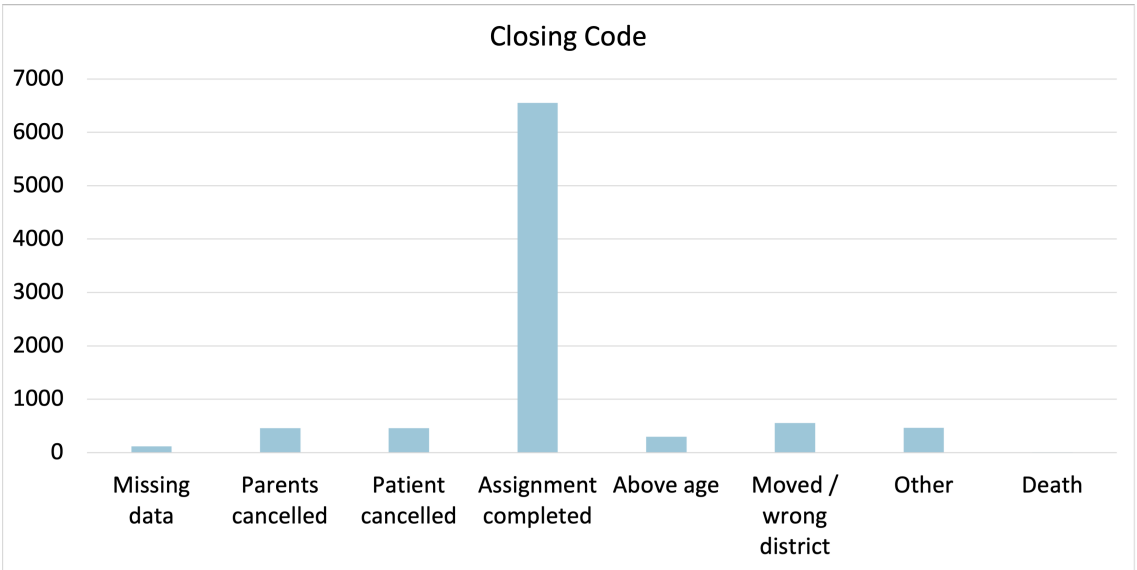


Figure 6.52: The distributions of the different closing codes.

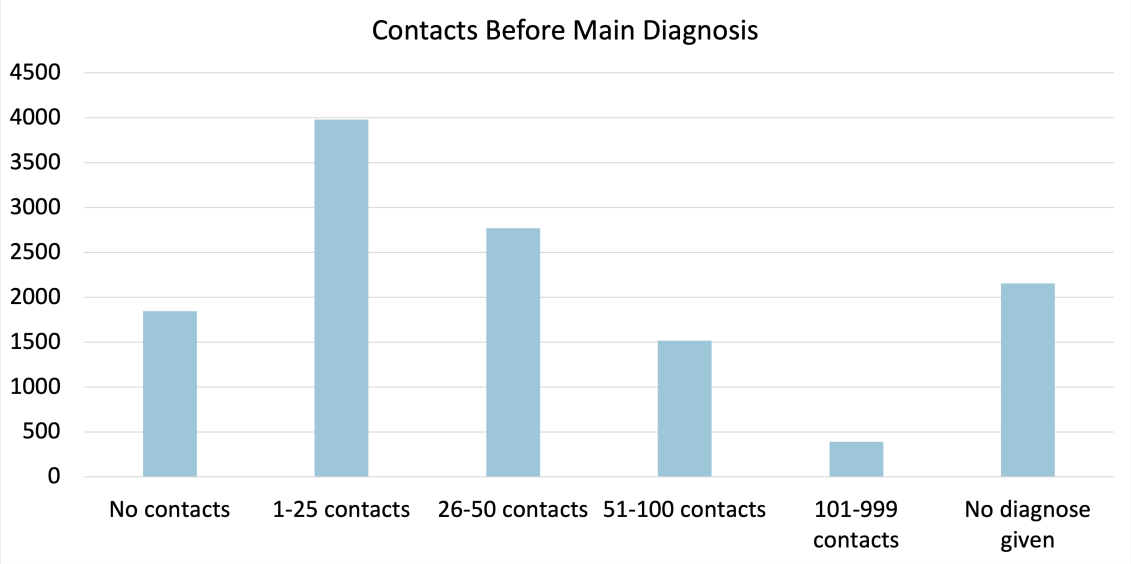


Figure 6.53: The distribution of the number of contacts a patient had before getting a main diagnosis on Axis 1.

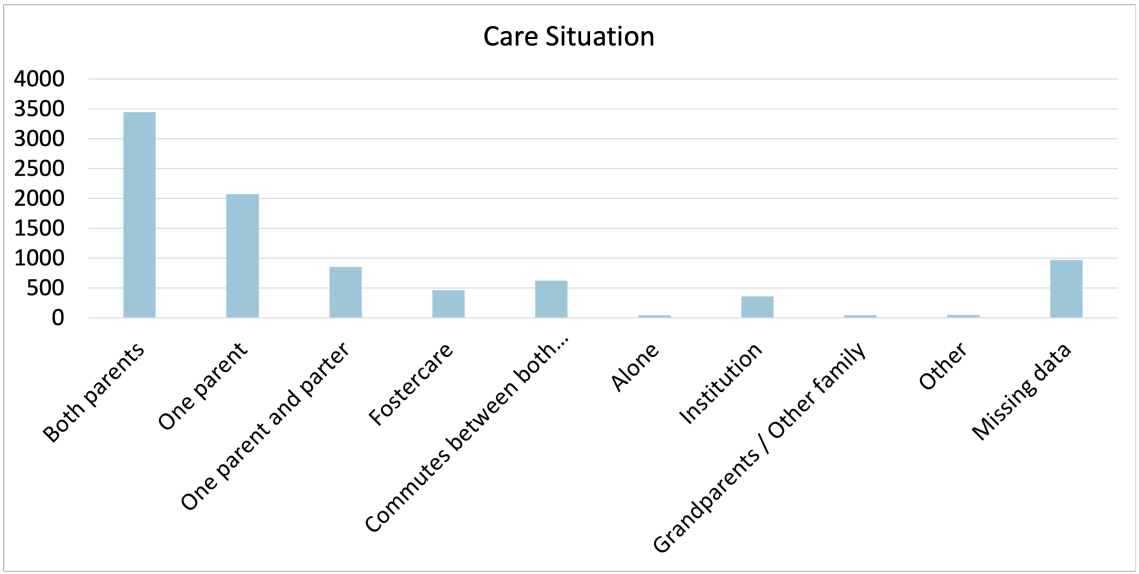


Figure 6.54: The distribution of the different care situations.

Tables 6.28 and 6.29 present the chosen features for the third iteration.

EoC Table - Third Iteration	
Feature	**Description of addition/change/removal**
EoC length	
Care level	
Immediacy level	
Nr. of therapy per day	
Nr. of examination per day	
Nr. of indirect contact per day	
Nr. of planning per day	
Nr. of no-shows per day	
Percentage of primary axis diagnoses	
Nr. of diagnoses on Axis 1 per day	
Nr. of diagnoses on Axis 2 per day	
Nr. of diagnoses on Axis 3 per day	
Nr. of diagnoses on Axis 4 per day	
Nr. of diagnoses on Axis 5 per day	
Nr. of diagnoses on Axis 6 per day	
Nr. of contacts before the main diagnosis is given	The number of contacts a patient had before the first primary diagnosis is given on Axis 1.

Table 6.28: EoC table features and feature description for the third iteration.

EoC Bundle Table - Third Iteration	
Feature	**Description of addition/change/removal**
Age at EoC Bundle start	
Gender	
Care situation	The patient's care situation, being one of the following: - Missing data - Both parents - Commutes between both parents - One parent - One parent and partner - Grandparents / other family - Fostercare - Institution - Alone - Other
EoC Bundle length	
Diagnosis Axis 1	
Diagnosis Axis 2	
Diagnosis Axis 3	
Diagnosis Axis 4	
Diagnosis Axis 5	
Diagnosis Axis 6	
Closing code	The reasons for closing a patient's EoC Bundle: - Missing data - Assignment completed - Patient cancelled - Parent(s) cancelled - Guardian(s) cancelled - Above age - Moved / wrong district - Death - Other

Table 6.29: EoC Bundle table features and feature description for the third iteration.

The selected features for the last iteration are then standardised using the same approach as the former two iterations. Code written for this data preparation is given in Appendix B.1.1.

Clustering

Using a similar approach as the two previous clustering iterations, the prepared data for the third iteration is clustered. The code written for the clustering is elaborated in Appendix B.1.2. The elbow plot in Figure 6.55 is generated by employing the Cao method as the initialisation technique for the k-prototypes algorithm and considering k values ranging from 1-10. The plot reveals that the elbow point is observed when k equals 3, indicating an optimal number of clusters for the third iteration's EoC data is 3.

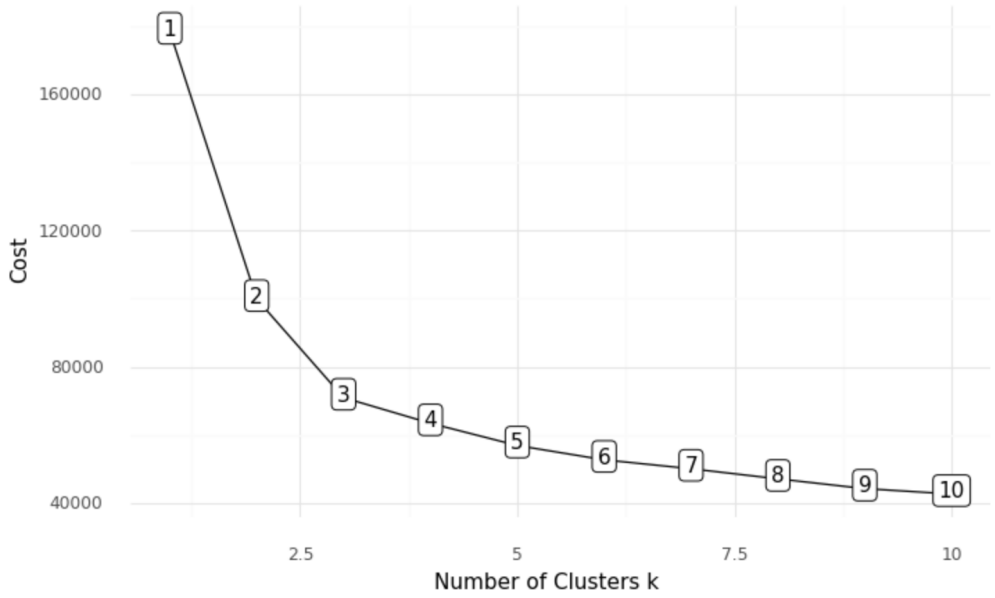


Figure 6.55: Third iteration's elbow method for the EoCs.

Using the k-prototypes algorithm, the EoC data is clustered into three groups: *EoC Type 0*, *EoC Type 1*, and *EoC Type 2*. These clusters consist of 872, 1 340 and 10 448 EoCs.

A SHAP plot explaining the impact and the importance of the EoC features on the EoC clustering result is visualised in Figure 6.56.

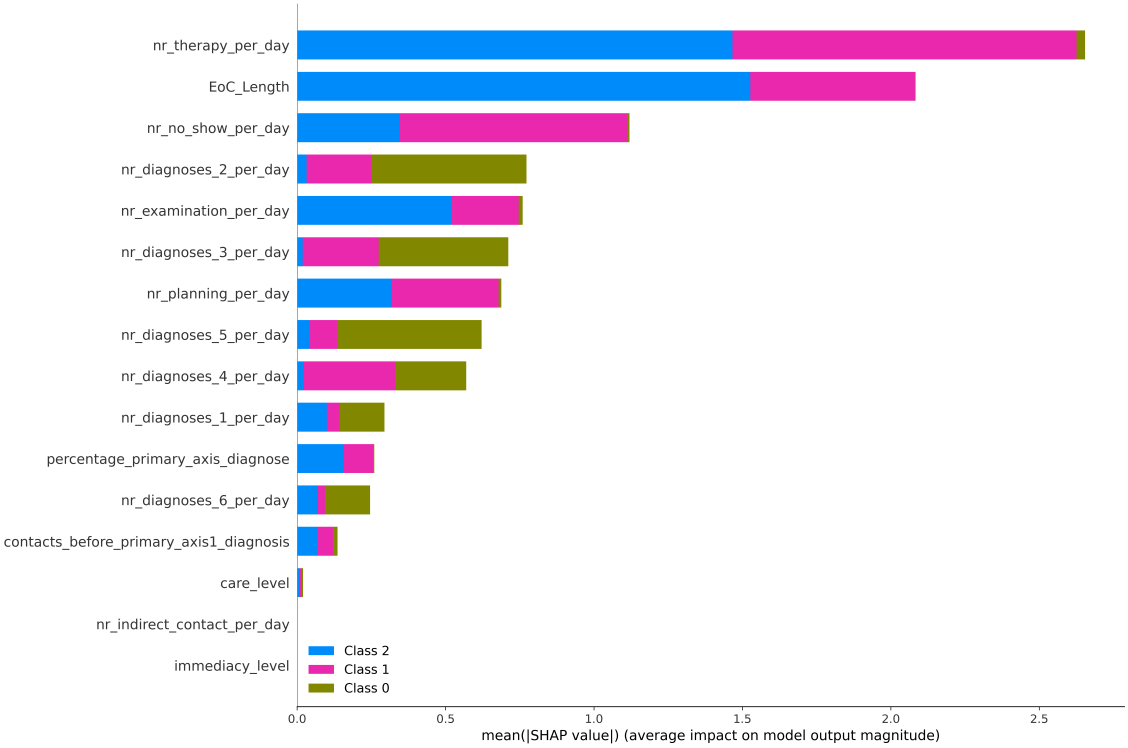


Figure 6.56: SHAP plot of the third iteration's EoC features.

Once the three subgroups of EoCs have been identified, the subgroups are incorporated into the EoC Bundle clustering and the Elbow method is used for the last time to find the optimal number of clusters. As indicated in the elbow plot in Figure 6.57, the optimal number of clusters for k-prototypes is when k equals 4.

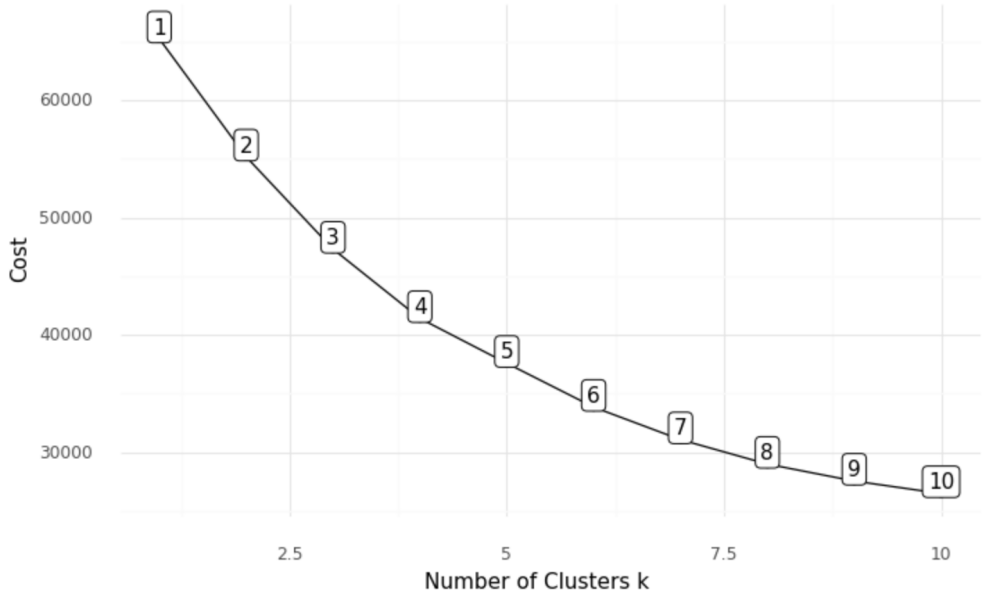


Figure 6.57: Third iteration's elbow method for the EoC Bundles.

Applying the k-prototypes algorithm to the EoC Bundle data, it is clustered into four groups: *EoC Bundle Type 0*, *EoC Bundle Type 1*, *EoC Bundle Type 2*, and *EoC Bundle Type 3*. These clusters comprise 4 031, 733, 3 503 and 636 EoC Bundles.

Finally, Figure 6.58 is included to inform the impact and the importance of the different features on the EoC Bundle clustering results.

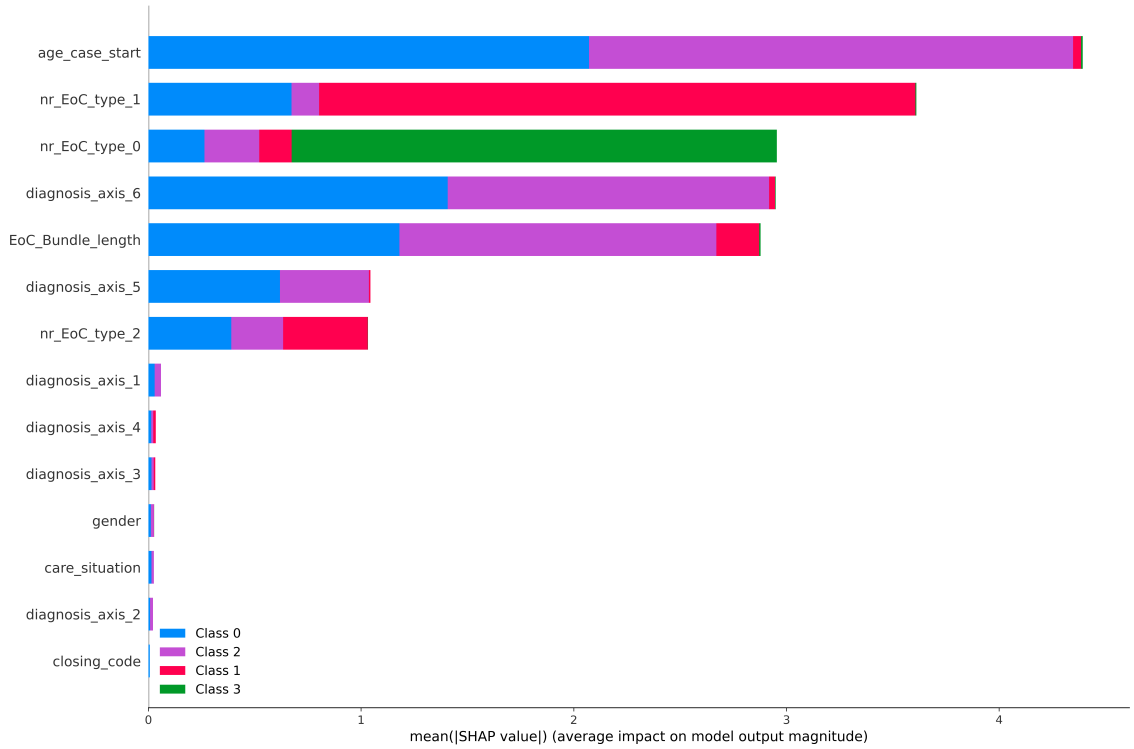


Figure 6.58: SHAP plot of the third iteration's EoC Bundle features.

For the third iteration, the findings are presented as the final experiment results in Chapter 7. These results are then evaluated more in detail, as presented in 8.

Chapter 7

Results

This chapter presents the results obtained from the third and final iteration, following the same approach as the previous two iterations. The code written to present these results and additional bar charts are presented in Appendices B.1.3 and C.3. The presentation of results begins with the EoC data clustering, followed by the results pertaining to the EoC Bundle clustering.

7.1 EoC Clustering Results

For the last iteration, the EoCs distribution in the identified clusters is presented in Table 7.1. Figures 7.1, and 7.2, 7.3 present the distributions of EoC lengths and care and immediacy levels

Clusters	Nr. Data Points
EoC Type 0	1 340
EoC Type 1	872
EoC Type 2	10 448

Table 7.1: Third iteration's distribution of EoCs in the EoC clusters.

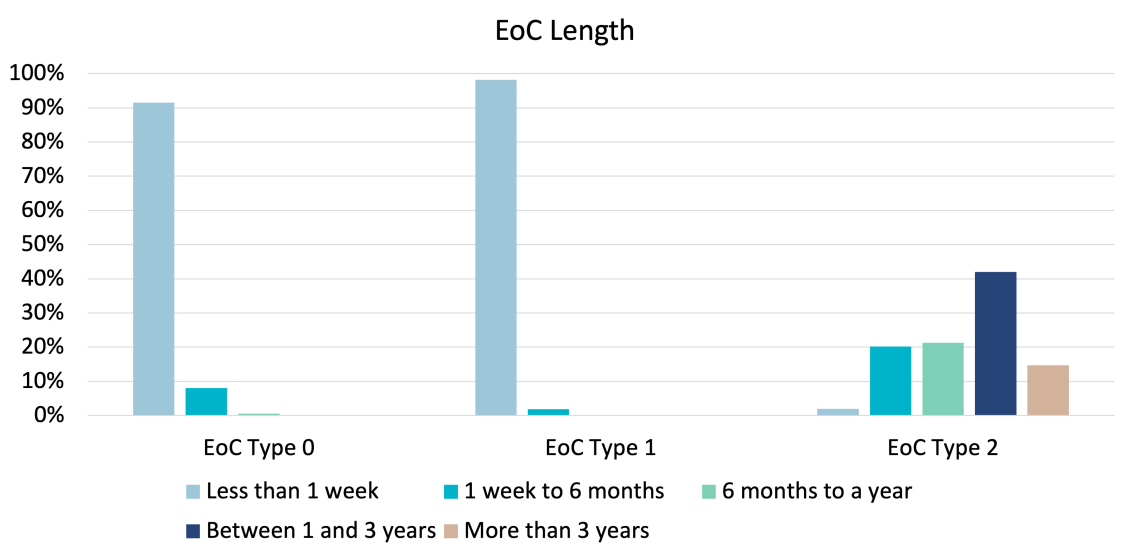


Figure 7.1: Third iteration's distribution of EoC lengths.

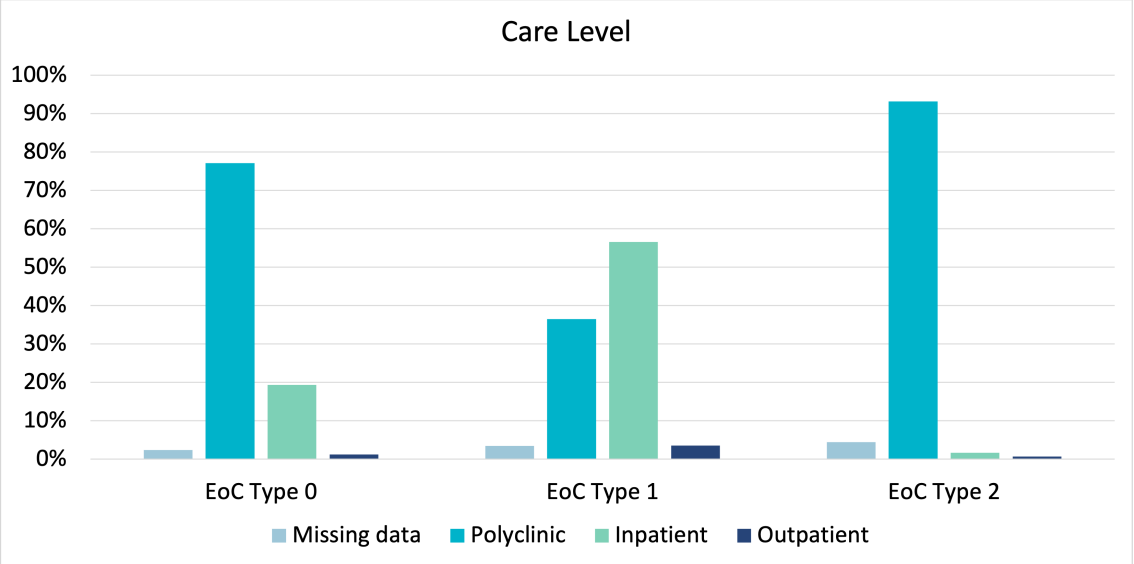


Figure 7.2: Third iteration's distribution of care levels.

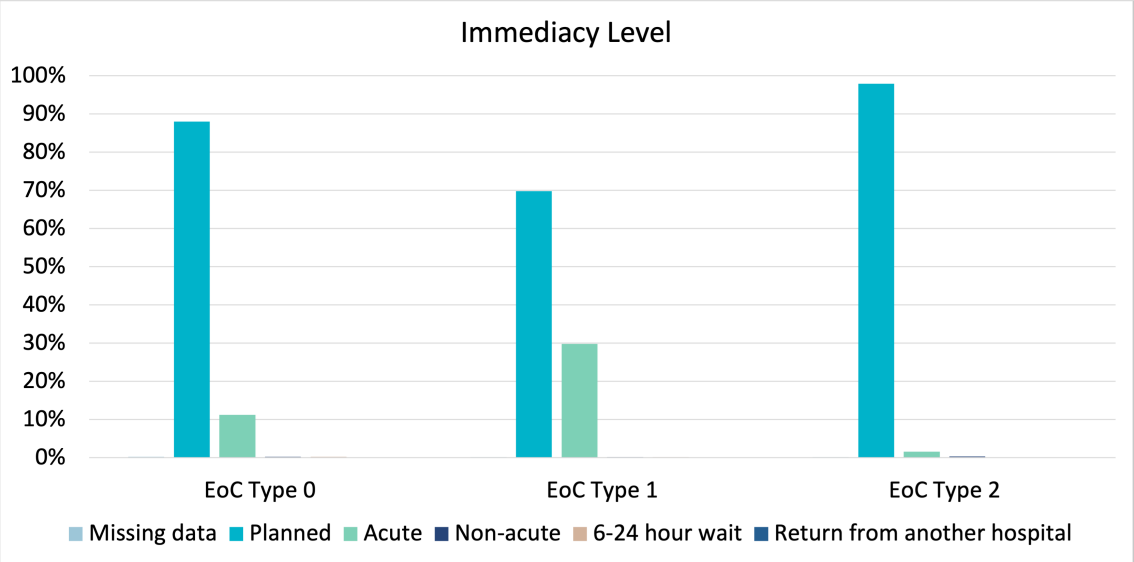


Figure 7.3: Third iteration's distribution of immediacy levels.

Table 7.2 presents the value distribution of the five different contact types for each EoC cluster.

Contact Feature	**Values**	**EoC Type 0**	**EoC Type 1**	**EoC Type 2**
Frequency of therapy contacts	Never.	14%	13%	5%
	Less than once a year.	0%	0%	2%
	Between once a year and once a month.	0%	0%	33%
	Between once a month and once a week.	0%	0%	**54%**
	Between once a week and once a day.	36%	21%	6%
	More than once a day.	**50%**	**66%**	0%
Frequency of planning contacts	Never.	13%	13%	5%
	Less than once a year.	0%	0%	2%
	Between once a year and once a month.	1%	1%	**47%**
	Between once a month and once a week.	3%	1%	43%
	Once a week and once a day.	34%	32%	3%
	More than once a day.	**49%**	**53%**	0%
Frequency of examination contacts	Never.	37%	23%	12%
	Less than once a year.	0%	0%	4%
	Between once a year and once a month.	0%	0%	**64%**
	Between once a month and once a week.	3%	1%	19%
	Between once a week and once a day.	**32%**	32%	1%
	More than once a day.	28%	**44%**	0%
Frequency of no-show contacts	Never.	21%	34%	9%
	Less than once a year.	0%	0%	5%
	Between once a year and once a month.	0%	0%	**74%**
	Between once a month and once a week.	7%	1%	11%
	Between once a week and once a day.	**41%**	**39%**	0%
	More than once a day.	31%	26%	0%
Frequency of indirect contacts	Never.	14%	13%	5%
	Less than once a year.	0%	0%	1%
	Between once a year and once a month.	0%	0%	34%
	Between once a month and once a week.	0%	0%	**54%**
	Between once a week and once a day.	36%	21%	6%
	More than once a day.	**50%**	**66%**	0%

Table 7.2: Third iteration's distribution of the frequency of the contact types.

Figure 7.5 visualises the feature *Nr contacts before the main diagnosis* added in the third iteration. Successive, Figure 7.4 presents the value distribution of the percentage of primary axis diagnoses. The frequency distribution of diagnoses on the six axes during an EoC is presented in 7.3.

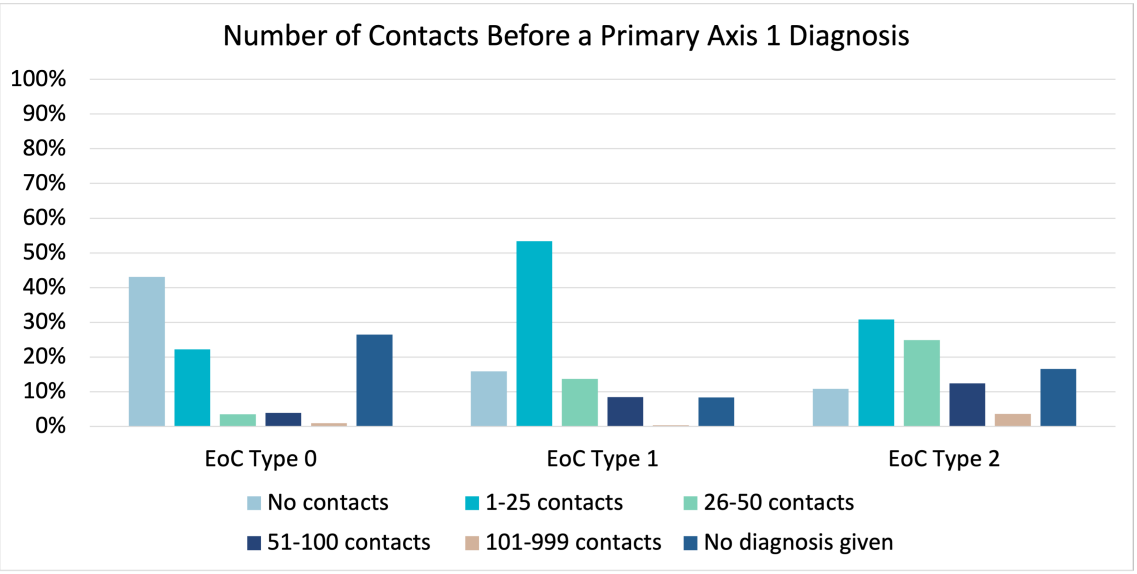


Figure 7.4: Third iteration's distribution of the number of contacts had before a primary diagnosis is given on Axis 1.

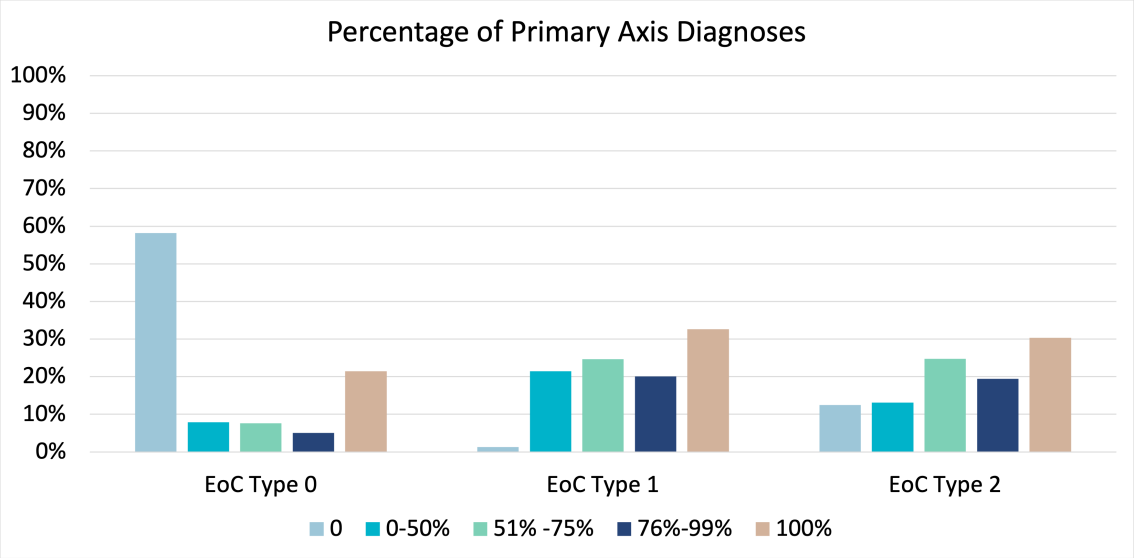


Figure 7.5: Third iteration's distribution of the percentage of diagnoses given as the primary diagnosis on one of the six axes.

Diagnostic Feature	Values	EoC Type 0	EoC Type 1	EoC Type 2
Frequency of diagnoses on Axis 1	Never.	**84%**	0%	13%
	Less than once a year.	0%	0%	35%
	Between once a year and once a month.	6%	0%	**51%**
	Between once a month and once a week.	2%	2%	1%
	Between once a week and once a day.	7%	**80%**	0%
	More than once a day.	1%	18%	0%
Frequency of diagnoses on Axis 2	Never.	**93%**	1%	18%
	Less than once a year.	0%	0%	**45%**
	Between once a year and once a month.	7%	1%	37%
	Between once a month and once a week.	0%	1%	0%
	Between once a week and once a day.	0%	**90%**	0%
	More than once a day.	0%	7%	0%
Frequency of diagnoses on Axis 3	Never.	**93%**	1%	18%
	Less than once a year.	0%	0%	**46%**
	Between once a year and once a month.	7%	1%	36%
	Between once a month and once a week.	0%	1%	0%
	Between once a week and once a day.	0%	**92%**	0%
	More than once a day.	0%	5%	0%
Frequency of diagnoses on Axis 4	Never.	**84%**	0%	12%
	Less than once a year.	0%	0%	35%
	Between once a year and once a month.	6%	0%	**51%**
	Between once a month and once a week.	2%	2%	1%
	Between once a week and once a day.	8%	**80%**	0%
	More than once a day.	1%	19%	0%
Frequency of diagnoses on Axis 5	Never.	**92%**	2%	19%
	Less than once a year.	0%	0%	**40%**
	Between once a year and once a month.	6%	0%	**40%**
	Between once a month and once a week.	1%	2%	1%
	Between once a week and once a day.	1%	**78%**	0%
	More than once a day.	0%	18%	0%
Frequency of diagnoses on Axis 6	Never.	**92%**	6%	20%
	Less than once a year.	0%	0%	**42%**
	Between once a year and once a month.	7%	1%	38%
	Between once a month and once a week.	0%	1%	0%
	Between once a week and once a day.	1%	**85%**	0%
	More than once a day.	0%	7%	0%

Table 7.3: Third iteration's distribution of the frequency of diagnoses on the different axes.

Table 7.4 presents the categorical features' modes and the numerical features' medians and means on the EoC level.

Feature	Meassure	EoC Type 0	EoC Type 1	EoC Type 2
EoC lenght	Mean	10	4	602
	Median	0	3	455
Care level	Mode	Polyclinic	Inpatient	Polyclinic
Immediacy level	Mode	Planned	Planned	Planned
Nr. of therapy contacts per day	Mean	4.00	4.47	0.05
	Median	1.02	2.5	0.04
Nr. of planning contacts per day	Mean	2.22	2.34	0.04
	Median	1.0	1.29	0.03
Nr. of examination contacts per day	Mean	1.40	2.07	0.03
	Median	0.51	1.0	0.01
Nr. of no-show contacts per day	Mean	1.28	0.97	0.032
	Median	1.0	0.3	0.01
Nr. of indirect contacts per day	Mean	3.0	4.47	0.05
	Median	1.03	2.5	0.04
Nr. contacts before primary Axis 1 diagnosis	Mean	272	99	193
	Median	4	10	32
Percentage of primary axis diagnoses	Mean	34%	73%	69%
	Median	0%	83%	75%
Nr. of diagnoses on Axis 1 per day	Mean	0.07	1.04	0.01
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 2 per day	Mean	0.0	0.63	0.0
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 3 per day	Mean	0.0	0.65	0.0
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 4 per day	Mean	0.0	0.65	0.0
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 5 per day	Mean	0.01	1.06	0.0
	Median	0.0	0.5	0.0
Nr. of diagnoses on Axis 6 per day	Mean	0.01	0.63	0.0
	Median	0.0	0.5	0.0

Table 7.4: Third iteration's EoC feature measurements.

An overview of the dominating feature values for each of the identified EoC clusters is presented in Table 7.5.

EoC Type 0	**EoC Type 1**	**EoC Type 2**
872 EoCs	1 340 EoCs	10 448 EoCs
• Shorter than a week. • Polyclinic. • Planned. • All contacts mostly multiple times per day, or some weekly to daily. • Either starting with a diagnosis or never being given one. • Seldom given a diagnosis on any axis.	• Shorter than a week. • Inpatient. • Planned or acute. • All contacts (except no-show) multiple times a day. • Main diagnosis is normally given with between 1-25 contacts. • Diagnoses given on all axes weekly to daily.	• Longer than a week. • Polyclinc. • Planned. • Mostly contacts yearly to monthly. • Varying number of contacts before the first main diagnosis is set. • Diagnoses given on all axes less than once a year or between once a year and once a month.

Table 7.5: Third iteration's EoC clusters summary.

7.2 EoC Bundle Clustering Results

The four identified EoC Bundle clusters for the third iteration are presented in Table 7.6.

EoC Bundle Cluster	**Nr. Data Points**
EoC Bundle Type 0	636
EoC Bundle Type 1	733
EoC Bundle Type 2	3 503
EoC Bundle Type 3	4 031

Table 7.6: Third iteration's distribution of EoC Bundles in the EoC Bundle clusters.

Figure 7.6 presents the distribution of EoC Bundle lengths. Then figure 7.7, 7.8, and 7.9 visualises the value distributions of the features *age*, *gender*, and *care situation*.

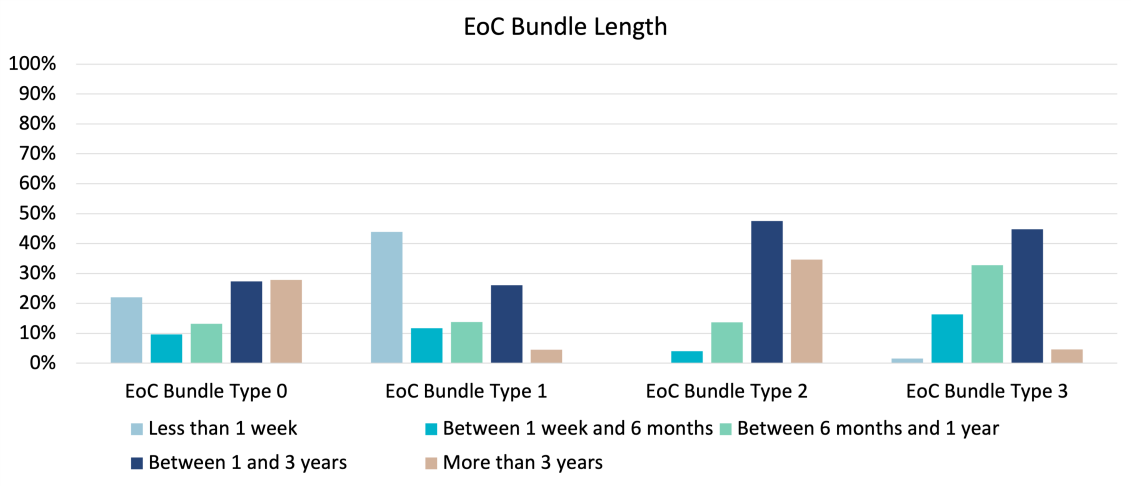


Figure 7.6: Third iteration's distribution of EoC Bundle lengths.

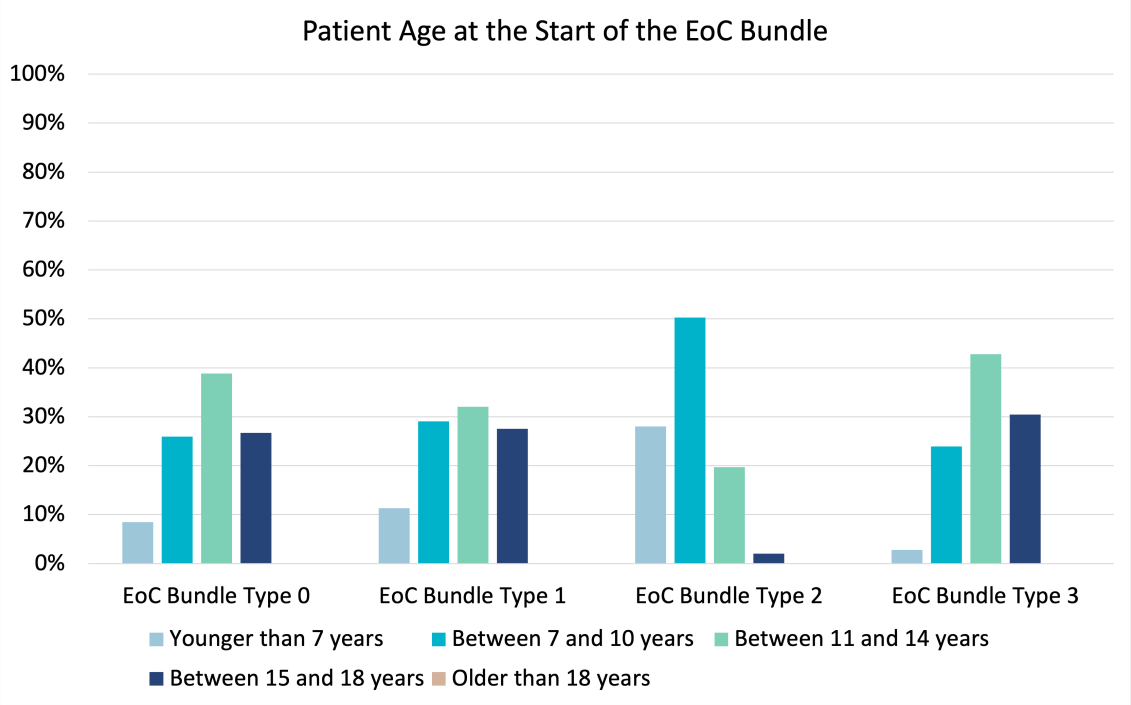


Figure 7.7: Third iteration's distribution of patients' age.

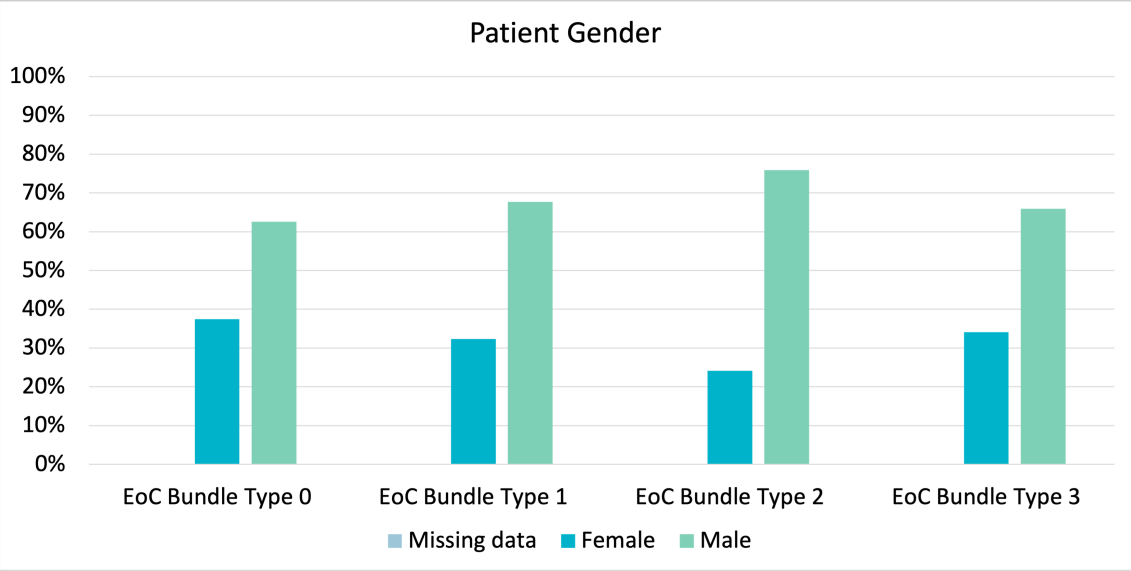


Figure 7.8: Third iteration's distribution of patients' gender.

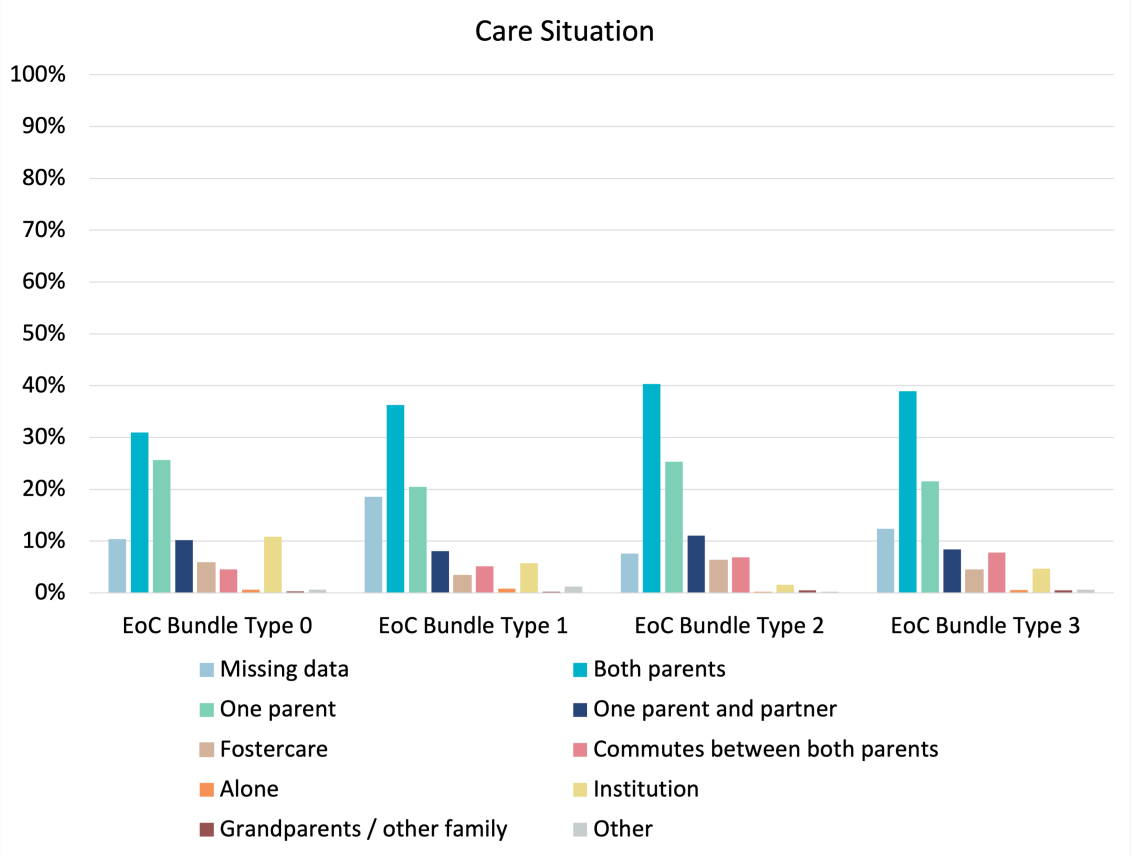


Figure 7.9: Third iteration's distribution of patients' care situation.

The diagnostic information on the EoC Bundle level is visualised in Figures 7.10, 7.11, and 7.12.

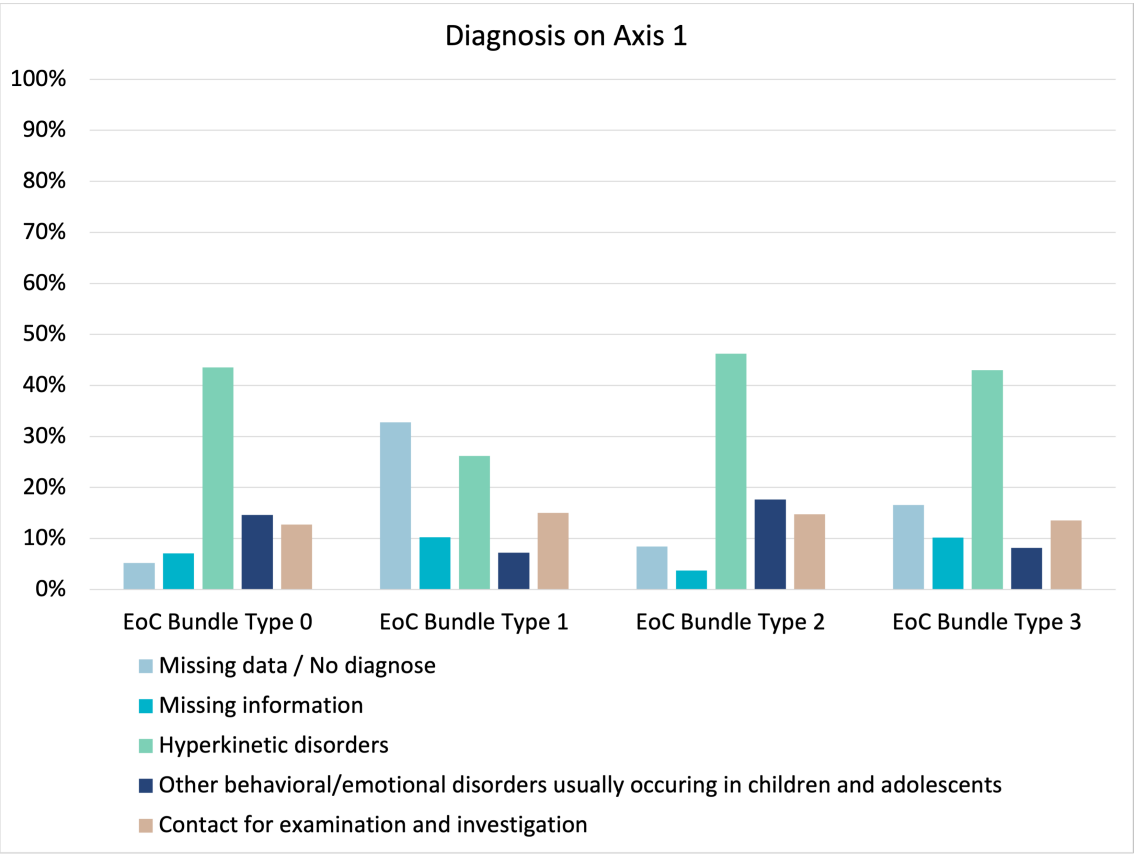


Figure 7.10: Third iteration's distribution of diagnoses on Axis 1 at the beginning of an EoC Bundle.

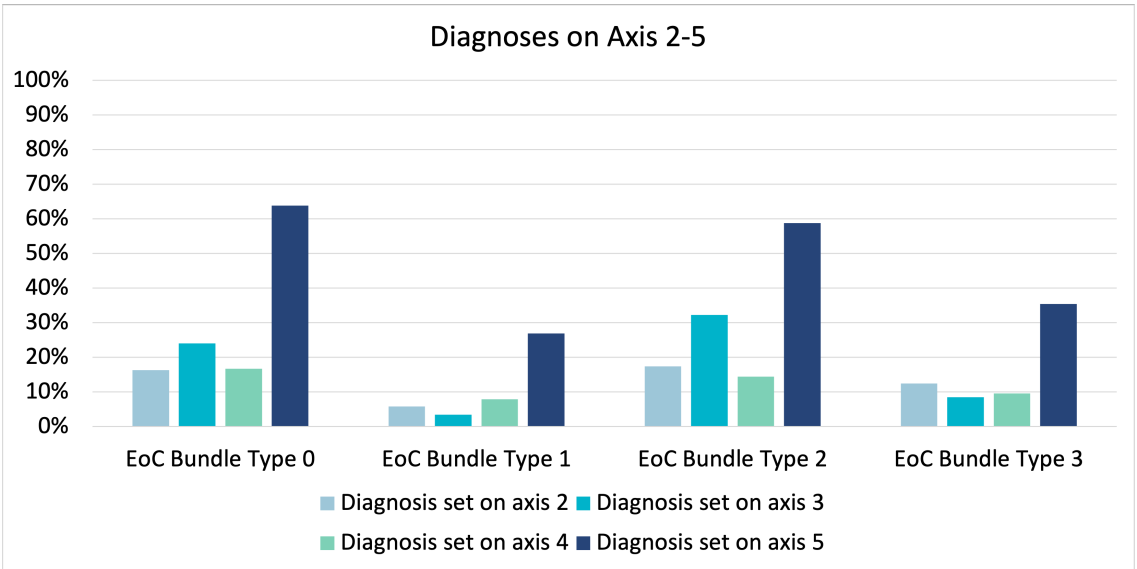


Figure 7.11: Third iteration's distribution of diagnoses on axes 2-5 at the beginning of an EoC Bundle.

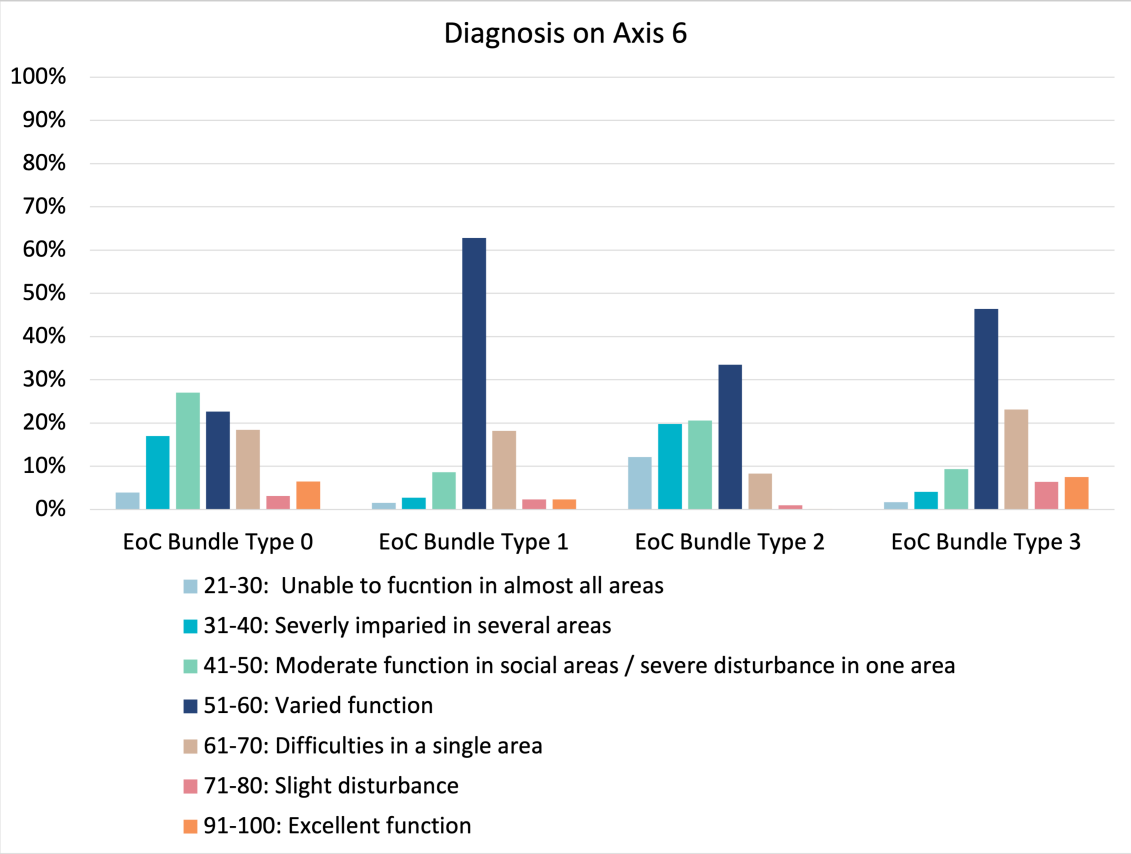


Figure 7.12: Third iteration's distribution of diagnoses on Axis 6 at the beginning of an EoC Bundle.

The value distribution of the number of EoCs of the three EoC types identified in the third iteration is presented in Figures 7.13, 7.14, and 7.15.

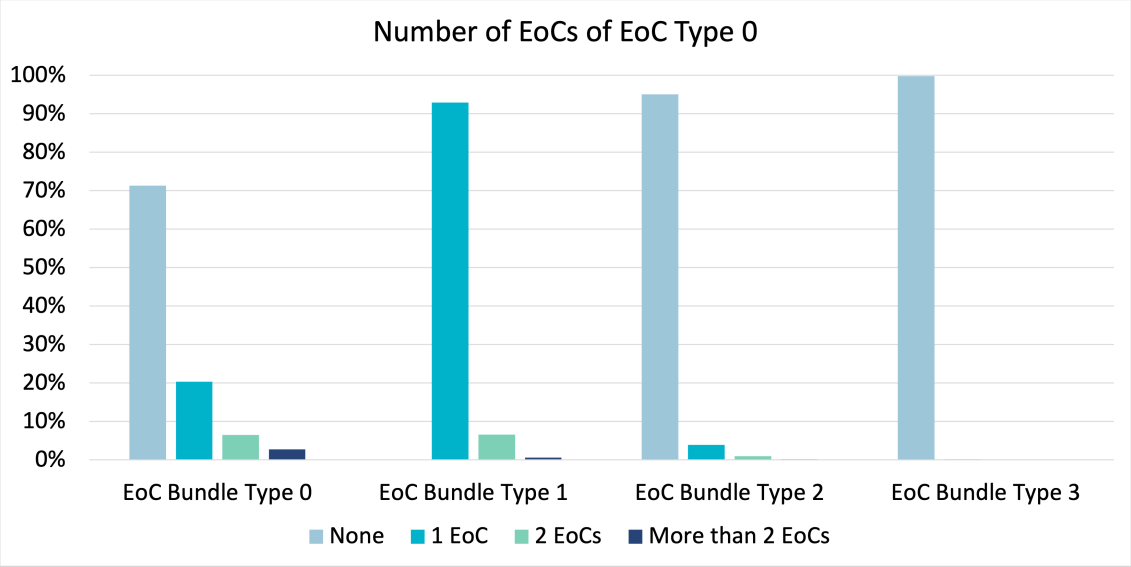


Figure 7.13: Third iteration's distribution of the number of EoCs of Type 0.

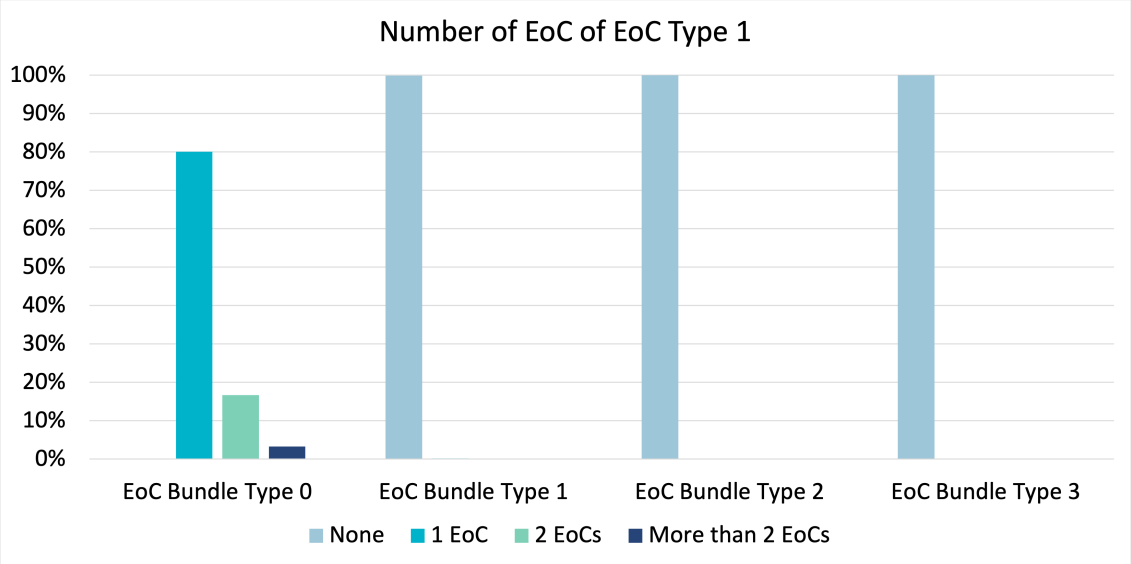


Figure 7.14: Third iteration's distribution of the number of EoCs of Type 1.

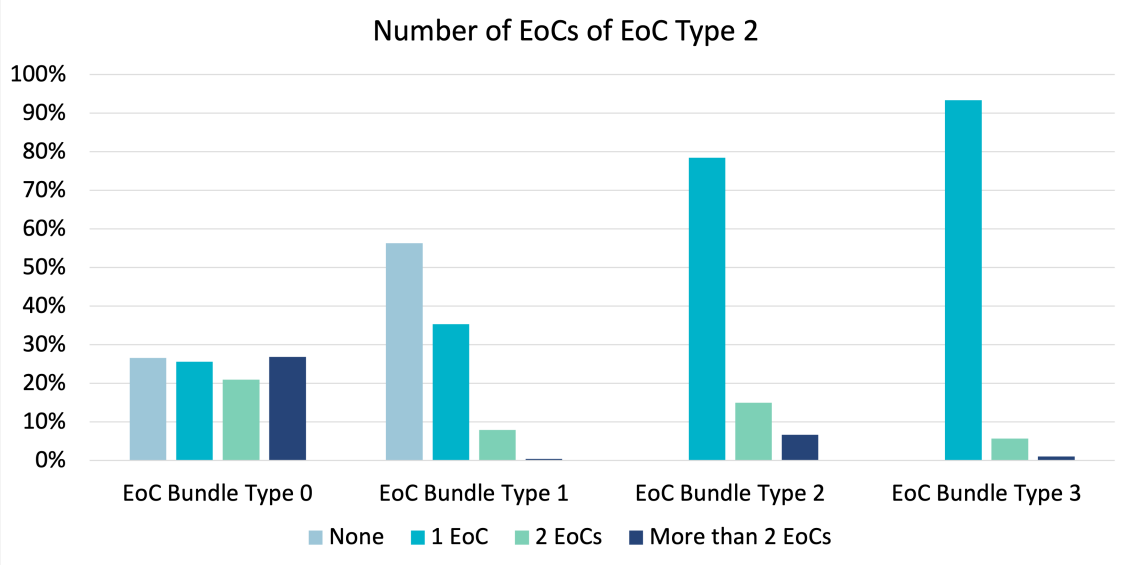


Figure 7.15: Third iteration's distribution of the number of EoCs of Type 2.

Lastly, the distribution of closing codes of the EoC Bundles is presented in Figure 7.16

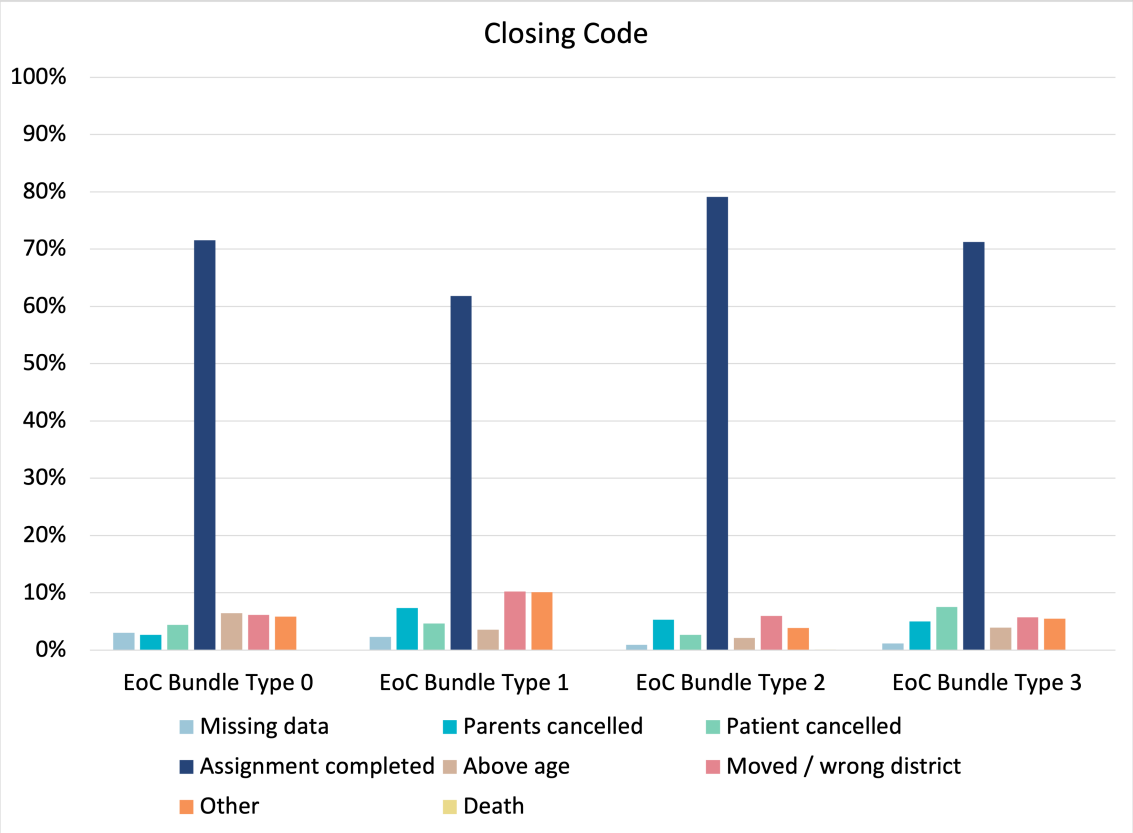


Figure 7.16: Third iteration's distribution of the closing codes.

For all EoC Bundle features the modes, medians, and means are calculated and presented in Table 7.7.

Feature	**Meassure**	**EoC Bundle Type 0**	**EoC Bundle Type 1**	**EoC Bundle Type 2**	**EoC Bundle Type 2**
EoC Bundle lenght	Mean	760	275	997	437
	Median	465	92	821	365
Age at EoC Bundle start	Mean	12	11	8	12
	Median	12	12	8	13
Patient Gender	Mode	Male	Male	Male	Male
Care situation	Mode	Both parents	Both parents	Both parents	Both parents
Diagnosis on Axis 1	Mode	Hyperkinetic disorders	Missing data	Hyperkinetic disorders	Hyperkinetic disorders
Diagnosis on Axis 2	Mode	No	No	No	No
Diagnosis on Axis 3	Mode	No	No	No	No
Diagnosis on Axis 4	Mode	No	No	No	No
Diagnosis on Axis 5	Mode	Yes	No	Yes	No
Diagnosis on Axis 6	Mean	4.7	5.1	3.9	5.5
	Median	5	5	4	5
Nr. EoC of EoC type 0	Mean	1.4	0.0	0.0	0.0
	Median	1	0	0	0
Nr. EoC of EoC type 1	Mean	0.5	1.1	0.1	0.0
	Median	0	1	0	0
Nr. EoC of EoC type 2	Mean	1.8	0.5	1.3	1.1
	Median	1	0	1	1
Closing code	Mode	Assignment completed	Assignment completed	Assignment completed	Assignment completed

Table 7.7: Third iteration's EoC Bundle feature measurements.

A summarising table describing the key feature distributions in the third iteration's four EoC Bundle clusters is presented in Table 7.8. Following the table is a reminder of important factors in the identified EoC types.

EoC Bundle Type 0 636 EoC Bundles	**EoC Bundle Type 1** 733 EoC Bundles
• All lengths. • Mostly older than 7 years. • 65:35 male-to-female ratio. • Both or one parent. • Hyperkinetic disorders on Axis 1. • Mostly no diagnoses on axes 2-5. • Moderate function in social areas / severe disturbance in one area. • Few have an EoC of type 0. • One or sometimes more EoC of type 1. • One or often more EoC of type 2. • Assignment completed.	• All lengths. • Mostly older than 7 years • 70:30 male-to-female ratio. • Both or one parent. • Missing data / no diagnoses or hyperkinetic disorders on Axis 1. • Mostly no diagnoses on axes 2-5. • Varied function. • One EoC of EoC type 0. • No EoCs of EoC type 1. • Some have an EoC of type 2. • Assignment completed.
EoC Bundle Type 2 3 503 EoC Bundles	**EoC Bundle Type 3** 4 031 EoC Bundles
• Mostly longer than a year. • Mostly younger than 10 years. • 75:25 male-to-female ratio. • Both or one parent. • Hyperkinetic disorders on Axis 1. • Mostly no diagnoses on axes 2-5. • Moderate function in social areas / severe disturbance in one area. • No EoC of type 0. • No EoC of type 1. • One or sometimes more EoCs of type 2. • Assignment completed.	• Mostly longer than six months. • Mostly older than 7 years. • 65:35 male-to-female ratio. • Both or one parent. • Hyperkinetic disorders. • Mostly no diagnoses on axes 2-4 and only some on Axis 5. • Varied function. • No EoC of type 0. • No EoC of type 1. • One EoC of type 2. • Assignment completed.

Table 7.8: Third iteration's EoC Bundle clusters summary.

- **EoC type 0:** Shorter than a week, planned polyclinic EoCs with contacts weekly to daily. Either starting with or never getting a main diagnosis and seldom diagnoses given on the axes.
- **EoC type 1:** Shorter than a week, planned or acute inpatient EoCs with many contacts daily. The main diagnosis is given before 25 contacts are had, and diagnoses are given weekly to daily.
- **EoC type 2:** Longer than a week, planned polyclinic EoCs with contacts yearly to monthly and diagnoses are given less than once a year.

Chapter 8

Evaluation

This chapter aims at evaluating the clustering performed in this project, assessing the effectiveness, validity, and reliability of the conducted research. The evaluation process begins by analysing the cluster partitioning of the final results using quantitative and objective measures. Subsequently, a result evaluation is presented to determine the degree to which the experimental aims have been achieved. Finally, a clinical evaluation is conducted to complement the quantitative analysis and provide a more thorough evaluation of the clustering performance.

8.1 Clustering Validation

Cluster validation refers to formal procedures that evaluate cluster analysis results quantitatively and objectively, and it is recognised as a vital issue essential to the success of the clustering application (Liu et al., 2010; Vazirgiannis, 2009). For this project, clustering validity is crucial for two reasons. Firstly, the absence of prior knowledge of partitions and structures in this project's data means the absence of a solid reference point to judge the quality of the clustering model (Burkov, 2019). Secondly, most clustering algorithms behave significantly differently depending on the selected features and the initial assumptions for defining the partitions (Halkidi et al., 2002). Therefore, validating the completed clustering process is essential to secure the quality of the results.

To validate the clustering, one separates between *external*, *internal*, and *relative* cluster validity. External validation implies comparing a clustering algorithm's results with a pre-specified data structure (Halkidi et al., 2002). Since the obtained clustering structure was unknown before the experiment was conducted, this validation approach was irrelevant to this project. Internal cluster validation evaluates whether the clustering structure produced by a clustering algorithm fits the data without reference to any external information. Relative validation compares different clustering structures resulting from the same algorithm but with different parameter values (Theodoridis and Koutroumbas, 2008). For this project, internal and relative clustering validation is completed.

An internal cluster validation technique used was measuring the data's clustering tendency. This was done to ensure that the data were predisposed to cluster into natural groups without identifying the groups themselves. This precaution was made because clustering algorithms tend to find clusters in the data irrespective of whether or not any clusters are present. To measure the clustering tendency, *Hopkins statistics* was used. Hopkins statistics is a well-known estimator of randomness in a data set (Banerjee and Davé, 2004). To this date, no Python implementation of Hopkins statistics directly supports mixed data. Therefore, the *get dummies* function provided by *Pandas* was leveraged to convert categorical values into dummy or indicator variables (Pandas, n.d.). From this function, the Python toolkit *pyclustertend* was utilised to calculate the data's Hopkins scores for the EoC and EoC Bundle data across all three iterations (pyclustertend, n.d.). These scores are presented in Table 8.1 and Appendix B.2 presents the implementation. A score gravitating to 0 indicates that the data has a high cluster tendency, while a score above 0.3 indicates that the data has a low cluster tendency (Banerjee and Davé, 2004). Given the low Hopkins scores obtained in all three iterations for both the EoC and EoC Bundle data, it can be inferred that the data exhibits natural clusters. Hence, using clustering to facilitate the predictive analysis based on groupings is reasonable from a technical perspective.

Iteration	**EoC Data's Hopkins Score**	**EoC Bundle Data's Hopkins Score**
First iteration	0.018	0.008
Second iteration	0.008	0.003
Third iteration	0.013	0.012

Table 8.1: Assessing the clusterability of the data by calculating the Hopkins scores.

Internal and relative clustering validation may be used separately or in combination to evaluate the optimal number of clusters. This evaluation is important since the number of clusters (k) largely impacts the quality of the clustering results. If the wrong number of clusters is assigned to the algorithm, the clustering results in a partitioning scheme that is not optimal, which may lead to the wrong grouping of objects (Halkidi et al., 2002). This project used the well-known Elbow method to find and evaluate the optimal number of clusters (Burkov, 2019). The Elbow method evaluates the sum of the square distances between the data points and the cluster centroids as a function of the total number of clusters. As the number of clusters increased, the percentage of within-cluster dispersion decreased. This was visualised using plots, and the optimal cluster number was chosen where the graphs changed from rapidly decreasing to more or less being parallel with the x-axis (Tibshirani et al., 2001). The different elbow representations of finding an optimal k in the experiment (referring to the Figures 6.13, 6.15, 6.35, 6.37, 6.55, and 6.57) are shaped more or less as elbows, verifying the clustering tendency indicated by the Hopkins score (Theodoridis and Koutroumbas, 2008).

Various internal clustering validation techniques can be used to validate obtained cluster partitioning. However, since these are mostly based on statistical tests, the mathematical operations cannot be applied directly to mixed data (Aschenbruck and Szepannek, 2020). Hence, these internal validation techniques were not applied in this project.

8.2 Result Evaluation

The following section assesses the experimental aims to evaluate the findings from the three clustering iterations. Further details regarding the findings' interpretation, implications, and limitations are presented in Section 9.2.4.

1. Assess the feasibility of clustering for identifying patient trajectory subgroups

This first experimental aim entails the ability to use a clustering approach to identify subgroups. As detailed in the subsequent evaluation, the application of clustering successfully yielded subgroups of patient trajectories highlighting both similarities and differences. Additionally, as described in Section 8.1, the evaluation confirmed the data's distinguishability.

Through the clustering process, subgroups were identified, indicating the achievement of the first aim. After three iterations, the features exhibited reduced redundancy, leading to more insightful results. Notable, altering certain categorical features to numerical features and modifying the categorisation of some categorical features resulted in a clearer clustering outcome. This demonstrates that the iterative approach and the repetitive feature selection enhanced the feasibility of identifying patient trajectory subgroups through clustering.

2. Identify subgroups of EoCs that have similar characteristics

The three iterations could all distinguish the EoCs into three EoC subgroups. In the first iteration, the clustering process grouped the EoCs into three relatively equal-sized subgroups compared to the last two iterations. However, the EoCs in the first iteration were impacted by the EoC lengths, which resulted in less informative results. The last two iterations grouped EoCs in three unequally sized subgroups. However, these results were more insightful when interpreted by clinicians. Compared to the results from the first iteration, the results obtained in the last two iterations can be considered better as unnecessary features were removed.

In the third iteration, the clustering identified three subgroups that consisted of 1 340, 872 and 10 448 EoCs. The distributions of various feature values distinguish these subgroups. In short, the EoC subgroups identified had the following characteristics. The first subgroup, EoC Type 0, is characterised by shorter, planned, polyclinic EoCs dominated by high-frequency contacts but few diagnoses. The second subgroup, EoC Type 1, is also characterised by shorter lengths. However, these EoCs are often inpatient, and for some EoCs, the immediacy level is "acute". These EoCs also include frequent contacts, and diagnoses are given on each axis weekly to daily. The third subgroup, EoC Type 2, largely varies from the prior two by being longer, planned, polyclinic EoCs. The frequency of contacts and diagnoses for these EoCs is much lower. Further interpretations of the EoC clusters are presented in the result discussion in Section 9.2.4.

Similarities can be observed within the three EoC subgroups identified in the final results. However, it is noteworthy that one of these subgroups was larger than the other two. This led to a comparatively less comprehensive representation of this larger subgroup's details. To address this limitation, feature modifications were made during the final iteration to obtain a more comprehensive outcome differentiating the EoCs within this subgroup. Unfortunately, this objective was not accomplished. Further insights regarding the reason for this outcome are presented in the result discussion.

3. Identity subgroups of EoC Bundles that have similar characteristics

Throughout all three iterations, distinguishable subgroups of EoC Bundles were successfully identified. Three EoC Bundle subgroups were identified in the first iteration, but the subgroups did not reveal many informative differences. More interesting results emerged in the last two iterations, with four distinct EoC Bundle subgroups identified. The outputs in these two iterations were quite similar.

In the final iteration, the resulting EoC Bundle subgroups comprise 636, 733, 3 503 and 4 031 EoC Bundles. Several features can distinguish these subgroups. In short, the key distinguishing features in the cluster subgroups are the following:

- **EoC Bundle Type 0:** These EoC Bundles exhibit varying lengths, predominantly involving patients older than 7 years with "Moderate function in social areas". This subgroup has the highest percentage of females. The EoC Bundles within this subgroup typically include at least one EoC of type 1 (short, inpatient and sometimes acute EoCs with frequent contacts and diagnoses) and often one or more EoCs of type 2 (longer, planned polyclinic EoCs with low frequency of contacts and diagnoses).
- **EoC Bundle Type 1:** This subgroup primarily consists of patients older than 7, with many EoC Bundles where "Missing data" or "No diagnosis" on the first axis. The EoC Bundles in this subgroup always include at least one EoC of type 0 (short, planned polyclinic EoCs with frequent contacts but low frequency of diagnoses) and sometimes an EoC of type 2.
- **EoC Bundle Type 2:** This subgroup includes mostly patients younger than 14 and exhibits the highest percentage of males. This subgroup also has the highest amount of "Severely impaired patients in many social areas". Almost all EoC Bundles of type 2 only include one EoC of type 2.
- **EoC Bundle Type 3:** This subgroup includes the largest percentage of patients older than 11 and a CGAS score above 60. These EoC Bundles exclusively include EoCs of type 2.

Considering these distinguishing features, it is possible to differentiate between the various EoC Bundle subgroups. However, one should note that the EoC Bundle clusters are also affected by the large EoC subgroup, making it challenging to distinguish EoC Bundle subgroups, including this EoC type, on a detailed level. Despite this, it is evident that subgroups of EoC Bundles with similar characteristics have been identified. Therefore, the overall findings demonstrate the successful identification of EoC Bundle subgroups exhibiting similar characteristics.

4. Identify similarities in patient characteristics.

The patient characteristics, age and gender, were included in all three iterations. In the initial iteration, the gender and age distribution among the cluster were almost identical. The distributions were similar to the overall cohort description provided in Section 6.3. However, in the second and third iterations, the age and gender distribution were more varied.

Upon consultation with clinicians, the initial assumption was that patient trajectories would be noticeably differentiated based on gender and age. However, this assumption did not hold, which is an interesting result. It suggests that the patients' age and gender do not substantially indicate distinct trajectories.

The *Care situation* feature was added in the last iteration in the hope of detailing more patient information. However, the value distribution of this feature is not showing large differences. This result is also intriguing, as it reveals that the care situation perhaps is not key in identifying patient trajectories.

Overall, while similarities regarding patient characteristics were found, they emerged differently than anticipated, and there were limited distinguishing differences between the EoC Bundle subgroups. These findings highlight that age, gender, and care situation might not hold significant predictive power when identifying patient trajectories.

5. Identify commonalities based on key characteristics defining the EoCs and EoC Bundles

Key characteristics defining EoCs and EoC Bundles include EoC length, care level, immediacy level, EoC Bundle length, and diagnoses on the six axes at the beginning of an EoC Bundle. The EoC length significantly impacted the results in the first iteration, rendering the other features less influential. This can be visualised in the first iteration's EoC SHAP plot in Figure 6.14. The commonalities regarding care and immediacy level were much clearer in the second and third iterations. One subgroup had a much higher percentage of acute inpatient EoCs, while the other two consisted of planned polyclinic EoCs.

The EoC Bundle lengths were distinguishable in all three iterations. The average EoC Bundle length in the final iteration varied from 365 to 997 days. There might be a relation between the EoC lengths and the CGAS score given on Axis 6. Upon examining the means, the EoC Bundle subgroup with the longest EoC Bundles tends to have the EoC Bundles with the lowest CGAS score. However, evaluating the EoC Bundle length in conjunction with the diagnoses becomes challenging. Referencing the discussion section will provide further insights into the potential implications of EoC length.

In the first iteration, all diagnoses given on the six axes were included as categorical values. Although this approach provided more detailed diagnostic findings than the subsequent iterations, the details lacked informative value, and several diagnostic codes had a low occurrence. Therefore, a modification in the second iteration specifies whether axes 2-5 have diagnoses rather than which specific diagnoses are given. This adjustment resulted in a clearer separation between EoC Bundles with or without diagnoses and revealed a trend of EoC Bundles not having diagnoses on axes 2-4. The diagnosis on Axis 6 at the beginning of an EoC Bundle was converted to a numerical value after the first iteration, facilitating a more realistic comparison. In iterations two and three, these features greatly impacted the model output. The *Diagnosis Axis 6* feature's impact on each of the three iterations' model output can be visualised in the following plots: Figures 6.16, 6.38, and 6.58. However, it is worth noting that changing all missing values to "5" in these iterations made it more challenging to find true commonalities and differences among EoC Bundles, as many values ended up being "5" on Axis 6.

6. Identify similarities related to trajectory actions

Actions within a patient trajectory are captured by including features related to contacts and diagnoses. The total count of diagnoses and contacts and a breakdown of the individual contact types and diagnoses across the different axes were included in the first iteration. It was observed from the SHAP plot in Figure 6.14 that these features impacted the results. Due to a strong correlation between the total count of diagnoses and contacts and the individual types of these, the two total count features were removed. Additionally, due to an identified correlation between the action features and the length of the EoCs, the features were changed to be presented as frequencies instead of counts. This made it possible to identify more distinct similarities between the individual types of contacts and diagnoses.

Comparable similarities concerning the trajectory actions are observed in the last two iterations. Looking at the final iteration, EoC Type 0 has frequent contacts but a lower frequency of diagnoses, EoC Type 1 has both frequent contacts and diagnoses, and EoC Type 2 has a low frequency of both. Additionally, the last two iterations include the *Percentage of primary axis diagnoses* feature. This feature demonstrates no clear relation between the frequency of diagnoses given and the percentage of these being primary diagnoses on one of the axes. Notably, EoC Type 0 and EoC Type 2, which have low frequencies of diagnoses set, differ largely in the percentage of primary diagnoses. This discrepancy suggests the possibility that no diagnoses are given in EoC Type 0, although further exploration is necessary.

The feature *Number of contacts before primary Axis 1 diagnosis* is introduced in the third iteration. This feature shows a difference between the subgroups ranging from EoC type 0 that either has a diagnosis at the beginning of the EoC or never gets one, to EoC type 1, getting a diagnosis before 25 contacts and EoC type 2 displaying a more varied amount of contacts before diagnosis.

8.3 Clinical Evaluation

The results from the final cluster iteration were presented to clinicians to obtain a professional evaluation of the experiment conducted. Recognising that evaluation solely based on data can be challenging, the objective was to gather insights and feedback based on human judgement (Burkov, 2019). During the presentations, no interpretations were provided. This was done to ensure that the professionals' evaluations remained unbiased. The individuals involved were encouraged to express their thoughts, provide remarks, and raise concerns from their professional standpoint, considering clinical practice and their knowledge derived from previous research. The aim of involving clinicians and other domain experts was to gain valuable insights and perspectives that could further enhance the evaluation. The feedback obtained is also utilised in Section 9.2.4 to discuss the final results.

The expert evaluation was conducted through the following series of meetings:

- May 16th, 2023: Meeting with Odd-Sverre Westbye.
- May 26th, 2023: Meeting with the IDDEAS team.
- May 26th, 2023: Meeting with Birgit Kleinau.

The specific structure of the meetings varied depending on the expertise and preference of the individuals involved. However, a consistent aspect was the visualisations presented, detailed in Chapter 7. This evaluation will present key elements derived from the meetings, starting with evaluating the EoC subgroups and the associated features. It then delves into the identified EoC Bundle subgroups and discusses features involved on this level. Finally, some concluding clinical evaluation thoughts are presented.

The initial feedback regarding the EoC clustering from all clinicians was the meaningfulness of the division of three EoC clusters. The EoC Type 2 cluster was considered representative of "typical" hyperkinetic EoCs, characterised by being longer, planned polyclinic. Separating these EoCs from the shorter EoCs in EoC Type 0 and EoC Type 1 was deemed clinically logical. According to the experts, typical hyperkinetic EoCs would include initial diagnoses on the six axes and then regular appointments for therapy sessions and examinations to re-evaluate the diagnoses. They also note that it is reasonable that these EoCs run over many years.

An important aspect discussed by all experts was the desire for more detailed information regarding these typical hyperkinetic EoCs. This aligns with the feedback obtained during the second iteration of the clustering process. The experts noted in this final feedback round that it is intriguing that after removing rejected patients, the clustering process did not succeed in further differentiating the subgroups. This finding suggests adding more features to achieve a more detailed EoC separation.

Looking at specific features, experts deem the correlation between acute and inpatient EoCs reasonable, and categorising them into a separate cluster confirms that these EoCs are distinct and treated differently. The EoCs in the EoC Type 1 group exhibit the highest frequency of contacts and diagnoses compared to the other EoC types. This finding is considered logical by clinicians since these EoCs are more severe, and inpatient clinics often follow a more systematic approach involving multiple contacts and numerous diagnoses. The experts highlight that admitting a patient to an inpatient EoC is rare. Hence, it is expected that they require more resources.

It is also interesting to note that EoCs shorter than a week but not classified as inpatient or acute, often identified in EoC Type 0, have fewer diagnoses. According to the experts, this could be attributed to the EoCs being polyclinic visits without the same strict, systematic approach or that these are patients who may not meet the diagnostic requirements. Furthermore, the experts observe that if it is true that many of the patients in this group never received diagnoses on the six axes, it is surprising that they are not rejected from further treatment. They speculate that this may be attributed to human or system errors or possibly differences in approaches over the years. Clinical experts emphasise that it is seldom for a patient without any diagnosis to continue their treatment.

The experts acknowledge that the contact frequency being mostly yearly to monthly for EoCs belonging to EoC Type 2 is reasonable. However, they expressed an interest in more detailed information regarding the timing of the contacts. They believe that most contacts occur during the initial period of an EoC to diagnose a patient. The clinicians suggest that relying solely on the frequency of contacts for the whole EoC may provide a misleading indication for longer EoCs. Experts also state that they have experienced that polyclinic EoCs have similar frequencies of contacts during the initial phase, but this cannot be discerned from the current results.

The final EoC feature presented is the number of contacts had before a main diagnosis is assigned. The experts find this feature interesting but are concerned that many EoCs have a “No main diagnosis set” value. “No main diagnosis set” is represented by the numerical value “1 000”, assigned during the preprocessing when a patient does not have an Axis 1 diagnosis at the EoC Bundle level and never receives a main diagnosis throughout the EoC. The experts question giving this value since most EoC Bundles should have a main diagnosis, and the error may lay elsewhere. However, the experts acknowledge that for certain EoCs, the missing value may be because the patients have gotten a main diagnosis in a different EoC Bundle than the one this EoC is a part of.

During the evaluation of the *Contacts before a main diagnosis* feature, an error was identified in which Z-codes were included as a main diagnosis. The experts communicated that these codes should not be considered as a main diagnosis. Consequently, the method for determining the timing of a main diagnosis should be revised to exclude Z-codes.

When presenting the overall EoC Bundle cluster distribution, the clinicians express that the division into four EoC Bundle subgroups is logical. They find it particularly interesting that the EoC Bundle Type 2 and 3 clusters exclusively consist of typical hyperkinetic EoCs. Here, the experts are interested in determining which other features distinguish the two EoC Bundle clusters. Additionally, they express curiosity about obtaining more information regarding the two smaller groups: EoC Bundle Type 0 and EoC Bundle Type 1, to investigate how valuable these clusters are for further research into trajectories related to hyperkinetic disorders.

Among the EoC Bundle features, the experts find the gender and age distributions intriguing, as they do not demonstrate any apparent differences. However, they note that small differences between the four EoC Bundle clusters warrant discussion. EoC Bundle Type 0 have the highest proportion of females, followed by EoC Bundle Type 3. EoC Bundle Type 2 has the lowest ratio. The experts point out that EoC Bundle Type 0 and EoC Bundle Type 3 primarily comprise patients older than 7. EoC Bundle Type 2 mainly includes younger patients, aligning with previous clinical experiences stating that males are treated for hyperkinetic disorders at a younger age.

Furthermore, the experts note a relationship between gender, age, and CGAS score assigned at the beginning of an EoC Bundle. They observe that the EoC Bundle subgroup with the highest ratio of younger males (EoC Bundle 2) also has the lowest CGAS scores, which suggests a higher level of disability. This finding aligns with clinicians' previous experiences and provides data analytic evidence supporting these.

When further evaluating the diagnoses on the six axes, the experts note that setting the CGAS score to "5" when no value is given may introduce a data bias. However, they acknowledge that using numerical values when comparing CGAS scores is more informative than categorical ones. Thus, they understand the need to remove the missing values and agree it was a good solution. Additionally, the experts state that although it may be misinformative to analyse the distribution of CGAS value "5", examining other values provides more insightful information. By focusing on the extreme values, they can better understand the patient trajectories.

The experts expected that the care situation would have a greater influence on the patient trajectories than what was uncovered. The results are intriguing, prompting a reevaluation of their assumption and a need to reassess the relationship between care situations and observed trajectories.

Lastly, the distribution of EoCs within the EoC Bundle clusters is evaluated. First, the experts observe that EoC Type 0 and 1 only occur in the first two EoC Bundle clusters, while EoC Type 2 is present in varying frequencies across all four EoC Bundle clusters, either in combination with EoC Type 0 and 1 or alone. They note the significance of evaluating EoC Bundle Type 2 and 3 together since both exclusively contain EoC Type 2 EoCs. Compared with other findings, the experts emphasise that these clusters are distinguishable. EoC Bundle Type 2 include more males, younger patients, lower CGAS scores, and longer EoCs than the ones comprising EoC Bundle Type 3.

The experts state some interesting thoughts, focusing on EoC Bundle Type 0, which contains most of the EoC of type 1 (short, inpatient, and acute EoCs). First, they state that when investigating the EoC types within these EoC Bundles, the low CGAS scores concur with the EoC being the acute ones. The experts also state that this coincides with their experience that EoCs of type 1 are more severe and that this impact the severity of the EoC Bundles (as seen by the CGAS scores). Furthermore, the clinicians state that it is interesting to note that the EoC Bundles of Type 0 often include both an EoC of type 1 and one or more "typical" hyperkinetic EoCs. This is interesting to the clinicians because it shows that changes occur within trajectories resulting in more acute and severe EoCs.

Looking at EoC Bundle Type 1, the EoC Bundles mainly include EoCs of type 0 (short, planned and polyclinic EoCs with few diagnoses). Additionally, almost half of the EoC Bundles of type 1 also include a "typical" hyperkinetic EoC. This surprises the clinicians because it may contradict the suggestion that patients with an EoC of type 1 should have been rejected. However, further detailed information is needed to evaluate the reason behind this combination of EoCs and understand the contributing factors.

Ending the meetings, the clinicians all state that they find the results interesting. They express curiosity regarding certain similarities found and distributions that emerge within the results. The experts note that distinguishable features have shown intriguing combinations, either confirming previous experiences or raising new ideas and questions. Overall, their keen interest and exploration of various aspects indicate that they find the results engaging and thought-provoking. They also note that many similarities and differences are starting points for new questions and areas for further research. Lastly, the clinicians highlight their enthusiasm for using EHR data to obtain insight regarding complete patient trajectories.

Chapter 9

Discussion

The discussion critically analyses the research methodology and results, considering the choices made during this project. It begins by presenting the methodology discussion, followed by a detailed examination of the results obtained. This chapter highlights the challenges encountered and suggests areas for improvement.

9.1 Method Discussion

This project adhered to the research method outlined in Section 1.3. A key component of this method was the clustering methodology followed in the experiment. To discuss this methodology, the implications of the k-prototypes algorithm are presented. Furthermore, the challenges encountered during the clustering validation, conducted as a part of the evaluation, are discussed. This method discussion also addresses the implications of specific clustering techniques employed in the experiment. Furthermore, another important factor of this method was the incorporation of clinical evaluation throughout the experiment and in the concluding evaluation. The final part of this section elaborates on the implications of this inclusion.

9.1.1 Implications of the Clustering Algorithm

The choice of the clustering algorithm in this project was made considering the factors presented in Section 3.2.1. As presented, this choice mainly depended on this project's data. To cluster mixed data, there were two possible options; either clustering the mixed data directly or converting the values to either numerical or categorical values and choosing a suitable algorithm for the chosen type. For this project, the mixed data was directly clustered to avoid potential information loss associated with data conversion. Consequently, the chosen clustering algorithm for this project was k-prototypes. Using k-prototypes, the project resulted in distinct clusters identifying similarities and differences in patient trajectories. However, one should note that other approaches also could have been used.

The choice of directly clustering mixed data differentiates from the data conversion performed in the reviewed papers presented in Section 4.1. These research converted the data and then applied k-means algorithm. Considering internal validation, converting the data is advantageous since indices can be more easily applied. However, the data conversion techniques used in these papers are complex and time-consuming. Therefore, in this project, it is considered beneficial to have chosen a less complex approach due to the limited time scope. This decision allowed for more time to interpret and evaluate intermediate findings and results in collaboration with clinicians.

The choice of directly clustering the mixed data using the k-prototypes algorithm has influenced the validation of this clustering project and the selection of specific techniques employed during the clustering process.

9.1.2 Implications of the Clustering Validation

The clustering validation was conducted as outlined in Section 8.1. Both the internal and relative validations were challenging due to the choice to cluster mixed data directly. The discussion that follows delves into the difficulties encountered due to this approach.

An important step in the clustering validation was assessing whether the available data possesses a clusterable structure. This assessment was carried out using Hopkins statistics to evaluate the clustering tendency. However, since the data used for this validation first underwent a temporary conversion, it is important to consider the extent to which these Hopkins scores accurately reflect the original data. It can be argued that the temporarily converted data deviates from the original data used in the clustering process, raising concerns about its representativeness for drawing accurate conclusions about the clustering tendency of the original data. However, gaining insight into the clustering tendency of slightly modified data may still be considered sufficient to confirm the presence of a clusterable structure in this project's data. While the converted data may not perfectly reflect the original data, it can provide valuable indications and insights into its clustering behaviour. Therefore, even with the conversion-induced modifications, the analysis of the converted data can still offer meaningful information regarding the clusterability of the original data.

In an ideal clustering scenario, the cluster partitioning should demonstrate similarity among data points within the same cluster and dissimilarity to data points in other clusters. To assess the extent to which this characteristic was achieved in the clustering structures of this project, internal validation indices should have been employed to evaluate the clustering results. It is worth noting that many existing cluster validation indices are not suitable for handling mixed data. The possible adoption of clustering validation techniques to work on mixed data was investigated to address this challenge. However, it was determined upon closer examination that these adaptations were not incorporated into the project.

The subsequent discussion aims to provide insight into the reasons behind these adaptations' exclusion and demonstrate the potential impact such adoptions could have had on the validation process and the overall validity of the project's results.

The following indices were investigated to evaluate an adaption to work on mixed data by adjusting the distance metric presented in Section 3.2.2, Equation 3.2 (Aschenbruck and Szepannek, 2020):

- *Gamma index*, *Gplus index* and *Tau index*: Indices where every within-cluster distance is compared with every between-cluster distance. These indices have shown promise in being adapted to work with mixed data. However, it is important to note that these indices require high computational costs, as each within-cluster distance must be compared with every between-cluster distance.
- *Dunn index*: Index adaptable to mixed data as its only requirement is the distance between two clusters.
- *Silhouette index*: Index transferable to mixed data as it considers the average within-cluster distance for each cluster. Compared to the abovementioned indices, this performed the best.

(Aschenbruck and Szepannek, 2020)

A search for pre-existing code or libraries was done to adapt these indices for validating k-prototypes clustering on mixed data. The R package *clustMixType* was discovered in this search (Szepannek, 2019). This R package enables utilising the mentioned internal validation indices on a k-prototype object and getting an index value rating the cluster partition in return. This package can also be used to find an optimal value of k by specifying the search range for the optimal number of clusters (Aschenbruck and Szepannek, 2020). Unfortunately, no equivalent package implemented in Python was found during the search. Therefore, the possibility of implementing these indices from scratch as an extension to the Python implementation of the k-prototypes algorithms was evaluated. However, due to the project's limited time scope, it was not feasible to undertake this implementation from scratch.

Including similar validation indices in the Python implementation of the k-prototypes algorithm could have impacted the performance and analysis of this project's clustering results. These validation indices would have provided a quantitative and objective measure of the resulting cluster partitions, enabling a comparison of the effects of various decisions on the clustering outcomes. However, the unavailability of a Python implementation of these validation indices prevented their incorporation into this project's clustering validation.

9.1.3 Implications of the Clustering Techniques

When completing this clustering process, many techniques that impacted the results were used (Theodoridis and Koutroumbas, 2008). The techniques considered to have the most important impact on the clustering results are presented in the following discussion.

Feature Selection

Due to the high-dimensional real-world data used in this project, feature selection was a crucial part of the clustering process. The feature selection aimed to select the features that encode as much relevant information regarding patient trajectories as possible. However, including more features did not necessarily make the clustering outcome more informative and made the results harder to interpret. Irrelevant features blurred the clusters, making it increasingly difficult to pinpoint the features responsible for the observed differences. Balancing the trade-off between including informative features and limiting their number for interpretability was crucial for this project. It required consideration of which features could comprehensively describe the distinguishing factors among EoCs and EoC Bundles.

The feature selection included evaluating the features from a domain expert and a technological perspective to obtain optimal results across all three experiments. The SHAP plots, visualised in Figures 6.14, 6.16, 6.36, 6.38, 6.56, and 6.58, played a crucial role in analysing the impact and importance of the various features on the decision-making processes of the clustering models. By exploiting the insights provided by the SHAP method and incorporating the feedback from domain experts, adjustments were made to the selected features throughout the clustering iterations. Notably, significant changes were made to the features used in iteration 1, which substantially impacted the clustering results obtained in iteration 2.

Feature Scaling

The feature scaling method is another aspect that impacted the results. Features scaling was an important part of the data preparation since the numerical data utilised had varying ranges of units and measurements. Before clustering, the numerical data had to be converted into a uniform scale because the similarity between data points was determined by their distance from each other. This data transformation was also beneficial in reducing the impact of the outliers and improving the model's performance (Shalabi et al., 2006).

Both normalisation and standardisation were considered when choosing the feature scaling method since no definitive answer exists on when to use one over the other (Burkov, 2019). To choose between the two methods, considering outliers was particularly important. Standardisation scaled the features to a common scale without altering the value range, while normalisation compressed the data into a smaller range, making distinguishing between the values challenging. This aligned with the recommendation that unsupervised learning algorithms, like clustering, typically benefit more from standardisation than normalisation (Burkov, 2019). Therefore, the preferred feature scaler for this project was standardisation, and specifically, Power Transformer was utilised to transform the numerical data. This feature scaler was selected because it is well-suited for data sets containing outliers (Pedregosa et al., 2011).

After selecting standardisation as the preferred feature scaling technique and identifying Power Transformer as a promising method, a logical next step would have been to compare it with other standardisation techniques, such as Standard Scaler. This comparison would aim to observe and evaluate the data implications of each technique. However, without any straightforward internal validation index to rate the clustering outcomes, it was difficult to determine which technique produced the best results. Thus, Power Transformer was chosen without delving further into its impact on the project's specific data, which may have limited the clustering results.

Initialisation Method for k-prototypes

Another important aspect of the clustering process was the initialisation method. The initialisation method was used to find the initial centres for the clusters and directly impacted the formation of the final clusters.

Therefore, the initialisation method had to be carefully chosen. Since there is limited research on selecting initial cluster centres for mixed data, and no universally accepted method exists, the choice was not straightforward (Cao et al., 2009). The initialisation methods considered for this project's clustering algorithm were random initialisation, Huang, and Cao.

First, random initialisation was considered. k-prototypes can be initialised by randomly selecting an initial set of cluster centres and then iteratively refining this set. Although this method is commonly used for its simplicity, it required multiple reruns of the clustering algorithm with different initial prototypes to identify a good starting point (Cao et al., 2009). Therefore, this initialisation method was decided early not to be used in this project.

The next initialisation method examined was the Huang method. This method was the proposed initialisation method when Huang developed k-modes and k-prototypes. Initialising k-prototypes with the Huang method entails using the most frequent categorical data as initial prototypes to increase diversity among the prototypes. Huang conducted an experiment where he compared initialising k-modes with the Huang method and with random initialisation using real-world data. The results showed that initialising k-modes with Huang yielded significantly better results than random initialisation. However, when Huang was used with k-prototypes on the same real-world data set, the results were not enhanced much from randomly initialising the cluster centres (Huang, 1998). The difference in using the Huang method with k-modes and k-prototypes lies in the initialisation approach, which fails to account for both numerical and categorical features when selecting the most common categories as initial prototypes (Huang, 1998).

The third initialisation method examined was the Cao method. Cao, presented in Section 6.9, is a frequency-based initialisation method proven superior to the random initialisation method. The Cao method, similar to the Huang method, chooses the first initial cluster centre based on the assumption that the more objects around a data point, the more possible it is for this data point to be a cluster centre. When selecting the rest of the initial cluster centres, the Cao method distinguishes from Huang by considering both the density of objects and the distance between them (Cao et al., 2009).

An explicit evaluation comparing Huang and Cao as the initialisation method for k-prototypes when clustering a real-world data set lacks. Therefore, the first iteration of the EoC clustering was conducted twice to decide whether Huang or Cao was the optimal initialisation method for this experiment. Once again, the issue of not having any internal validation indices to rate the performance of the two clustering outcomes was raised. However, while determining the optimal number of clusters and subsequently clustering the data using this number, both methods identified the same cluster count, and the resulting clustering centroids were quite similar. Given this resemblance, Cao was selected as the preferred initialisation method. Refer to Appendix B.3.1 to see the code written when deciding the initialisation method and the comparison of the resulting cluster centroids obtained when using Huang and Cao.

Finding an Optimal *k*

Since the number of clusters (k) in the data set was now known beforehand, finding the optimal number of clusters was crucial for accurate clustering results. This process relied on making an "educated guess" based on visualisation or metrics to determine the appropriate value for k (Burkov, 2019). While various techniques are available for finding k when clustering numerical data, the number of techniques suitable for mixed data is limited and is considered a challenging problem (Ahmad and Khan, 2019). The following discussion presents the exploration of methods for finding k to contextualise the "educated guess" made and demonstrates how identifying an optimal k was complicated with multiple possible solutions.

First, visualisation techniques were examined and tested to find an optimal number of clusters. *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP) was the first to be tested. UMAP reduced the dimensionality of the features selected for the first clustering iteration to two dimensions and visually represented the data structure (McInnes et al., 2018). The visualisations did not indicate clear EoC or EoC Bundle data groupings. Thus, UMAP was not further used to determine an optimal number of clusters in this project. Refer to Appendix B.3.2 to see the implementation of the UMAP technique using the first iteration's EoC and EoC Bundle data and the resulting visual representations.

Another visualisation technique that was tested is the Elbow method. Two plots displaying an elbow-shaped graph were generated by applying the Elbow method to the EoC and EoC Bundle data from the first iteration (referring to Figure 6.13 and Figure 6.15). This outcome suggested that the Elbow method performed well with the available data in this project. Consequently, the Elbow method was utilised to determine the optimal number of clusters for both the EoC and EoC Bundle data in each of the three iterations conducted throughout the experiment (referring to the Plots 6.13, 6.15, 6.35, 6.37, 6.55, and 6.57).

Then, an investigation of methods that determine the optimal number of clusters based on a validation index was conducted. By utilising both visualisation and a validation index, the goal was to ensure that the number of clusters identified was optimal. Although numerous established distance-based clustering validation indices are available for numerical data, the options are limited for mixed data. However, research has shown that numerical validation indices can be adapted for mixed data by modifying the distance metric (Aschenbruck and Szepannek, 2020). Nonetheless, these modified validation indices have not yet been implemented in Python and thus were not employed for this project.

9.1.4 Implications of Clincans Involvment

Choosing the features to cluster, the scaling method, the initialisation method, and determining the optimal number of clusters are all choices that could have led to significantly different clustering outcomes. According to the book *Pattern Recognition* in the context of clustering as a data mining tool, it is stated that *Subjectivity is a reality we have to live with from now on.* (Theodoridis and Koutroumbas, 2008, p. 597). The book emphasises that in terms of clustering, multiple clustering results might exist that can all be considered valid. In light of this, involving domain experts has been crucial to obtaining meaningful clustering results. While expert evaluation provides valuable insights, including quantitative validation indices could have added an extra layer of objectivity and rigour to the evaluation process.

The selection of clinicians in this experiment has impacted the project. Incorporating feedback and evaluations from domain experts influenced the modifications made throughout the experiment and the subsequent evaluation. It is important to acknowledge that opinions are subjective and can be influenced by personal biases, leading to individual variations. To mitigate the subjectivity of the feedback and the evaluation, an effort was made to include clinicians with extensive experiences and diverse backgrounds. The aim was to minimise individual biases and provide a more comprehensive perspective. However, it should be recognised that other clinicians may have offered different opinions. This could have influenced the experimental choices and potentially led to different results. This aspect should be considered when presenting the interpretations in the next section.

9.2 Result Discussion

Interpretations, implications, limitations, and recommendations are presented to discuss the results obtained. The interpretations utilise the evaluation done and the knowledge gained. Then, the limitations found are summarised to provide a critical project analysis. The implications will detail the significance of the experiment performed. Lastly, the recommendations highlight potential directions for further research, suggested modifications or improvements and unanswered questions requiring additional exploration. This discussion aims to contribute to understanding the experiment performed and guide future research.

9.2.1 Interpretations

The initial interesting interpretation from the results suggests that most patient EoCs can be categorised as "typical" EoCs for individuals with hyperkinetic disorders. The evaluation revealed that this outcome did not surprise the clinicians. However, it provides empirical evidence through data analysis supporting existing experiences within the field. Furthermore, these results could be attributed to hyperkinetic treatment guidelines followed by CAMHS in Norway. The objective of these guidelines, presented in Section 2.4, is to minimise undesired variation in patient treatment, uphold treatment quality, and assist in resource allocation prioritisation (Helsedirektoratet, n.d.). Based on the clustering results, one can infer that the prevalence of EoC Type 2 may be due to the procedure guidelines fostering a similar treatment and diagnostic approach for most hyperkinetic patients. While it is impossible to prove this from the results, the similarity among EoCs could indicate an effective procedure.

EoCs belonging to EoC Type 1 are observed in all four EoC Bundle clusters. However, EoC Bundle Type 2 and 3 predominantly consist of one or more EoCs of type 2. Interpreting the difference between these two EoC Bundle clusters is intriguing. EoC Bundle Type 2 comprises younger patients and a higher percentage of males than EoC Bundle Type 3, concurring with the clinicians' previous experiences of females often being treated for hyperkinetic disorders later in life. Additionally, EoC Bundles of type 2 are normally longer. This can be interpreted as a result of regular follow-up until the patients reach a certain age where they are no longer eligible for treatment. Therefore, when hyperkinetic disorders are identified earlier, the treatment period tends to be longer. Another notable difference between the two EoC Bundle clusters is that the cluster with a higher ratio of younger male patients exhibits lower CGAS scores. This discrepancy warrants further analysis to understand if this difference is gender- or age-related.

Upon consulting with clinicians, it has been established that there are many requirements to fulfil for a patient's EoC to be classified as acute and inpatient, which is uncommon for hyperkinetic patients. Multiple criteria are often met to classify an EoC as an EoC Type 1. This EoC type's frequency of contacts can be an indicator of a more severe EoC. Comparing the EoC clusters, EoC Type 1 has the highest frequency of diagnoses on all six axes. Furthermore, these EoCs also have the highest percentage of diagnoses given as the primary axis diagnosis. This diagnostic profile indicates that EoCs of type 1 involve more severely impacted patients requiring more extensive treatments. However, it should be noted that the high frequency of diagnoses may also be influenced by other procedures implemented in an inpatient clinic.

When observing the EoC Bundles of type 0, including the most EoCs of type 1, these EoC Bundles often have significantly lower CGAS scores. This indicates a higher level of disability, which aligns with these EoC Bundles having at least one inpatient, acute EoC during the EoC Bundle. Additionally, it is interesting to note that EoC Bundle Type 0 is the EoC Bundle cluster with the highest percentage of females. This observation may imply that females diagnosed with hyperkinetic disorders experience severe issues or face other challenges in combination with hyperkinetic disorders requiring greater attention. However, this interpretation needs more detailed findings to validate.

Further interpreting the EoC Bundles including EoC Type 1, it becomes apparent that these EoC Bundles frequently include one or more of the EoCs of Type 2 referred to as the "typical" hyperkinetic disorders EoCs. This observation suggests an occurrence within an EoC Bundle, leading to a modification in the immediacy and care level of the consecutive EoCs. Due to limited information regarding the timing of these changes, no definitive interpretations have been made. However, it would be interesting to investigate the underlying cause of this alteration.

An interpretation made by the experts was that the EoCs of type 1 include patients that should have been rejected. This interpretation stems from the diagnostic profile of these EoCs. Most of these EoCs rarely include diagnoses given on the six axes. Furthermore, these EoCs often start with a main diagnosis on the EoC Bundle level that never changes during the EoC, or they start without a diagnosis on the EoC Bundle level and never receive one during the EoC Bundle. Iteration three excluded all the EoC Bundles and belonging EoC entries where the EoC Bundle had been assessed to be rejected or its closing code was either "Rejected" or "Did not get started". By doing this, the aim was to eliminate all patients rejected. However, the interpretation that some patients still should be classified as rejected implies that the data might not be logged correctly or that further investigation into other features is necessary to eliminate all rejected patients. Another input from the clinicians was that some patients do not receive diagnoses during the EoCs of type 0 because they already have gotten a diagnosis on Axis 1 in a previous EoC Bundle. This interpretation and clinicians' comments suggest a need to explore this EoC subgroup further to obtain a clearer picture.

Examining the EoC Bundles, including EoCs of type 0, it is evident that these predominantly are represented in EoC Bundle Type 1. Many of these EoC Bundles last less than a week, suggesting that these EoC Bundles consist of a single EoC of type 0, where the patients do not receive diagnoses. No further treatment is likely determined to be necessary for these patients, as suggested by the clinicians. However, almost half of the EoC Bundles of type 1 are longer than six months, indicating that the treatments included more EoCs than a single EoC of type 0. Many EoC Bundles of type 1 also include a typical hyperkinetic EoC. Once again, further information is necessary to interpret the factors contributing to the separation of EoCs within these EoC Bundles.

9.2.2 Limitations

Numerous limitations have been highlighted in the experiment, interpretation, and evaluation. Primarily, the limited amount of data introduced multiple challenges. Firstly, one can note that the limited data and considerations done because of this limitation may have led to the uneven partitioning of clusters. The resulting clusters consist of two smaller and one larger EoC cluster and two smaller and two larger EoC Bundle clusters. These uneven cluster divisions limited the detailed information one could extract from the EoC and EoC Bundle clustering.

Another potential limitation related to the limited amount of data is that the clustering analysis might have captured random patterns in the data resulting from outliers instead of actual meaningful patterns (Theodoridis and Koutroumbas, 2008). This may be evident in the result, as the clusters were unevenly divided into EoC and EoC Bundle subgroups. With this limited amount of data, such outliers might have impacted the reliability of the clustering results. With more data, the clustering would have been more robust, and the impact of outliers and noise would have been reduced.

Another limitation regarding the limited amount of data is that the experiment is based on data collected solely from one hospital in Norway. This limitation implies that the results may not represent the entire country's full spectrum of patient trajectories within CAMHS. Therefore, caution should be exercised when attempting to generalise this project's findings beyond the context of the St. Olavs Hospital.

The options to exclude entries were also limited when conducting the clustering process with this limited data set. Recognising the importance of preserving information, the aim was to retain as many entries as possible without including error-prone values. The goal was to have the largest data set feasible to obtain the deepest exploration of clusters and relationships between data points, resulting in meaningful and interpretable results. Wanting to include many entries impacted the choice of features and the data preparation. The feature selection was constrained to features with minimal occurrence of error-prone values. However, handling missing values was necessary for the included features. The ratio of missing data for each feature varied, but for many features, this made up a significant portion. The missing values were replaced with "Missing data" for categorical values. Regarding numerical features, the missing values had to be transformed into specific numerical values. As evaluated in the clinical valuation, this might have introduced data bias.

In the third iteration, a decision was made to exclude all rejected patients from the data set after careful consideration and consultation with experts. This was done after considering the pros and cons of reducing the data set size. For further research, this will continue being a limitation impacting the choice of, for instance, investigating only the EoCs including "typical" hyperkinetic EoCs. Additionally, the experts suggested exploring the clustering analysis separately for a data set that includes only females and only males, as well as dividing the data set based on age groups. While these suggestions are intriguing, the current limitations may restrict the feasibility of conducting these analyses.

The clustering process was completed using a data set from numerous legacy systems spanning an extended period. This posed challenges for several reasons. Firstly, heterogeneous system usage may limit consistency. Each legacy system has been designed and implemented differently, impacting how data is collected. Not having direct access to the systems made investigating their functionality and data entered difficult. Additionally, the terminology employed in the system might have varied, leading to inconsistencies in the data interpretation. This was experienced when consulting with experts who interpreted the feature values differently. The feature understanding was a time-consuming process that included eliminating features from further analysis due to a lack of understanding. Furthermore, due to the systems' long duration of use, inconsistent data quality, including errors and missing data, and discrepancies were prevalent. This largely impacted the data preparation part of this project. Overall, working to understand these systems and their resulting data required us to be "data archaeologists". This involved meticulous investigation, clinical investigation to acquire domain knowledge, and consultations with experts and systems users to understand the data.

A timeline of actions was condensed into count-based features to capture the activity level regarding contacts and diagnoses during patient EoCs. By doing this, data dimensionality was reduced, making the data easier to interpret. The count-based features also facilitated analysis of the intensity of contacts and diagnoses within an EoC. This helped distinguish EoCs with frequent and infrequent contacts and diagnoses. However, this simplification has certain limitations. The expert evaluation suggests that temporal information may be lost as it becomes difficult to identify specific timing patterns regarding contacts and diagnoses. To address this, dividing the count into smaller, time-specific counts could capture different aspects of the data and enhance the clustering analysis. Still, there is always a possibility that a simplification may lead to a loss in valuable temporal patterns and dynamics present in the data set.

9.2.3 Implications

The experiment conducted to identify clusters of EoC and EoC Bundles within the context of hyperkinetic disorders in CAMHS in Norway carries several implications. Identifying distinct clusters provides insights into the characteristics of patient trajectories, shedding light on different patterns and variations within the data set. In short, the results provide information regarding patients' characteristics, contacts, diagnoses, and overall trajectories related to hyperkinetic disorders. The project contributes to understanding the disorder's treatment and potential differences separating the trajectories. Furthermore, the insights lay a foundation for further research and raise multiple questions to evaluate.

Considering the theoretical implications of this Master's Thesis, the experiment conducted is the first to consider these features and CAMHS patient trajectory aspects within Norwegian research. A previous IDDEAS research has used the same data but limited the analysis to patients' referral to CAMHS in Norway (Solheim, 2022). As the first to analyse patient trajectory similarities using a cluster-analytic approach, this research holds the potential to serve as a stepping stone for future research within the same field. The hope is that this work establishes a solid foundation for subsequent research by encompassing comprehensive data preprocessing and analysis, evaluation and raised important questions.

Furthermore, using clustering, the experiment resulted in subgroups of EoCs and EoC Bundles. By doing this, this master thesis has demonstrated how clustering mixed data retrieved from an EHR can identify similarities and differences in a data set. The thesis has also shed light on challenging aspects of clustering mixed data and proposed possible ways to handle these. By evaluating and discussing the choices made, one can get information regarding potential successful outcomes of the clustering and limitations experienced during the experiment. All this can be useful to consider in coming clustering research.

9.2.4 Recommendations

From the Discussion and evaluation, several avenues for research have emerged. These recommendations encompass both areas that require further detailed analysis and new approaches to investigating patient trajectories. Following is a summary of the main recommendations:

1. **Condict separate analyses on different subsets of the data set:** One suggestion is to analyse the "typical" hyperkinetic EoCs and the more unconventional EoCs separately. This approach could cluster these subsets of data into new subgroups identifying more details regarding each subgroup and revealing new insights about patient trajectories. Similarly, dividing the data set based on age and gender and performing individual clustering processes for each group can help identify their similarities and differences. For this recommendation, it is important to note the limitation of the data set size.

2. **Explore features capturing differences within a trajectory:** As identified in this Master's Thesis, many EoCs come in a sequence in an EoC Bundle. Identifying the order of these EoCs and the changes occurring between them could yield future interest. Here one must be prepared to handle missing data, making comparing subsequent EoCs more difficult.

3. **Break down count-based features into smaller time periods:** Analysing contacts and diagnoses in smaller intervals is recommended to gain more insights into the patient trajectories. For example, separating an EoC's first week or month from the remaining time period may help identify differences. This approach allows for identifying patterns or spontaneous changes within the EoC. A limitation impacting this avenue of research is the missing and error-prone date values of both contacts and diagnoses.

In short, these recommendations provide a starting point for further investigation into patient trajectories. All relevant code used in the experiment is mentioned throughout the Master's Thesis and included in the appendices, which can help guide future research. By building upon the theoretical background, preprocessing steps, clustering experimental setup, evaluation, and discussion conducted in this Master's Thesis, future research can benefit from a stronger foundation.

Chapter 10

Conclusion and Contributions

This chapter first concludes the work done in this Master's Thesis in light of the research questions introduced in Chapter 1. Then the chapter concludes with a presentation of the contributions to the field of clustering within CAMHS in Norway.

10.1 Conclusion

The following conclusion presents final thoughts on how the work done in this project answers the research questions derived in Section 1.2.

> **Research Question 1** How can hyperkinetic patient trajectories in an electronic health record be identified?

The theoretical investigation gave insight into the patient treatment guidelines for hyperkinetic disorders and highlighted important factors within patient trajectories. Specifically, the breakdown of patient trajectories into individual *Episode of Care Bundles* (EoC Bundles) containing one or more *Episodes of Care* (EoCs) was demonstrated. Then, the clustering methodology showcased the potential of employing a cluster analytic approach to identify natural subgroups within the data, drawing from relevant prior research that similarly utilised Electronic Health Record (EHR) data.

Using EHR data collected at St. Olavs Hospital, important characteristics regarding EoC Bundles and EoCs were extracted. The data was prepared for clustering analysis after carefully selecting data based on the data quality and feature importance. Subsequently, the clustering process was conducted stepwise using the k-prototypes algorithm. The first cluster step identified subgroups of the EoC data. Then, clustering of the EoC Bundle data, including the compressed EoC data, was performed to characterise higher-level subgroups. Consequently, patient trajectory subgroups were successfully identified, each characterised by distinguishing factors. In short, one could distinguish the EoCs based on duration, care level, and immediacy level. Furthermore, the frequency of contacts and diagnoses helped to distinguish the three EoC subgroups. These clusters were evaluated in collaboration with clinicians, who confirmed that one subgroup identified the "typical" hyper-

kinetic EoCs. The other two subgroups were differentiated by their shorter duration, with one subgroup including most of the inpatient and acute EoCs with frequent contacts and diagnoses, while the other subgroup consisted of planned polyclinic EoCs with frequent contacts but a low frequency of diagnoses. By including the identified EoC subgroups when clustering the EoC Bundles, four EoC Bundles subgroups were identified. The EoC Bundles were also distinguished based on duration. Furthermore, gender, age, and CGAS score showed distinguishing factors between the clusters. The EoC Bundle subgroups were further differentiated by the EoCs included within the EoC Bundles. Two EoC Bundle subgroups were dominated by including "typical" hyperkinetic EoCs, either in combination with other EoCs or alone. The other two EoC Bundle subgroups included either the shorter, acute inpatient EoCs or the shorter, planned polyclinic EoCs.

To conclude, this master thesis addressed the first research question by demonstrating how clustering the St. Olavs EHR data could identify subgroups of hyperkinetic patient trajectories. The theoretical background, methodology, and analysis of the EHR data provided valuable insights and enabled differentiation of the subgroups based on key characteristics and clinical evaluation.

> **Research Question 2** How can patient trajectory clusters be made meaningful to clinicians?

Important factors within patient trajectories were identified by official guidelines within *Child and Adolescent Mental Health Services* (CAMHS) in Norway and by involving clinicians throughout the experiment. This ensured that the features used in the clustering process were clinically relevant and informative. Continuously presenting the findings and modifying, adding, and removing features based on clinical feedback allowed for continuous improvement of the clustering outcomes. These modifications involved changing categorical features to numerical, when this was more informative, changing from sum to frequency of contacts and diagnoses, and adapting the presentation of the different diagnoses on the six axes to reflect the amount of data registered on the axes. These changes all made the clustering findings more meaningful for clinicians.

The findings could be interpreted by visualising the clustering results, showing these to clinicians, and involving clinicians in the evaluation. Clinicans' experiences could be confirmed from the EoC and EoC Bundle subgroups identified by clustering patient trajectories. These experiences confirm the differentiation between longer, planned polyclinic EoCs and shorter ones. Furthermore, the EoC Bundle subgroups could be distinguished by patients' gender, age, and CGAS scores set. This separation between the EoC Bundle clusters confirmed that more females are starting their EoC Bundle at an older age. It also showed that younger boys often have a lower CGAS score, indicating a more severe level of disability. From the evaluation, the clinicians also confirmed that the clusters showcased the resources used for different patient trajectories by stating the difference in frequency of contacts and diagnoses. Lastly, the clusters identified clinical questions and areas for further research that could provide information to clinicians. This includes separating the trajectories to gain even more insightful results, investigating the sequence of EoCs within an EoC Bundle, and breaking down the count-based features.

From this, one can state that the clusters were made meaningful by involving clinicians and guidelines when deciding the features and by clinically evaluating and interpreting the results.

10.2 Contributions

This thesis has presented the work done to analyse patient trajectories of patients related to hyperkinetic disorders in CAMHS in Norway by using clustering of EHR data to identify subgroups. Through three iterations, the clustering process was conducted using the k-prototypes algorithm, identifying distinct clusters of patient trajectories. These clusters exhibited features that could be further analysed by examining the distribution of relevant characteristics within each subgroup.

Using the k-prototypes algorithm to compress information and identify natural subgroups demonstrated its potential in effectively clustering patient trajectories using EHR data. The thesis discussed all the considerations made throughout the clustering experiment and highlighted the potential impact of these choices. The methodology employed in the research successfully clusters patient trajectories into meaningful subgroups, thus showcasing the applicability of using k-prototypes to cluster mixed data directly. Furthermore, the acknowledged limitations of cluster validity and methodology choices shed light on crucial considerations and areas of improvement.

Involving clinicians in this data analysis process enhanced the interpretability and relevance of the findings. Their expertise and evaluation provided valuable insights throughout the experiment iterations, confirmed the results, and aligned findings with their experiences within the field. Collaborating with clinicians improved the overall analysis and identified potential avenues for future research.

Based on the aforementioned contributions, it can be affirmed that this work accomplishes the primary goal of the Master's Thesis:

> **Goal** Analyse patient trajectories of hyperkinetic disorders in child and adolescent mental health using clustering of electronic health record data to identify subgroups.

Bibliography

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data and Knowledge Engineering*, *63*(2), 503–527. https://doi.org/10.1016/J.DATAK.2007.03.016

Ahmad, A., & Khan, S. S. (2019). Survey of State-of-the-Art Mixed Data Clustering Algorithms. *IEEE Access*, *7*, 31883–31902. https://doi.org/10.1109/ACCESS.2019.2903568

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. *SIGMOD Rec.*, *28*(2), 49–60. https://dl.acm.org/doi/pdf/10.1145/304181.304187

Aschenbruck, R., & Szepannek, G. (2020). Cluster Validation for Mixed-Type Data. *Archives of Data Science, Series A*, *6*(1), 02. https://doi.org/10.5445/KSP/1000098011/02

Bakken, V., Koposov, R., Røst, T. B., Clausen, C., Nytrø, Ø., Leventhal, B., Westbye, O. S., Koochakpour, K., Mandahl, A., Hafstad, H., & Skokauskas, N. (2022). Attitudes of Mental Health Service Users Toward Storage and Use of Electronic Health Records. *Psychiatric Services*, *73*(9), 1013–1018. https://doi.org/10.1176/APPI.PS.202100477

Banerjee, A., & Davé, R. N. (2004). Validating clusters using the Hopkins statistic. *IEEE International Conference on Fuzzy Systems*, *1*, 149–153. https://doi.org/10.1109/FUZZY.2004.1375706

Berner, E. S. (2016). *Clinical Decision Support Systems Theory and Practice* (2nd ed., Vol. 1). Springer New York, NY.

Berry, M. W., Mohamed, A., & Yap, B. W. (2020). *Supervised and Unsupervised Learning for Data Science* (1st ed.). Springer Cham.

Boutsidis, C., Zouzias, A., Mahoney, M. W., & Drineas, P. (2015). Randomized Dimensionality Reduction for k-Means Clustering. *IEEE Transactions on Information Theory*, *61*(2), 1045–1062. https://doi.org/10.1109/TIT.2014.2375327

Breivik, M. (2020). *Jubileumsskrift BUP 50år* (tech. rep.). https://stolav.no/Documents/BUP/BUP_50_%C3%A5r_jubileumshefte_LR.pdf

Brocke, J. v., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. In J. v. Brocke, A. Hevner & Maedche Alexander (Eds.), *Design science research. cases* (pp. 1–13). Springer.

Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burvok.

Cao, F., Liang, J., & Bai, L. (2009). A new initialization method for categorical data clustering. *Expert Systems with Applications*, *36*(7), 10223–10228. https://doi.org/10.1016/J.ESWA.2009.01.060

Chen, C.-h., Härdle, W., & Unwin, A. (2008). *Handbook of Data Visualization* (1st ed.). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-33037-0

Clausen, C. E., Leventhal, L., B., Nytrø, Ø., Koposov, R., Westbye, O. S., Røst, T. B., Bakken, V., Koochakpour, K., Thorvik, K., & Skokauskas, N. (2020). Testing an individualized digital decision assist system for the diagnosis and management of mental and behavior disorders in children and adolescents. *BMC Medical Informatics and Decision Making*, *20*, 232. https://doi.org/10.1186/s12911-020-01239-2.

Conda. (n.d.). Conda Documentation. https://docs.conda.io/en/latest/

Dash, M., & Liu, H. (2000). Feature selection for clustering. *Lecture Notes in Computer Science*, *1805*, 110–121. https://doi.org/10.1007/3-540-45571-X{_}13/COVER

DBeaver. (2021). DBeaver Documentation. https://dbeaver.io/

Direktoratet for e-helse. (2001). Volven. https://volven.no/produkt.asp?id=363&catID=1&subID=&oid=

Direktoratet for e-helse. (2022). FinnKode - Direktoratet for e-helse medisinske kodeverk. https://finnkode.ehelse.no/#icd10/0/0/0/2599550

Direktoratet for e-helse. (2023). Retningslinjer for koding : Multiaksial klassifikasjon i psykisk helsevern for barn og unge (PHBUP). https://www.ehelse.no/kodeverk-og-terminologi/Multiaksial-klassifikasjon-i-psykisk-helsevern-for-barn-og-unge-(PHBU)

Evans, R. S. (2016). Electronic Health Records: Then, Now, and in the Future. *Yearbook of medical informatics*, *1*(1), 48–61. https://doi.org/10.15265/IYS-2016-s006

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *Knowledge discovery and data mining*.

Gårdvik, K. S. (2007). *Kortversjon av ICD- 10 for bruk ved barne- og ungdomspsykiatrisk klinikk, St. Olavs hospital HF* (tech. rep.). St.%20Olavs%20Hospital

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001a). Clustering algorithms and validity measures. *IEEE Es*, 3–22. https://doi.org/10.1109/SSDM.2001.938534

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001b). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2-3), 107–145. https://doi.org/10.1023/A:1012801612483/METRICS

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Cluster Validity Methods : Part I. *ACM SIGMOD Record*, *31*(2), 40–45. https://doi.org/10.1145/565117.565124

He, Z., Xu, X., & Deng, S. (2005). Scalable Algorithms for Clustering Large Datasets with Mixed Type Attributes. *Int. J. Intell. Syst.*, *20*, 1077–1089. https://doi.org/10.1002/int.20108

Helsedirektoratet. (n.d.). Om Helsedirektoratets normerende produkter. https://www.helsedirektoratet.no/produkter/om-helsedirektoratets-normerende-produkter

Helsedirektoratet. (2021). *Samarbeid mellom kommune og BUP - Helsedirektoratet* (tech. rep.). https://www.helsedirektoratet.no/rapporter/psykisk-helsearbeid-for-barn-og-unge/samarbeid-mellom-kommune-og-bup

Helsedirektoratet. (2022). Behandling og oppfølging av ADHD/ Hyperkinetisk forstyrrelse - Helsedirektoratet. https://www.helsedirektoratet.no/retningslinjer/adhd/behandling-og-oppfolging-av-adhd-hyperkinetisk-forstyrrelse

Huang, Z. (1997a). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In *Workshop on research issues on data mining and knowledge discovery*.

Huang, Z. (1997b). CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES *.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, *12*, 283–304. https://doi.org/https://doi.org/10.1023/A:1009769707641

HUNT Cloud. (n.d.-a). FAQ on security. https://docs.hdc.ntnu.no/do-science/faq/security/#incident-reporting

HUNT Cloud. (n.d.-b). HUNT Workbench. https://docs.hdc.ntnu.no/

Hunter, J., Dale, D., Firing, E., & Droettboom, M. (n.d.). Matplotlib documentation. https://matplotlib.org/stable/index.html

IDDEAS. (n.d.). IDDEAS – Individual Digital DEcision Assist System — Regionalt kunnskapssenter for barn og unge - Institutt for psykisk helse - NTNU. https://www.ntnu.no/rkbu/iddeas#/view/about

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/J.PATREC.2009.09.011

Johor Bahru, U., Darul Ta, J., Bin Mohamad, I., Usman, D., & Bahru, J. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Article in Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303. https://doi.org/10.19026/rjaset.6.3638

Koochakpour, K., Nytrø, Ø., Westbye, S., Leventhal, B., Koposov, R., Bakken, V., Clausen, C., Brox Røst, T., & Skokauskas, N. (2022). *Success Factors of an Early EHR System for Lessons Learned for Future Practice Data-Driven Decision Aids* (tech. rep.). https://doi.org/10.3233/SHTI220057

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. In *2010 ieee international conference*. https://doi.org/10.1109/ICDM.2010.35

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://github.com/slundberg/shap

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. L. Cam & J. Neyman (Eds.), *Proceedings of the fifth symposium on mathematical statistics and probability* (pp. 281–297).

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://arxiv.org/abs/1802.03426v3

Mladenić, D. (2006). Feature selection for dimensionality reduction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3940 LNCS*, 84–102. https://doi.org/10.1007/11752790{_}5/COVER

N-BUP. (2009). *Bupdata brukerhåndbok* (tech. rep.).

Negi, N., & Chawla, G. (2021). Clustering Algorithms in Healthcare. In *Intelligent healthcare: Applications of ai in ehealth* (pp. 211–224). Springer.

NTNU. (n.d.). Collection of personal data for research projects - Knowledge base - NTNU. https://i.ntnu.no/wiki/-/wiki/English/Collection+of+personal+data+for+research+projects

NumPy. (n.d.). NumPy documentation. https://numpy.org/doc/stable/

Page, M. J., Mckenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *MBJ*, *372*(72). https://doi.org/10.1136/bmj.n71

Pandas. (n.d.). pandas.get_dummies. https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html

Pandas. (2023). pandas documentation. https://pandas.pydata.org/docs/index.html

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

Peiffer-Smadja, N., Rawson, T. M., Ahmad, R., Buchard, A., Pantelis, G., Lescure, F. X., Birgand, G., & Holmes, A. H. (2020). Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, *26*(5), 584–595. https://doi.org/10.1016/J.CMI.2019.09.009

pyclustertend. (n.d.). Welcome to pyclustertend's documentation! — pyclustertend 1.4.0 documentation. https://pyclustertend.readthedocs.io/en/latest/

Raballo, A., Ramsey, L. B., Skokauskas, N., Røst, T. B., Clausen, C., Nytrø, Ø., Koposov, R., Leventhal, B., Bakken, V., Helen, L., Flygel, K., & Koochakpour, K. (2020). Local, Early, and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental Health Services. *11*. https://doi.org/10.3389/fpsyt.2020.564205

Satopää, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. *Proceedings - International Conference on Distributed Computing Systems*, 166–171. https://doi.org/10.1109/ICDCSW.2011.20

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Prakash Patel, O., Tiwari, A., Joo Er, M., Ding, W., & Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053

Shalabi, L. A., Shaaban, Z., & Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine. *Journal of Computer Science*, *2*(9), 735–739. https://doi.org/AlShalabi,L.,Shaaban,Z.,{\&}Kasasbeh,B.(2006).DataMining:APreprocessingEngine.JournalofComputerScience,

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80717–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Solheim, F. (2022). *Characterising Patients Referred on Suspicion of ADHD and Behavioral Difficulties* (tech. rep.). Norwegian University of Science and Technology.

Soni Madhulatha, T. (2012). An overview on clustering methods. *IOSR Journal of Engineering*, *2*(4), 719–725. https://doi.org/Anoverviewonclusteringmethods

Szepannek, G. (2019). ClustMixType: User-friendly clustering of mixed-type data in R. *R Journal*, *10*(2), 200–208. https://doi.org/10.32614/RJ-2018-048

Theodoridis, S., & Koutroumbas, K. (2008). *Pattern Recognition* (4th ed.). Elsevier.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

van de Velden, M., D'Enza, A. I., & Markos, A. (2019). Distance-based clustering of mixed data. *WIREs Comput Stat.*, *11*. http://arxiv.org/abs/1411.4911

Vazirgiannis, M. (2009). Clustering Validity. *Encyclopedia of Database Systems*, 388–393. https://doi.org/10.1007/978-0-387-39940-9{_}616

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021

WHO. (n.d.). International Classification of Diseases (ICD). https://www.who.int/standards/classifications/classification-of-diseases

Young, S., Fitzgerald, M., & Postma, M. J. (2013). *ADHD: making the invisible visible* (tech. rep.). http://www.russellbarkley.org/factsheets/ADHD_MakingTheInvisibleVisible.pdf

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artifical Intelligence*, *17*, 375–381. https://doi.org/10.1080/713827180

Appendix A

CAMHS in Norway Code Mappings

This appendix presents all data features from the St. Olavs database used in the experiment. Each feature is presented with the corresponding mapping to informative values. The mappings are based on BUPdata to NPR code mappings described in internal system documentation by AS, later Visma Unique (last updated 24.03.2010), the table *Koder* in the database and dialogue with RKBU Midt-Norge Odd-Sverre Wetbye and psychologist at *BUP poliklinikk Klostergata* Sanja Prodanovic. In this appendix, the features are presented in the order described in the *Koder* table.

A.1 Gender

Code	**Henvisningsgrunn**	**Referral Reason**
1	Jente	Female
2	Gutt	Male

Table A.1: Mapping between code and patient gender (referring to *Koder 1*).

A.2 Care Situation

Code	**Omsorgssituasjon**	**Care Situation**
1	Hos begge foreldrene	Both parents
2	Pendler mellom mor og far	Commutes between both parents
3	Bor hos en av foreldrene	One parent
4	En foreldre og samboer	One parent and partner
5	Hos besteforeldre/andre	Grandparents / other family
6	Bor i fosterhjem	Fostercare
7	Bor på institusjon	Institution
8	Alene	Alone
9	Annet	Other

Table A.2: Mapping between code and patients' care situation (referring to *Koder 7*).

A.3 Referral Reason

Note that all codes above 20 are old codes for referral reasons. These have been mapped to new codes when possible.

Code	Henvisningsgrunn	Referral Reason
1	Alvorlig bekymring for barn under 6 år	Serious concern for children under 6 years
2	Mistanke om gjennomgripende utviklingsforstyrrelse (autimse)	Suspicion of Autism
21	Autistiske trekk	
3	Mistanke om trasslidelse/adferdsforstyrrelse	Suspicion of defiance/conduct disorder
29	Atferdsvansker	
4	Mistanke om hyperkinetisk forstyrrelse (ADHD)	Suspicion of hyperkinetic disorder (ADHD)
30	Hyperaktiv/konsentrasjonsvansker	
5	Mistanke om Tourettes syndrom	Suspicion of Tourette's syndrome
6	Skolevegring	School refusal
7	Mistanke om angstlidelse	Suspicion of anxiety)
25	Angst/fobi	
8	Mistanke om tvangstanker-/tvangshandlinger	Suspicion of obsession)
26	Tvangstrekk	
9	Mistanke om spiseforstyrrelse	Suspicion of eating disorder)
36	Spiseproblemer	
10	Mistanke om depresjon	Suspicion of depression)
27	Tristhet/Depresjon/sorg	
11	Mistanke om bipolar lidelse	Suspicion of bipolar disroder
12	Vedvardende of alvorlig selskading	Suspicion of severe self harm
13	Mistanke om psykose	Suspicion of psychosis)
22	Psykotiske trekk	
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
16	Annet	Other reasons
31	Rusmiddelmisbruk	
32	Asosial/kriminalitet	
34	Språk/talevansker	
35	Syn/hørselsproblem	
37	Andre somatiske symptomer	
38	Annet	
20	Ikke fylt ut av henviser	Not set by referrer
39	Ingen	
23	Suicidalfare	Suicide risk
24	Hemmet atferd	Inhibited behavior
28	Skolefravær	Absence from school
33	Lærevansker	Learning difficulties

Table A.3: Mapping between code and referral reason (referring to *Koder 11*).

A.4 Immediacy Level

Code	Type øyeblikkelig hjelp	Assessment
1	Akutt = uten opphold / venting	Acute
2	Ikke Akutt (behandling innen 6 timer)	Non-acute
3	Venting mellom 6 og 24 timer	6-24 hours
4	Planlagt	Planned
4	Tilbakeføring av pasient fra annet sykehus	Return from another level

Table A.4: Mapping between code and EoC immediacy level (referring to *Koder 13*).

A.5 Assessment

Code	Tatt imot	Assessment
1	Tatt imot	Accepted
2	Avlsag pga kapasitet	Rejected due to capacity
3	Avslag faglig grunn	Rejected due to professional reasons
4	Foreløpig vurdering	Assessment so far

Table A.5: Mapping between code and EoC Bundle assesment (referring to *Koder 19*).

A.6 Closing Code

Code	Avslutningskode	Closing code
1	Fullførst oppdrag	Assignment completed
2	Pasienten avbrød	Patient cancelled
3	Foresatte avbrød	Guardians cancelled
4	Over aldersgrensen	Above age
5	Flyttet / feil disrtrikt	Moved / wrong district
6	Død	Death
7	Avslag	Rejected
8	Kom ikke i gang	Did not get started
9	Annet	Other

Table A.6: Mapping between code and EoC Bundle closing code (referring to *Koder 22*).

A.7 Contact Type

Code	Kontakt type	Contact type
1	Terapi/samtale	Therapy
2	Undersøkelse/observasjon	Examination
3	Indirekte pasient arbeid / rådgivning	Indirect contact
4	Intern beh. planlegging	Planning
5	Ikke møt	No-show

Table A.7: Mapping between code and contact type (referring to *Koder 31*).

A.8 Care Level

Code	**Omsorgsnivå**	**Care level**
1	Døgnopphold	Therapy
2	Undersøkelse/observasjon	Examination
3	Indirekte pasient arbeid / rådgivning	Indirect contact
4	Intern beh. planlegging	Planning
5	Ikke møt	No-show

Table A.8: Mapping between code and contact type (referring to *NPR kodeverk 8406* and mapping from Westbye).

Appendix B

Code

B.1 Experiment Code

To ensure conciseness in the appendix, only the code corresponding to the final iteration of the clustering process will be included. This code contains all the necessary implementation details to understand the clustering process. By focusing on the final iteration, the essential information is captured without unnecessary duplication of code.

B.1.1 Data Preparation Code

The following sections present the data preparation code written for the final iteration.

Third Iteration EoC PostgreSQL Query

The following code presents the PostgreSQL Query for the EoC data in the third iteration.

```
SELECT
    opphold.nr  AS eoc_id,
    opphold.igangdato AS start_date,
    opphold.avsldato  AS end_date,
    min_journal_date,
    max_journal_date,
    opphold.omsniva AS care_level,
    opphold.ohjelp AS immediacy_level,
    nr_contacts,
    nr_therapy,
    nr_planning,
    nr_examination ,
    nr_no_show,
    nr_indirect_contact,
    nr_diagnoses_1,
    nr_diagnoses_2,
    nr_diagnoses_3,
    nr_diagnoses_4,
    nr_diagnoses_5,
```

```
    nr_diagnoses_6,
    nr_main_diagnoses,
    nr_diagnoses,
    sak.icd1 AS axis1_start,
    beforediagnose.count AS contacts_Before_Primary_axis1_diagnosis
FROM
    sak
INNER JOIN  (
SELECT
    *
FROM
    opphold
)AS opphold
ON  opphold.sak = sak.nr
LEFT JOIN (
SELECT
    journal.opphold,
    COUNT(distinct journal.nr) AS nr_contacts ,
    SUM(CASE WHEN journal.type1 = 1 THEN 1 ELSE 0 end) AS nr_therapy,
    SUM(CASE WHEN journal.type1 = 2 THEN 1 ELSE 0 end) AS nr_planning,
    SUM(CASE WHEN journal.type1 = 3 THEN 1 ELSE 0 end) AS nr_examination,
    SUM(CASE WHEN journal.type1 = 4 THEN 1 ELSE 0 end) AS nr_no_show,
    SUM(CASE WHEN journal.type1 = 5 THEN 1 ELSE 0 end) AS nr_indirect_contact,
    min(journal.dato1) AS min_journal_date,
    max(journal.dato1) AS max_journal_date
FROM
    journal
WHERE
    journal.type1=1
    OR journal.type1=2
    OR journal.type1=3
    OR journal.type1=4
    OR journal.type1=5
GROUP BY
    journal.opphold) AS journal ON
opphold.nr = journal.opphold
LEFT JOIN  (
SELECT
    diagnose.opphold,
    COUNT(CASE WHEN diagnose.akse = 1  THEN diagnose.diagnose end) AS nr_diagnoses_1,
    COUNT(CASE WHEN diagnose.akse = 2  THEN diagnose.diagnose end) AS nr_diagnoses_2,
    COUNT(CASE WHEN diagnose.akse = 3  THEN diagnose.diagnose end) AS nr_diagnoses_3,
    COUNT(CASE WHEN diagnose.akse = 4  THEN diagnose.diagnose end) AS nr_diagnoses_4,
    COUNT(CASE WHEN diagnose.akse = 5  THEN diagnose.diagnose end) AS nr_diagnoses_5,
    COUNT(CASE WHEN diagnose.akse = 6  THEN diagnose.diagnose end) AS nr_diagnoses_6,
    SUM(CASE WHEN diagnose.hoved = 1 THEN 1 ELSE 0 end) AS nr_main_diagnoses,
    COUNT(distinct diagnose.nr) AS nr_diagnoses
FROM
    diagnose
WHERE
    diagnose.diagnose is NOT null
    AND ( diagnose.akse=1
    OR diagnose.akse=2
```

```
    OR diagnose.akse=3
    OR diagnose.akse=4
    OR diagnose.akse=5
    OR diagnose.akse=6)
GROUP BY diagnose.opphold
)AS diagnose
ON  opphold.nr = diagnose.opphold
/*Adding the nr of contacts before diagnose on axis one WHERE hoved=1*/
LEFT JOIN (
SELECT
    opphold.nr AS opphold,
    d.dato,
    COUNT(distinct j.nr)
FROM
    sak
INNER JOIN  (
    SELECT
        *
    FROM
        opphold
)AS opphold
    ON  opphold.sak = sak.nr
LEFT JOIN (
    SELECT
        t.dato,
        t.diagnose,
        t.opphold,
        t.sak
        FROM (
         SELECT
             *,
             ROW_NUMBER() OVER (PARTITION  BY diagnose.opphold
             ORDER BY diagnose.dato AS C) AS row_number
         FROM diagnose
         WHERE akse = 1 AND hoved = 1
        ) t
            WHERE
                t.row_number = 1
                /*Main diagnose should not be any of the following codes*/
                AND NOT ( t.diagnose ='999'
                 OR t.diagnose ='f99'
                 OR t.diagnose ='1999'
                 OR t.diagnose ='1000'
                 OR t.diagnose ='000')
                ) AS d ON  d.opphold= opphold.nr
LEFT JOIN (
    SELECT
        journal.nr,
        journal.dato1,
        journal.opphold
    FROM
        journal
    WHERE
```

```
        (journal.type1=1
        OR journal.type1=2
        OR journal.type1=3
        OR journal.type1=4
        OR journal.type1=5) )AS j ON  j.opphold = opphold.nr
WHERE
/*Adding the same WHERE clauses here to optimise the code*/
    ((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 ='F900'
            OR sak.icd1 ='F901'
            OR sak.icd1 ='F908'
            OR sak.icd1 ='F909'))
        AND NOT (sak.avslkode = 0 AND sak.avsldato is null)
        AND NOT (sak.tattimot = 2 OR sak.tattimot = 3)
        AND j.dato1 <= d.dato
    GROUP BY
        opphold.nr,
        d.dato
) AS beforediagnose ON  beforediagnose.opphold = opphold.nr
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 ='F900'
        OR sak.icd1 ='F901'
        OR sak.icd1 ='F908'
        OR sak.icd1 ='F909'))
    AND NOT (sak.avslkode = 0 AND sak.avsldato is null)
    AND NOT (sak.tattimot = 2 OR sak.tattimot = 3)
    AND NOT (sak.avslkode = 7 OR sak.avslkode = 8)
ORDER BY
    sak.nr,
    opphold.nr
```

Third Iteration EoC Bundle PostgreSQL Query

The following code presents the PostgreSQL Query for the EoC Bundle data in the third iteration.

```
SELECT
    sak.nr AS eoc_bundle_id,
    sak.igangdato  AS eoc_bundle_start_date,
    sak.avsldato AS eoc_bundle_end_date,
    pasient.fdt AS birth_date,
    pasient.omsorg1 AS care_situation ,
    min_journal_date,
    max_journal_date,
    pasient.kjon n AS gender,
    sak.icd1 AS diagnose_axis_1,
    sak.icd2 AS diagnose_axis_2,
    sak.icd3 AS diagnose_axis_3,
    sak.icd4 AS diagnose_axis_4,
    sak.icd5 AS diagnose_axis_5,
    sak.icd6 AS diagnose_axis_6,
    sak.avslkode AS closing_code,
    opphold.nr AS eoc_id
    FROM pasient
    INNER JOIN (
    SELECT
        *
    FROM
        sak
    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 ='F900'
            OR sak.icd1 ='F901'
            OR sak.icd1 ='F908'
            OR sak.icd1 ='F909'))
        and not (sak.avslkode = 0 and sak.avsldato is null)
        and not (sak.tattimot = 2 OR sak.tattimot = 3)
        and not (sak.avslkode = 7 OR sak.avslkode = 8)
    )AS sak ON  pasient.nr = sak.pasient
    INNER JOIN  (
    SELECT
        *
    FROM
```

```
        opphold
    )AS opphold
        ON  opphold.sak = sak.nr
    LEFT JOIN  (
    SELECT
        journal.sak,
        min(journal.dato1) AS min_journal_date,
        max(journal.dato1) AS max_journal_date
   FROM
        journal
    WHERE
        journal.type1=1
        OR journal.type1=2
        OR journal.type1=3
        OR journal.type1=4
        OR journal.type1=5
    GROUP BY
        journal.sak) AS journal ON
        sak.nr = journal.sak
    ORDER BY
        sak.nr,
        opphold.nr
```

Third Iteration's EoC Preprocessing

The following code presents the preprocessing of the third iteration's EoC Data.

```
# Import necessary packages
import pandas as pd
import numpy as np
pd.set_option('display.max_columns', 30)
df = pd.read_csv("../../Data/Third_iteration_vol2/EoC.csv")

df = df.drop(['nr_contacts'], axis = 1)#removing total number of contacts

# Changing to categorical values, adding 0 = "missing data" to handle errorprone values
values_list = [1,2,3]
df.loc[~df["care_level"].isin(values_list), "care_level"] = "Missing_data"
df.loc[ df["care_level"] == 1, "care_level"] = "Polyclinic"
df.loc[ df["care_level"] == 2, "care_level"] = "Outpatient"
df.loc[ df["care_level"] == 3, "care_level"] = "Inpatient"

values_list = [1,2,3,4,5]
df.loc[~df["immediacy_level"].isin(values_list), "immediacy_level"] = "Missing_data"
df.loc[ df["immediacy_level"] == 1, "immediacy_level"] = "Acute"
df.loc[ df["immediacy_level"] == 2, "immediacy_level"] = "Non_acute"
df.loc[ df["immediacy_level"] == 3, "immediacy_level"] = "6-24_hour_wait"
df.loc[ df["immediacy_level"] == 4, "immediacy_level"] = "Planned"
df.loc[ df["immediacy_level"] == 5, "immediacy_level"] = "Return_from_another_hospital"

# Cleaning the icd1 codes set on axis 1 at beginning of the coresponding EoC_Bundle to determine the change og NULL values for the nr_contacts_before_primary_axis1_diagnosis

df['axis1_start'] = df['axis1_start'].astype(str)
```

```
27
28 def conditionsICD1F(f):
29     x = (f[1:])
30     if ((x == '21') | (x == '28') | (x == '29')):
31         return 'Schizophrenia/schizotypy/other_mental_disorders'
32     elif (x =='89'):
33         return 'Missing_data'
34     elif (x=='54'):
35         return 'Behavioral_syndromes_associated_with_physiological_disturbances/
    physical_factors'
36     elif (x=='99'):
37         return 'Unspecified'
38     else:
39         y = int(x[:2])
40         if (y <=9):
41             return 'Organic_including_symptomatic_psychological_disorders'
42         elif ((y >= 10) & (y <= 19)):
43             return 'Mental/behavioral_disorders_caused_by_psychoactive_substances'
44         elif ((y >= 20) & (y<= 29)):
45             return 'Schizophrenia/schizotypy/other_mental_disorders'
46         elif ((y >= 30) & (y <= 39)):
47             return 'Mood_disorders'
48         elif ((y >= 40) & (y<= 48)):
49             return 'Neurotic/stress-related/somatoform_disorders'
50         elif ((y >= 50) & (y <= 59)):
51             return 'Behavioral_syndromes_associated_with_physiological_disturbances
    /physical_factors'
52         elif ((y >= 60) & (y<= 69)):
53             return 'Personality_and_behavioral_disorders_in_adults'
54         elif ((y >= 70) & (y<= 79)):
55             return 'Missing_data'
56         elif ((y >= 80) & (y<= 89) &(y!=84) ):
57             return 'Missing_data'
58         elif (y==84):
59             return 'Intellectual_disability'
60         elif (y==90):
61             return 'Hyperkinetic_disorders'
62         elif ((y >= 91) & (y <= 98)):
63             return 'Other_behavioral/
    emotional_disorders_usually_occuring_in_children_and_adolescents'
64         else:
65             return f
66
67 def conditionsICD1(icd1):
68     if (icd1 =='nan'):
69         return 'Missing_data'
70     elif ((icd1 =='1999') | (icd1 =='999')):
71           return 'Missing_information'
72     elif ((icd1 =='1000') | (icd1 =='000')):
73           return 'No_diagnosis'
74     elif (icd1 =='Z00.4'):
75         return 'Contact_for_examination_and_investigation'
76     else:
77         if(icd1[:1]=='R'):
78             #Only interested in if a diagnosis is set or not
79            return icd1
80         elif(icd1[:1]=='Z'):
81              #Only interested in if a diagnosis is set or not
82            return icd1
83         elif(icd1[:1]=='F'):
84             #investigate if a diagnosis is set or if it is F999
85            return conditionsICD1F(icd1)
86         else:
87             return icd1
88
89 func = np.vectorize(conditionsICD1)
```

```
axis1 = func(df['axis1_start'])
df["axis1_start"] = axis1

# If the code on axis 1 at the beginning of the EoC Bundle is missing_data,
    missing_information, unspecified or no_diagnosis and no diagnosis is set as the
     primary axis diagnosis during the EoC (therefore the value
    contacts_before_primary_axis1_diagnosis is NULL),
    contacts_before_primary_axis1_diagnosisshould be set to 1000

# If there is a code given on axis 1 at the beginning of the EoC Bundle, but no
    code is set during the EoC, the contacts_before_primary_axis1_diagnosis should
    be set to 0

values_list = ['Missing_data', 'Missing_information','No_diagnosis','Unspecified']
df.loc[df["axis1_start"].isin(values_list) & df["
    contacts_before_primary_axis1_diagnosis"].isnull() , "
    contacts_before_primary_axis1_diagnosis"] = "1000"
df.loc[~df["axis1_start"].isin(values_list) & df["
    contacts_before_primary_axis1_diagnosis"].isnull(), "
    contacts_before_primary_axis1_diagnosis"] = "0"

df[['nr_therapy', 'nr_planning', 'nr_examination', 'nr_no_show', '
    nr_indirect_contact', 'nr_diagnoses_1', 'nr_diagnoses_2','nr_diagnoses_3', '
    nr_diagnoses_4', 'nr_diagnoses_5', 'nr_diagnoses_6','nr_main_diagnoses', '
    nr_diagnoses']] = df[['nr_therapy', 'nr_planning', 'nr_examination', '
    nr_no_show', 'nr_indirect_contact', 'nr_diagnoses_1', 'nr_diagnoses_2', '
    nr_diagnoses_3', 'nr_diagnoses_4', 'nr_diagnoses_5', 'nr_diagnoses_6','
    nr_main_diagnoses', 'nr_diagnoses']].fillna(0)


# Setting the EoC length based on start date and enddate or journal date1 based on
    missing values
df[['max_journal_date']] = df[['max_journal_date']].replace(dict.fromkeys(['
    2916-03-30'], '2017-04-06')) # Changing an out out bounce error value to the
    end date of this  specific EoC

# Set all dates to datetime to enable calculations with them
df['start_date'] = pd.to_datetime(df['start_date'])
df['end_date'] = pd.to_datetime(df['end_date'])
df['min_journal_date'] = pd.to_datetime(df['min_journal_date'])
df['max_journal_date'] = pd.to_datetime(df['max_journal_date'])


# Calculating the EoC length based on the four dates, depending on NULL values
def conditionsLength(start, end, minJournal, maxJournal):
    if (str(start) != 'NaT') & (str(end)!='NaT') :
        if (int(str(pd.Timedelta(end - start)).split(' ',1)[0]) < 0) & (str(
    minJournal) != 'NaT') & (str(maxJournal)!='NaT'):
            return pd.Timedelta(maxJournal - minJournal).days
        else : return str(pd.Timedelta(end - start)).split(' ',1)[0]
    elif (str(start) == 'NaT') & (str(end)=='NaT') & (str(minJournal)!='NaT')& (str
    (maxJournal)!='NaT') :
        return pd.Timedelta(maxJournal - minJournal).days
    elif (str(start) == 'NaT') & (str(end)!='NaT') & (str(minJournal)!='NaT') :
        return pd.Timedelta(end - minJournal).days
    elif (str(start) != 'NaT') & (str(end)=='NaT') & (str(minJournal)!='NaT'):
        return pd.Timedelta(maxJournal - start).days
    else:  return pd.NaT


func = np.vectorize(conditionsLength)
LengthNew = func(df["start_date"], df["end_date"], df["min_journal_date"], df["
    max_journal_date"])
df["EoC_length"] = LengthNew

# Remove values that are NULL after calculations
```

```
df = df[df.EoC_length != "NaT"]

# Changing all EoC lengths to integer values
df = df.astype({'EoC_length':'int'})

# Removing all dates used to derive the EoC lengths
df = df.drop(['start_date', 'end_date', 'min_journal_date', 'max_journal_date'],
    axis = 1)

# Contacts per day
def conditionsContact(contact, EoClength):
    if (contact == 0):
        return float(0)
    elif (EoClength == 0 ):
        return float(contact)
    else:
        return float(contact/EoClength)

func = np.vectorize(conditionsContact)

therapyNew = func(df["nr_therapy"], df["EoC_length"])
nrplanningNew = func(df["nr_planning"], df["EoC_length"])
nrexaminationNew = func(df["nr_examination"], df["EoC_length"])
nrnoshow = func(df["nr_no_show"], df["EoC_length"])
nrindirectcontactNew = func(df["nr_therapy"], df["EoC_length"])

df["nr_therapy_per_day"] = therapyNew
df["nr_planning_per_day"] = nrplanningNew
df["nr_examination_per_day"] = nrexaminationNew
df["nr_no_show_per_day"] = nrnoshow
df["nr_indirect_contact_per_day"] = nrindirectcontactNew

# Diagnoses per day
def conditionsDiagnose(diagnose, EoClength):
    if (diagnose == 0):
        return float(0)
    elif (EoClength == 0 ):
        return float(EoClength)
    else:
        return float(diagnose/EoClength)

func = np.vectorize(conditionsDiagnose)

nrdiagnoses1New = func(df["nr_diagnoses_1"], df["EoC_length"])
nrdiagnoses2New = func(df["nr_diagnoses_2"], df["EoC_length"])
nrdiagnoses3New = func(df["nr_diagnoses_3"], df["EoC_length"])
nrdiagnoses4New = func(df["nr_diagnoses_4"], df["EoC_length"])
nrdiagnoses5New = func(df["nr_diagnoses_5"], df["EoC_length"])
nrdiagnoses6New = func(df["nr_diagnoses_6"], df["EoC_length"])

df["nr_diagnoses_1_per_day"] = nrdiagnoses1New
df["nr_diagnoses_2_per_day"] = nrdiagnoses2New
df["nr_diagnoses_3_per_day"] = nrdiagnoses3New
df["nr_diagnoses_4_per_day"] = nrdiagnoses4New
df["nr_diagnoses_5_per_day"] = nrdiagnoses5New
df["nr_diagnoses_6_per_day"] = nrdiagnoses6New

# Percentage primary axis diagnose
def conditionsMainDiagnose(diagnosetotal, main):
    if (main == 0):
        return float(0)
    elif (diagnosetotal == 0 ):
        return float(main)
    else:
        return float(main/diagnosetotal)

```

```
func = np.vectorize(conditionsMainDiagnose)

mainNew = func(df["nr_diagnoses"], df["nr_main_diagnoses"])

df["percentage_primary_axis_diagnose"] = mainNew

# Removing the columns not to be included in the experiment
df = df.drop(['nr_therapy','nr_planning', 'nr_examination', 'nr_no_show','
    nr_indirect_contact', 'nr_diagnoses_1', 'nr_diagnoses_2', 'nr_diagnoses_3', '
    nr_diagnoses_4', 'nr_diagnoses_5', 'nr_diagnoses_6',  'nr_main_diagnoses', '
    nr_diagnoses','axis1_start' ], axis = 1 )

# Saving the preprocessed data to file
df.to_csv('EoC_preprocessed.csv', index=False)
```

Listing B.1: EoC Preprossesing (Third Iteration).

Third Iteration's EoC Bundle Preprocessing

The following code presents the preprocessing of the third iteration's EoC Bundle Data.

```
# Import necessary packages
import pandas as pd
import numpy as np
df = pd.read_csv("../../Data/Third_iteration_vol2/EoC_Bundle.csv")
pd.set_option('display.max_columns', 30)
pd.set_option('display.max_rows', 200)

#EoC Bundle Length and Age at EoC Bundle Start
#Setting the EoC Bundle length based on the start date and end date or journal
    date1 based on missing values
Set all dates to datetime to enable calculations with them

# Changing an out-of-bounce error value to the end date of this  specific EoC
df[['max_journal_date']] = df[['max_journal_date']].replace(dict.fromkeys(['
    2916-03-30'], '2017-04-06'))

df['EoC_Bundle_start_date'] = pd.to_datetime(df['EoC_Bundle_start_date'])
df['EoC_Bundle_end_date'] = pd.to_datetime(df['EoC_Bundle_end_date'])
df['min_journal_date'] = pd.to_datetime(df['min_journal_date'])
df['max_journal_date'] = pd.to_datetime(df['max_journal_date'])
df['birth_date'] = pd.to_datetime(df['birth_date'])

#Calculating the EoC length based on the four dates, depending on NULL values
def conditionsLength(start, end, minJournal, maxJournal):
    if (str(start) != 'NaT') & (str(end)!='NaT') :
        if (int(str(pd.Timedelta(end - start)).split(' ',1)[0]) < 0) & (str(
    minJournal) != 'NaT') & (str(maxJournal)!='NaT'):
            return pd.Timedelta(maxJournal - minJournal).days
        else : return str(pd.Timedelta(end - start)).split(' ',1)[0]
    elif (str(start) == 'NaT') & (str(end)=='NaT') & (str(minJournal)!='NaT')& (str
    (maxJournal)!='NaT') :
        return pd.Timedelta(maxJournal - minJournal).days
    elif (str(start) == 'NaT') & (str(end)!='NaT') & (str(minJournal)!='NaT') :
        return pd.Timedelta(end - minJournal).days
    elif (str(start) != 'NaT') & (str(end)=='NaT') & (str(maxJournal)!='NaT'):
        return pd.Timedelta(maxJournal - start).days
    else:  return pd.NaT

func = np.vectorize(conditionsLength)
Length = func(df["EoC_Bundle_start_date"], df["EoC_Bundle_end_date"], df["
    min_journal_date"], df["max_journal_date"])
df["EoC_Bundle_length"] = Length

#Remove values that are NULL after calculations
```

```
df = df[df.EoC_Bundle_length != "NaT"]

#Changing all EoC Bundle lengths to integer values
df = df.astype({'EoC_Bundle_length':'int'})

#Removing all negative lengths
df = df[df.EoC_Bundle_length >= 0]

#Age at EoC Bundle Start
def conditionsAge(start, minJournal, birth):
    if (str(start) != 'NaT') & (str(birth)!='NaT') :
        return (pd.Timedelta(start - birth).days / 365.25)//1
    elif (str(start) == 'NaT') & (str(minJournal)!='NaT')& (str(birth)!='NaT') :
        return (pd.Timedelta(minJournal - birth).days / 365.25)//1
    else:  return pd.NaT

func = np.vectorize(conditionsAge)
Age = func(df["EoC_Bundle_start_date"], df["min_journal_date"], df["birth_date"])
df["age_EoC_Bundle_start"] = Age
df = df.astype({'age_EoC_Bundle_start':'int'})

df.loc[df['age_EoC_Bundle_start']==47] #Investigating the one age outlier
df = df[df.age_EoC_Bundle_start != 47] #removing this outlier

#Remove data values
df = df.drop(['EoC_Bundle_start_date', 'EoC_Bundle_end_date', 'min_journal_date', '
    max_journal_date', 'birth_date'], axis = 1)

#Gender
values_list = [1,2]
df.loc[~df["gender"].isin(values_list), "gender"] = "Missing_data"
df.loc[ df["gender"] == 1, "gender"] = "Female"
df.loc[ df["gender"] == 2, "gender"] = "Male"

#Closing code
values_list = [1,2,3,4,5,6,9]
df.loc[~df["closing_code"].isin(values_list), "closing_code"] = "Missing_data"
df.loc[ df["closing_code"] == 1, "closing_code"] = "Assignment_completed"
df.loc[ df["closing_code"] == 2, "closing_code"] = "Patient_cancelled"
df.loc[ df["closing_code"] == 3, "closing_code"] = "Parents_cancelled"
df.loc[ df["closing_code"] == 4, "closing_code"] = "Above_age"
df.loc[ df["closing_code"] == 5, "closing_code"] = "Moved/wrong_district"
df.loc[ df["closing_code"] == 6, "closing_code"] = "Death"
df.loc[ df["closing_code"] == 9, "closing_code"] = "Other"


#Diagnoses

#Axis 1
df['diagnosis_axis_1'] = df['diagnosis_axis_1'].astype(str)
def conditionsICD1R(r):
    x = (r[1:])
    y = int(x[:2])
    if (y ==40):
        return 'Somnolence_stupor_coma'
    elif (y ==41):
        return 'Symptoms_assosiated_with_cognitive_functions'
    elif (y ==42):
        return 'Dizziness'
    elif (y ==43):
        return 'Disturbances_smell_and_taste'
    elif (y ==44):
        return 'Symptoms_assosiated_general_sensations_and_perseptions'
    elif (y ==45):
        return 'Symptoms_assosiated_with_emotional_state'
    elif (y ==46):
```

```
105         return 'Symptoms_assosiated_with_looks'
106     else:
107         return r
108 def conditionsICD1Z(z):
109     x = (z[1:])
110     y = int(x[:2])
111     if (y <=13):
112         return 'Contact_for_examination_and_investigation'
113     elif ((y >= 55) & (y <= 65)):
114         return 'Contact_due_to_potential_health_risk_socio-
    economic_and_psychosocial_conditions'
115     elif ((y >= 70 )&( y <= 76)):
116         return 'Contact_for_other_circumstances'
117     elif ((y >= 80) & (y <= 99)):
118         return 'Contact_due_to_information_regarding_potential_health_risk_family/
    personal_history'
119     else:
120         return z
121 def conditionsICD1F(f):
122     x = (f[1:])
123     if ((x == '21') | (x == '28') | (x == '29')):
124         return 'Schizophrenia/schizotypy/other_mental_disorders'
125     elif (x =='89'):
126         return 'Missing_data'
127     elif (x=='54'):
128         return 'Behavioral_syndromes_associated_with_physiological_disturbances/
    physical_factors'
129     elif (x=='99'):
130         return 'Unspecified'
131     else:
132         y = int(x[:2])
133         if (y <=9):
134             return 'Organic_including_symptomatic_psychological_disorders'
135         elif ((y >= 10) & (y <= 19)):
136             return 'Mental/behavioral_disorders_caused_by_psychoactive_substances'
137         elif ((y >= 20) & (y<= 29)):
138             return 'Schizophrenia/schizotypy/other_mental_disorders'
139         elif ((y >= 30) & (y <= 39)):
140             return 'Mood_disorders'
141         elif ((y >= 40) & (y<= 48)):
142             return 'Neurotic/stress-related/somatoform_disorders'
143         elif ((y >= 50) & (y <= 59)):
144             return 'Behavioral_syndromes_associated_with_physiological_disturbances
    /physical_factors'
145         elif ((y >= 60) & (y<= 69)):
146             return 'Personality_and_behavioral_disorders_in_adults'
147         elif ((y >= 70) & (y<= 79)):
148             return 'Missing_data'
149         elif ((y >= 80) & (y<= 89) &(y!=84) ):
150             return 'Missing_data'
151         elif (y==84):
152             return 'Intellectual_disability'
153         elif (y==90):
154             return 'Hyperkinetic_disorders'
155         elif ((y >= 91) & (y <= 98)):
156             return 'Other_behavioral/
    emotional_disorders_usually_occuring_in_children_and_adolescents'
157         else:
158             return f
159
160 def conditionsICD1(icd1):
161     if (icd1 =='nan'):
162         return 'Missing_data'
163     elif ((icd1 =='1999') | (icd1 =='999')):
164           return 'Missing_information'
165     elif ((icd1 =='1000') | (icd1 =='000')):
```

```
166             return 'No_diagnosis'
167     elif (icd1 =='Z00.4'):
168         return 'Contact_for_examination_and_investigation'
169     else:
170         if(icd1[:1]=='R'):
171             #Alternative
172             #return 'Symptoms/signs/
    abnormal_clinical_findings_and_laboratory_findings'
173             return conditionsICD1R(icd1)
174         elif(icd1[:1]=='Z'):
175             #Alternative
176             #return '
    Factors_impacting_health_status_and_contact_with_health_service'
177             return conditionsICD1Z(icd1)
178         elif(icd1[:1]=='F'):
179             return conditionsICD1F(icd1)
180         else:
181             return icd1
182
183 func = np.vectorize(conditionsICD1)
184 axis1 = func(df['diagnosis_axis_1'])
185 df["diagnosis_axis_1"] = axis1
186
187 df["diagnosis_axis_1"].fillna('Missing_data', inplace=True)
188
189 #Axis 2
190 df['diagnosis_axis_2'] = df['diagnosis_axis_2'].astype(str)
191 def conditionsICD2(icd2):
192     #Missing information
193     if ((icd2 == '999') | (icd2 == '2999')):
194         return "No"
195     #No diagnosis
196     elif ((icd2 == '000') | (icd2 =='2000')):
197         return "No"
198     #Nan or invalid diagnosis
199     elif ((icd2 == 'nan') | (icd2 == 'F84')):
200         return "No"
201     else:
202         return "Yes"
203
204 func = np.vectorize(conditionsICD2)
205 icd2 = func(df['diagnosis_axis_2'])
206 df["diagnosis_axis_2"] = icd2
207 df["diagnosis_axis_2"].fillna('No', inplace=True)
208
209 #Axis 3
210 df['diagnosis_axis_3'] = df['diagnosis_axis_3'].astype(str)
211 def conditionsICD3(icd3):
212     #Missing information
213     if ((icd3 == '99') | (icd3 == '39')):
214         return "No"
215     #No diagnosis
216     elif ((icd3 == '30')):
217         return "No"
218     #Nan or invalid diagnosis
219     elif ((icd3 == 'nan')):
220         return "No"
221     else:
222         return "Yes"
223
224 func = np.vectorize(conditionsICD3)
225 icd3 = func(df['diagnosis_axis_3'])
226 df["diagnosis_axis_3"] = icd3
227 df["diagnosis_axis_3"].fillna('No', inplace=True)
228
229 #Axis 4
```

```
df['diagnosis_axis_4'] = df['diagnosis_axis_4'].astype(str)
def conditionsICD4(icd4):
    #Missing information
    if ((icd4 == '4999') | (icd4 == '999')):
        return "No"
    #No diagnosis
    elif ((icd4 == '4000') | (icd4 =='000')):
        return "No"
    #Nan or invalid diagnosis
    elif ((icd4 == 'nan')):
        return "No"
    else:
        return "Yes"

func = np.vectorize(conditionsICD4)
icd4 = func(df['diagnosis_axis_4'])
df["diagnosis_axis_4"] = icd4
df["diagnosis_axis_4"].fillna('No', inplace=True)

#Axis 5
df['diagnosis_axis_5'] = df['diagnosis_axis_5'].astype(str)
def conditionsICD5(icd5):
    #Missing information
    if ((icd5 == '99.0') | (icd5 == '599.0')  | (icd5 == '0.0')):
        return "No"
    #No diagnosis
    elif ((icd5 == '000') | (icd5 =='500.0')):
        return "No"
    #Nan or invalid diagnosis
    elif ((icd5 == 'nan') | (icd5 =='1.5')):
        return "No"
    else:
        return "Yes"

func = np.vectorize(conditionsICD5)
icd5 = func(df['diagnosis_axis_5'])
df["diagnosis_axis_5"] = icd5
df["diagnosis_axis_5"].fillna('No', inplace=True)

#Axis 6
df['diagnosis_axis_6'].median()
df['diagnosis_axis_6'].mean()
df['diagnosis_axis_6'].fillna(5, inplace=True)

#Save the preprocessed data
df.to_csv('EoC_Bundle_preprocessed.csv', index=False)
```

Listing B.2: EoC Bundle Preprossesing (Third Iteration).

B.1.2 Clustering

Finding k for the Third Iteration's EoC Clustering

The following code finds an optimal number of clusters, k, for the third iteration's EoC data.

```
# Import necessary packages
import pandas as pd
import numpy as np
from kmodes.kprototypes import KPrototypes
from sklearn.preprocessing import PowerTransformer
from tqdm import tqdm
import plotly.graph_objs as go
from plotnine import *
import plotnine
```

```
from kneed import KneeLocator

# Format scientific notation from Pandas
pd.set_option('display.float_format', lambda x: '%.3f' % x)

# Importing data from a CSV file and saving it as a data frame
df = pd.read_csv('EoC_preprocessed.csv')

# Remove EoC ID in order to prepare for the clustering
df_cluster = df.drop(['EoC_id'], axis = 1)
df_cluster.head()

# Find the optimal number of clusters using the elbow method
kprot_data = df_cluster.copy()
for c in df_cluster.select_dtypes(exclude='object').columns:
    pt = PowerTransformer()
    kprot_data[c] =  pt.fit_transform(np.array(kprot_data[c]).reshape(-1, 1))

# Get the position of categorical columns
categorical_columns = [df_cluster.columns.get_loc(col) for col in list(df_cluster.
    select_dtypes('object').columns)]
print('Categorical columns           : {}'.format(list(df_cluster.select_dtypes('
    object').columns)))
print('Categorical columns position  : {}'.format(categorical_columns))

# Finding k using the Elbow method, using the k-prototypes algorithm initialised
    with Cao
costs = []
n_clusters = []
clusters_assigned = []

for i in tqdm(range(1, 11)):
    try:
        kproto = KPrototypes(n_clusters= i, init='Cao', verbose=2)
        clusters = kproto.fit_predict(kprot_data, categorical=categorical_columns)
        costs.append(kproto.cost_)
        n_clusters.append(i)
        clusters_assigned.append(clusters)
    except:
        print(f"Can't cluster with {i} clusters")

fig = go.Figure(data=go.Scatter(x=n_clusters, y=costs ))
fig.show()

# Converting the results into a data frame and plotting them
df_cost = pd.DataFrame({'Cluster':range(1, 11), 'Cost':costs})

plotnine.options.figure_size = (8, 4.8)
(
    ggplot(data = df_cost)+
    geom_line(aes(x = 'Cluster',
                  y = 'Cost'))+
    geom_point(aes(x = 'Cluster',
                   y = 'Cost'))+
    geom_label(aes(x = 'Cluster',
                   y = 'Cost',
                   label = 'Cluster'),
               size = 11,
               nudge_y = 1000) +
    labs(title = 'Optimal number of cluster with Elbow Method')+
    xlab('Number of Clusters k')+
    ylab('Cost')+
    theme_minimal()
)

```

```
# Confirm visual clue of the elbow plot using the KneeLocator class to detect
    elbows if the curve is convex

cost_knee_c3  = KneeLocator(
        range(1,11),
        costs,
        S=0.1, curve="convex", direction="decreasing", online=True)

K_inertia_b3 = cost_knee_c3 .elbow
print("elbow at k =", f'{K_inertia_b3:.0f} clusters')
```

Listing B.3: Find k for the EoC clustering (third iteration).

Cluster the EoC data

The following code cluster the third iteration's EoC data using k-prototypes with the identified optimal number of clusters.

```
# Import necessary packages
import pandas as pd
import numpy as np
from kmodes.kprototypes import KPrototypes
from sklearn.preprocessing import PowerTransformer
import matplotlib.pyplot as plt
import seaborn as sns
from tqdm import tqdm
import plotly.graph_objs as go
import plotnine
from lightgbm import LGBMClassifier
import shap
from sklearn.model_selection import cross_val_score

# Format scientific notation from Pandas
pd.set_option('display.float_format', lambda x: '%.3f' % x)

# Importing data from a CSV file and saving it as a data frame
df = pd.read_csv('EoC_preprocessed.csv')

# Prepare the data for clustering
df_cluster = df.drop(['EoC_id'], axis = 1)
df_cluster.head()

# Transform the continuous data to prepare for the clustering
kprot_data = df_cluster.copy()
for c in df_cluster.select_dtypes(exclude='object').columns:
    pt = PowerTransformer()
    kprot_data[c] =  pt.fit_transform(np.array(kprot_data[c]).reshape(-1, 1))

# Get the position of categorical columns
categorical_columns = [df_cluster.columns.get_loc(col) for col in list(df_cluster.
    select_dtypes('object').columns)]
print('Categorical columns           : {}'.format(list(df_cluster.select_dtypes('
    object').columns)))
print('Categorical columns position  : {}'.format(categorical_columns))

# Cluster using the k-prototypes algorithm with k=3 and 'Cao' as initialisation
    method
kprototype = KPrototypes(n_jobs = -1, n_clusters = 3, init = 'Cao', random_state =
    0)
kprototype.fit_predict(kprot_data, categorical = categorical_columns)

# Print the cluster centroids
kprototype.cluster_centroids_

```

```
# Check the iteration of the clusters created
kprototype.n_iter_

# Check the cost of the clusters created
kprototype.cost_

# Add the labels resulting from the clustering to the data frame
df_clustered = df.copy()
# Add the cluster to the dataframe
df_clustered['EoC_cluster'] = kprototype.labels_

# Save the clustered EoC data to a CSV file to be used for the visualisation
df_clustered.to_csv('EoC_clustered.csv', index=False)

# Visualise the clusters
clusters = pd.DataFrame(df_clustered['EoC_cluster'].value_counts())
clusters

# Plot the three clusters to illustrate the distribution of data points in the
    different clusters
sns.barplot(x=clusters.index, y=clusters['EoC_cluster'])

# To see how the different EoC features affect the clustering result, this can be
    visualised using a SHAP summary plot
data = kprot_data.copy()

for i in data.select_dtypes(include='object'):
    data[i] = data[i].astype('category')

clf_kp = LGBMClassifier(colsample_by_tree=0.8)
cv_scores_kp = cross_val_score(clf_kp , data, df_clustered['EoC_cluster'], scoring=
    'f1_weighted')
print(f'CV F1 score for K-Prototypes clusters is {np.mean(cv_scores_kp)}')

clf_kp.fit(data, df_clustered['EoC_cluster'])

explainer_kp = shap.TreeExplainer(clf_kp)
shap_values_kp = explainer_kp.shap_values(data)

shap.summary_plot(shap_values_kp, data, plot_type="bar", plot_size =(15, 10), show=
    False)
```

Listing B.4: Using the identified clustering number to cluster the EoC data (third iteration).

Finding k for the Third Iteration's EoC Bundle Clustering

The following code finds an optimal number of clusters, k, for the third iteration's EoC Bundle data.

```
# Import necessary packages
import pandas as pd
import numpy as np
from kmodes.kprototypes import KPrototypes
from sklearn.preprocessing import PowerTransformer
from tqdm import tqdm
import plotly.graph_objs as go
from plotnine import *
import plotnine
from kneed import KneeLocator

# Format scientific notation from Pandas
pd.set_option('display.float_format', lambda x: '%.3f' % x)

# Importing the EoC Bundle data from a CSV file and saving it as a data frame
```

```
df_EoC_Bundle = pd.read_csv('EoC_clustered.csv')

# Importing the clustered EoC data from a CSV file and saving it as a data frame
df = pd.read_csv('EoC_preprocessed.csv')

# Joining the EoC data and the EoC Bundle data to get the EoC clusters together with the EoC Bundle data
df_joined = df_EoC_Bundle.set_index('EoC_id').join(df_EoC.set_index('EoC_id'), how='inner')
df_joined

# Removing most of the columns related to EoC since we are only interested in the EoC Bundle data and the EoC clusters
df_joined = df_joined.drop(["care_level" ,"immediacy_level", "contacts_before_primary_axis1_diagnosis", "EoC_length", "nr_therapy_per_day", "nr_planning_per_day", "nr_examination_per_day", "nr_no_show_per_day", "nr_indirect_contact_per_day", "nr_diagnoses_1_per_day", "nr_diagnoses_2_per_day", "nr_diagnoses_3_per_day", "nr_diagnoses_4_per_day", "nr_diagnoses_5_per_day", "nr_diagnoses_6_per_day", "percentage_primary_axis_diagnoses"], axis = 1)

df_joined = df_joined.reset_index('EoC_id')

# Since one EoC Bundle can have multiple EoCs, we first group by EoC Bundle ID and then count the number of EoC types for each of the three EoC types each EoC Bundle has
df_count_EoC_cluster0 = df_joined.groupby('EoC_Bundle_id')['EoC_cluster'].apply(lambda x: (x==0).sum()).reset_index(name='nr_EoC_type_0')
df_count_EoC_cluster1 = df_joined.groupby('EoC_Bundle_id')['EoC_cluster'].apply(lambda x: (x==1).sum()).reset_index(name='nr_EoC_type_1')
df_count_EoC_cluster2 = df_joined.groupby('EoC_Bundle_id')['EoC_cluster'].apply(lambda x: (x==2).sum()).reset_index(name='nr_EoC_type_2')

# Create a new data frame where each row has a unique EoC Bundle ID
df_unique_EoC_Bundles = df_joined.copy()

# Remove the EoC cluster feature from the data frame and the EoC id to change the data frame to only include unique EoC Bundles and a count of each EoC cluster. Then drop the rows with duplicated EoC Bundle IDs.
df_unique_EoC_Bundles = df_unique_EoC_Bundles.drop(["EoC_id" ,"EoC_cluster"], axis = 1)

df_unique_EoC_Bundles = df_unique_EoC_Bundles.drop_duplicates(subset=["EoC_Bundle_id"], keep='last')

# Integrate the number of EoC clusters in the data frame
df_cluster = df_count_EoC_cluster1.set_index('EoC_Bundle_id').join(df_cluster.set_index('EoC_Bundle_id'), how='inner')

df_cluster = df_count_EoC_cluster0.set_index('EoC_Bundle_id').join(df_unique_EoC_Bundles.set_index('EoC_Bundle_id'), how='inner')

df_cluster = df_count_EoC_cluster3.set_index('EoC_Bundle_id').join(df_cluster, how='inner')

df_cluster = df_cluster.reset_index('EoC_Bundle_id')

# Save the preprocessed data in a CSV file
df_cluster.to_csv('EoC_Bundle_ready_clustering.csv', index=False)

# Remove "EoC_Bundle_id" to prepare for the EoC Bundle clustering
df_cluster = df_cluster.drop(['EoC_Bundle_id'], axis = 1)

df_EoC_Bundle_cluster = df_cluster.copy()

# Making sure the different columns have the correct data type before clustering
```

```
df_EoC_Bundle_cluster['nr_EoC_type_2'] = df_EoC_Bundle_cluster['nr_EoC_type_2'].
    astype(float)
df_EoC_Bundle_cluster['nr_EoC_type_1'] = df_EoC_Bundle_cluster['nr_EoC_type_1'].
    astype(float)
df_EoC_Bundle_cluster['nr_EoC_type_0'] = df_EoC_Bundle_cluster['nr_EoC_type_0'].
    astype(float)
df_EoC_Bundle_cluster['age_EoC_Bundle_start'] = df_EoC_Bundle_cluster['
    age_EoC_Bundle_start'].astype(float)
df_EoC_Bundle_cluster['care_situation'] = df_EoC_Bundle_cluster['care_situation'].
    astype(str)
df_EoC_Bundle_cluster['closing_code'] = df_EoC_Bundle_cluster['closing_code'].
    astype(str)
df_EoC_Bundle_cluster['gender'] = df_EoC_Bundle_cluster['gender'].astype(str)
df_EoC_Bundle_cluster['diagnose_axis_1'] = df_EoC_Bundle_cluster['diagnose_axis_1'
    ].astype(str)
df_EoC_Bundle_cluster['diagnose_axis_2'] = df_EoC_Bundle_cluster['diagnose_axis_2'
    ].astype(str)
df_EoC_Bundle_cluster['diagnose_axis_3'] = df_EoC_Bundle_cluster['diagnose_axis_3'
    ].astype(str)
df_EoC_Bundle_cluster['diagnose_axis_4'] = df_EoC_Bundle_cluster['diagnose_axis_4'
    ].astype(str)
df_EoC_Bundle_cluster['diagnose_axis_5'] = df_EoC_Bundle_cluster['diagnose_axis_5'
    ].astype(str)
df_EoC_Bundle_cluster['diagnose_axis_6'] = df_EoC_Bundle_cluster['diagnose_axis_6'
    ].astype(float)
df_EoC_Bundle_cluster['EoC_Bundle_length'] = df_EoC_Bundle_cluster['
    EoC_Bundle_length'].astype(float)

# Find the optimal number of clusters using the Elbow method
kprot_data = df_cluster.copy()
for c in df_cluster.select_dtypes(exclude='object').columns:
    pt = PowerTransformer()
    kprot_data[c] =  pt.fit_transform(np.array(kprot_data[c]).reshape(-1, 1))

# Get the position of categorical columns
categorical_columns = [df_cluster.columns.get_loc(col) for col in list(df_cluster.
    select_dtypes('object').columns)]
print('Categorical columns           : {}'.format(list(df_cluster.select_dtypes('
    object').columns)))
print('Categorical columns position  : {}'.format(categorical_columns))

# Finding k using the elbow method, using the k-prototypes algorithm initialised
    with Cao
costs = []
n_clusters = []
clusters_assigned = []

for i in tqdm(range(1, 11)):
    try:
        kproto = KPrototypes(n_clusters= i, init='Cao', verbose=2)
        clusters = kproto.fit_predict(kprot_data, categorical=categorical_columns)
        costs.append(kproto.cost_)
        n_clusters.append(i)
        clusters_assigned.append(clusters)
    except:
        print(f"Can't cluster with {i} clusters")

fig = go.Figure(data=go.Scatter(x=n_clusters, y=costs ))
fig.show()

# Converting the results into a dataframe and plotting them
df_cost = pd.DataFrame({'Cluster':range(1, 11), 'Cost':costs})

plotnine.options.figure_size = (8, 4.8)
(
    ggplot(data = df_cost)+
```

```
    geom_line(aes(x = 'Cluster',
                  y = 'Cost'))+
    geom_point(aes(x = 'Cluster',
                   y = 'Cost'))+
    geom_label(aes(x = 'Cluster',
                   y = 'Cost',
                   label = 'Cluster'),
               size = 11,
               nudge_y = 1000) +
    labs(title = 'Optimal number of cluster with Elbow Method')+
    xlab('Number of Clusters k')+
    ylab('Cost')+
    theme_minimal()
)

# Confirm visual clue of the elbow plot using the KneeLocator class to detect
    elbows if the curve is convex

from kneed import KneeLocator
cost_knee_c3  = KneeLocator(
        range(1,11),
        costs,
        S=0.1, curve="convex", direction="decreasing", online=True)

K_inertia_b3 = cost_knee_c3 .elbow
print("elbow at k =", f'{K_inertia_b3:.0f} clusters')
```

Listing B.5: Find k for the EoC Bundle clustering (third iteration).

Cluster the EoC Bundle data

The following code cluster the third iteration's EoC Bundle data using k-prototypes with the identified optimal number of clusters.

```
# Import necessary packages
import pandas as pd
import numpy as np
from kmodes.kprototypes import KPrototypes
from sklearn.preprocessing import PowerTransformer
import matplotlib.pyplot as plt
import seaborn as sns
from tqdm import tqdm
import plotly.graph_objs as go
import plotnine
from lightgbm import LGBMClassifier
import shap
from sklearn.model_selection import cross_val_score

# Format scientific notation from Pandas
pd.set_option('display.float_format', lambda x: '%.3f' % x)

# Importing data from a CSV file and saving it as a data frame
df = pd.read_csv('EoC_Bundle_ready_clustering.csv')

# Prepare the data for clustering
df_cluster = df.drop(['EoC_id'], axis = 1)
df_cluster.head()

# Transform the continuous data to prepare for the clustering
kprot_data = df_cluster.copy()
for c in df_cluster.select_dtypes(exclude='object').columns:
    pt = PowerTransformer()
    kprot_data[c] =  pt.fit_transform(np.array(kprot_data[c]).reshape(-1, 1))

```

```
# Get the position of categorical columns
categorical_columns = [df_cluster.columns.get_loc(col) for col in list(df_cluster.select_dtypes('object').columns)]
print('Categorical columns           : {}'.format(list(df_cluster.select_dtypes('object').columns)))
print('Categorical columns position  : {}'.format(categorical_columns))

# Cluster using the k-prototypes algorithm with k=4 and 'Cao' as initialisation method
kprototype = KPrototypes(n_jobs = -1, n_clusters = 4, init = 'Cao', random_state = 0)
kprototype.fit_predict(kprot_data, categorical = categorical_columns)

# Print the cluster centroids
kprototype.cluster_centroids_

# Check the iteration of the clusters created
kprototype.n_iter_

# Check the cost of the clusters created
kprototype.cost_

# Add the labels resulting from the clustering to the data frame
df_clustered = df.copy()
df_clustered['EoC_Bundle_cluster'] = kprototype.labels_

# Save the clustered EoC Bundle data to a CSV file to be used for the visualisation
df_clustered.to_csv('EoC_Bundle_clustered.csv', index=False)

# Visualise the clusters
clusters = pd.DataFrame(df_clustered['EoC_Bundle_cluster'].value_counts())
clusters

# Plot the three clusters to illustrate the distribution of data points in the different clusters
sns.barplot(x=clusters.index, y=clusters['EoC_Bundle_cluster'])

# To see how the different EoC Bundle features affect the clustering result, this can be visualised using a SHAP summary plot
data = kprot_data.copy()

for i in data.select_dtypes(include='object'):
    data[i] = data[i].astype('category')

clf_kp = LGBMClassifier(colsample_by_tree=0.8)
cv_scores_kp = cross_val_score(clf_kp , data, df_clustered['EoC_Bundle_cluster'], scoring='f1_weighted')
print(f'CV F1 score for K-Prototypes clusters is {np.mean(cv_scores_kp)}')

clf_kp.fit(data, df_clustered['EoC_Bundle_cluster'])

explainer_kp = shap.TreeExplainer(clf_kp)
shap_values_kp = explainer_kp.shap_values(data)

shap.summary_plot(shap_values_kp, data, plot_type="bar", plot_size =(15, 10), show=False)
```

Listing B.6: Using the identified clustering number to cluster the EoC Bundle data (third iteration).

B.1.3 Visualisation

Visualising the identified EoC clusters

The following code visualises the obtained EoC results from the third clustering iteration.

```
# Import necessary packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Importing the already clustered EoC data
df_EoC = pd.read_csv('EoC_clustered.csv')

# Copy the data frame to visualise the results without changing the original data frame
df_result = df_EoC.copy()

# Visualise the EoC length for each of the EoC clusters
# less than a week, a week to half a year, half a year to a year, 1-3 years, more than 3 years
bins = pd.IntervalIndex.from_tuples([(-1, 7), (7, 182), (182, 365), (365, 1095), (1095, 6500)])
df_result["EoC_length"] = pd.cut(df_result["EoC_length"], bins=bins)
df_result["EoC_length"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("EoC_length"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the distribution of care level for each of the EoC clusters
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("care_level"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the distribution of immediacy level for each of the EoC clusters
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("immediacy_level"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of therapy contacts per day for each of the EoC clusters
# none, none to once a month, once a month to twice a month, twice a month to once a week, once a week to multiple times a week, multiple times a week to multiple times a day
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_therapy_per_day"] = pd.cut(df_result["nr_therapy_per_day"], bins=bins)
df_result["nr_therapy_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_therapy_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of planning contacts per day for each of the EoC clusters
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_planning_per_day"] = pd.cut(df_result["nr_planning_per_day"], bins=bins)
df_result["nr_planning_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_planning_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of examination contacts per day for each of the EoC clusters
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_examination_per_day"] = pd.cut(df_result["nr_examination_per_day"], bins=bins)
```

```
df_result["nr_examination_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_examination_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of no-show contacts up per day for each of the EoC clusters
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_no_show_per_day"] = pd.cut(df_result["nr_no_show_per_day"], bins=bins)
df_result["nr_no_show_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_no_show_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of indirect contacts per day for each of the EoC clusters
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_indirect_contact_per_day"] = pd.cut(df_result["nr_indirect_contact_per_day"], bins=bins)
df_result["nr_indirect_contact_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_indirect_contact_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of contacts before a primary Axis 1 diagnosis is set for each of the EoC clusters
# no contacts, 1-25, 26-50, 51-100, 101-999, no diagnosis given
bins = pd.IntervalIndex.from_tuples([(-1, 0), (0, 25), (25, 50), (50, 100), (100, 999), (999, 1000)])
df_result["contacts_before_primary_axis1_diagnosis"] = pd.cut(df_result["contacts_before_primary_axis1_diagnosis"], bins=bins)
df_result["contacts_before_primary_axis1_diagnosis"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("contacts_before_primary_axis1_diagnosis"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the percentage of diagnoses set as primary axis diagnosis
# 0%, 0-50%, 51-75%, 76-99%, 100%
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.5), (0.5, 0.75), (0.75, 0.99), (0.99, 1)])
df_result["percentage_primary_axis_diagnose"] = pd.cut(df_result["percentage_primary_axis_diagnosis"], bins=bins)
df_result["percentage_primary_axis_diagnose"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("percentage_primary_axis_diagnosis"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of diagnoses set on Axis 1 per day for each of the EoC clusters
# none, none to once a month, once a month to twice a month, twice a month to once a week, once a week to multiple times a week, multiple times a week to multiple times a day
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_diagnoses_1_per_day"] = pd.cut(df_result["nr_diagnoses_1_per_day"], bins=bins)
df_result["nr_diagnoses_1_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_diagnoses_1_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of diagnoses set on Axis 2 per day for each of the EoC clusters
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
```

```
df_result["nr_diagnoses_2_per_day"] = pd.cut(df_result["nr_diagnoses_2_per_day"], bins=bins)
df_result["nr_diagnoses_2_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_diagnoses_2_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of diagnoses set on Axis 3 per day for each of the EoC clusters
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_diagnoses_3_per_day"] = pd.cut(df_result["nr_diagnoses_3_per_day"], bins=bins)
df_result["nr_diagnoses_3_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_diagnoses_3_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of diagnoses set on Axis 4 per day for each of the EoC clusters
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_diagnoses_4_per_day"] = pd.cut(df_result["nr_diagnoses_4_per_day"], bins=bins)
df_result["nr_diagnoses_4_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_diagnoses_4_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of diagnoses set on Axis 5 per day for each of the EoC clusters
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_diagnoses_5_per_day"] = pd.cut(df_result["nr_diagnoses_5_per_day"], bins=bins)
df_result["nr_diagnoses_5_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_diagnoses_5_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of diagnoses set on Axis 6 per day for each of the EoC clusters
bins = pd.IntervalIndex.from_tuples([(-1,0), (0,0.0027), (0.0027,0.03288), (0.03288, 0.14247), (0.14247, 1), (1, 1000)])
df_result["nr_diagnoses_6_per_day"] = pd.cut(df_result["nr_diagnoses_6_per_day"], bins=bins)
df_result["nr_diagnoses_6_per_day"].value_counts()
ax = sns.countplot(x='EoC_cluster', hue='{}'.format("nr_diagnoses_6_per_day"), data=df_result)
# ax.bar_label(ax.containers[0])
```

Listing B.7: Visualisation of the identified EoC clusters (third iteration).

Visualising the identified EoC Bundle clusters

The following code visualises the obtained EoC Bundle results from the third clustering iteration.

```
# Import necessary packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Importing the already clustered EoC data
df_EoC_Bundle = pd.read_csv('EoC_Bundle_clustered.csv')
```

```
9
10 # Copy the data frame to visualise the results without changing the original data
       frame
11 df_result = df_EoC_Bundle.copy()
12
13 # Visualise the EoC Bundle length for each of the EoC Bundle clusters
14 # less than a week, a week to half a year, half a year to a year, 1-3 years, more
       than 3 years
15 bins = pd.IntervalIndex.from_tuples([(-1, 7), (7, 182), (182, 365), (365, 1095),
       (1095, 6500)])
16 df_result["EoC_Bundle_Length"] = pd.cut(df_result["EoC_Bundle_Length"], bins=bins)
17 df_result["EoC_Bundle_Length"].value_counts()
18 ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("EoC_Bundle_Length"),
       data=df_result)
19 # ax.bar_label(ax.containers[0])
20
21 # Visualise the age distribution at EoC Bundle start for each of the EoC Bundle
       clusters
22 bins = pd.IntervalIndex.from_tuples([(-1, 6), (6, 10), (10, 14), (14, 18), (18, 22)
       ])
23 df_result["age_EoC_Bundle_start"] = pd.cut(df_result["age_EoC_Bundle_start"], bins=
       bins)
24 df_result["age_EoC_Bundle_start"].value_counts()
25 ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("age_EoC_Bundle_start"),
        data=df_result)
26 # ax.bar_label(ax.containers[0])
27
28 # Visualise the gender distribution for each of the EoC Bundle clusters
29 ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("gender"), data=
       df_result)
30 # ax.bar_label(ax.containers[0])
31
32 # Visualise the distributions of care situations for each of the EoC Bundle
       clusters
33 ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("care_situation"), data=
       df_result)
34 # ax.bar_label(ax.containers[0])
35
36 # Visualise the diagnosis set on Axis 1 at the beginning of an EoC Bundle for each
       of the EoC Bundle clusters
37 ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("diagnoses_axis_1"),
       data=df_result)
38 # ax.bar_label(ax.containers[0])
39
40 # Visualise the number that has a diagnosis on Axis 2 at the beginning of an EoC
       Bundle for each of the EoC Bundle clusters
41 ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("diagnosis_axis_2"),
       data=df_result)
42 # ax.bar_label(ax.containers[0])
43
44 # Visualise the number that has a diagnosis on Axis 3 at the beginning of an EoC
       Bundle for each of the EoC Bundle clusters
45 ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("diagnosis_axis_3"),
       data=df_result)
46 # ax.bar_label(ax.containers[0])
47
48 # Visualise the number that has a diagnosis on Axis 4 at the beginning of an EoC
       Bundle for each of the EoC Bundle clusters
49 ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("diagnosis_axis_4"),
       data=df_result)
50 # ax.bar_label(ax.containers[0])
51
52 # Visualise the number that has a diagnosis on Axis 5 at the beginning of an EoC
       Bundle for each of the EoC Bundle clusters
53 ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("diagnosis_axis_5"),
       data=df_result)
```

```
# ax.bar_label(ax.containers[0])

# Visualise the distribution of different CGAS scores set on Axis 6 at the
    beginning of an EoC Bundle for each of the EoC Bundle clusters
ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("diagnoses_axis_6"),
    data=df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of EoC Type 0 for each of the EoC Bundle clusters
bins = pd.IntervalIndex.from_tuples([(-1, 0), (0, 1), (1, 3), (3, 15)])
df_result["nr_EoC_type_0"] = pd.cut(df_result["nr_EoC_type_0"], bins=bins)
ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("nr_EoC_type_0"), data=
    df_result)
# ax.bar_label(ax.containers[0])

# Visualise the number of EoC Type 1 for each of the EoC Bundle clusters
bins = pd.IntervalIndex.from_tuples([(-1, 0), (0, 1), (1, 3), (3, 15)])
df_result["nr_EoC_type_1"] = pd.cut(df_result["nr_EoC_type_1"], bins=bins)
ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("nr_EoC_type_1"), data=
    df_result)
# ax.bar_label(ax.containers[1])

# Visualise the number of EoC Type 2 for each of the EoC Bundle clusters
bins = pd.IntervalIndex.from_tuples([(-1, 0), (0, 1), (1, 3), (3, 15)])
df_result["nr_EoC_type_2"] = pd.cut(df_result["nr_EoC_type_2"], bins=bins)
ax = sns.countplot(x='EoC_Bundle_cluster', hue='{}'.format("nr_EoC_type_2"), data=
    df_result)
# ax.bar_label(ax.containers[2])
```

Listing B.8: Visualisation of the identified EoC clusters (third iteration).

B.2 Evaluation Code

The Hopkins score is calculated for the EoC and EoC Bundle data for all three iterations in this appendix section.

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import PowerTransformer
from pyclustertend import hopkins

# Importing the finished preprocessed EoC data from the tree iterations

df_EoC1 = pd.read_csv('1_EoC_preprocessed.csv')
df_EoC2 = pd.read_csv('2_EoC_preprocessed.csv')
df_EoC3 = pd.read_csv('3_EoC_preprocessed.csv')

# Preparing the EoC data by removing EoC_id. Then, temporarily convert the
    categorical features to numerical features. It is preferable to scale the data
    before calculating the Hopkins score, as the distance between the data points
    is used. The scaled data must be converted from a Pandas Dataframe to a Numpy
    Array.

def prepareEoCForHopkins(df_EoC):
    df_EoC = df_EoC.drop(['EoC_id'], axis=1)
    df_converted = pd.get_dummies(
        df_EoC, columns=['care_level', 'immediacy_level'])

    scaled_data = df_converted.copy()
    for c in df_converted.select_dtypes(exclude='object').columns:
        pt = PowerTransformer()
        scaled_data[c] = pt.fit_transform(
            np.array(scaled_data[c]).reshape(-1, 1))
```

```
    return scaled_data.values

df_EoC_converted1 = prepareEoCForHopkins(df_EoC1)
df_EoC_converted2 = prepareEoCForHopkins(df_EoC2)
df_EoC_converted3 = prepareEoCForHopkins(df_EoC3)

# Assess the clusterability of the data using the Hopkins test. The sampling size
    is set to approximately 10% of the contained data points, as this is
    recommended to avoid any small sample problems with the distributions of the
    statistics.

hopkins_EoC1 = hopkins(df_EoC_converted1, 1500)
hopkins_EoC2 = hopkins(df_EoC_converted2, 1500)
hopkins_EoC3 = hopkins(df_EoC_converted3, 1500)

# Importing the preprocessed EoC Bundle data from the three iterations
df_EoC_Bundle1 = pd.read_csv('1_EoC_Bundle_preprocessed.csv')
df_EoC_Bundle2 = pd.read_csv('2_EoC_Bundle_preprocessed.csv')
df_EoC_Bundle3 = pd.read_csv('3_EoC_Bundle_preprocessed.csv')

# Preparing the EoC Bundle data by removing EoC_id and EoC_Bundle_id. Then,
    temporarily convert the categorical features to numerical features. It is
    preferable to scale the data before calculating the Hopkins score, as the
    distance between the data points is used. The scaled data must be converted
    from a Pandas Dataframe to a Numpy Array.

def prepareEoC_BundleForHopkins1(df_EoC_Bundle):
    df_EoC_Bundle = df_EoC_Bundle.drop(['EoC_Bundle_id', 'EoC_id'], axis=1)
    df_converted = pd.get_dummies(
        df_EoC_Bundle, columns=['gender', 'diagnose_axis_1', 'diagnose_axis_2', '
    diagnose_axis_3', 'diagnose_axis_4', 'diagnose_axis_5', 'diagnose_axis_6'])

    scaled_data = df_converted.copy()
    for c in df_converted.select_dtypes(exclude='object').columns:
        pt = PowerTransformer()
        scaled_data[c] = pt.fit_transform(
            np.array(scaled_data[c]).reshape(-1, 1))

    return scaled_data.values

def prepareEoC_BundleForHopkins2(df_EoC_Bundle):
    df_EoC_Bundle = df_EoC_Bundle.drop(['EoC_Bundle_id', 'EoC_id'], axis=1)
    df_converted = pd.get_dummies(
        df_EoC_Bundle, columns=['gender', 'diagnose_axis_1', 'diagnose_axis_2', '
    diagnose_axis_3', 'diagnose_axis_4', 'diagnose_axis_5'])

    scaled_data = df_converted.copy()
    for c in df_converted.select_dtypes(exclude='object').columns:
        pt = PowerTransformer()
        scaled_data[c] = pt.fit_transform(
            np.array(scaled_data[c]).reshape(-1, 1))

    return scaled_data.values

def prepareEoC_BundleForHopkins3(df_EoC_Bundle):
    df_EoC_Bundle = df_EoC_Bundle.drop(['EoC_Bundle_id', 'EoC_id'], axis=1)
    df_converted = pd.get_dummies(
        df_EoC_Bundle, columns=['care_situation', 'gender', 'diagnose_axis_1', '
    diagnose_axis_2', 'diagnose_axis_3', 'diagnose_axis_4', 'diagnose_axis_5', '
    diagnose_axis_6', 'closing_code'])

    scaled_data = df_converted.copy()
    for c in df_converted.select_dtypes(exclude='object').columns:
        pt = PowerTransformer()
        scaled_data[c] = pt.fit_transform(
            np.array(scaled_data[c]).reshape(-1, 1))
```

```
79
80      return scaled_data.values
81
82  df_EoC_Bundle_converted1 = prepareEoC_BundleForHopkins1(df_EoC_Bundle1)
83  df_EoC_Bundle_converted2 = prepareEoC_BundleForHopkins2(df_EoC_Bundle2)
84  df_EoC_Bundle_converted3 = prepareEoC_BundleForHopkins3(df_EoC_Bundle3)
85
86  hopkins_EoC_Bundle1 = hopkins(df_EoC_Bundle_converted1, 1500)
87  hopkins_EoC_Bundle2 = hopkins(df_EoC_Bundle_converted2, 1500)
88  hopkins_EoC_Bundle3 = hopkins(df_EoC_Bundle_converted3, 1500)
```

Listing B.9: Exploring the data's cluster tendency using the Hopkins statistics

The Hopkins score is calculated for the EoC and EoC Bundle data for all three iterations in this appendix section.

```
1  import pandas as pd
2  import numpy as np
3  from sklearn.preprocessing import PowerTransformer
4  from pyclustertend import hopkins
5
6  # Importing the finished preprocessed EoC data from the tree iterations
7
8  df_EoC1 = pd.read_csv('1_EoC_preprocessed.csv')
9  df_EoC2 = pd.read_csv('2_EoC_preprocessed.csv')
10 df_EoC3 = pd.read_csv('3_EoC_preprocessed.csv')
11
12 # Preparing the EoC data by removing EoC_id. Then, temporarily convert the
       categorical features to numerical features. It is preferable to scale the data
       before calculating the Hopkins score, as the distance between the data points
       is used. The scaled data must be converted from a Pandas Dataframe to a Numpy
       Array.
13
14 def prepareEoCForHopkins(df_EoC):
15     df_EoC = df_EoC.drop(['EoC_id'], axis=1)
16     df_converted = pd.get_dummies(
17         df_EoC, columns=['care_level', 'immediacy_level'])
18
19     scaled_data = df_converted.copy()
20     for c in df_converted.select_dtypes(exclude='object').columns:
21         pt = PowerTransformer()
22         scaled_data[c] = pt.fit_transform(
23             np.array(scaled_data[c]).reshape(-1, 1))
24     return scaled_data.values
25
26 df_EoC_converted1 = prepareEoCForHopkins(df_EoC1)
27 df_EoC_converted2 = prepareEoCForHopkins(df_EoC2)
28 df_EoC_converted3 = prepareEoCForHopkins(df_EoC3)
29
30 # Assess the clusterability of the data using the Hopkins test. The sampling size
       is set to approximately 10% of the contained data points, as this is
       recommended to avoid any small sample problems with the distributions of the
       statistics.
31
32 hopkins_EoC1 = hopkins(df_EoC_converted1, 1500)
33 hopkins_EoC2 = hopkins(df_EoC_converted2, 1500)
34 hopkins_EoC3 = hopkins(df_EoC_converted3, 1500)
35
36 # Importing the preprocessed EoC Bundle data from the three iterations
37 df_EoC_Bundle1 = pd.read_csv('1_EoC_Bundle_preprocessed.csv')
38 df_EoC_Bundle2 = pd.read_csv('2_EoC_Bundle_preprocessed.csv')
39 df_EoC_Bundle3 = pd.read_csv('3_EoC_Bundle_preprocessed.csv')
40
```

```
# Preparing the EoC Bundle data by removing EoC_id and EoC_Bundle_id. Then, temporarily convert the categorical features to numerical features. It is preferable to scale the data before calculating the Hopkins score, as the distance between the data points is used. The scaled data must be converted from a Pandas Dataframe to a Numpy Array.

def prepareEoC_BundleForHopkins1(df_EoC_Bundle):
    df_EoC_Bundle = df_EoC_Bundle.drop(['EoC_Bundle_id', 'EoC_id'], axis=1)
    df_converted = pd.get_dummies(
        df_EoC_Bundle, columns=['gender', 'diagnose_axis_1', 'diagnose_axis_2', 'diagnose_axis_3', 'diagnose_axis_4', 'diagnose_axis_5', 'diagnose_axis_6'])

    scaled_data = df_converted.copy()
    for c in df_converted.select_dtypes(exclude='object').columns:
        pt = PowerTransformer()
        scaled_data[c] = pt.fit_transform(
            np.array(scaled_data[c]).reshape(-1, 1))

    return scaled_data.values

def prepareEoC_BundleForHopkins2(df_EoC_Bundle):
    df_EoC_Bundle = df_EoC_Bundle.drop(['EoC_Bundle_id', 'EoC_id'], axis=1)
    df_converted = pd.get_dummies(
        df_EoC_Bundle, columns=['gender', 'diagnose_axis_1', 'diagnose_axis_2', 'diagnose_axis_3', 'diagnose_axis_4', 'diagnose_axis_5'])

    scaled_data = df_converted.copy()
    for c in df_converted.select_dtypes(exclude='object').columns:
        pt = PowerTransformer()
        scaled_data[c] = pt.fit_transform(
            np.array(scaled_data[c]).reshape(-1, 1))

    return scaled_data.values

def prepareEoC_BundleForHopkins3(df_EoC_Bundle):
    df_EoC_Bundle = df_EoC_Bundle.drop(['EoC_Bundle_id', 'EoC_id'], axis=1)
    df_converted = pd.get_dummies(
        df_EoC_Bundle, columns=['care_situation', 'gender', 'diagnose_axis_1', 'diagnose_axis_2', 'diagnose_axis_3', 'diagnose_axis_4', 'diagnose_axis_5', 'diagnose_axis_6', 'closing_code'])

    scaled_data = df_converted.copy()
    for c in df_converted.select_dtypes(exclude='object').columns:
        pt = PowerTransformer()
        scaled_data[c] = pt.fit_transform(
            np.array(scaled_data[c]).reshape(-1, 1))

    return scaled_data.values

df_EoC_Bundle_converted1 = prepareEoC_BundleForHopkins1(df_EoC_Bundle1)
df_EoC_Bundle_converted2 = prepareEoC_BundleForHopkins2(df_EoC_Bundle2)
df_EoC_Bundle_converted3 = prepareEoC_BundleForHopkins3(df_EoC_Bundle3)

hopkins_EoC_Bundle1 = hopkins(df_EoC_Bundle_converted1, 1500)
hopkins_EoC_Bundle2 = hopkins(df_EoC_Bundle_converted2, 1500)
hopkins_EoC_Bundle3 = hopkins(df_EoC_Bundle_converted3, 1500)
```

Listing B.10: Exploring the data's cluster tendency using the Hopkins statistics.

B.3 Discussion Code

B.3.1 Comparing Initialisation Method

In this Appendix, the initialisation methods Huang and Cao are compared. First, the optimal number of k is identified and used to cluster the first iteration's EoC Data with k-prototypes initialised with Huang. Then, the same approach is completed again, with Cao as the initialisation method. Finally, the clustering centroids obtained from the two approaches are presented.

k-prototypes initialised with Huang on First Iteration's EoC Data

```
# Importing the necessary packages
import pandas as pd
import numpy as np
from kmodes.kprototypes import KPrototypes
from sklearn.preprocessing import PowerTransformer
import plotly.graph_objs as go
from plotnine import *
import plotnine
from kneed import KneeLocator

# Format scientific notation from Pandas
pd.set_option('display.float_format', lambda x: '%.3f' % x)

# Importing the finished preprocessed EoC data
df = pd.read_csv('EoC_preprocessed.csv')

# Remove EoC ID to prepare for the clustering
df_cluster = df.drop(['EoC_id'], axis = 1)
df_cluster.head()

# Transform the continous features
kprot_data = df_cluster.copy()
for c in df_cluster.select_dtypes(exclude='object').columns:
    pt = PowerTransformer()
    kprot_data[c] =  pt.fit_transform(np.array(kprot_data[c]).reshape(-1, 1))

# Get the position of categorical columns
categorical_columns = [df_cluster.columns.get_loc(col) for col in list(df_cluster.
    select_dtypes('object').columns)]
print('Categorical columns           : {}'.format(list(df_cluster.select_dtypes('
    object').columns)))
print('Categorical columns position  : {}'.format(categorical_columns))

# Converting the data frame to a matrix
dfMatrix = kprot_data.to_numpy()

# Finding k using the elbow method with Huang as the initialisation method
cost_huang = []
for cluster in range(1, 11):
    try:
        kprototype_huang = KPrototypes(n_jobs = -1, n_clusters = cluster, init = '
    Huang', random_state = 0)
        kprototype_huang.fit_predict(dfMatrix, categorical = categorical_columns)
        cost_huang.append(kprototype_huang.cost_)
        print('Cluster initiation: {}'.format(cluster))
    except:
        break

# Converting the results into a data frame and plotting them
df_cost_huang = pd.DataFrame({'Cluster':range(1, 11), 'Cost':cost_huang})
df_cost_huang.head(10)
```

```

# Visualise the elbow plot using Huang as initialisation method
plotnine.options.figure_size = (8, 4.8)
(
    ggplot(data = df_cost_huang)+
    geom_line(aes(x = 'Cluster',
                  y = 'Cost'))+
    geom_point(aes(x = 'Cluster',
                   y = 'Cost'))+
    geom_label(aes(x = 'Cluster',
                   y = 'Cost',
                   label = 'Cluster'),
               size = 10,
               nudge_y = 1000) +
    labs(title = 'Optimal number of cluster with Elbow Method')+
    xlab('Number of Clusters k')+
    ylab('Cost')+
    theme_minimal()
)

# Confirm visual clue of elbow plot
# KneeLocator class will detect elbows if curve is convex; if concavem will detect knees
cost_knee_c3  = KneeLocator(
        range(1,10),
        cost_huang,
        S=0.1, curve="convex", direction="decreasing", online=True)

K_inertia_b3 = cost_knee_c3 .elbow
print("elbow at k =", f'{K_inertia_b3:.0f} clusters')


# Using the identified optimal number of clusters to cluster the EoC data
kprototype = KPrototypes(n_jobs = -1, n_clusters = 3, init = 'Huang', random_state = 0)
kprototype.fit_predict(kprot_data, categorical = categorical_columns)

# Print the obtained clustering centroids
kprototype.cluster_centroids_

# Check the iteration of the clusters created
kprototype.n_iter_

# Check the cost of the clusters created
kprototype.cost_
```

Listing B.11: Cluster the first iteration's EoC data using k-prototypes with Huang as initialisation method.

k-prototypes initialised with Cao on First Iteration's EoC Data

```
# Importing the necessary packages
import pandas as pd
import numpy as np
from kmodes.kprototypes import KPrototypes
from sklearn.preprocessing import PowerTransformer
import plotly.graph_objs as go
from plotnine import *
import plotnine
from kneed import KneeLocator

# Format scientific notation from Pandas
pd.set_option('display.float_format', lambda x: '%.3f' % x)

```

```
# Importing the finished preprocessed EoC data
df = pd.read_csv('EoC_preprocessed.csv')

# Remove EoC ID to prepare for the clustering
df_cluster = df.drop(['EoC_id'], axis = 1)
df_cluster.head()

# Transform the continous features
kprot_data = df_cluster.copy()
for c in df_cluster.select_dtypes(exclude='object').columns:
    pt = PowerTransformer()
    kprot_data[c] =  pt.fit_transform(np.array(kprot_data[c]).reshape(-1, 1))

# Get the position of categorical columns
categorical_columns = [df_cluster.columns.get_loc(col) for col in list(df_cluster.
    select_dtypes('object').columns)]
print('Categorical columns           : {}'.format(list(df_cluster.select_dtypes('
    object').columns)))
print('Categorical columns position  : {}'.format(categorical_columns))

# Finding k using the elbow method
costs = []
n_clusters = []
clusters_assigned = []

for i in tqdm(range(1, 11)):
    try:
        kproto = KPrototypes(n_clusters= i, init='Cao', verbose=2)
        clusters = kproto.fit_predict(kprot_data, categorical=categorical_columns)
        costs.append(kproto.cost_)
        n_clusters.append(i)
        clusters_assigned.append(clusters)
    except:
        print(f"Can't cluster with {i} clusters")

fig = go.Figure(data=go.Scatter(x=n_clusters, y=costs ))
fig.show()

# Converting the results into a dataframe and plotting them
df_cost = pd.DataFrame({'Cluster':range(1, 11), 'Cost':costs})

plotnine.options.figure_size = (8, 4.8)
(
    ggplot(data = df_cost)+
    geom_line(aes(x = 'Cluster',
                  y = 'Cost'))+
    geom_point(aes(x = 'Cluster',
                   y = 'Cost'))+
    geom_label(aes(x = 'Cluster',
                   y = 'Cost',
                   label = 'Cluster'),
               size = 11,
               nudge_y = 1000) +
    labs(title = 'Optimal number of cluster with Elbow Method')+
    xlab('Number of Clusters k')+
    ylab('Cost')+
    theme_minimal()
)

# Confirm visual clue of elbow plot KneeLocator class will detect elbows if curve
    is convex; if concavem will detect knees

from kneed import KneeLocator
cost_knee_c3  = KneeLocator(
        range(1,11),
        costs,
```

```
        S=0.1, curve="convex", direction="decreasing", online=True)

K_inertia_b3 = cost_knee_c3 .elbow
print("elbow at k =", f'{K_inertia_b3:.0f} clusters')

# Using the identified optimal number of clusters to cluster the EoC data
kprototype = KPrototypes(n_jobs = -1, n_clusters = 3, init = 'Cao', random_state =
    0)
kprototype.fit_predict(kprot_data, categorical = categorical_columns)

# Print the obtained clustering centroids
kprototype.cluster_centroids_

# Check the iteration of the clusters created
kprototype.n_iter_

# Check the cost of the clusters created
kprototype.cost_
```

Listing B.12: Cluster the first iteration's EoC data using k-prototypes with Cao as initialisation method.

Clustering Centroids obtained using Huang

```
array([['0.8994676158198726', '0.9278223948675354', '0.8336980087794877',
    '0.8611961619179367', '0.7336897373909423', '0.6348944446497196',
    '0.6354331670513417', '0.643940303325123', '0.6135411307586467',
    '0.598455477861984', '0.6572174978938633', '0.6552908284289688',
    '0.7597795837626988', '0.9534380138590786', '0.6385446656256897',
    'Polyclinic', 'Planned'],
   ['-0.40614895872001316', '-0.43134634021099616',
    '-0.3679464314854653', '-0.5569999737571355',
    '-0.4494214207871815', '0.3319810207858752', '0.530689543935265',
    '0.5257514065527068', '0.5092460209346316', '0.5287054547097749',
    '0.4133598110473474', '0.5178438727433178',
    '-0.25297598895286366', '-0.42654342622900615',
    '0.5800664848785415', 'Polyclinic', 'Planned'],
   ['-0.8496839464078503', '-0.8653686946048286',
    '-0.7951717642241827', '-0.6629430442289482',
    '-0.587279776970226', '-1.153860590144894',
    '-1.3325607094654397', '-1.339614650095579',
    '-1.2838207625965203', '-1.2808978972761622',
    '-1.2568630501335798', '-1.3478450014402688',
    '-0.7984234000276745', '-0.9042278766197227',
    '-1.380982831822819', 'Polyclinic', 'Planned']], dtype='<U32')
```

Clustering Centroids obtained using Cao

```
array([['-0.40614895872001333', '-0.431346340210998',
    '-0.36794643148546874', '-0.5569999737571358',
    '-0.449421420787183', '0.3319810207858775', '0.5306895439352637',
    '0.5257514065527', '0.5092460209346313', '0.5287054547097861',
    '0.4133598110473469', '0.5178438727433112',
    '-0.25297598895286477', '-0.42654342622900626',
    '0.5800664848785491', 'Polyclinic', 'Planned'],
   ['-0.8496839464078747', '-0.8653686946048276',
    '-0.7951717642242988', '-0.6629430442289285',
    '-0.5872797769702602', '-1.153860590144448',
    '-1.3325607094655538', '-1.3396146500956951',
    '-1.2838207625966618', '-1.28089789727641',
    '-1.2568630501331124', '-1.3478450014404793',
    '-0.7984234000276443', '-0.9042278766197224',
    '-1.3809828318228976', 'Polyclinic', 'Planned'],
   ['0.8994676158198069', '0.9278223948675327', '0.8336980087793757',
    '0.8611961619179515', '0.7336897373909547', '0.6348944446495994',
    '0.6354331670513493', '0.643940303325143', '0.6135411307586524',
    '0.5984554778620194', '0.6572174978937421', '0.6552908284290205',
    '0.7597795837626956', '0.9534380138590732', '0.6385446656256867',
    'Polyclinic', 'Planned']], dtype='<U32')
```

B.3.2 Finding the Optimal Number of Clusters Using UMAP

This appendix section used UMAP to find the optimal number of clusters for the first iteration's EoC and EoC data. First, the code used to implement UMAP on the EoC and EoC Bundle data is provided before presenting the obtained results.

Implementation of UMAP on the EoC and EoC Bundle Data

```
# Importing the necessary packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from kmodes.kprototypes import KPrototypes
from sklearn.preprocessing import PowerTransformer
import umap

# Format scientific notation from Pandas
pd.set_option('display.float_format', lambda x: '%.3f' % x)

# Importing the finished preprocessed EoC data
df_EoC = pd.read_csv('EoC_preprocessed.csv')

# Prepare the data by removing the EoC ID
df_cluster = df.drop(['EoC_id'], axis = 1)
df_cluster.head()

# Transforming the numerical EoC features to be on the same scale
numerical = df_EoC.select_dtypes(exclude='object')
for c in numerical.columns:
    pt = PowerTransformer()
    numerical.loc[:, c] = pt.fit_transform(np.array(numerical[c]).reshape(-1, 1))

# Convert categorical EoC features into dummy/indicator variables
categorical = df_EoC.select_dtypes(include='object')
categorical = pd.get_dummies(categorical)

# The percentage of categorical columns is used as a weight parameter in embeddings later
categorical_weight = len(df_EoC.select_dtypes(include='object').columns) / df_EoC.shape[1]

# Embedding the preprocessed numerical & categorical features
fit1 = umap.UMAP(metric='l2').fit(numerical)
fit2 = umap.UMAP(metric='dice').fit(categorical)

# Printing the categorical weight
categorical_weight

# Augmenting the numerical embedding with categorical
intersection = umap.umap_.general_simplicial_set_intersection(fit1.graph_, fit2.graph_, weight=categorical_weight)
intersection = umap.umap_.reset_local_connectivity(intersection)
embedding = umap.umap_.simplicial_set_embedding(fit1._raw_data, intersection, fit1.n_components,
                                                fit1._initial_alpha, fit1._a, fit1._b,
                                                fit1.repulsion_strength, fit1.negative_sample_rate,
                                                200, 'random', np.random, fit1.metric,
                                                fit1._metric_kwds, False,
    densmap_kwds={}, output_dens=False)

print(embedding)
```

```
plt.figure(figsize=(20, 10))
plt.scatter(*np.array(embedding)[0].T, s=2, cmap='Spectral', alpha=1.0)
plt.show()

# Importing the finished preprocessed EoC data
df_EoC_Bundle = pd.read_csv('EoC_Bundle_preprocessed.csv')

# Prepare the data by removing EoC Bundle ID
df_EoC_Bundle = df_EoC_Bundle.drop(['EoC_Bundle_id'], axis = 1)
df_EoC_Bundle.head()

# Transforming the numerical EoC Bundle features to be on the same scale
numerical_EoC_Bundle = df_EoC_Bundle.select_dtypes(exclude='object')
for c in numerical_EoC_Bundle.columns:
    pt = PowerTransformer()
    numerical_EoC_Bundle.loc[:, c] = pt.fit_transform(np.array(numerical_EoC_Bundle
    [c]).reshape(-1, 1))

# Convert categorical EoC_Bundle features into dummy/indicator variables
categorical_EoC_Bundle = df_EoC_Bundle.select_dtypes(include='object')
categorical_EoC_Bundle = pd.get_dummies(categorical_EoC_Bundle)

# The percentage of categorical columns is used as a weight parameter in embeddings
     later
categorical_weight_EoC_Bundle = len(df_EoC_Bundle.select_dtypes(include='object').
    columns) / df_EoC_Bundle.shape[1]

# Embedding the preprocessed numerical & categorical features
fit1_EoC_Bundle = umap.UMAP(metric='l2').fit(numerical_EoC_Bundle)
fit2_EoC_Bundle = umap.UMAP(metric='dice').fit(categorical_EoC_Bundle)

# Printing the categorical weight
categorical_weight_EoC_Bundle

# Augmenting the numerical embedding with categorical
intersection_EoC_Bundle = umap.umap_.general_simplicial_set_intersection(
    fit1_EoC_Bundle.graph_, fit2_EoC_Bundle.graph_, weight=
    categorical_weight_EoC_Bundle)
intersection_EoC_Bundle = umap.umap_.reset_local_connectivity(
    intersection_EoC_Bundle)
embedding_EoC_Bundle = umap.umap_.simplicial_set_embedding(fit1_EoC_Bundle.
    _raw_data, intersection_EoC_Bundle, fit1_EoC_Bundle.n_components,
                                                fit1_EoC_Bundle._initial_alpha,
    fit1_EoC_Bundle._a, fit1_EoC_Bundle._b,
                                                fit1_EoC_Bundle.repulsion_strength,
     fit1_EoC_Bundle.negative_sample_rate,
                                                200, 'random', np.random,
    fit1_EoC_Bundle.metric,
                                                fit1_EoC_Bundle._metric_kwds, False
    , densmap_kwds={}, output_dens=False)

print(embedding_EoC_Bundle)
plt.figure(figsize=(20, 10))
plt.scatter(*np.array(embedding_EoC_Bundle)[0].T, s=2, cmap='Spectral', alpha=1.0)
plt.show()
```

Listing B.13: Using UMAP to find the optimal number of clusters for the first iteration's EoC and EoC Bundle data.

The Resulting UMAP Plots

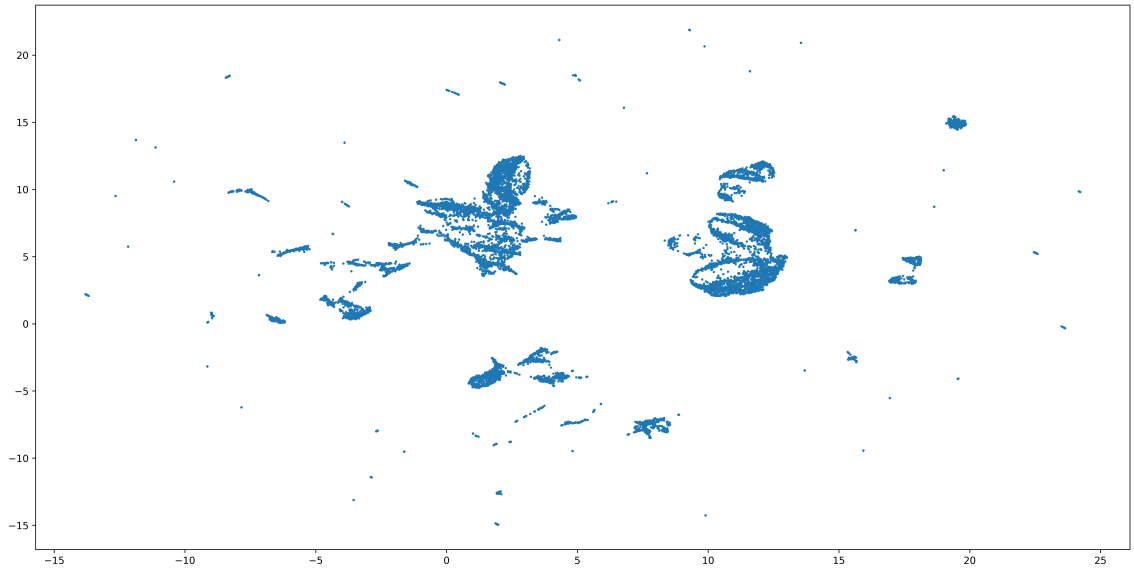


Figure B.1: Dimentionality reduction of EoC data using UMAP.

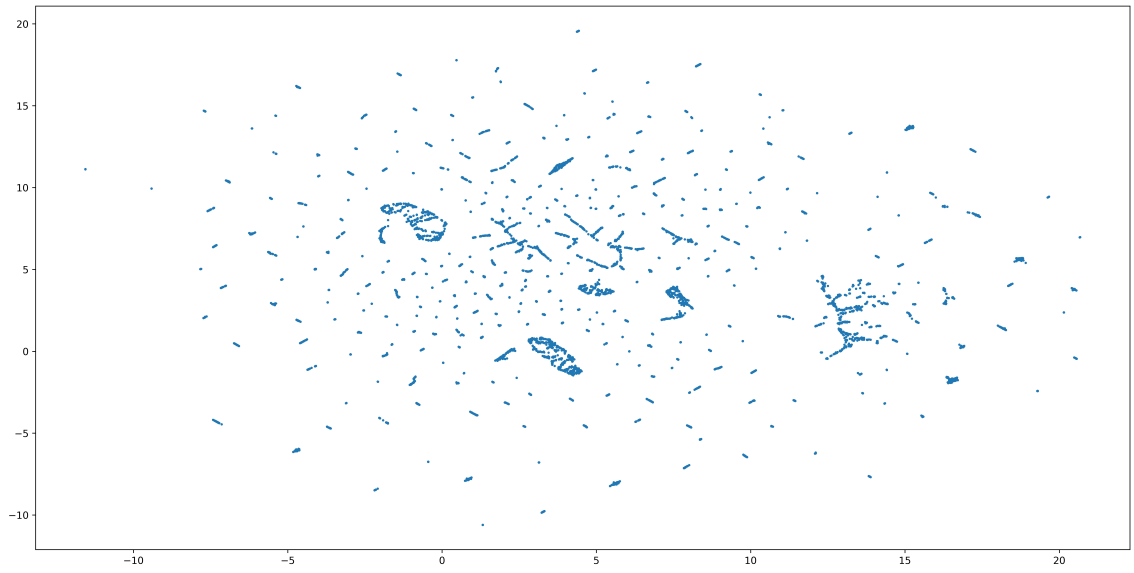


Figure B.2: Dimentionality reduction of EoC Bundle data using UMAP.

Appendix C

Additional Visualisations

C.1 First Iteration Visualisations

The following present additional visualisations presented to the clinicians when they asked for more details regarding contacts of diagnoses.

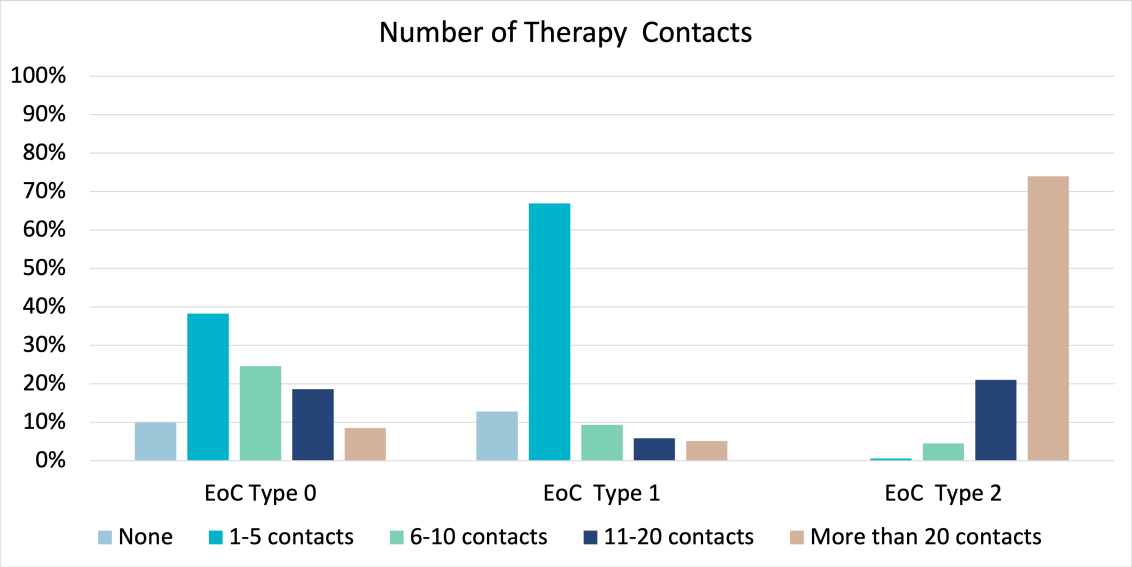


Figure C.1: First iteration's distribution of therapy contacts.

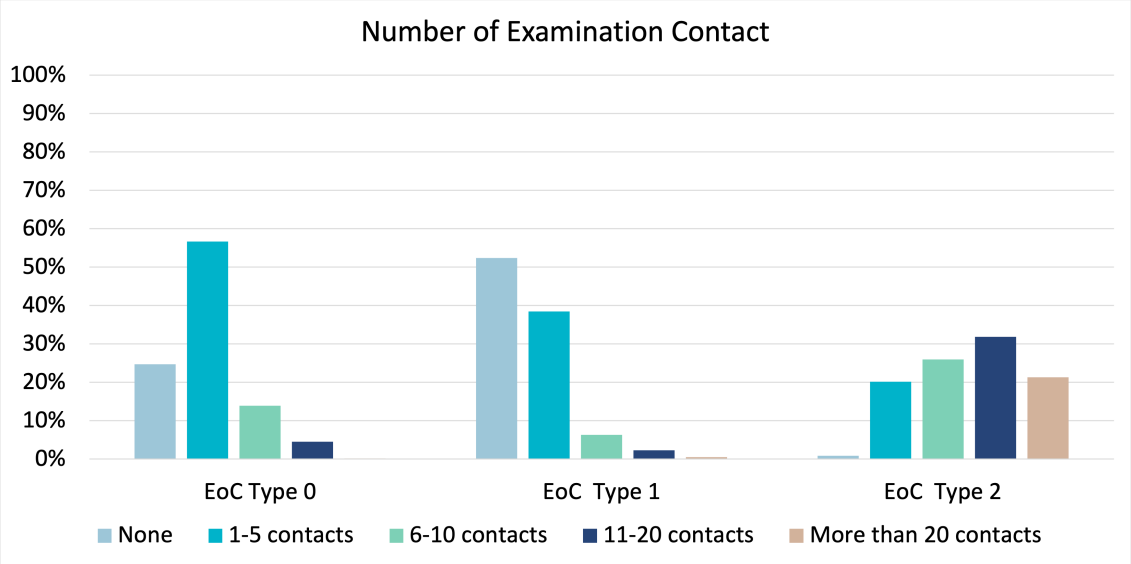


Figure C.2: First iteration's distribution of examination contacts.

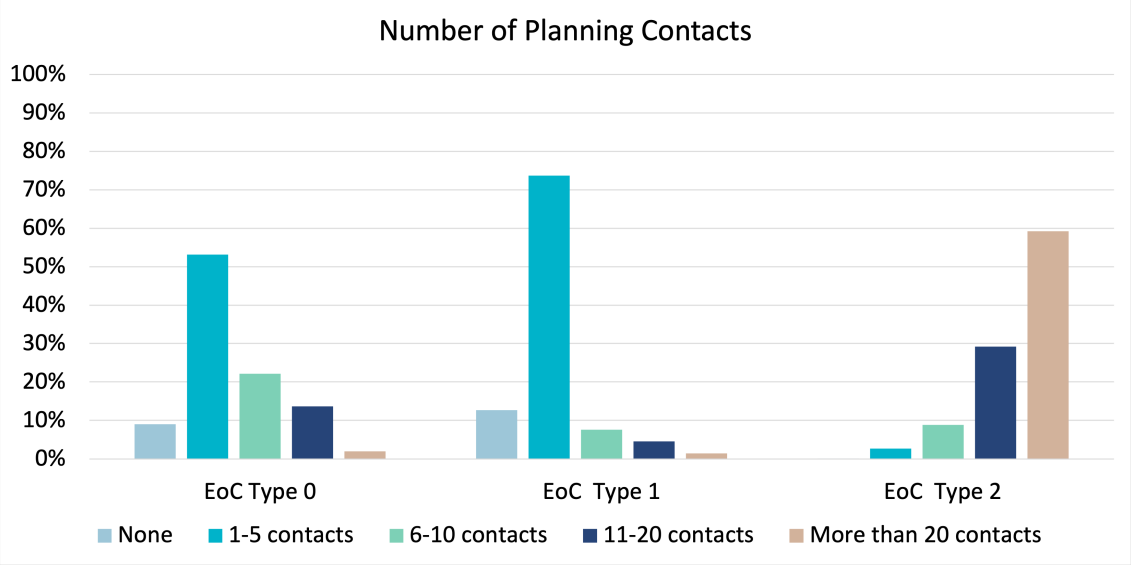


Figure C.3: First iteration's distribution of planning contacts.

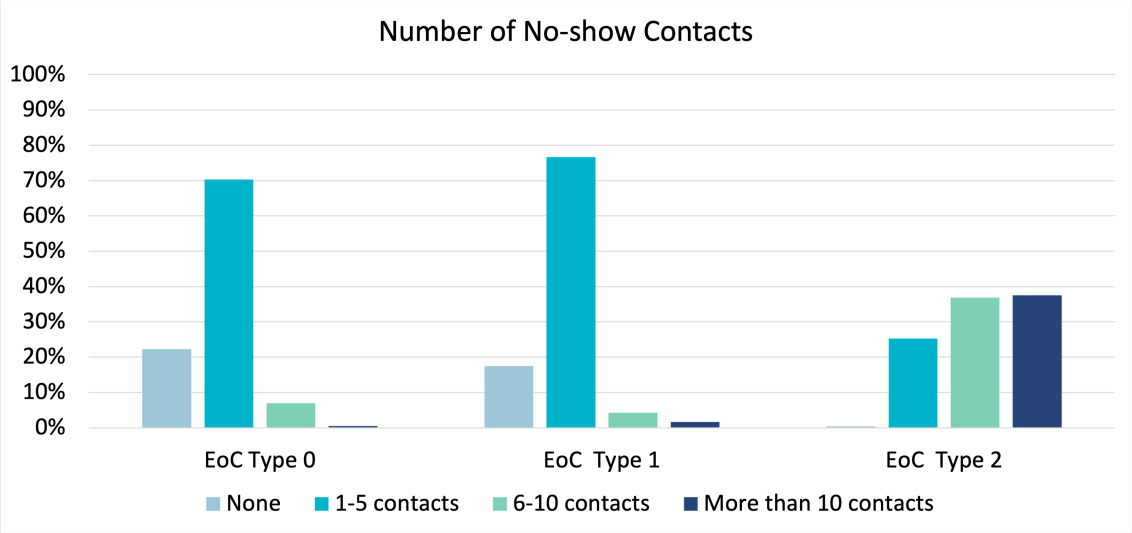


Figure C.4: First iteration's distribution of no-show contacts.

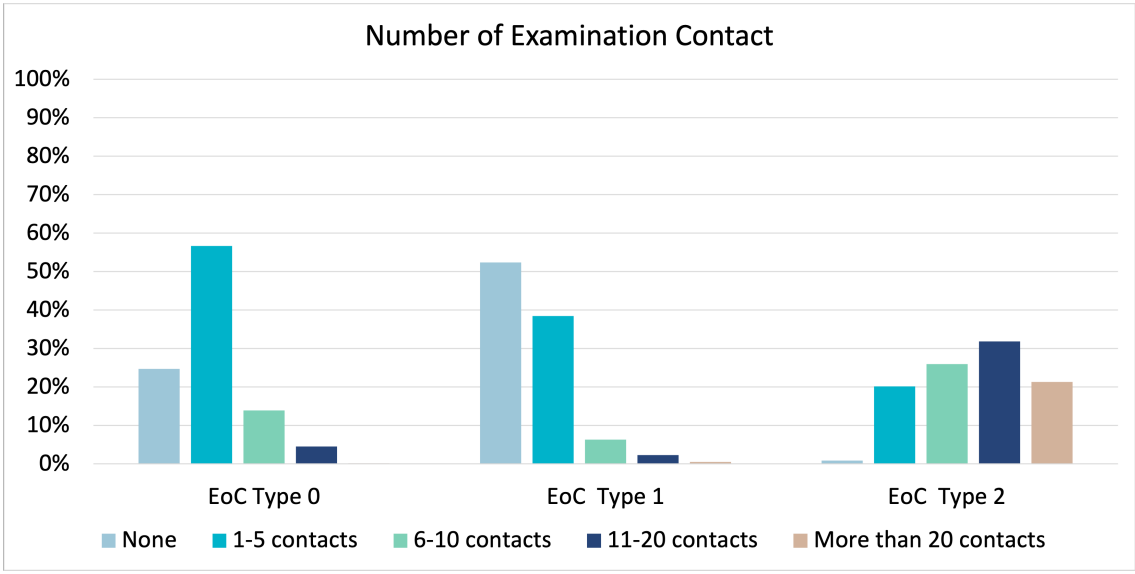


Figure C.5: First iteration's distribution of examination contacts.

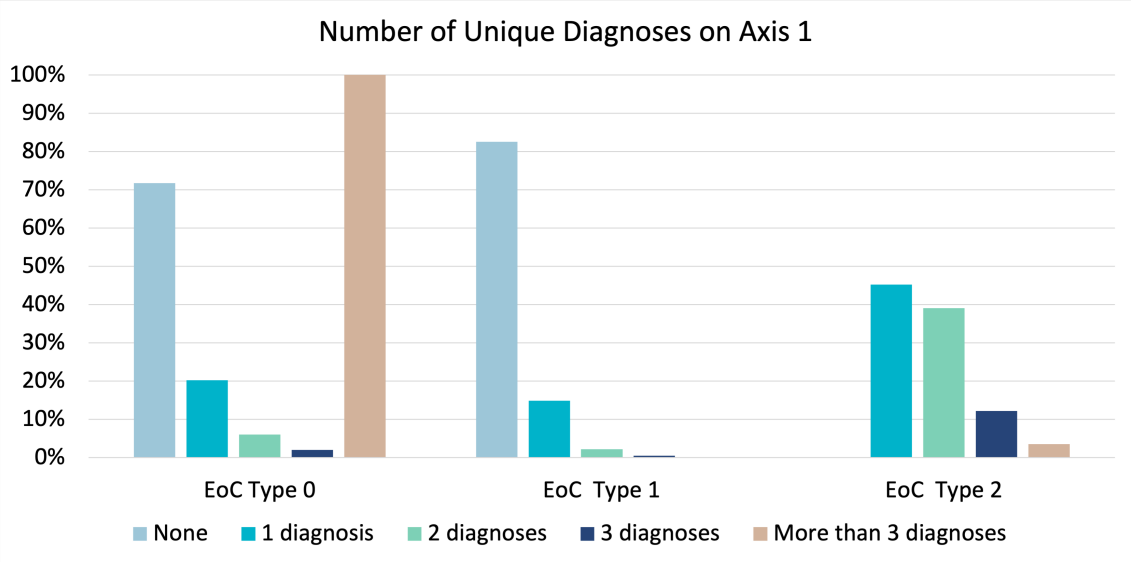


Figure C.6: First iteration's distribution of the unique number of diagnoses on Axis 1.

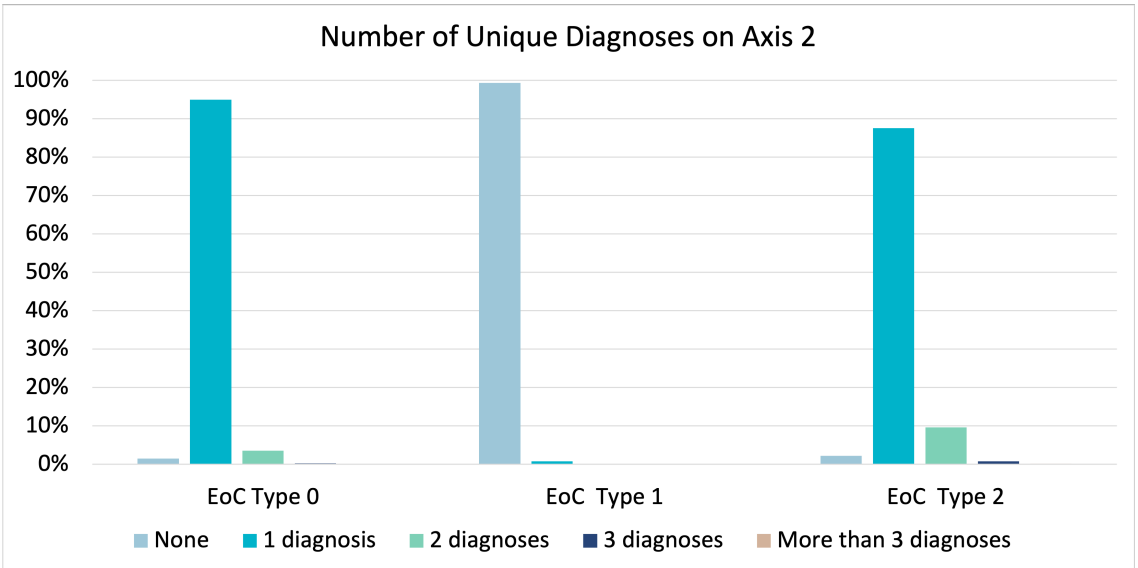


Figure C.7: First iteration's distribution of the unique number of diagnoses on Axis 2.

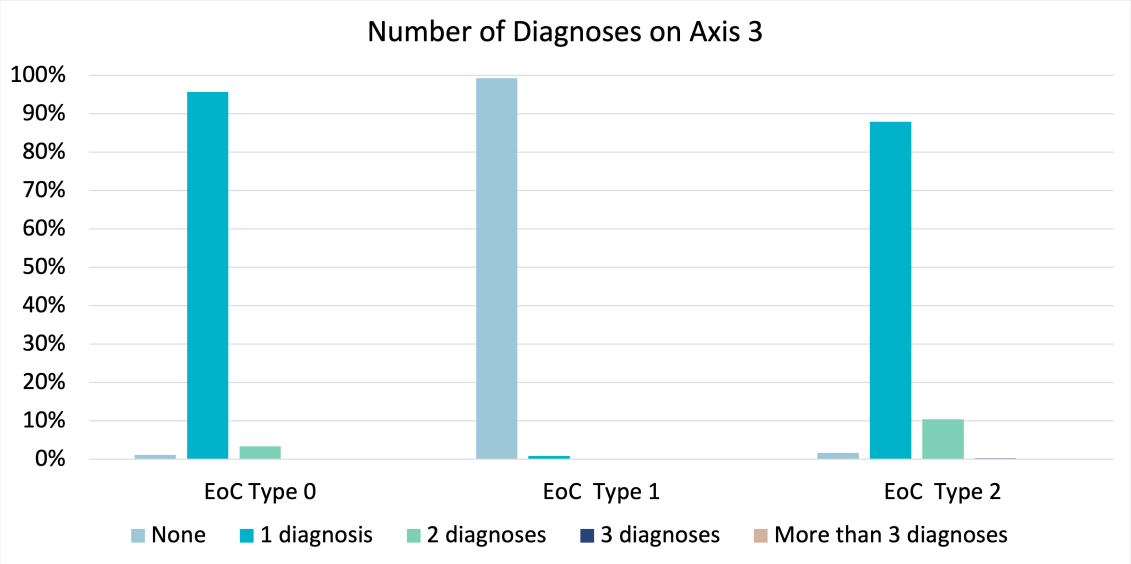


Figure C.8: First iteration's distribution of the unique number of diagnoses on Axis 3.

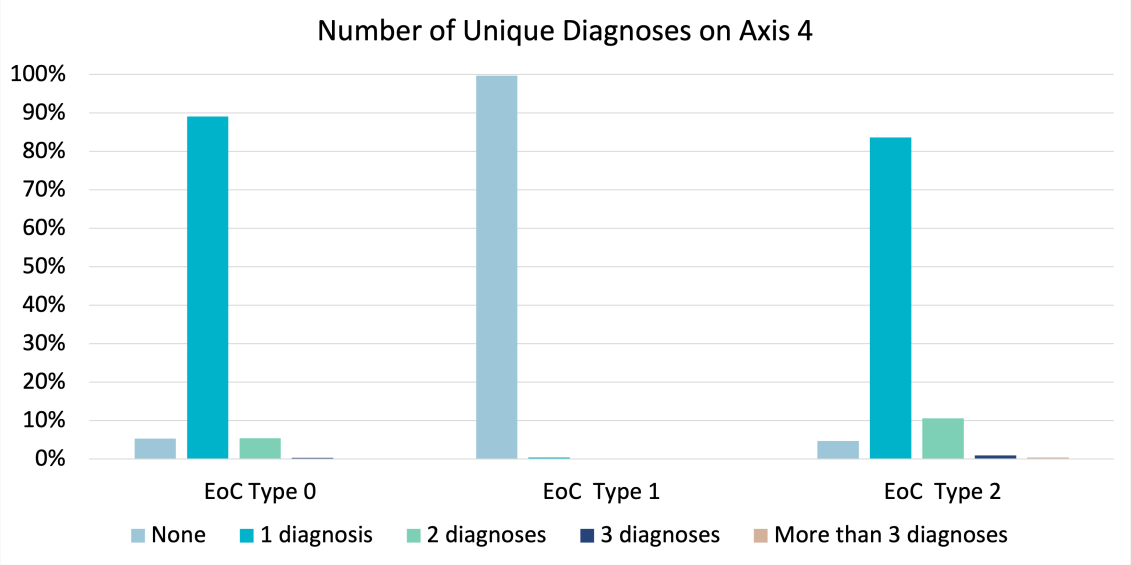


Figure C.9: First iteration's distribution of the unique number of diagnoses on Axis 4.

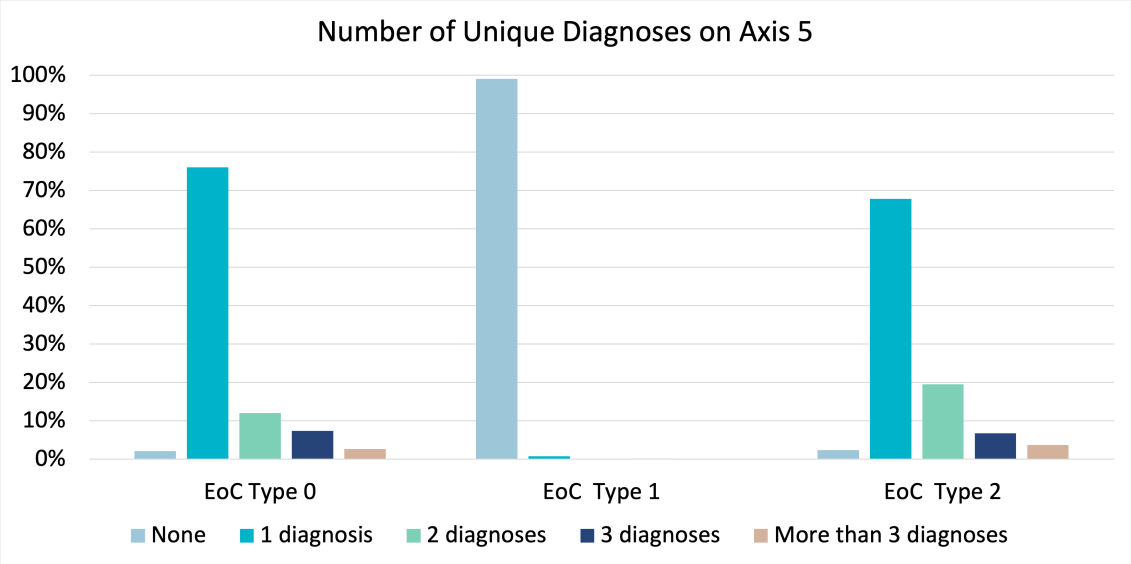


Figure C.10: First iteration's distribution of the unique number of diagnoses on Axis 5.

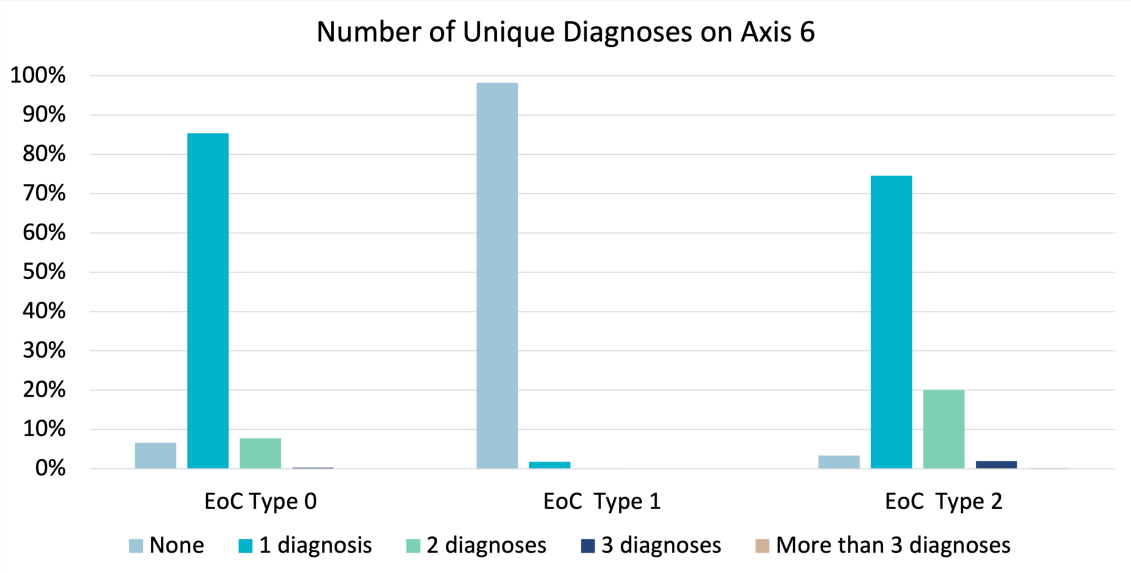


Figure C.11: First iteration's distribution of the unique number of diagnoses on Axis 6.

C.2 Second Iteration Visualisations

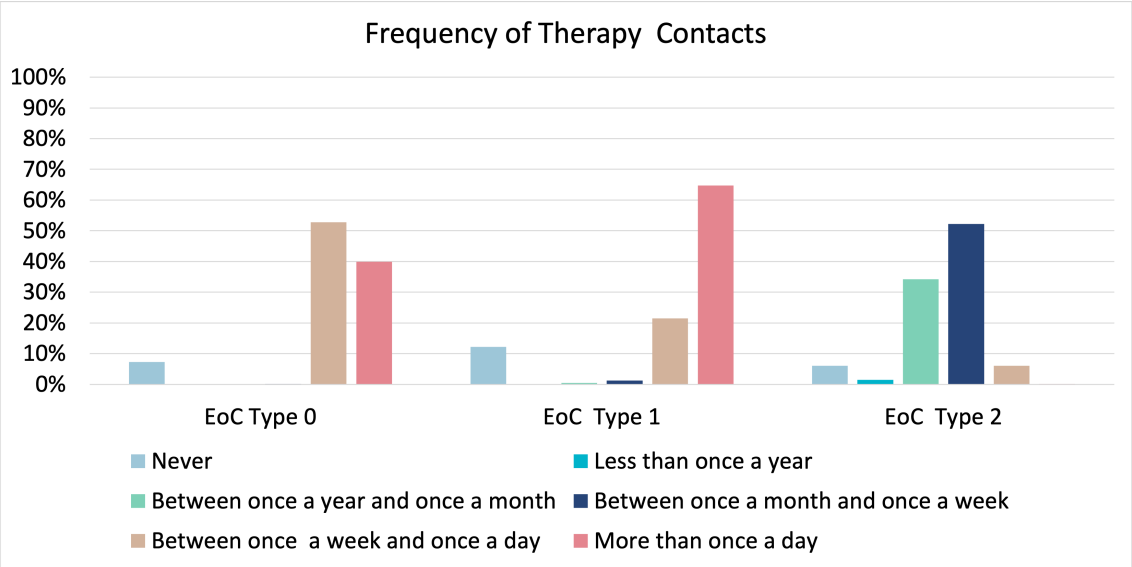


Figure C.12: Second iteration's distribution of therapy contacts.

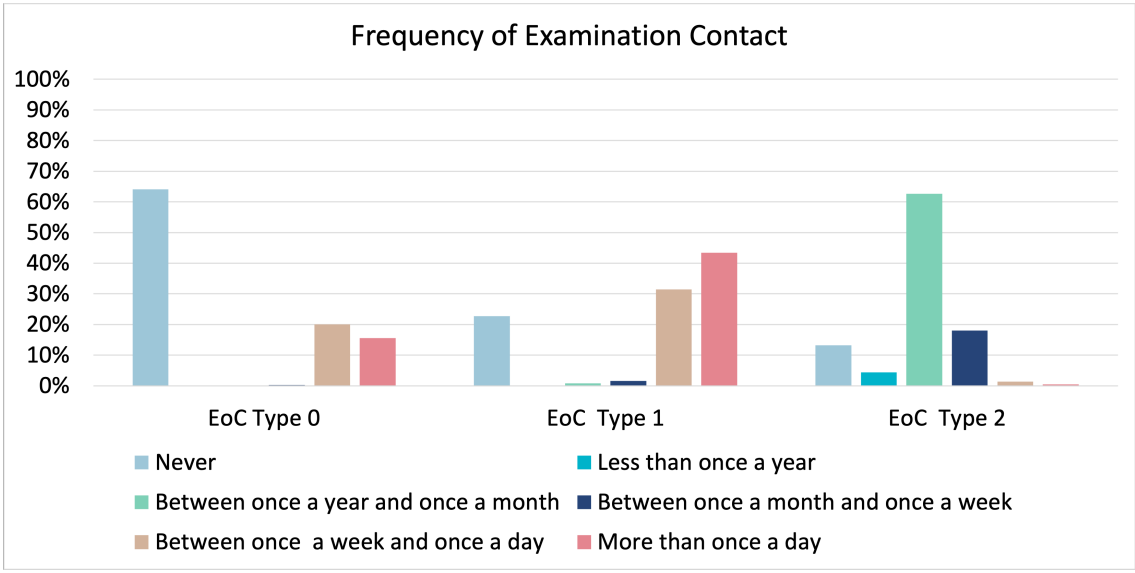


Figure C.13: Second iteration's distribution of examination contacts.

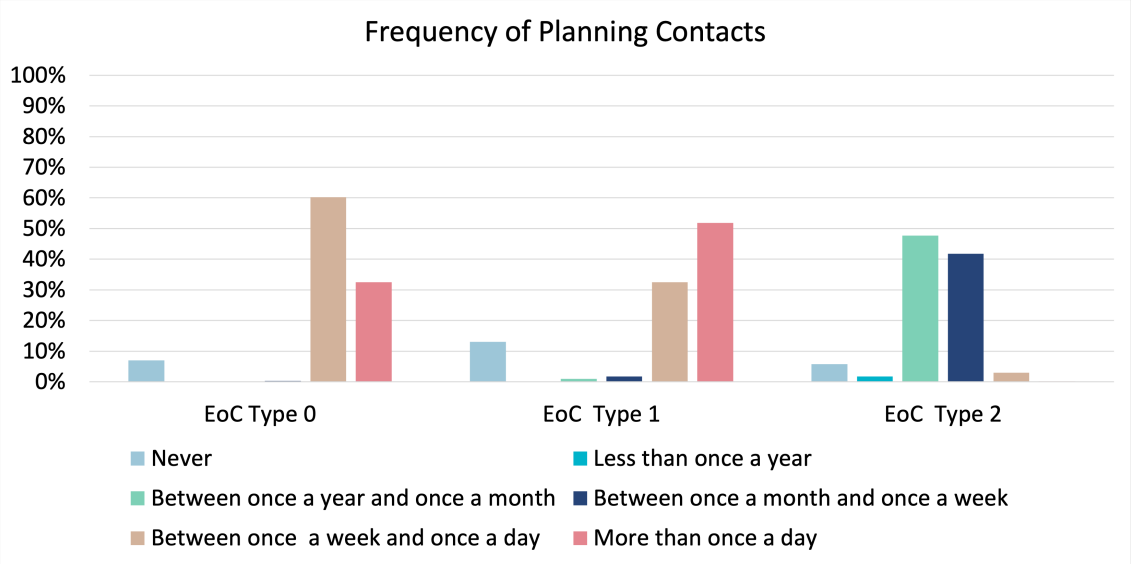


Figure C.14: Second iteration's distribution of planning contacts.

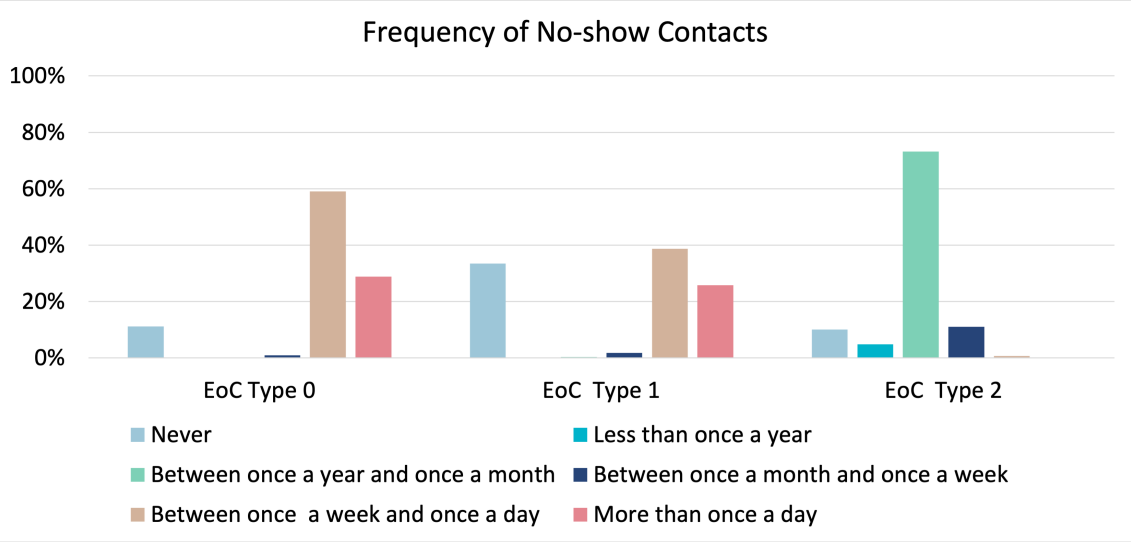


Figure C.15: Second iteration's distribution of no-show contacts.

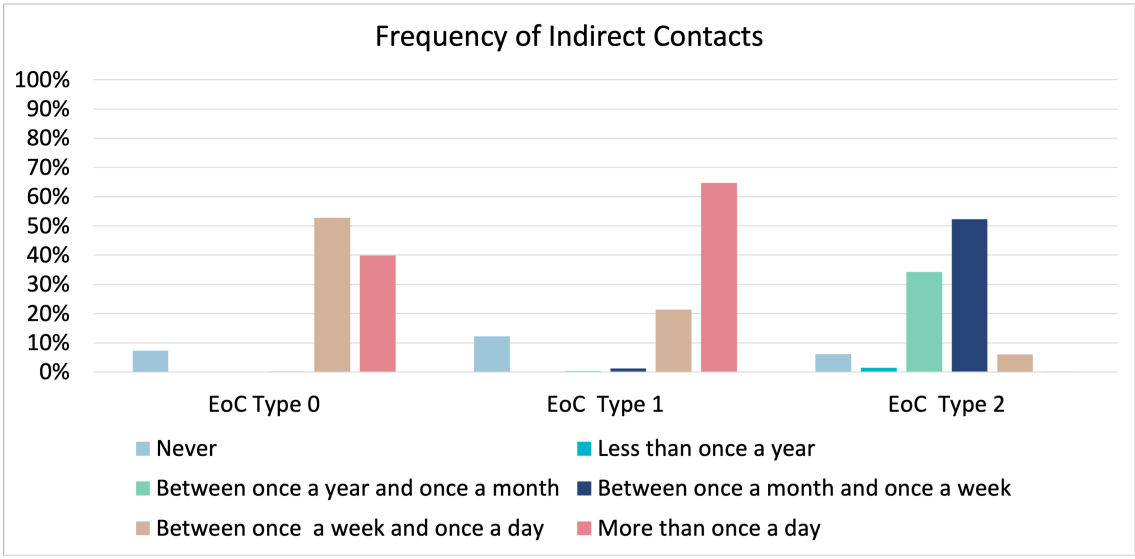


Figure C.16: Second iteration's distribution of examination contacts.

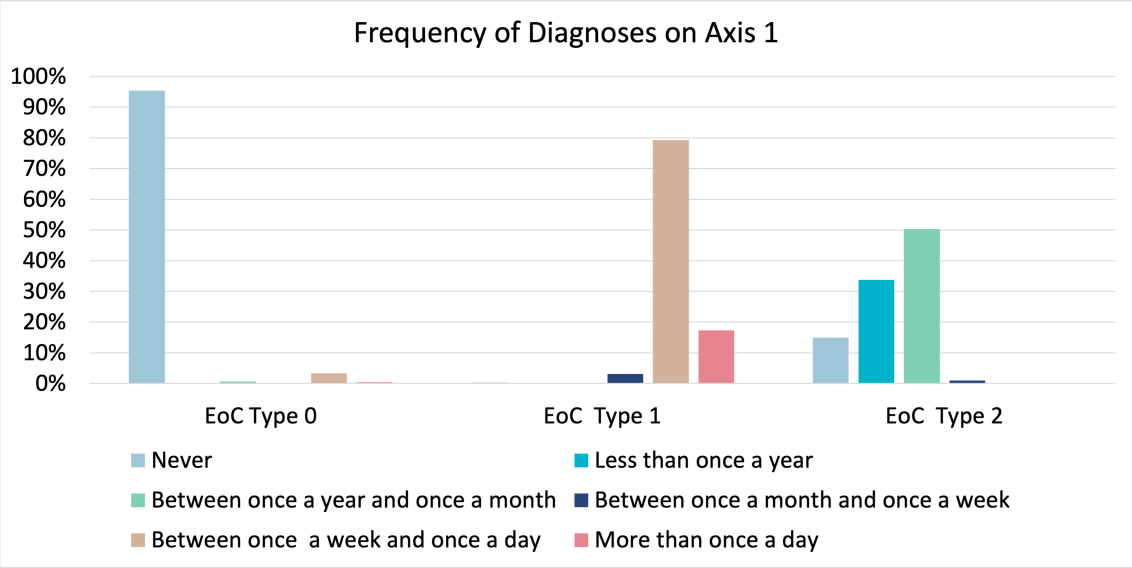


Figure C.17: Second iteration's distribution of the frequency of diagnoses set on Axis 1.

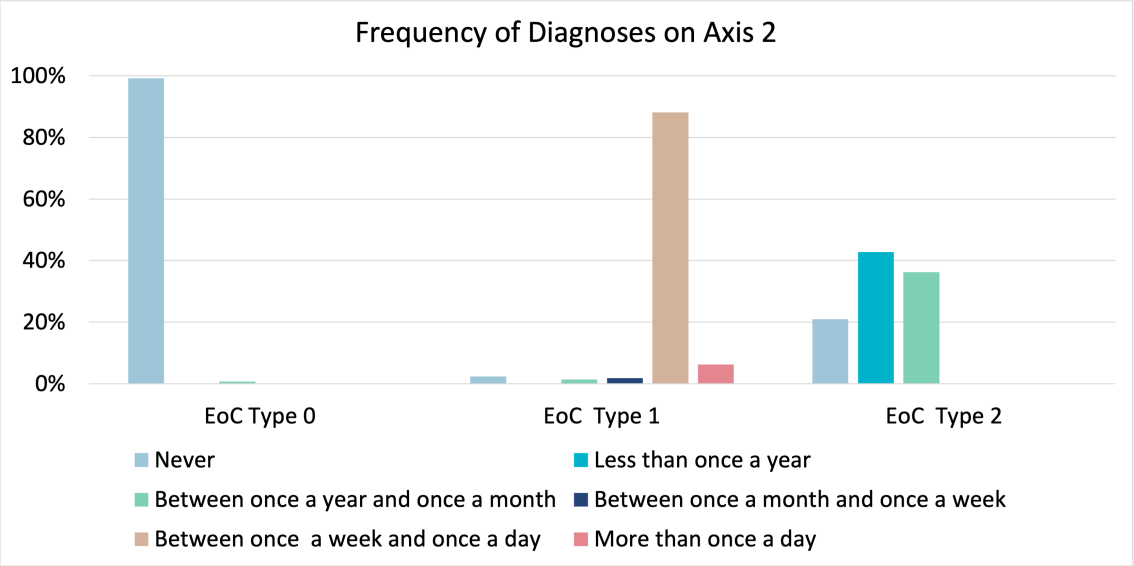


Figure C.18: Second iteration's distribution of the frequency of diagnoses set on Axis 2.

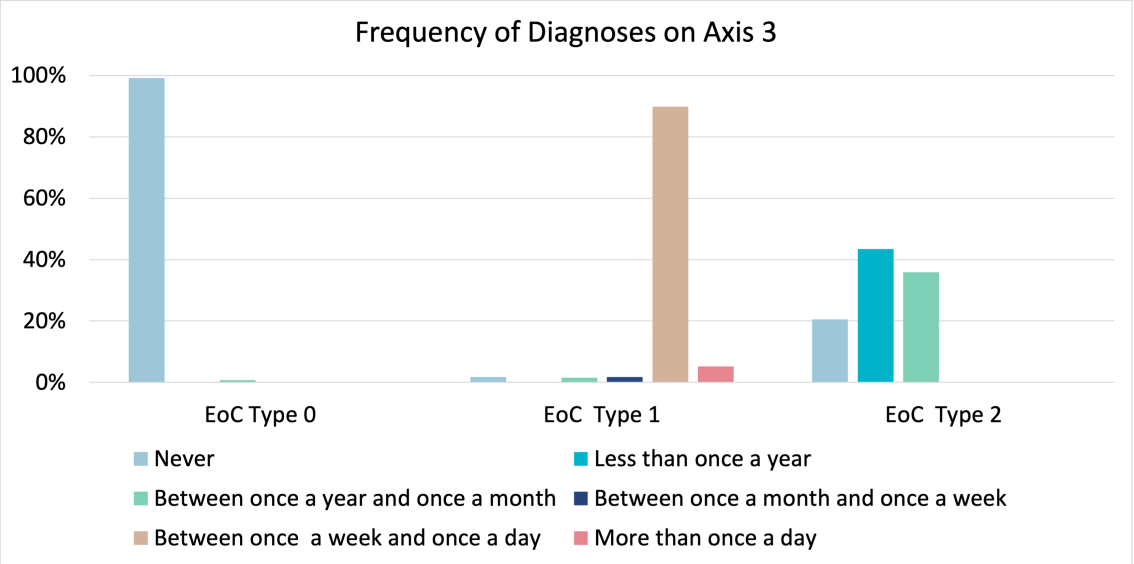


Figure C.19: Second iteration's distribution of the frequency of diagnoses set on Axis 3.

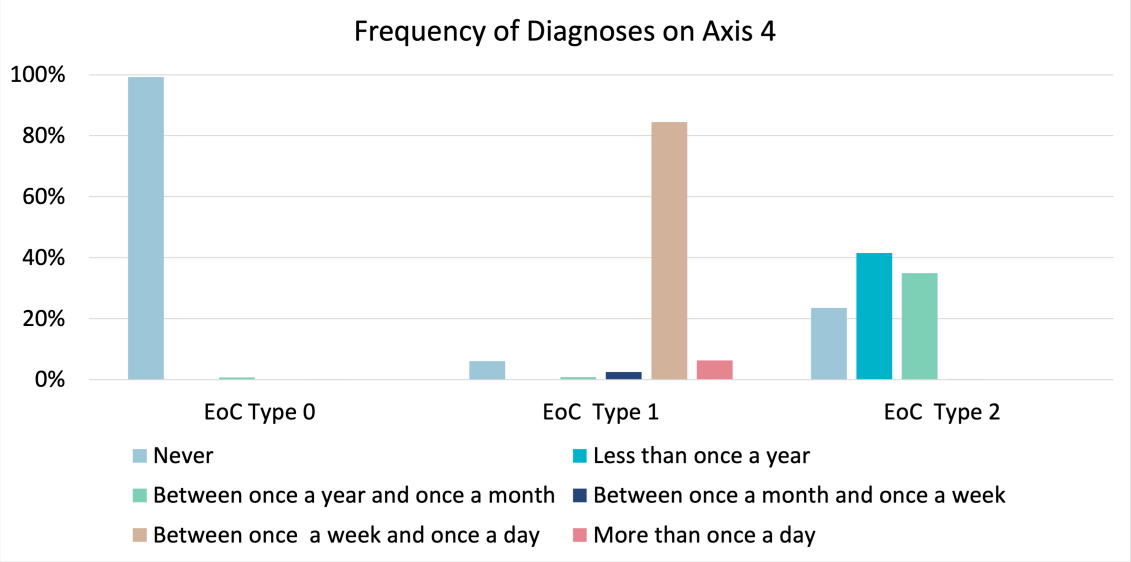


Figure C.20: Second iteration's distribution of the frequency of diagnoses set on Axis 4.

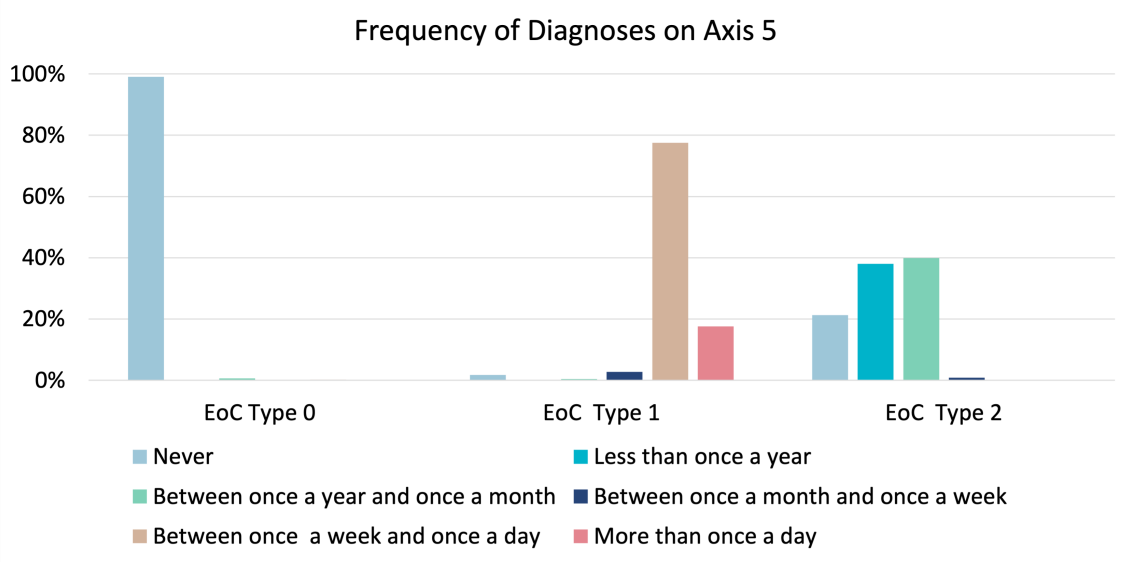


Figure C.21: Second iteration's distribution of the frequency of diagnoses set on Axis 5.

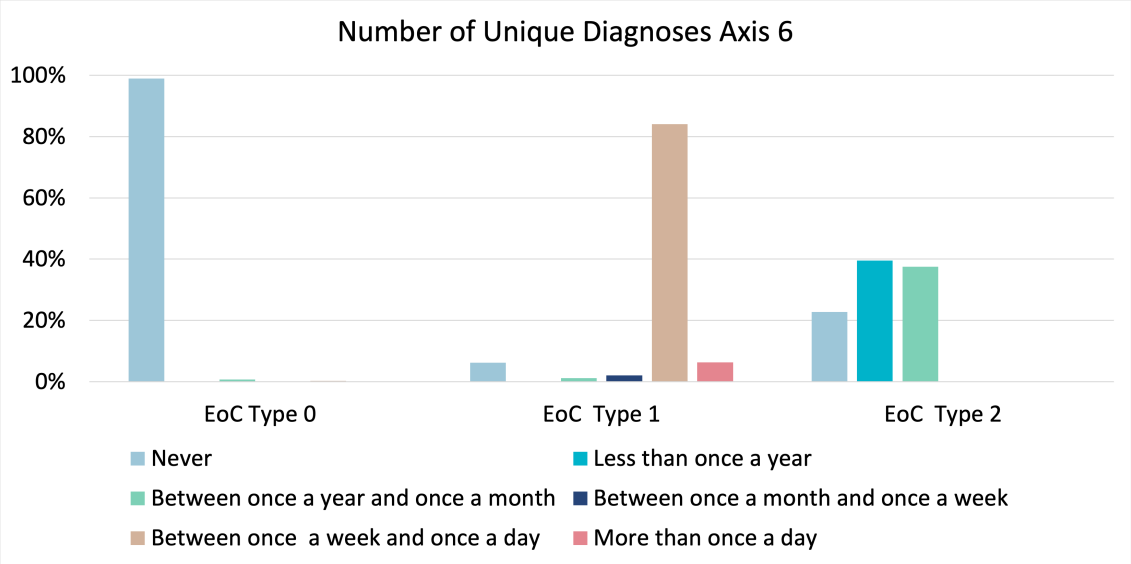


Figure C.22: Second iteration's distribution of the frequency of diagnoses set on Axis 6.

C.3 Third Iteration Visualisations

This appendix includes all the additional visualisations shown to clinicians if they asked for further details regarding contacts or diagnoses.

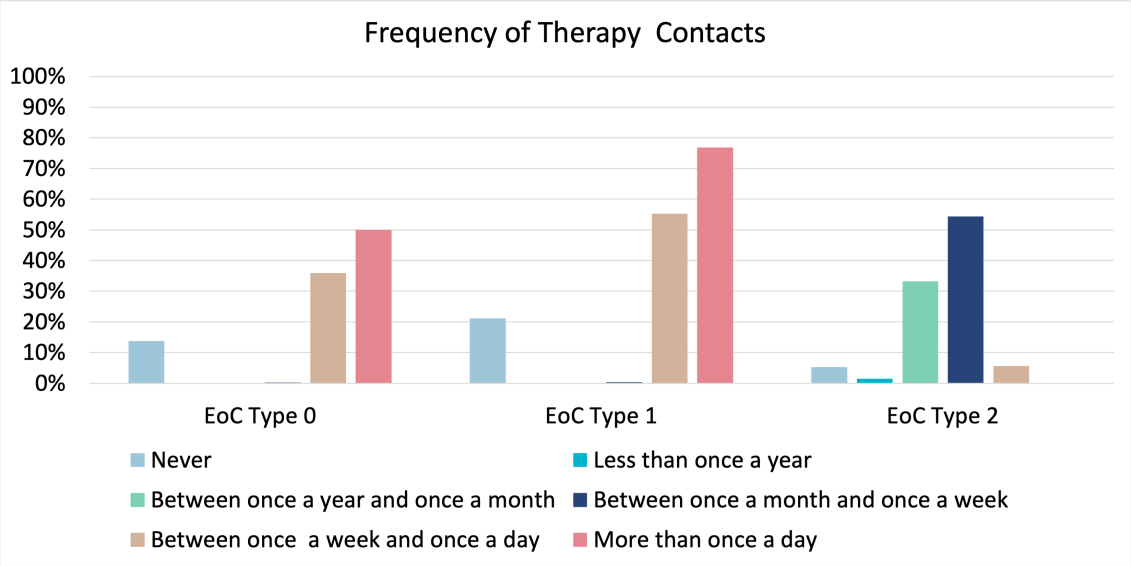


Figure C.23: Third iteration's distribution of therapy contacts.

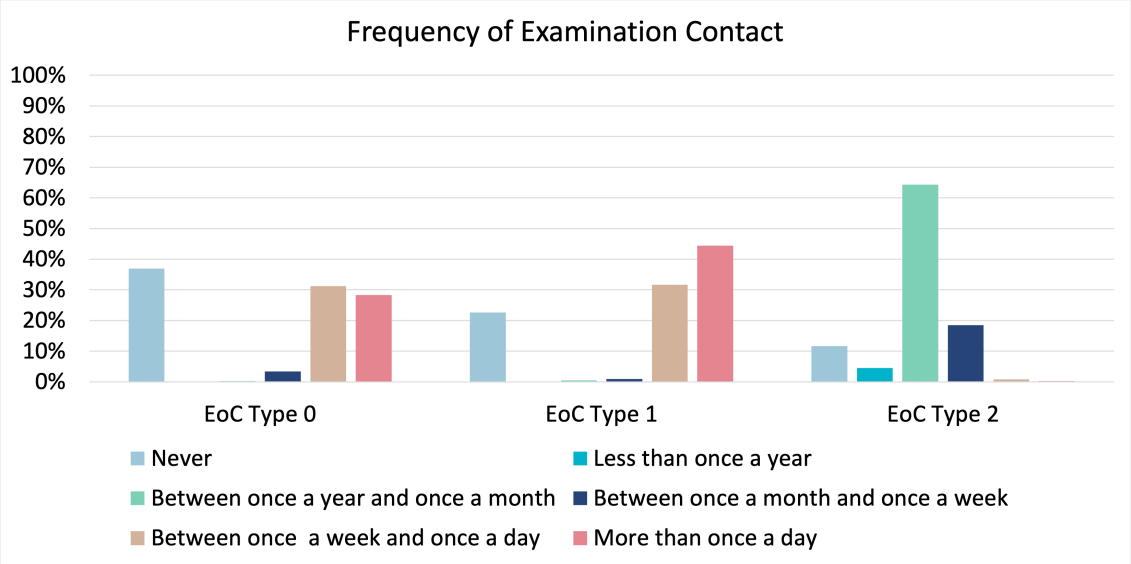


Figure C.24: Third iteration's distribution of examination contacts.

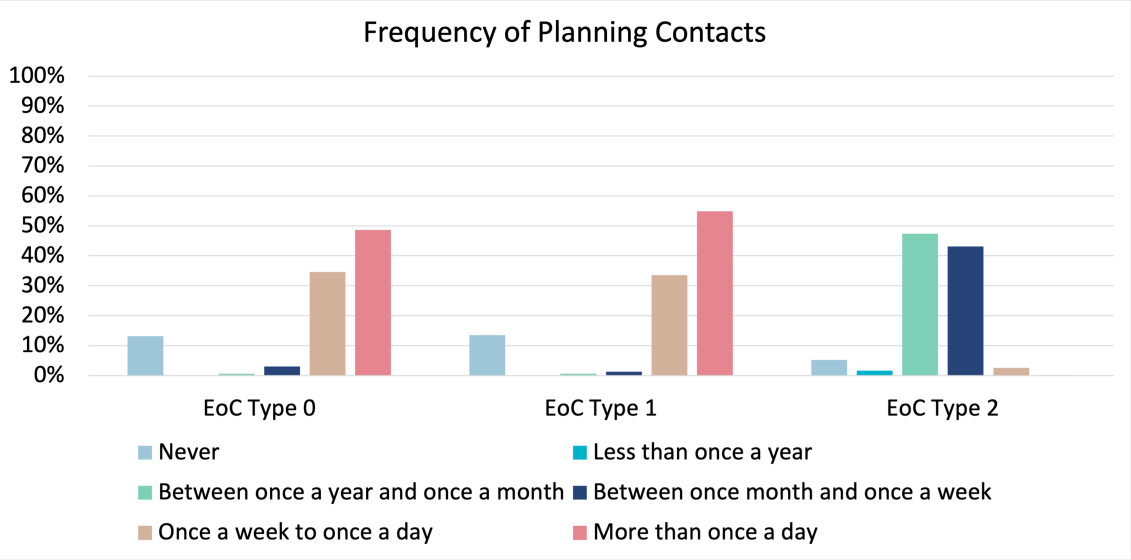


Figure C.25: Third iteration's distribution of planning contacts.

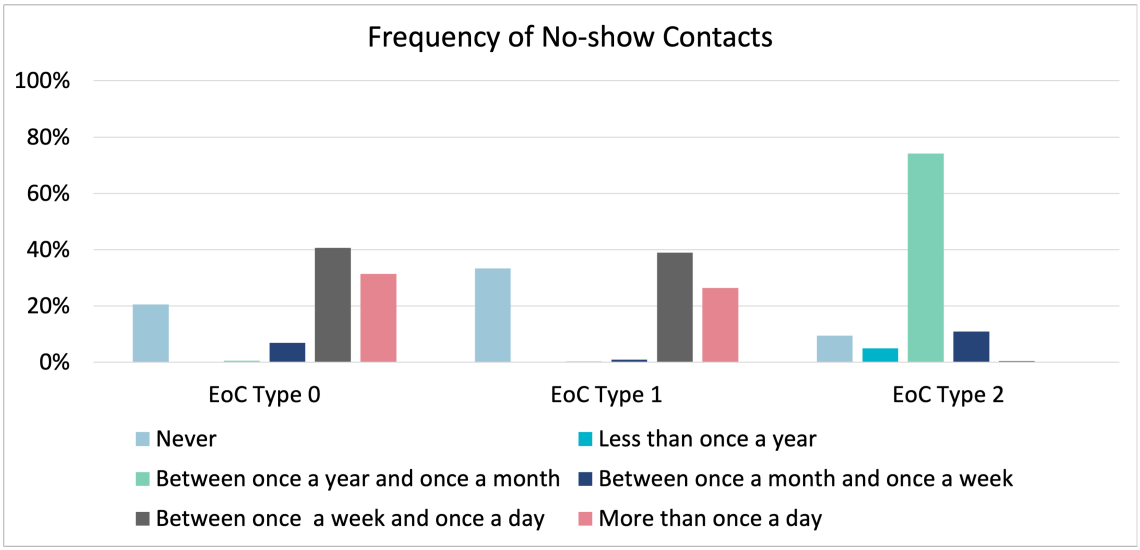


Figure C.26: Third iteration's distribution of no-show contacts.

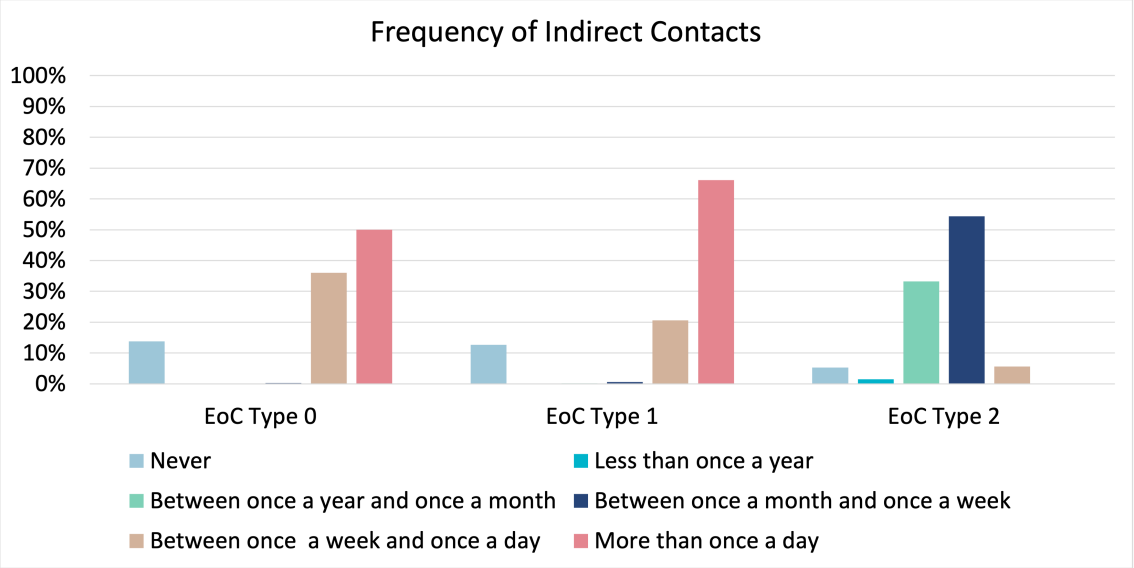


Figure C.27: Third iteration's distribution of examination contacts.

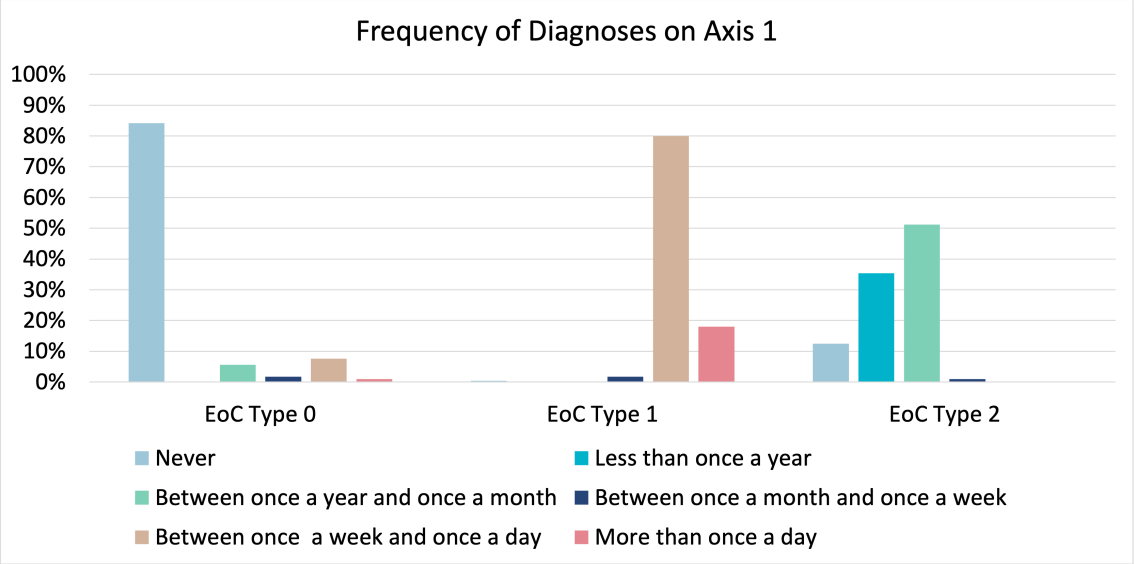


Figure C.28: Third iteration's distribution of the frequency of diagnoses set on Axis 1.

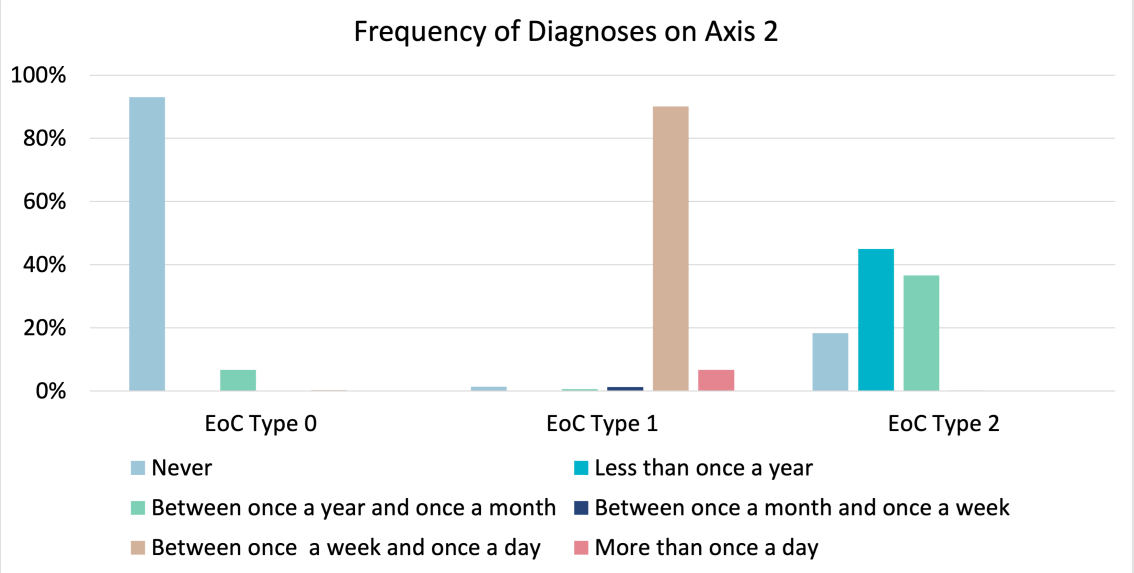


Figure C.29: Third iteration's distribution of the frequency of diagnoses set on Axis 2.

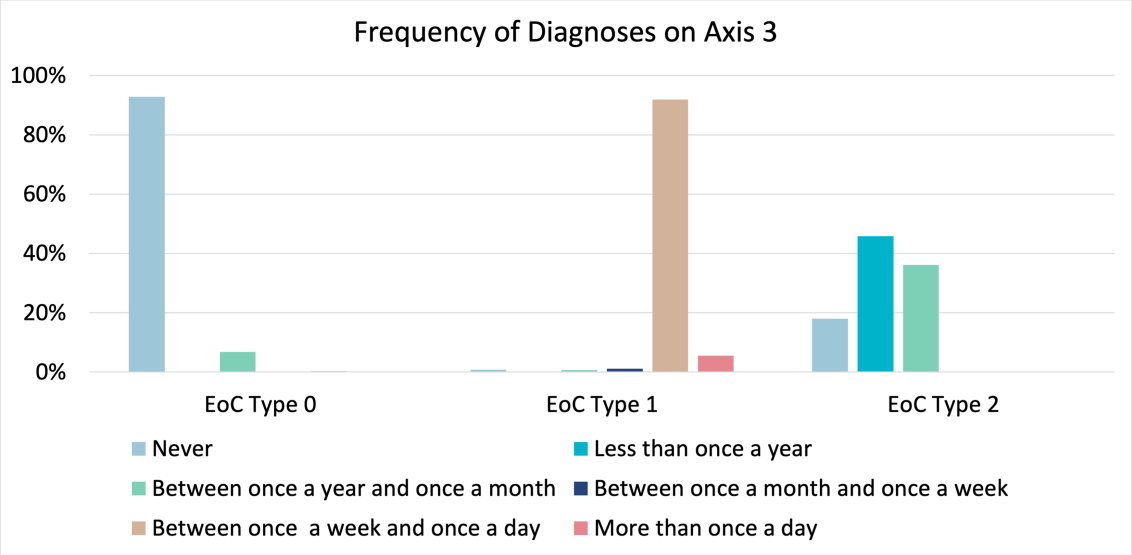


Figure C.30: Third iteration's distribution of the frequency of diagnoses set on Axis 3.

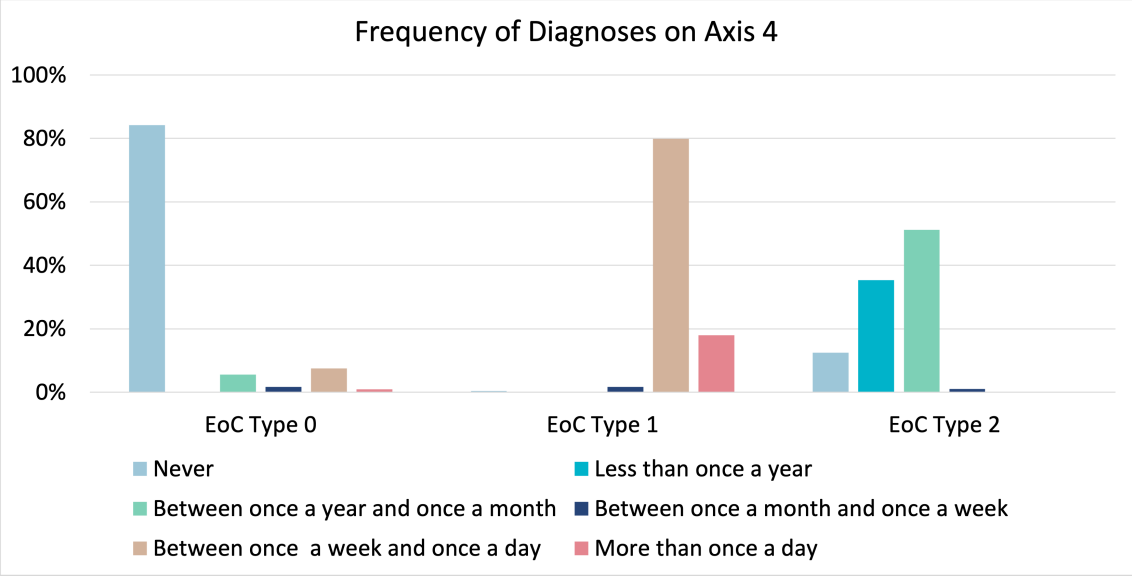


Figure C.31: Third iteration's distribution of the frequency of diagnoses set on Axis 4.

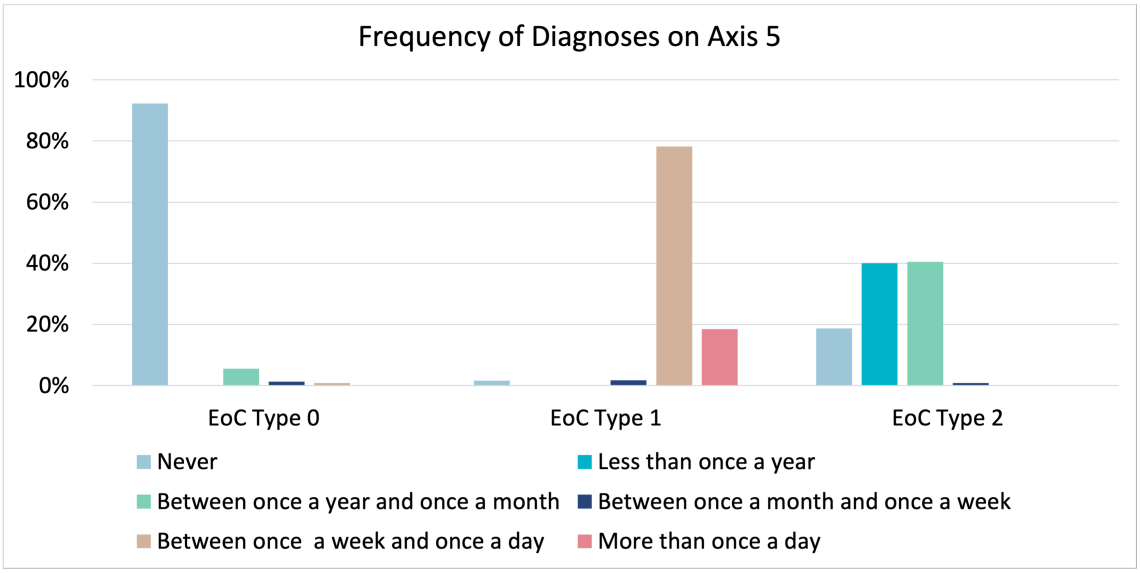


Figure C.32: Third iteration's distribution of the frequency of diagnoses set on Axis 5.

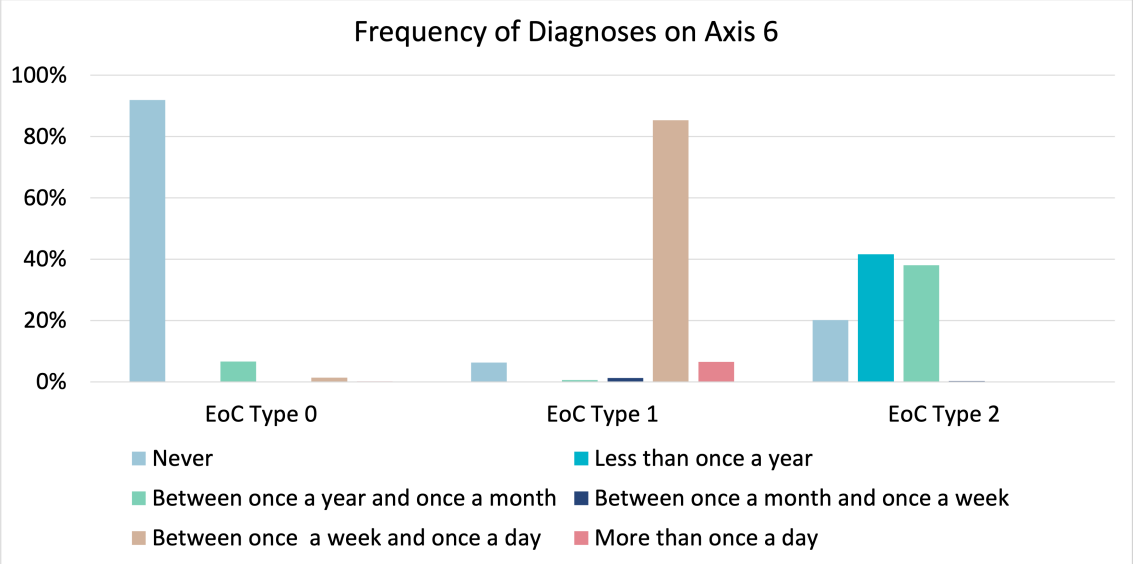
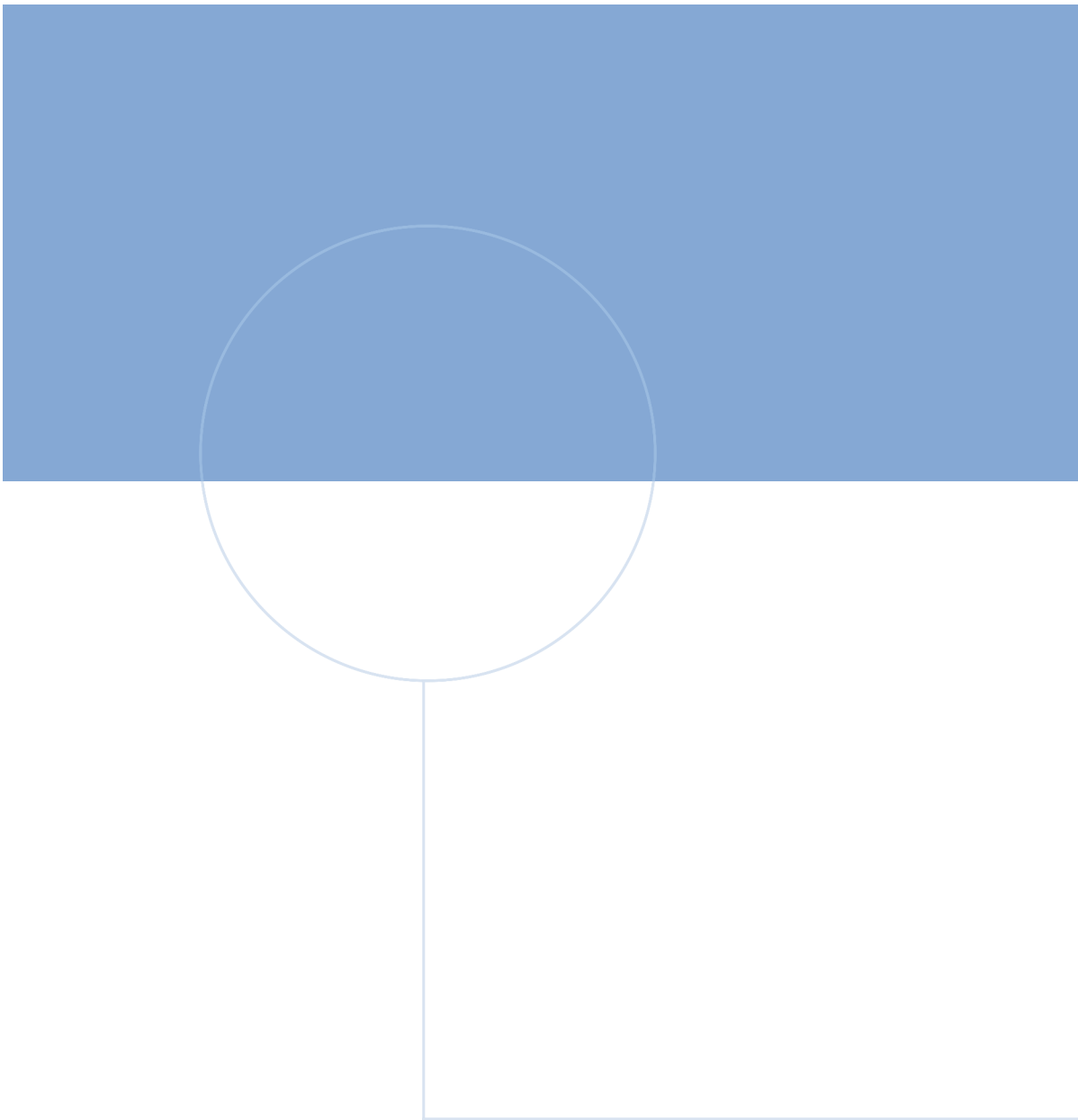


Figure C.33: Third iteration's distribution of the frequency of diagnoses set on Axis 6.



NTNU

Norwegian University of
Science and Technology