Aalto University
School of Science

Master’s programme in Life Science Technologies

Cluster analysis of cancer patients’ initial session patterns in a patient portal application

Michaela Weiss

Master’s Thesis
2024

Aalto University
School of Science

Author Michaela Weiss

Title Cluster analysis of cancer patients' initial session patterns in a patient portal application

Degree programme Life Science Technologies

Major Bioinformatics and Digital Health

Supervisor D.Sc. Talayeh Aledavood

Advisor M.Sc. Anni Parkkila

Date 2024 **Number of pages** 57 **Language** English

Abstract

The rising number of cancer patients creates a need for more efficient follow-up of cancer treatment as well as communication between cancer patients and their healthcare providers. Digital patient portals can provide improved care by giving patients access to their health records and communication with their care team. However, actual usage of different digital patient portals has shown variability, and the initial perception of a patient portal may affect further continued use. This thesis examines whether the user patterns of patients' initial sessions on the Kaiku Health patient portal, a portal dedicated to patients under cancer treatment, can be divided into groups and interpreted using clustering. The clustering task aims to increase the understanding of the patient users.

In this thesis, transactional log data are used to identify the initial sessions of 444 patients and clustered using four different approaches: Hierarchical clustering based on Gower's distance, hierarchical clustering based on cosine distance, dimensionality reduction using factorial analysis of mixed data (FAMD) and k-means clustering as well as uniform manifold approximation and projection (UMAP) dimensionality reduction and k-means clustering. According to internal validation indices, the best result is the UMAP-based k-means clustering result. All four clustering results are analyzed and validated using decision trees, statistical summation, and visual inspection.

The FAMD-based k-means clustering result is determined to produce the best clusters using domain knowledge. These clusters are transformed into user groups by summarizing the cluster centroids. Five distinct user groups are observed: fast mobile browser users, no-feature users, notification enablers and result viewers, message-using multi-feature users, and late registered super users. The user groups show distinct differences in the activity after the initial sign-in. Still, there are no signs of unfocused and nonstrategic use when the Kaiku Health patient portal is used for the first time. Hence, findings like this increase the understanding of the patient users and contribute to developing the portal.

Keywords clustering, digital patient portal, user pattern, initial sign-in , cancer treatment

Aalto-universitetet
Högskolan för
teknikvetenskaper

Författare Michaela Weiss

Titel Klusteranalys av patienters användarmönster vid första sessionen i en patient portal applikation

Utbildningsprogram Life Science Technologies

Huvudämne Bioinformatics and Digital Health

Övervakare D.Sc. Talayeh Aledavood

Handledare M.Sc. Anni Parkkila

Datum 2024 **Sidantal** 57 **Språk** engelska

Sammandrag

Det ökade antalet cancerpatienter skapar ett behov av en mer effektiv uppföljning av cancerbehandling samt kommunikation mellan cancerpatienter och deras vårdgivare. Digitala patientportaler kan erbjuda förbättrad vård genom att ge patienten tillgång till sina journaler och möjliggöra kommunikation mellan patienten och patientens vårdteam. Dock har den verkliga användningen av patientportaler visat sig vara varierande, och det första intrycket av en patientportal kan ha en inverkan på fortsatt användning av portalen. Denna avhandling undersöker huruvida patienter kan grupperas baserat på sina användarmönster vid första inloggningen på Kaiku Health portalen, en portal för uppföljning av patienter under cancerbehandling. Dessutom tolkas vad dessa grupper representerar. Syftet med arbetet är att öka förståelsen för patientanvändare genom analysen.

I det här arbetet analyseras transaktionsloggsdata för att identifiera den första inloggnings-sessionen för 444 patientanvändare. Inloggnings-sessionerna grupperas med hjälp av fyra olika klustringsmetoder: Hierarkisk klusteranalys baserad på Gowers distans, hierarkisk klusteranalys baserad på cosinusdistans samt k-means algoritmen kombinerad med FAMD-dimensionalitetsreduktion och k-means algoritmen kombinerad med UMAP-dimensionalitetsreduktion. Enligt interna valideringsmått är den bästa klustringen UMAP-baserad k-means klustringen. Alla fyra klustringsresultat analyseras och valideras med beslutsträd, statistisk sammanfattning och visuell inspektion.

Klustringsresultatet baserat på FAMD-baserad k-means klustring är bäst enligt expertkunskap. Denna metods kluster omvandlas till användargrupper baserat på en sammanfattning av klustercentroiderna. Fem olika användargrupper identifieras: snabba mobil-webbläsare användare, inga-funktioners användare, notifikations-aktiverande och resultat-uppföljande användare, meddelande-användande flerfunktions användare samt sent registrerade superanvändare. Användargrupperna uppvisar tydliga skillnader i aktivitet efter första inloggningen, men det finns inga tecken på ofokuserad och ostrategisk användning när Kaiku Health portalen används för första gången. Dessa resultat kan öka förståelsen för hur patienter upplever patientportalen och bidra till utvecklingen av denna.

Nyckelord klustring, digital patient portal, användarmönster, första inloggning, cancerbehandling

Preface

I would like to thank my employer Elekta, for giving me this research project opportunity that resulted in this thesis. At Elekta, I would also like to thank my advisor Anni Parkkila, for taking the time to advise on this thesis. In addition, I would like to thank Idamaria Lehtinen for her thoughts and advice on the research problem and Riku Lehtomäki for his input in forming the research question. Your knowledge of the topic has been an important part of the thesis.

Thanks to my supervisor, Talayeh Aledavood from Aalto University, for your advice on writing a good academic text and for giving me valuable feedback throughout the whole thesis process.

Finally, I like to thank friends and family, who have encouraged me while working on the thesis. Thanks to all my friends, especially the community at Teknologföreningen, who made my years studying at Aalto fun and inspiring. Without you, there would have been a lot less laughter in my everyday student life.

Helsinki, 20 March 2024

Michaela Weiss

Contents

Abbreviations and symbols

Abbreviations

CH	Calinski-Harabasz index
FAMD	factor analysis of mixed data
MCA	multiple correspondence analysis
PCA	principal component analysis
SC	Silhouette coefficient
UMAP	uniform manifold approximation and projection

Symbols

$\in$	is an element of
$\sum$	summation

Operators

$\mathbf{A} \cdot \mathbf{B}$	dot product of vectors **A** and **B**
$\|\| \cdot \|\|$	Euclidean norm

1 Introduction

Cancer is a disease that many people will come into contact with during their lifetime, either as a patient or as a relative of one. In Europe, the burden of the disease has increased since the 1990s, and with an aging population, the trend seems to continue [1]. Although the cancer death rate has declined by 33% since 1991 due to the development of treatments, nearly 2 million new cancer cases and 600 000 cancer deaths are estimated to occur during 2023 in the United States [2]. Consequently, there is a need for more efficient follow-up of cancer patients and their treatment, as well as more efficient communication between cancer patients and their healthcare providers [3]. Increased use of digital patient portals can be part of the solution to the need for more efficient cancer care [4].

Digital or electronic patient portals provide patients access to their health records and an opportunity to communicate with their health care providers by a messaging function [4]. Hence, the digital patient portal can improve the care the patients receive [5]. When used in cancer care, digital patient portals can also offer patients a way to receive help in managing the side effects of cancer treatment. Furthermore, digital patient portals can engage cancer patients in their care and make the patients communicate with their care team between appointments for improved cancer care [6, 7].

An example of a digital patient portal provider is Kaiku Health, a digital health intervention platform in cancer care. When used as an electronic patient portal for cancer patients, Kaiku Health provides a messaging function, an appointment calendar, and a platform for documents and records. Used both for symptom monitoring and as an electronic patient portal, Kaiku Health has been reported as easy to use for patients [8]. In addition, the study found that some symptoms get detected early in the treatment when using Kaiku Health. Phone calls related to chemotherapy treatment have also been reduced when using Kaiku Health [8]. These results demonstrate a way to use Kaiku Health to improve cancer care, and patients have also reported a feeling of improvement in cancer care when using Kaiku Health in their care [8]. This thesis will focus on Kaiku Health as a patient portal.

Tracking the usage of an electronic patient portal is critical. To assume that technology usage, such as electronic patient portal usage, will be beneficial in caregiving without monitoring the actual usage seldom leads to improved care [9]. It may even lead to undetected weaknesses of the patient portal if results of the electronic patient portal usage are not collected [9]. The use of transactional log data from the electronic patient portal provides a reliable base for data analytics to improve the tool and get to know the users of the tool [10]. Understanding user behavior and interaction with the electronic patient portal is a key issue when developing the electronic patient portal and its features [5]. User behavior understanding is also essential when incorporating the electronic patient portal into the workflows of the health care providers and the patients [5]. Perceptions of the users and the workflows are created by analyzing transactional log data from the electronic patient portal.

Descriptive statistical studies and more complex user pattern studies of electronic patient portals use transactional log data from electronic patient portals. Several descriptive studies examine electronic patient portal enrollment [4] and different kinds of usage over time [10, 11, 12] together with various patient characteristics. In addition, cancer patients as patient portal users have been studied [6]. Also, the increased use of smartphones to access patient portals has been the focus of descriptive research, as smartphone ownership has become increasingly widespread in all age groups [7]. Still, performing more complex investigations on user patterns is important since descriptively examining the amount of use might not give the whole truth. Greater use of an electronic patient portal will not necessarily demonstrate improved health care outcomes and instead might reveal unfocused use and unsuitable systems [13]. These types of user pattern studies have been performed in closely related fields, for instance, on types and user patterns among patients with cardiac disease or diabetes in an electronic patient portal [5]. In addition, studies on patients' actual over-time usage of the electronic patient portals have shown variability in the uptake. Yet the main drivers for electronic patient portal implementation are transparency and increased access to health care [12]. Therefore, further research on the initial use of an electronic patient portal and an increased understanding of patient user patterns and patient portal users is needed.

This thesis will examine the user patterns of patients' initial sessions on the Kaiku Health patient portal and what device was used for the initial session to increase the understanding of the patient users. The questions to answer in the thesis are what happens during the session in the patient portal after the patient's initial sign-in and how long the patient will stay on the patient portal before sign-out. Furthermore, the thesis will examine the distribution of the initial sessions among accessing the patient portal by mobile app, mobile browser, and desktop browser. The goal of the thesis is to identify user groups based on patients' user patterns during the initial session and determine if there are groups of specific user types and which user types there are. The log data generated during the user's initial session and the length of the session reflect the user's initial perception of the patient portal. In addition, the relationship between the initial session length and the type and amount of generated log data also reflects the user's perception. This initial perception is significant as it may affect further uptake and use of the patient portal among patients [14]. Hence, the initial user experience perspective is critical for developing and improving the patient portal, eventually improving cancer care.

A cross-sectional cluster analysis of log data from the initial sessions of 444 anonymized cancer patients form user groups based on initial user session patterns. The cluster analysis consists of hierarchical and k-means clustering after feature selection, where relevant features from the Kaiku Health usage and login data are selected. The interpretation of the resulting clusters can be used in the development of the Kaiku Health patient portal.

2 Background

2.1 Cancer and cancer treatment

Cancer is defined as a disease in the genes [15] leading to malignant or uncontrollable cell growth, which starts somewhere in the body and then spreads to other parts. The normal lifespan of a cell in the human body includes controlled growth, cell division where the cell splits into new cells, and programmed cell death when the cell becomes old or damaged [16]. However, sometimes cells do not follow this pattern due to changes in the genes, which causes them to malfunction [15]. These damaged and abnormal cells can grow and divide out of control and may form tumors, which can be cancerous. A cancerous tumor can then spread to tissues nearby and other distant parts of the body [16], a process which is known as metastasis [17]. Furthermore, cancerous tumors can reappear and start growing again after being removed. Most cancer types, except blood cancer, form solid tumors [16].

Cancer is one of the most common causes of death worldwide, leading to nearly 10 million deaths during 2020 [17]. The primary cause of death in cancer is widespread metastases. Risk factors related to cancer are the use of tobacco and alcohol, obesity, unhealthy diet as well as lack of physical activity [17]. In addition, the cancer incidence increases with age [17], and it is estimated that two in ten persons will get cancer during their lifetime, of which one of these two persons will die from cancer [18]. Still, early diagnosis and appropriate treatment have a high chance of cure for many cancers [17].

There are different categorization ways and systems to describe the stage of a cancer tumor. Most systems cover the location and size of the tumor and whether the cancer has spread to nearby lymph nodes or different parts of the body [19].

The dataset in this study consists of patients with various cancer diagnoses in the active phase of the treatment, which means that they are under active treatment. The patients are receiving both radiotherapy, a type of treatment where the patient are receiving ionizing radiation targeted to kill the cancer cells, as well as medical oncology treatments. Medical oncology treatment includes chemotherapy, immunotherapy, hormonal therapy, and targeted therapies. Chemotherapy uses drugs to kill fast-growing and dividing cancer cells, and immunotherapy aims to strengthen the immune system, e.g., by vaccines or specialized medications. Furthermore, hormonal therapy targets the hormones that help cancer cells grow and targeted therapies target the proteins that control how cancer cells grow and divide [20].

2.2 Digital patient portals and Kaiku Health

Digital patient portals, also called electronic patient portals, are secure websites or apps that give patients continuous access to personal health information. The patient portal contains different features, depending on the patient portal and the provider of the electronic patient portal. Numerous patient portals provide messaging with healthcare professionals, scheduling of appointments, and viewing health records on lab results or summaries of appointments [21]. The patient portal usage can improve the engagement between patients and their healthcare [4, 6, 10, 11]. Therefore, they are associated with enhanced patient-provider communication and patient satisfaction [7, 22]. Patient portals have also been reported to improve health status awareness and commitment to therapy [23]. In addition, the usage of patient portals accessed through mobile phones has increased significantly over the past years. A study investigating mobile device use of patient portals in oncology found that the log-ins by mobile device users averaged three times the log-ins of non-mobile users in 2017 [7].

Especially in oncology, results have shown an increase in patient portal use in recent years [6, 24], and some surveys show that the patient portal enrollment has even doubled [12]. Cancer patients often have longer treatment periods, including several appointments, side effects from treatment, and different medications, as well as communication and financial challenges. These longer treatment periods could explain the need for a tool such as a patient portal to manage cancer care-related work [6, 25, 26]. Cancer patients are more likely to use patient portals compared to patients free from cancer [27]. The use of patient portals in cancer care helps patients communicate with their care team and coordinate care. Cancer patients also feel better control of the disease and are more informed when accessing a patient portal [6].

Kaiku Health is a platform for digital health interventions in cancer care by Elekta Kaiku and can be used both for symptoms and quality of life reporting as well as a patient portal. Used only as a patient portal, Kaiku Health provides a messages function where patients can communicate with their healthcare professionals. In addition, Kaiku Health, as a patient portal, also provides appointment schedules and viewing and downloading patient records. Kaiku Health can be flexibly accessed on a mobile app or via a web browser, either on a phone or a computer for patients, only requiring internet access. For healthcare professionals, Kaiku Health is only accessible from a web browser.

2.2.1 Patient portal users and user patterns

Mutual characteristics for all digital patient portal users are access to computer or mobile devices and the internet. In addition, computer literacy and health literacy are needed knowledge for using and benefiting from digital patient portal usage [4]. Studies have shown that patient portal users tend to be older and more likely female [4, 11, 10]. However, more specified characteristics of the patient portal users are hard to determine since little is known about actual users. Still, the patient portal users are a widely heterogeneous group [5].

Cancer patient portal users tend to be more frequent portal users when compared to non-cancer patient portal users [6]. The frequency of portal usage is also linked to metastatic disease [6]. Furthermore, cancer patient portal users tend to use the messaging function and view medical records in the patient portal [25].

Patient portal users are also a heterogeneous group regarding user patterns. On a session level, patients tend to use similar features and functions when visiting the patient portal, hence repeating similar user patterns in several sessions [28]. Still, when looking at user patterns on a patient level, variability between individual patients' user patterns can be observed [28]. These results show that patients tend to form their own specific user patterns and repeat them over time.

Nevertheless, to be able to follow patient user patterns in the first place, patients need to enroll in the patient portal. Intended patient users of the Kaiku Health patient portal get an invitation to enroll in the portal. The clinic the patient is visiting provides the patient portal invitation.

2.3 Log data in patient portals

Transactional log data are records of behavioral data describing an interaction between a user of a system and that system. A logged interaction can include a mouse click, a browsing pattern, or opening or using some content in the system. As the actual use of a function generates log data, it represents actual usage behavior and not the user's recalled behavior [9, 14]. Log data is also generated without active participant involvement and hence called passive data [29]. As an analysis method, log data analysis provides a way to analyze the system's performance and acceptance. Log data analysis offers information on how the users use the technology and whether the system fulfills its purpose and fits the user. Therefore, log data analysis supports the development of not only providing information but also supporting the user [14]. Log data is used for studying behavioral and temporal patterns in communication [30], calls [31], emails [32], mobility [33] and sleep patterns [34]. In addition, log data analysis has provided great methods for research on health informatics [9, 14].

Within patient portals, log data analysis is used mainly for conducting descriptive statistics, but also for examining types and patterns of portal users [5, 28, 14]. Examination of patterns of portal users through log data analysis can reveal how patients use the patient portal [28] and provide a broader insight than descriptive statistics. A large amount of use does not have to imply better health outcomes. It can instead be due to unfocused and nonstrategic use, which may not indicate a good result of the patient portal use [9]. Hence, log data analysis in patient portals can provide a deeper insight into how patient portals work and give real-time insight into how different users react in similar situations. This knowledge can then be used for developing the portal and its long-term use [14].

However, using log data analysis for patient portals has limitations and requirements. Log data analysis always depends on the technical specifications of the portal or the system, and which functions and parts of the system are being tracked [28]. Hence, the log data analysis cannot cover actions not tracked. Log data neither tell whether users pay attention to the content shown in the portal [9]. Therefore, the analysis does not always imply why a specific user pattern occurs in the data [14]. Careful prepossessing of the log data is also necessary, as log data can contain artifacts [28]. Another requirement for log data analysis on patient portal log data is a carefully considered session definition, to obtain truthful results [14]. Defining a session as the time between sign in and sign out does not necessarily generate a complete session as users can take a break in using the portal and then continue later on where they left off to finish [14].

2.4 Clustering

Clustering is a classification technique to discover and form groups of unlabeled data points in a dataset. The classification is done without having a known prior knowledge of the resulting groups [35]. Therefore, clustering is also called unsupervised classification [35] and belongs to unsupervised machine learning techniques [36]. Clustering is often used for exploratory data analysis and data summarisation tasks of large multivariate datasets [37]. The goal of clustering is to gain new information about the data, learn about similarities and differences between the clusters, and derive conclusions about them [38].

Clustering aims to divide the data into groups or clusters that maximize the distances between the data points in different clusters while minimizing the distances between data points within a cluster [37]. In addition, the resulting clusters should contain all data points in the dataset [35]. Therefore, characteristics commonly preferred in clustering results are compact and isolated clusters [37].

Multivariate datasets consist of several individual data points described by numerous variables. These datasets can consist of either categorical variables, in which each data point takes a value belonging to one of a defined set of categories, or continuous variables, in which the data points can take any real value. Examples of categorical variables are sex and language, and examples of continuous variables are age and distance. A mixed dataset consists of categorical and continuous variables [37].

However, there are some challenges with clustering. One risk is that the clustering results are a consequence of the chosen clustering method. Hence, the clustering result depends only on the used clustering algorithm. This dependence means that the clustering algorithm forces a structure on the data that naturally does not occur in the data [35]. Furthermore, clustering mixed data comes with difficulties, as most mathematical algorithms do not suit categorical variables [36]. A solution is to divide the dataset into two and use different metrics for the categorical and the continuous variables. For instance, Hamming distance for the categorical variables and Euclidean distance for the continuous variables [36]. Still, also this dataset-dividing-by-variable-type approach for mixed data has its limitations. One problem is that it remains unknown whether both chosen distance metrics measure a comparable type of similarity. In addition, the comparability of the distance metrics scales remains unclear. Consequently, the proportions of the two distance metrics in a dataset-dividing-by-variable-type approach are ambiguous [36].

2.4.1 Clustering process

The clustering process consists of four steps [39], after defining the problem statement. Ordinary tasks to use clustering for are finding possible groups and determining the characteristics of the groups. However, the clustering process is not a one-shot process in general and often includes a feedback pathway [39], which is seen in Figure 1. The process usually contains a series of trials and repeated iterations [39].

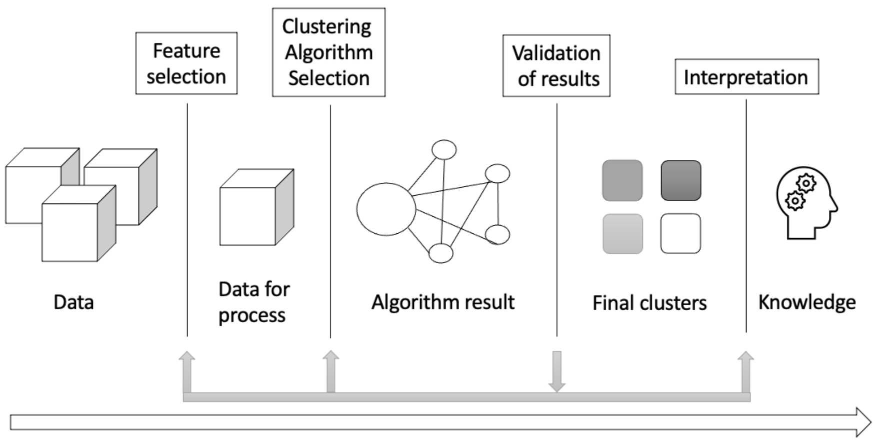
The first step is called *feature selection* or *extraction step*. In this step, either features are chosen directly from the provided feature candidates in the dataset, or appropriate and novel features are extracted and generated from the dataset [39]. The first step is crucial for the whole clustering process as one should select or extract the variables containing the most information regarding the defined problem statement [38, 40]. In addition, the chosen features will affect the effectiveness of the remaining process [39]. Therefore, data preprocessing is often needed before feature selection or extraction [38]. This step results in a dataset consisting of the chosen features.

The second step of the clustering process is *algorithm selection*. In this step, based on the dataset and problem statement, a suitable clustering algorithm is selected [38]. In addition, one chooses a proximity measure and a clustering criterion. A proximity measure is a defined measurement of how similar two data points in the dataset are according to their features [40]. A clustering criterion defines the rule in the clustering algorithm process, e.g., a linkage function describing how the distances between clusters are measured [38]. The result of this step is the suggested clusters after running the chosen algorithm on the features dataset.

The third step consists of the *validation* of the suggested clusters. In the validation step, the correctness of the clusters, produced by the algorithm chosen in the previous step, is validated [38]. The validation is done using applicable techniques and criteria. The goal of the validation step is to select the best clustering suggestion, whether the algorithm selection step uses several different clustering algorithms. The choice of the best clustering suggestion is done based on the validation results from distinct cluster algorithms [39]. The result of this step is the chosen set of final clusters.

The final step of the clustering process is the *interpretation* step. The final resulting clusters from the validation step are interpreted and evaluated concerning the problem statement and expert knowledge from the data field [38, 39, 40]. The goal of the interpretation step is to gain new knowledge and insights about the data. Therefore, other experimental evidence and analyses can be used to draw reliable conclusions [38, 39].

Figure 1: Steps of the clustering process, adapted from [38]



3 Research material and methods

3.1 Data description

The dataset used for this thesis consists of completely anonymized data from 444 patients using Kaiku Health at five different clinics in North America. The patients use Kaiku Health as a digital patient portal. The sex and age distribution can be seen in Table 1, and the device distribution is seen in Table 2.

Table 1: Sex and age distribution of the patients

	<20	20-39	40-65	65>	Total
Female	1	41	134	91	267
Male	1	18	65	93	177
Total	2	59	199	184	

All patients in the data set are in the active phase of their treatment and use English as the portal language. The data set is collected during a three months period, starting from the first of June and extending to the end of August 2023.

The dataset consists of three files. One file contains metadata such as user characteristics for each user. Another file contains login times, login dates, and the used device type at the login time for each user. The third file contains tracked log data from used features during the period for each user, but no clinical information is included. Furthermore, all available features and functions in the patient portal were consistent during the whole period. The dataset consists of continuous variables, such as the age of the patients, and categorical values, such as the used device type.

Table 2: Device distribution of the patients

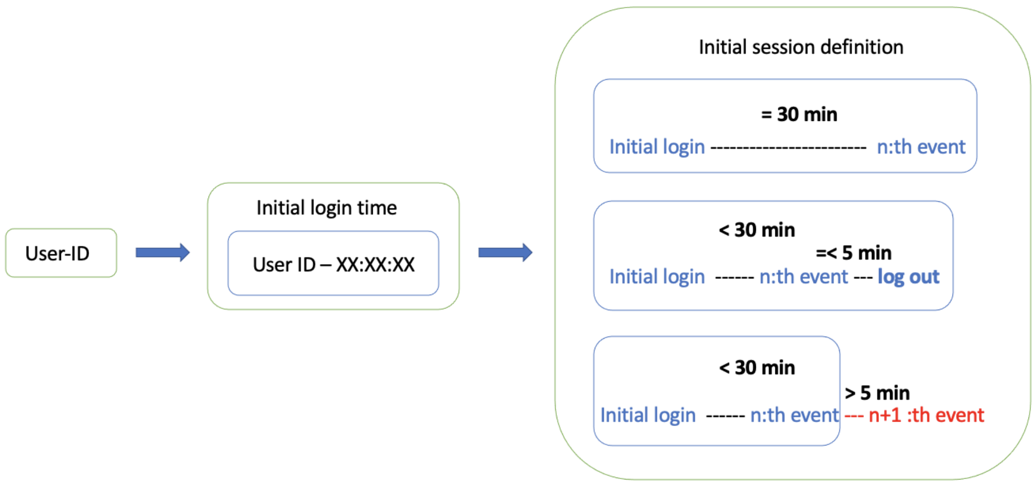
Device	Number of users
Desktop	160
Mobile app	108
Mobile browser	176

3.2 Data preprocessing

Defining the initial session for each user in the tracked log data is crucial for the following feature selection. Counting every log-in of a user as a unique session will not necessarily provide actual usage information. A user might leave the platform for some minutes to continue with platform activities later where they left off [14]. In addition, a fixed session measurement seldom performs well, as numerous different time-related levels can define a session [9]. A user session can, for instance, be ended by log-out or inactivity for a limited period [5]. Hence, a dynamic definition of a user session is needed.

The initial sessions in this work are defined in a process using a unique user ID, which is present in all dataset files. First, one specifies the time for the initial log-in for each user. The initial session is then defined from the tracked log data, using the user ID and the time for the initial log-in. An initial session is specified to begin at the initial log-in time and contains all tracked events in the log data until it reaches one of the session-ending definitions. The initial session ends immediately when 30 minutes from the initial log-in have passed. However, if the user logs out within 5 minutes since the previous tracked event and before 30 minutes have passed, the log-out event is set to be the last event of the session. In addition, if a time gap between the time stamps of any two following tracked events is greater than 5 minutes, the session is set to end after the first of the two events. In this case, the session ends, although it does not pass the absolute limit of 30 minutes from the initial log-in. For each user ID, the process of the initial session definition is visualized in Figure 2. The definitions of the session endings are established utilizing experts of the Elekta Kaiku employees, who are familiar with the log data of the product, and by exploring the dataset. The pre-processed dataset consists of the user ID and the defined initial session for each user, which is 444 initial sessions in total.

Figure 2: The process of defining an initial session



3.3 Feature selection

Feature selection can be done in two ways [41]. One of the feature selection approaches is automatic feature selection. In automatic feature selection, a search algorithm finds a set of features that optimally define the data from all available features in the dataset. Assumptions when using automatic feature selection are that the chosen features are uncorrelated with each other but highly correlated with the classes or clusters, which will be derived later in the process [41]. The other feature selection method approach is knowledge-based expert judgment. In feature selection using expert knowledge, a person's gained knowledge of the topic and clustering task domain and intuition are used to determine which features are relevant for the clustering task. Hence, the uncorrelatedness of the data might have less impact on feature selection using expert knowledge [41]. Still, both approaches have their unique biases, and both approaches might fail to select important features. This failure can happen either because of the importance of a feature not being recognized in previous knowledge or not being mathematically favorable. This study uses the expert judgment of Kaiku Health's employees, who are familiar with the product and the generated data for performing the feature selection.

The selected features are chosen and defined based on the tracked log data events of the initial sessions. In addition, general information about the initial sessions is utilized. The selected features can be seen in Table 3. The initial sign-in is not one of the selected features, as the initial sign-in has to take place for the initial session to occur, hence providing no significant information in the research. Although the tracked log data events in each initial session are a time series, the ordering of the tracked events and the repeated occurrence of a specific event in an initial session are ignored. Instead, this cross-sectional study focuses on which unique events occur in the initial sessions and whether a specific event is tracked in an initial session or not. Consequently, binary variables denote these features for each initial session. The original tracked events in the log data of an initial session are grouped into fewer selected features to avoid correlated features, as clustering models do not handle correlated data well. Therefore, each feature should contain unique information.

Table 3: Features for clustering

	Number of features	Features
General session information	3	Time of the initial sign in, Device, Days between the invite and the registration
Duration	1	Duration of the initial session in seconds
Comprehensiveness of use	2	Number of tracked log events, Number of unique features (tracked in the tracked log events)
Sign in & out	3	Additional sign in, Sign out manually, Session expired
Home page	1	Home page
Notifications	1	Receive notification
Settings	3	Changing of password, Profile editing, Notification enabling
Documents and results	3	View results, View documents, Download documents
Messaging	2	Read messages, Send message
Total number of features	19	

The total number of unique features defined in each initial session is noted in the *number of unique features*. In addition, *number of events* reports the total quantity of tracked log data events in each initial session. For each initial session, the duration of the session is calculated in seconds. These session-describing features are numerical and continuous variables. Outlier detection is performed using the number of events and session length features. Removed are initial sessions with either a duration longer than the 99 percentile of the sessions' session lengths or initial sessions containing more events than the 99 percentile of the initial sessions' number of events count.

Three features describe the general information about the initial sessions. The time of the initial session is converted into a numerical variable, with the hours of the day as possible values. Although the data is collected in different timezones, the logging of the timestamps is in the user's local timezone. Therefore, the sign-in time of each initial session corresponds to the time of day when the initial session has occurred. The device is a categorical variable with three categories: mobile app, mobile browser, and desktop. However, for some clustering algorithms in this study, the device feature is one-hot encoded into three binary values. The *days between invite and registration* feature notes the number of days between sending the portal invitation to a possible future patient user and the occurrence of the initial session. No user characteristics are present among the selected features, as this study focuses on the tracked events of the initial session. Instead, age and sex are added to the clustering results using the user ID before interpreting the results.

The preprocessing and feature selection are done using Python and Python libraries *pandas* [42] and *numpy* [43].

3.4 Clustering methods

All clustering problems' underlying assumption is that some grouping exists in the data. This underlying structure in the data is revealed by examining the similarity in the features of different data points. Furthermore, another assumption is that these discovered clusters have a meaning in the real world. In this cluster analysis, the assumption is that the dataset will consist of different patient user groups according to their user pattern in the initial session at the patient portal.

The pre-processed dataset used for clustering in this clustering analysis consists of categorical and continuous variables. Hence, the chosen clustering methods have to be suitable for mixed data. In addition, some of the selected features in the pre-processed might be more related to each other, as one can expect the features *read messages* and *send message*. Thus the chosen method needs to be suitable also for correlated data. The data context might also be crucial for interpreting the clusters. Therefore, four approaches based on two different clustering methods are chosen for the clustering task. The first and second approaches include a context-preserving clustering method, hierarchical clustering with Gower's distance, and hierarchical clustering with cosine distance. The third and fourth approaches include a non-context-preserving clustering method, factor analysis of mixed data combined with K-means clustering, and uniform manifold approximation and projection combined with K-means clustering.

3.4.1 Hierarchical clustering approaches

Hierarchical clustering aims to build a binary merge tree based on the distances or similarities between the data points. In the binary merge tree, every leaf represents a data point while the root of the tree represents the whole dataset [44]. The two main techniques of hierarchical clustering are *agglomerative hierarchical clustering* and *divisive hierarchical clustering* [45]. The idea in agglomerative hierarchical clustering is to start from the leaves in the binary merge three and merge two by two of the most similar or closest leaves into a cluster. The distances or similarities between the newly formed clusters are calculated, and the two most similar or closest clusters are merged. This process repeats until the root of the binary merge tree is reached. In divisive hierarchical clustering, the process goes the other way around, starting from the root and splitting the cluster into smaller clusters and repeating this until the leaves are reached [45]. The resulting graphical representation of the binary merge tree from both methods is called a *dendrogram* [45]. This clustering analysis uses the agglomerative hierarchical clustering technique. Moreover, the hierarchical clustering in this clustering analysis is context-preserving as the pre-processed dataset is used without further dimension reduction.

A distance metric has to be defined for measuring the distance or similarity between two data points according to their features [45]. If the data consists of continuous variables, an often-used distance metric is Euclidean distance after normalization of the data. However, since the dataset in this clustering analysis consists of mixed data, the chosen distance metric needs to be suitable for mixed data. Therefore, the following two different similarity measurements are chosen:

Gower's distance is a distance metric [46, 47]:

$$d_G(x_i, x_j) = \frac{\sum_{l=1}^{m} w_l f_l(x_{il}, x_{jl})}{\sum_{l=1}^{m} w_l} \tag{1}$$

In equation 1, w_l consists of weights for the variables but is set to $w_l = 1$ if no more specific information is available [47]. Furthermore, $f_l(x_{il}, x_{jl})$ is defined differently depending on if the variables x_{il} and x_{jl} are quantitative or qualitative. If x_{il} and x_{jl} are quantitative, then $f_l(x_{il}, x_{jl})$ is defined as $f_l(x_{il}, x_{jl}) = 1 - \frac{|x_{il} - x_{jl}|}{r_l}$, where r_l is the sample range of l. If x_{il} and x_{jl} are qualitative, then $f_l(x_{il}, x_{jl})$ is defined as in equation 2, which means that $f_l(x_{il}, x_{jl}) = 1$ if x_{il} and x_{jl} belongs to the same category and has the same value and $f_l(x_{il}, x_{jl}) = 0$ if x_{il} and x_{jl} differ [46].

$$f_l(x_{il}, x_{jl}) = \begin{cases} 1 \text{ if } x_{il} = x_{jl} \\ 0 \text{ if } x_{il} \neq x_{jl} \end{cases} \tag{2}$$

The similarity matrix for the first method is created based on Gower's distance using the Python library *gowers* [48].

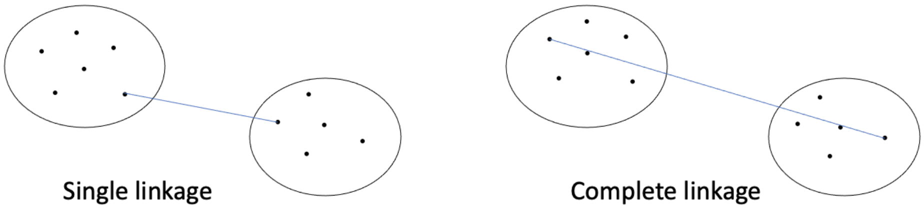
Cosine distance is a similarity metric [45]:

$$d_C(x_i, x_j) = 1 - \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|} = 1 - \frac{\sum_{l=1}^{m} x_{il} x_{jl}}{\sqrt{\sum_{l=1}^{m} x_{il}^2} * \sqrt{\sum_{l=1}^{m} x_{jl}^2}} \tag{3}$$

and measures the cosine angle between two vectors [44]. In equation 3, the dot product between the data points x_i and x_j is denoted by $x_i \cdot x_j$ and the Euclidean norm by $\| \cdot \|$. The similarity matrix based on cosine distance in the second clustering approach is created using Python library *scipy* [49].

A linkage function is needed when the clusters consist of two or more data points [44]. As long as the clusters consist of one data point each, the distance between any two clusters can be measured by the chosen distance metric $d(x_i, x_j)$. However, when the two clusters consist of several data points, the linkage function defines how the distance between the two clusters shall be measured. Commonly used linkage functions are single linkage and complete linkage [44]. Single linkage measures the minimum distance between two points in two different clusters, and complete linkage measures the maximum distance between two different clusters. The difference between single linkage and complete linkage is illustrated in Figure 3. The chosen linkage function influences the final clustering result.

Figure 3: Single and complete linkage, adapted from [44]



The agglomerative hierarchical clustering process is summarized in an algorithm consisting of three steps [44]:

1. Initialize all data points so that each data point forms a cluster.
2. Choose two clusters minimizing the calculated linkage function distance according to the distance metric.
3. Merge the two chosen clusters into one new cluster and remove the two earlier chosen clusters.

Repeat steps 2 and 3 until only one cluster remains and the root of the binary merge three is reached. The resulting dendrogram is now to be created and used to analyze clusters of different heights of the binary tree. However, the order of the original clusters in step 1 of the algorithm does make a difference. One cluster can only be part of one pair in step 2, even though several clusters might be at the same distance as the nearest cluster. Hence, one can get another set of clusters if one swaps the order of the original clusters [44].

In both hierarchical cluster approaches, the created similarity matrix is used to repeatedly determine and merge the closest clusters, thus constructing a dendrogram. Complete linkage is used as the linkage function in both approaches to avoid the chaining phenomenon in the dendrogram when using a single linkage [44].

3.4.2 Dimension reduction and K-means clustering approaches

Dimension reduction is a commonly used technique for analyzing and transforming large multivariate datasets into lower dimensional subspaces [50]. Hence, it is often applied to visualize and pre-process multivariate datasets in machine learning [51]. Dimension reduction algorithms are divided into two categories: Algorithms that maintain pairwise distance and algorithms that preserve the local distances between data points over global distance [51]. Examples of pairwise distance-maintaining algorithms are principal component analysis (PCA) and multidimensional scaling (MDS). In contrast, examples of local distance-favoring algorithms are t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) [51].

The choice of dimension reduction algorithm is affected by the data type. PCA is a common choice when the dataset consists of continuous variables [52]. PCA intends to extract the important information from the data and represent it as a new set of orthogonal variables called principal components. The principal components are required to maintain the maximum variance from the original dataset. Multiple correspondence analysis (MCA) is a corresponding dimension reduction algorithm suited for categorical data and occasionally mixed data as well [53]. When using MCA for mixed data, the continuous variables are transformed into categorical variables by breaking down the values of each continuous variable into classes and treating each class as a separate variable. However, when the number of continuous variables is large compared to the number of categorical variables, or there are only a small number of data points in the dataset, it is better to do no transformation of the continuous variables [53]. Therefore, the following two dimension reduction algorithms are chosen:

Factor analysis of mixed data (FAMD) is a generalization of PCA and MCA suited for datasets consisting of both continuous and categorical variables [53]. In FAMD, the continuous variables are standardized and hence represented by vectors of length 1, as they contain one number each. K denotes the set of continuous variables. The categorical variables are divided so that there is a separate vector for each category of a categorical variable. Thus, a categorical variable q is represented by k vectors creating a subspace with k - 1 dimensions. Q denotes the set of categorical variables. Correspondingly to PCA, one wants to preserve the maximum variance from the original dataset when applying FAMD, and hence the criterion in equation 4 is maximized [53]:

$$\sum_{k \in K} r^2(k, v) \sum_{q \in Q} \eta^2(q, v) \tag{4}$$

In equation 4, $r^2(k, v)$ denotes the squared correlation coefficient between variables k and v, and $\eta^2(q, v)$ denotes the squared correlation ratio between variables q and v. In addition, the context of the dataset is unpreserved when performing dimensional reduction with FAMD. The dimension reduction using FAMD in the third method is implemented using the Python library *prince* [54] and the reduced dataset consists of two dimensions.

Uniform manifold approximation (UMAP) is a non-linear dimension reduction technique [51]. Based on manifold approximations, UMAP creates a topological representation of the high dimensional data in a low dimensional space. When using UMAP, the data is assumed to be uniformly distributed on a manifold, which is locally connected. Another assumption is that one wants to preserve this topological local structure when performing dimension reduction. Computationally, UMAP is a type of k-neighbor-based graph. Hence, the UMAP algorithm consists of constructing a weighted k-neighbour graph and finding a low dimensional representation of the constructed weighted k-neighbour graph [51].

As UMAP is a non-linear technique, the interpretability of the results of UMAP is poor compared to linear PCA-related techniques such as FAMD. Hence, the components of UMAP have no meaning [51]. In addition, as UMAP favors local structures, it is harder to analyze the difference between individuals compared to FAMD.

Dimensional reduction using UMAP is implemented using the Python library *umap-learn* [55]. Parameters such as the number of neighbors and minimal distance between points in the low-dimensional representation can be adjusted using this package. Furthermore, a distance metric needs to be chosen. As UMAP supports no mixed data or categorical data [51], the data is divided into numerical and binary together with categorical data. The categorical data is one-hot encoded into binary values, where each categorical variable category is transformed into a binary value. As a result, two different UMAP representations are created. One representation is the numerical variables using Euclidean distance as the distance metric. The other representation, using Hamming distance, is based on the binary variables, as well as the formerly categorical variables. The intersection between these two representations represents the final transformed data. The final transformed data consists of two dimensions.

K-means clustering [56] is used to perform the clustering in both dimension reduction including approaches. The well-known K-means clustering aims to cluster the n data points, where each data point is described by a vector, into k clusters denoted by C_I, minimizing the within-sum of squares of distances in the cluster. The k-means algorithm by Lloyd 1982 [56] consists of the following steps after defining an initial set of randomly chosen means $c_1,...,c_k$ of k clusters:

- Assign each data point x_i to the cluster C_I with the closest cluster mean or centroid c_I, which is the cluster mean c_I is nearest to x_i concerning the least squared Euclidean distance
- Update and recalculate the cluster means c_I to be in the new center of mass for the clusters C_I by

$$c_I = \frac{1}{|C_I|} \sum_{x_i \in C_I} x_i \tag{5}$$

The two steps in the k-means algorithm repeat until the means or centroids c_I of the clusters C_I no longer change. Although a data point could be assigned to several clusters based on the least squared Euclidean distance in the first step, one point can only belong to one cluster. Furthermore, the number of clusters k always has to be determined before initializing the algorithm.

However, it may not be ideal to use dimension reduction and k-means for finding clusters true to real-world data [57]. Dimension reduction methods do not preserve the data's context. The principal components derived from the PCA-based dimension reduction methods only define the dimensions with the highest variance. Hence, the dimension reduction method may favor variables less significant in the real world, as they contribute more to the transformed data set.

The k-means generated clusters in both approaches are created using the k-means clustering algorithm implemented in the Python library *scikit-learn* [58]. The randomness of the initially chosen cluster means is controlled by assigning a fixed initial random state to the algorithm implementation, thus guaranteeing the same result on each run. To make visualization feasible, the vectors describing the data points from both FAMD and UMAP consist of two dimensions each. These vectors can be used as a pair of coordinates projecting the data points onto a plane. The resulting k-means generated clusters in both approaches are illustrated in two-dimensional graphs.

3.5 Cluster validation and chosen number of clusters

There is no universal rule on determining the true clusters corresponding to a real-world grouping and understanding what they are when doing cluster analysis [59]. Whether the resulting clusters are meaningful in the real world depends on the context and the aim of the clustering. Still, validation methods can guide the evaluation of the clusters' goodness.

The main cluster validation categories are external and internal validation [60]. External validation uses external information not present in the data for cluster validation. As an example, this can be predefined true cluster labels. Thus, external validation is often used to choose the optimal clustering algorithm. On the contrary, internal validation relies only on information available in the data. Internal validation is used when the optimal number of clusters and the optimal cluster algorithm have to be determined. In most cases with large multivariate datasets, internal validation is the only option as no external information is available [60]. In this cluster analysis internal validation is used as there are no predefined true cluster labels for the data points.

The process of internal validation consists of determining the set of cluster algorithms and applying each algorithm to the data several times using different combinations of parameters [60]. Numerous clustering results are hence generated, and the various clustering results for each clustering algorithm are then evaluated using an internal validation index. Finally, the best resulting clustering algorithm and optimal clustering number are chosen based on the criteria. This process can be performed using one or multiple of several internal validation measurements. Nearly all internal validation measurements measure the compactness and separation of the resulting clusters. Compactness measures how close the data points in a cluster are, while separation measures how well the clusters are separated. The following two different internal validation measurements are used in this analysis:

The **Silhouette coefficient (SC)** validates the goodness of the clustering based on the pairwise difference of between-and within-cluster distances [60]. The Silhouette coefficient ranges from -1 to +1. The optimal number of clusters is determined by maximizing the Silhouette coefficient, seen in equations 6 and 7:

$$SC = \max_k \frac{1}{n} \sum_i s(x_i) \tag{6}$$

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max[a(x_i), b(x_i)]} \tag{7}$$

where $a(x_i)$ describes the distance between x_i and all the other data points in the same cluster C_I while $b(x_i)$ describes the distance between x_i and all the data points in the another cluster C_J. $a(x_i)$ and $b(x_i)$ are defined in equations 8 and 9:

$$a(x_i) = \frac{1}{n_I - 1} \sum_{x_j \in C_I, x_j \neq x_i} d(x_i, x_j) \tag{8}$$

$$b(x_i) = \min_{J, J \neq I} \frac{1}{n_J} \sum_{x_j \in C_J} d(x_i, x_j) \tag{9}$$

In equations 6, 7, 8, 9, 10 and 11 k is the number of clusters, n is the total number of data points in all clusters, n_I is the number of data points in cluster C_I. c_I denotes the centroid of cluster C_I and c the the overall centroid of the data. Furthermore, $d(x_i, x_j)$ is the distance between x_i and x_j calculated using the distance metric for the clustering.

The **Calinski-Harabasz index (CH)** is a cluster validation index based on the average between-cluster and within-cluster sum of squares [60]. There are no boundaries for the Calinski-Harabasz index. Instead, a higher value indicates more dense and separated clusters. Hence, one wants to find the highest value among the different clustering options. The Calinski-Harabasz index is defined in equation 10:

$$CH = \frac{\sum_{I=1}^{k} n_I d^2(c_I, c)/(k-1)}{\sum_{I=1}^{k} \sum_{x_i \in C_I} d^2(x_i, c_I)/(n-k)} \tag{10}$$

In the hierarchical clustering-based approaches with Gower's distance and Cosine distance, the computed distance metric for the clustering is used in the validation measurements. In the dimension reduction and k-means clustering-based approaches, the validation measurements use Euclidean distance to calculate the distance between the data points in the lower dimensional space. Both the Silhouette coefficient in equation 6 and the Calinski-Harabasz index in equation 10 are implemented using the *scikit-learn* [58] implementations of each index.

However, both chosen internal validation measurements have their weaknesses. The Silhouette coefficient is impacted by subcluster [60], thus tending to group neighboring clusters into one big cluster. The Calinski-Harabasz index is easily affected by noise in data, which makes the Calinski-Harabasz index unstable [60]. Therefore, a compound [61] is created and used for determining the goodness of the clusters and the optimal number of clusters. Since the data is expected to be fairly noisy, the Silhouette coefficient has a greater impact on the compound. To count for the instability of the Calinski-Harabasz index, the compound checks if the Calinski-Harabasz index has reached a peak by subtracting the previous and the succeeding values from the current value. Moreover, to compare compound values from different clustering algorithms, the Calinski-Harabasz index is normalized using min-max scaling, thus ranging from 0 to 1. Finally, the compound is divided by three and ranges from -0.66 to 1, making the interpretation easier. The compound is defined in equation 11:

$$Compound_k = \frac{1}{3}(2SI_k + \frac{\frac{1}{2}(CH_k - CH_{k-1} + CH_k - CH_{k+1}) - \min(CH)}{(\max(CH) - \min(CH))}) \tag{11}$$

The optimal number of clusters for all four approaches is chosen based on the value of the compound and context knowledge. Since the research goal is to identify user groups, context knowledge determines that it is inconvenient to select a k, or a number of clusters for the hierarchical clustering, that approaches the total number of data points in the dataset. Similarly, choosing the number of clusters to be three or fewer will not consider differences between the data points sufficiently, as several significantly different data points will be included in the same cluster. Further cluster evaluation is done using expert knowledge, as no true labels of the clusters exist.

3.6 Cluster interpretation

The cluster interpretation goal is to gain new knowledge about the data. In this cluster analysis, this is done by confirming that the generated clusters are user groups and identifying the characteristics of these user groups. Approaches used to achieve this goal are statistical techniques applied to each cluster. These are using the cluster means or centroids as a representative for the cluster, visual inspection of the cluster features, and machine learning techniques as decision trees [62]. Cluster interpretation through decision trees belongs to *post-modeling*, as it aims to explain and interpret the decisions made by a black box model after the clusters are generated [63]. Hence, decision trees used for clustering results interpretation are also a type of supervised modeling, as the generated clusters are used as true labels for the decision tree classification task.

Statistical measurements such as mode, median, mean, and standard deviations portray the cluster centroids for interpreting the cluster results in this cluster analysis. In addition, visualizations of the cluster features through scatterplots, bar plots, boxplots, and histograms investigate the clusters further. Cluster feature visualization is also used to grasp the spread of data points within a cluster.

A decision tree can be described as a node-link diagram, where each node describes a set of data points [64]. All other nodes are either internal nodes or leaf nodes. An internal node represents a choice that has to be made at the node, and the links from the internal node represent the alternatives the decision node consists of. Leaf nodes, also called end nodes, represent the final combination of results [64]. When the decision tree builds on an existing clustering result, every leaf node represents one of the final clusters. The links between the nodes represent the variable-based rules used for the cluster creation. The links coming from the uppermost node, also called the root node, represent the feature used for the initial split of the data into two sets. Since a decision tree is created based on a specific dataset, it has a significant risk of over-fitting. Therefore, a decision tree cannot be used for interpreting other datasets.

The decision tree in this cluster analysis is generated through post-pruning [64]. Hence, a more complex tree consisting of more nodes than needed is created first. Then, removing some nodes creates a smaller tree that explains the data nearly as well as the original tree. A common method to execute pruning is cost-complexity pruning. In cost-complexity pruning, the pruning of the decision tree is an iterative process generating a set of possible pruned sub-trees [65]. The error for removing a specific sub-tree from the original decision tree is defined for each sub-tree. The sub-trees generating the lowest total error of the classification are pruned. The implementation of the decision tree is done using *scikit-learn* [58] implementation *DecisionTreeClassifier*, which uses minimal cost-complexity pruning.

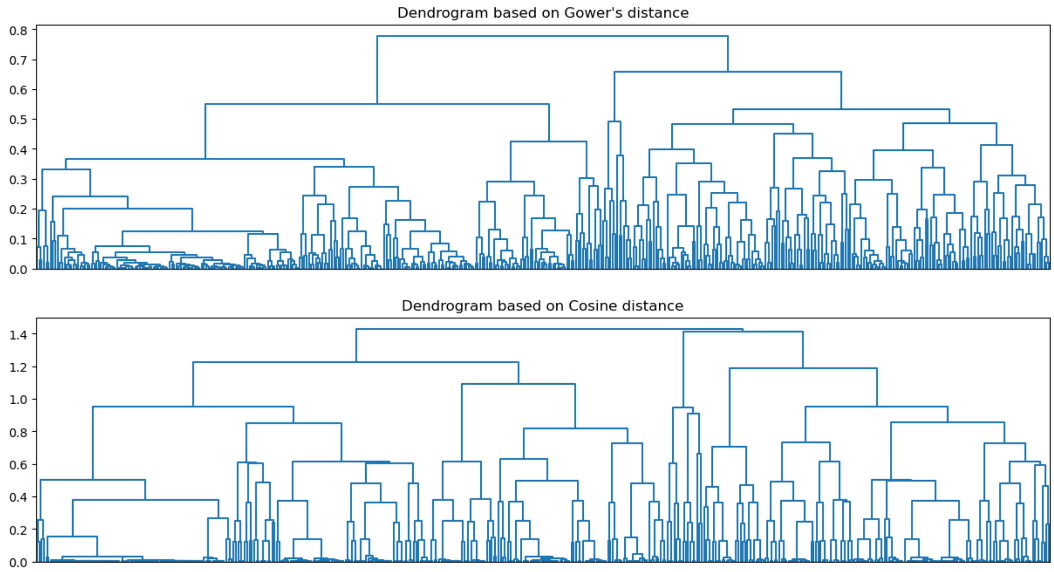
The decision tree interpretation is used to estimate the feature importance and understand which features have a significant role in forming the resulting clusters. However, domain knowledge is also needed to interpret the clustering results [38] as the decision tree can be very complex and contain excess data.

4 Results

4.1 Results from the hierarchical clustering approaches

A dendrogram with 444 leaves, each representing one data point or individual, is created in both hierarchical clustering approaches. The dendrograms can be seen in Figure 4. Complete linkage is used in both approaches for merging the leaves toward the top of the dendrograms. Furthermore, the *scipy* [49] function *fcluster* is used to flatten the respective dendrogram. A fixed number n of clusters is defined so that neighboring points in the respective original distance matrix are in the same cluster.

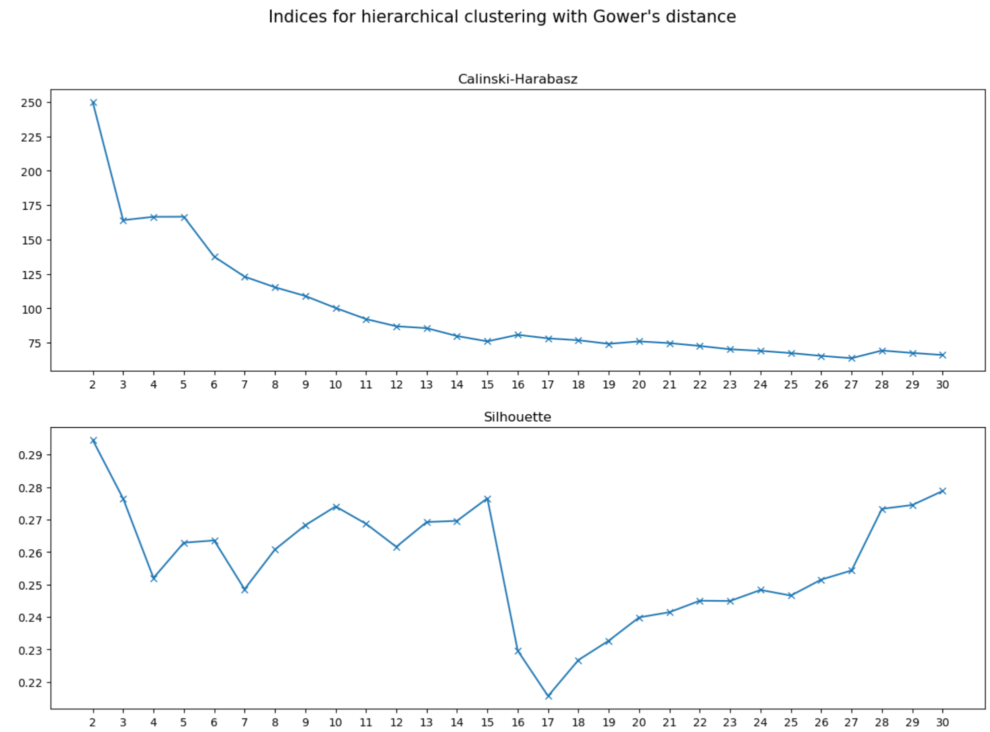
Figure 4: Dendrograms based on Gower's distance and Cosine distance



4.1.1 Gower's distance-based clusters

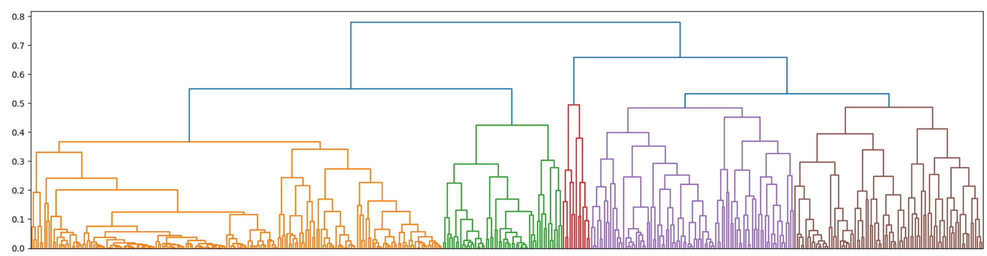
In Figure 5, visualizations of the Calinski-Harabasz index and Silhouette coefficient for 2 to 30 clusters using hierarchical clustering based on Gower's distance can be seen. Based on these results and the compound results in equation 11, the chosen number of clusters for the hierarchical clustering based on Gower's distance is n=5.

Figure 5: Calvinski-Harabasz index and Silhouette coefficient for hierarchical clustering with Gower's distance



The visualization of the chosen 5 clusters on the Gower's distance-based dendrogram is seen in Figure 6. The red-colored cluster seems smaller and contains significantly fewer data points than the other five. On the contrary, the yellow cluster seems to include several data points. The other three clusters seem to be roughly of the same size.

Figure 6: The resulting dendrogram for hierarchical clustering using Gower's distance colored according to n=5



4.1.2 Cosine distance-based clusters

The visualizations of the Calinski-Harabasz index and Silhouette coefficient for 2 to 30 clusters using hierarchical clustering based on cosine distance can be seen in Figure 7. Based on the compound in equation 11 and interpretation of the indices in Figure 7, the chosen number of clusters is n=5. For n=5, both the Calinski-Harabasz index and Silhouette coefficient generate good results respectively, and the chosen number of clusters is still large enough to take the differences between the data points into consideration.

In Figure 8, the chosen 5 clusters colored on the cosine-distance-based dendrogram can be seen. The yellow and brown clusters are significantly larger than the other clusters, and the red and lilac clusters contain only a few data points each.

Figure 7: Calvinski-Harabasz index and Silhouette coefficient for hierarchical clustering with cosine distance

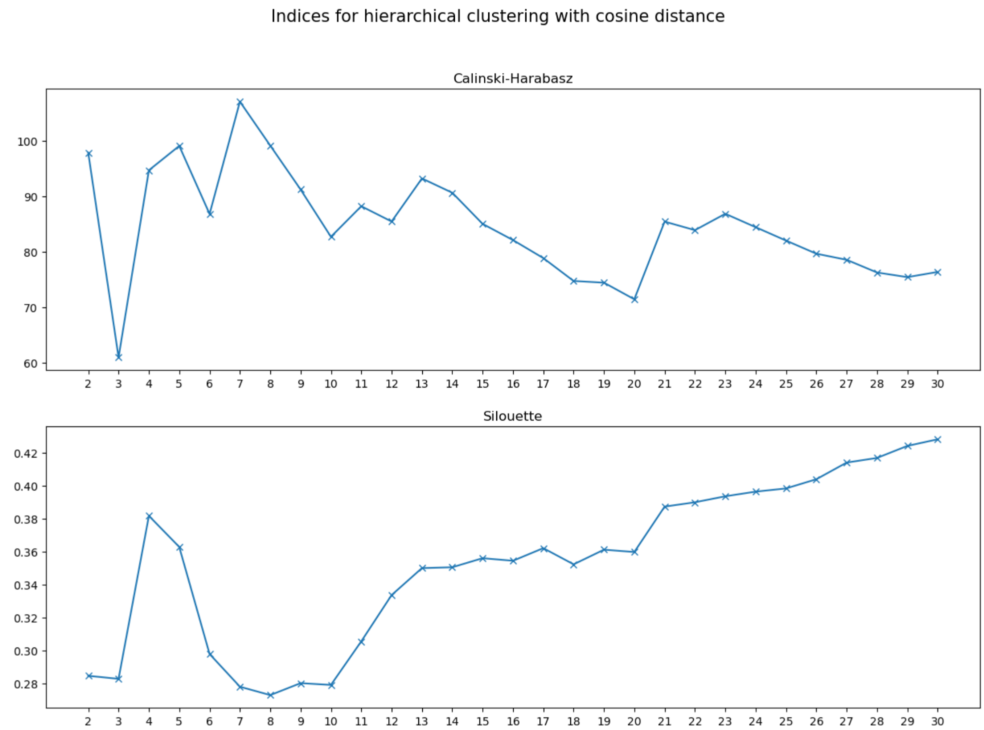
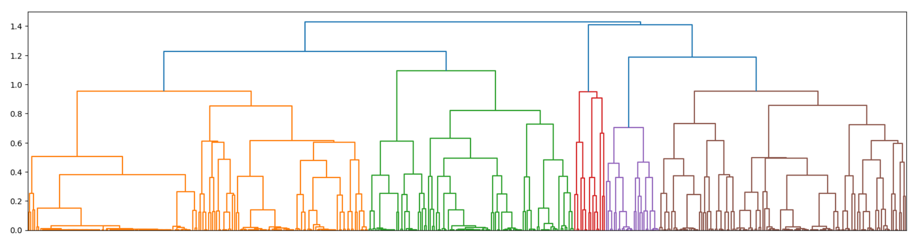


Figure 8: The resulting dendrogram for hierarchical clustering using cosine distance colored according to n=5



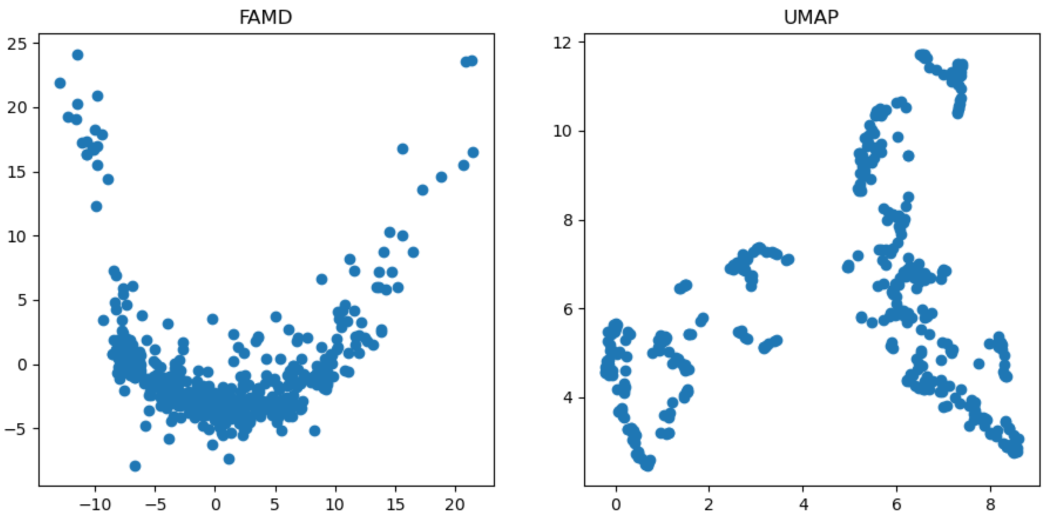
4.2 Results from the dimension reduction and K-means clustering approaches

FAMD and UMAP are used to perform dimension reduction on the pre-processed data set, after which the data points are visualized in two dimensions. Hence, the FAMD and UMAP visualizations are different representations of the same data. In the resulting visualizations of the dataset, each point denotes one individual.

The results of the respective dimension reduction of the data are shown in Figure 9. In the FAMD visualization, most data points are part of a parabolic shape in the lower part of the graph. In addition, several points are scattered around the parabolic, showing the noisyness of the data. Some points are located in the two upper corners, showing deviant points in the data set.

The UMAP visualization shows numerous smaller groups of data points spread in a triangle-shaped figure in the graph. This visualization indicates that there are several features describing the data points. While two points can be similar regarding one feature, they might be different and far from each other regarding some other feature. Hence, some groups are closer to each other than others.

Figure 9: Dimension reduction projections on two axes based on FAMD and UMAP



4.2.1 FAMD based clusters

Figure 10 shows the Calinski-Harabasz index and Silhouette coefficient for 2 to 30 clusters using k-means clustering based on FAMD dimension reduction of the preprocessed data set. Using the compound in equation 11 and visual inspection of the Figure 10, the chosen number of clusters based on FAMD dimension reduction is k=5. The Silhouette coefficient in Figure 10 reaches its highest values around k=5. Although the Calinski-Harabasz index takes its highest values at k bigger than 7, the index is easily affected by the noise in the data and therefore not as trustworthy when searching for the true number of clusters [60].

Figure 10: Calvinski-Harabasz index and Silhouette coefficient for k-means clustering based on FAMD dimension reduction

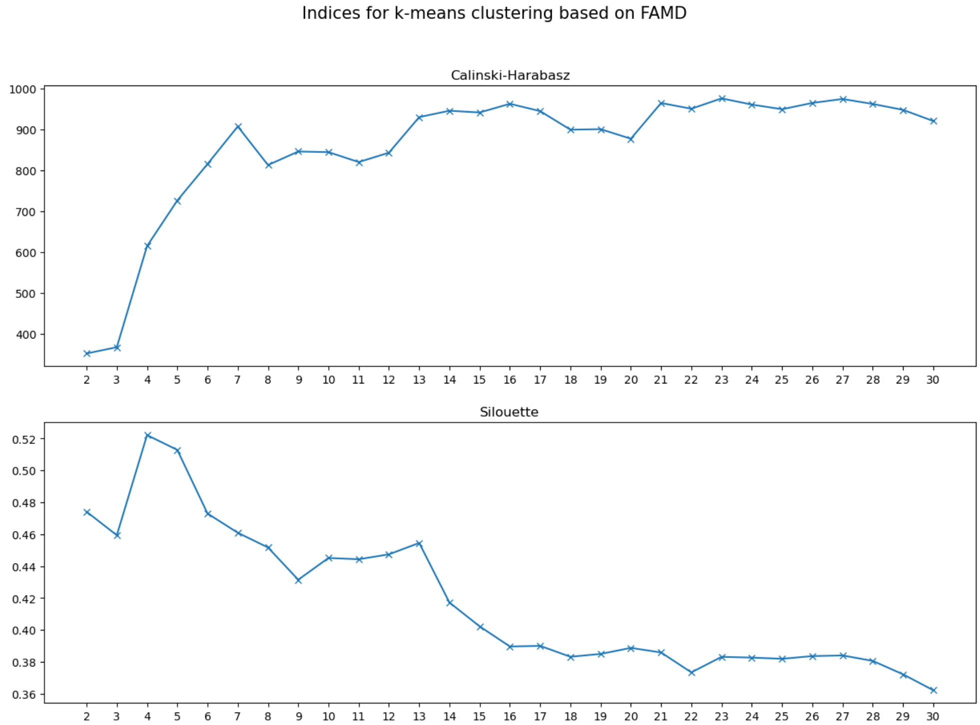
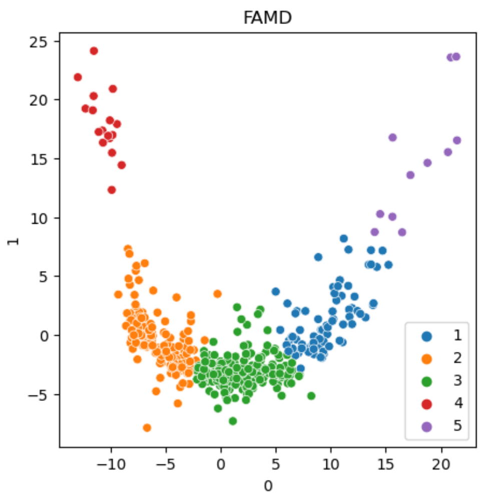


Figure 11 shows the visualization of the chosen 5 clusters on the k-means clustering based on FAMD dimension reduction. The red and the lilac clusters are significantly smaller than the other clusters, and the lilac cluster seems to be the least compact cluster. In addition, the red cluster is the best-separated cluster from the other clusters, while the orange and green clusters appear to be the most compact.

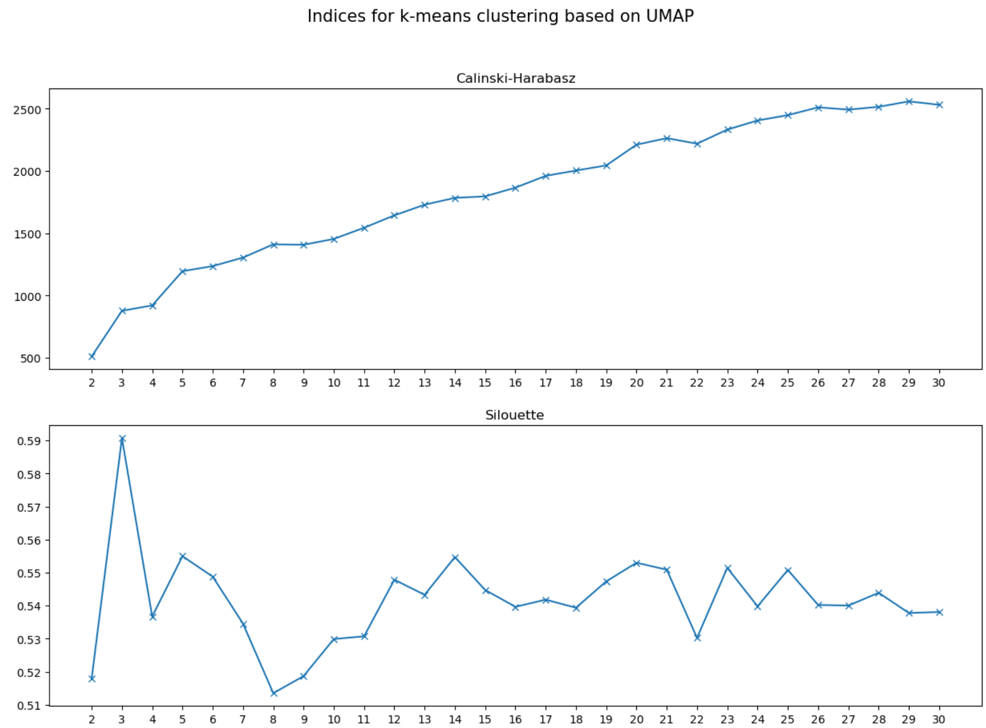
Figure 11: The resulting cluster division for k-means clustering based on FAMD dimension reduction colored according to k=5



4.2.2 UMAP based clusters

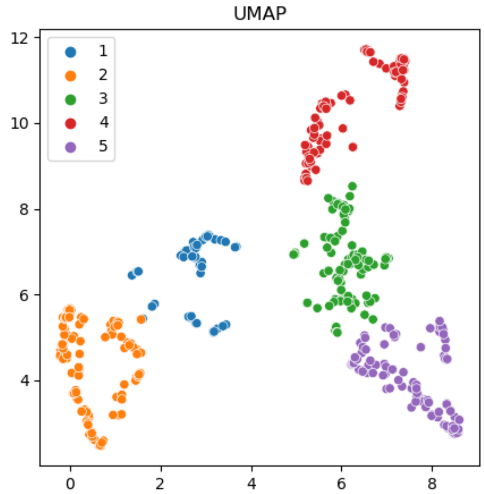
The Calinski-Harabasz index and the Silhouette coefficient for 2 to 30 clusters for k-means clustering based on UMAP dimension reduction of the preprocessed data set are shown in Figure 12. According to the Silhouette coefficient, three clusters appear to be the optimal choice. However, because differences between the data points are not sufficiently taken into account in a division into three clusters and the Calinski-Harabasz index reaches larger values for a greater number of clusters, the chosen number of clusters is k=5. This decision is based on the compound in equation 11 and the visual inspection of the Calinski-Harabasz index and the Silhouette coefficient graphs for k-means clustering based on the UMAP dimension.

Figure 12: Calvinski-Harabasz index and Silhouette coefficient for k-means clustering based on UMAP dimension reduction



The chosen 5 clusters on the k-means clustering based on UMAP dimension reduction are visualized in Figure 13. All five clusters consist of several subclusters of different sizes. The blue cluster is the most sparse and, in addition, the smallest. The other clusters appear to be about the same size. The orange and the blue clusters are considerably separated from the other three clusters. None of the clusters are compact.

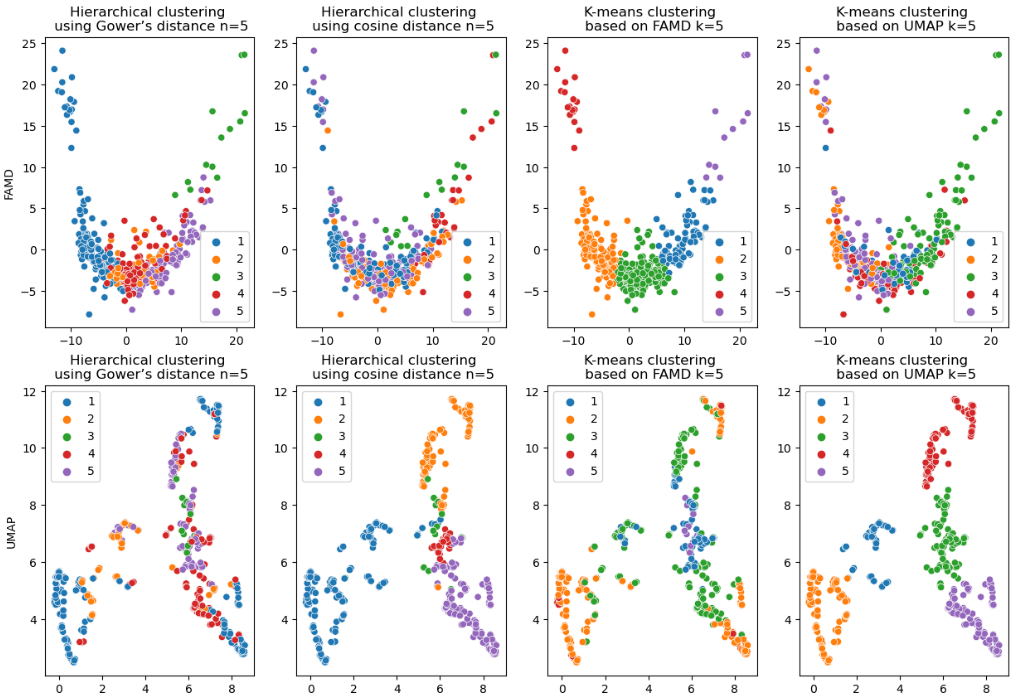
Figure 13: The resulting cluster division for k-means clustering based on UMAP dimension reduction colored according to k=5



4.3 Visual comparison of the cluster results

In Figure 14, the chosen clustering number and result for each investigated clustering method are visualized. On the upper row, all chosen clustering results are colored on the FAMD dimension reduction visualization. On the lower row, the results are colored on the UMAP dimension reduction visualization. All chosen methods partition the data points differently, indicating distinct features being used to perform the clustering. The cosine distance-based hierarchical clustering result and the UMAP-based k-means clustering result are fuzzy-looking on the FAMD visualizations. Likewise, the Gower's distance-based hierarchical clustering result and the FAMD-based k-means clustering result are fuzzy-looking on the UMAP visualizations. This result indicates that Gower's distance-based hierarchical clustering result and the FAMD-based k-means clustering result are based on partially similar features. A corresponding connection is seen between the cosine distance-based hierarchical clustering result and the UMAP-based k-means clustering result.

Figure 14: Comparison of the selected cluster results for each cluster algorithm on FAMD and UMAP dimension reduction visualizations



4.4 Evaluation of the clusters

No external information is available in this cluster analysis, and no true clustering labeling exists. Hence, internal validation and domain knowledge determine the number of clusters and choose the best clustering algorithm. A summation of the internal indices for the chosen number of clusters for each clustering method is seen in Table 4. In Table 4, SC denotes the Silhouette coefficient defined in equation 6, which ranges between -1 and 1, where a higher value indicates better clustering. CH denotes the Calinski-Harabasz index defined in equation 10, which has no upper bound or optimal value, but a higher value indicates better clustering. Therefore, CH is not comparable between different clustering algorithms. However, the CH index is normalized in the compound, defined in equation 11. Therefore, the compound ranges between -0.66 and 1, with a higher value indicating better clustering. Table 4 shows that the UMAP-based k-means clustering performed best, followed by the FAMD-based k-means clustering. Yet, the hierarchical clustering approaches did not produce outstanding results.

Still, internal validation only indicates the quality of the clusters, and domain knowledge is needed for evaluating and understanding the clusters [38]. The results of chosen internal validation methods are dependent on the data structure. The different distance metrics (Gower's distance and cosine distance in the hierarchical clustering approaches and Euclidean distance in the k-means-based approaches) might have influenced the results. Therefore, domain expertise is used to select the best clustering algorithm to identify groups, hence gaining new knowledge and information through the identified groups. By visual inspection combined with expert knowledge, the FAMD-based k-means clustering resulting clusters are determined to be the best result.

Table 4: Internal indices of the four clusterings

Clustering method	Number of clusters	SC	CH	Compound
Gower's distance-based hierarchical clustering	5	0.263	166.5	0.326
Cosine distance-based hierarchical clustering	5	0.363	99.1	0.443
FAMD based k-means clustering	5	0.513	726.4	0.544
UMAP based k-means clustering	5	0.555	1196.4	0.675

All four different clustering algorithm results show an underlying structure in the data. In addition, all methods found bigger and smaller groupings in the data. However, none of the dimension reduction methods and the clustering results projected on them produced both compact and isolated clusters. These results are shown in Figure 9 and Figure 14. The FAMD-based k-means clustering produced partly compact but noisy clusters close to each other. The UMAP-based k-means clustering produced several more isolated groups compared to FAMD-based k-means clustering. The compactness and the isolation of clusters by the hierarchical approaches are more challenging to evaluate. Yet similar behavior is expected as the data seem to be hard to divide into distinct groups.

The results in Table 4 also reflect the visual clustering results. The Silhouette coefficient value is not outstanding for any clustering algorithm result, but the FAMD and UMAP-based k-means clustering values are near each other. The pattern of the different clustering algorithms' compound results follows the pattern of the Silhouette coefficient results. Although the difference between the compound results of the FAMD and UMAP-based k-means clustering is greater than in the SC results, it is understood as an impact of the noisiness in the FAMD dimension reduction.

Hierarchical clustering is used in patient portal research [5, 14]. K-means clustering is a universal clustering algorithm, thus selected for the FAMD and UMAP dimension-reduced datasets clustering. As the internal validation results of the hierarchical clustering are not favorable and the k-means clustering works best on well-separated and spherical clusters, some other choice of clustering algorithm might have generated better results. For oddly shaped clusters, as in both graphs in Figure 9, the density-based DBSCAN [66] might have worked better. Another approach to generating clearer clusters could have been to one-hot encode all the numerical variables into categorical variables since the dataset consists of mixed data with only a few numerical values. Then hierarchical clustering or k-means clustering based on Hamming distance would have been applied to the dataset.

4.5 Interpretation of the clusters and defined user groups

The clusters are interpreted using statistical summation and decision trees fitted to the resulting clusters. Statistical summation is used to identify the typical member of a cluster, while the decision tree is used to explain the splitting features used to divide the dataset into clusters.

In the decision tree, the feature that is the cause of the initial split of the dataset is located in the root node [64] and can be interpreted as the rule of the root node. The internal nodes represent the further choices that divide the subsets of the data into smaller subsets. Hence, a link coming from a node represents one of the feature alternatives at the node. The leaf nodes represent the final grouping of the data, and the path between the root node and the leaf node represents all the conditions the points of the leaf node satisfy [64].

The number of unique features determines the splitting rule in the root node in the decision tree fitted to hierarchical clustering based on Gower's distance. Other splitting features of the internal nodes are additional sign-in, notification enabling, and viewing results. Similarly, the used device is the initial splitting feature of the cosine distance-based hierarchical clustering, and other internal node splitting features are the number of tracked log events and sending a message. For the k-means clustering approaches, the initial splitting of the FAMD-based clustering is the number of unique features, and other splitting features are the total number of tracked log events and sending a message. For UMAP-based clustering, the root node's feature is the device, and the initial nodes' features are the total number of events, the number of unique features, and the number of days between invitation and registration.

The feature importance of a decision tree tells which feature removal would cause the biggest reduction in the accuracy of the decision tree [64]. For cosine distance-based hierarchical clustering and UMAP-based k-means clustering, the most important features are the similar features used in the splitting rules of the data set in creating the decision trees. For Gower's distance-based hierarchical clustering and FAMD-based k-means clustering, session expiration is a feature impacting the accuracy of the decision trees besides the features used for splitting in each decision tree.

The statistical summation of the clusters is performed by calculating the median of the numerical features and the mode of the categorical features for each cluster. Hence, the centroid of each cluster is determined, revealing the typical member of the cluster. The user characteristics age, and sex are added to the clustering results before the statistical summation, determining the typical member of each cluster for these characteristics. In addition, a manual visual inspection of the value distribution for each cluster feature exposes possible deviations from the typical member within each cluster. The visual inspection also shows differences between clusters. The visual inspection is conducted by creating bar plots of the categorical features for each cluster. Box plots and scatterplots are created for the numerical features of each cluster. Consequently, the proportion of the patients in a cluster who, e.g., use the mobile app can be determined, although the desktop is the typical device for the cluster according to the statistical summation.

The typical member of each cluster based on the clustering approaches is summarized in Tables 5, 6, 7 and 8. In the Tables, *n* denotes the number of patients in the cluster and *login time (h)* the hour of the day when the initial sign-in takes place. Furthermore, the *unique features* denote the total number of unique features tracked, and *events* indicate the total number of tracked log data events in the initial session. In addition, the user characteristics age, and sex, added to the clustering results, are also shown in the tables.

Table 5 and Table 6 show that the Gower's and cosine-based hierarchical clustering methods produced fuzzy clusters. The respective table reveals several clusters with more or less defined characteristic features than the number of unique features defined for the cluster in question. The fuzziness of these clustering algorithm results is also estimated in the internal indices Table 4. Hence, this result is expected. The FAMD-based k-means clustering in Table 7 produces clusters of uneven size: two clusters contain over 150 patients, one cluster is 73 patients, and two clusters are less than 20 patients. UMAP-based k-means clustering in Table 8 produces the most even-sized clusters, which contain four clusters with around 100 patients and one with 56 patients.

Some trends in the data are seen in the patient clusters in all four clusterings when combining the results from the decision trees and the statistical summation of the respective clustering methods.

- The number of tracked log events estimates the number of unique features. This correlation can be seen in all clusters regarding the clustering algorithm.

- The most frequently occurring features characterizing the clusters in descending order are the home page, notification enabling and/or profile editing, and result viewing. Initial sessions containing a significant amount of used features tend to contain messaging-related features and an additional sign-in.

- Around 200 patients use only one or two features

- Mobile browser users tend to use fewer features independent of the clustering algorithm

- The days between the invitation and the registration impacts the number of unique features and session length, especially when the days between the invitation and the registration are 30 or more.

- Patients seem to do their initial login during office hours.

Two different clustering strategies can be identified from the four clustering results. The Gowers distance-based hierarchical clustering and the FAMD-based k-means clustering group the data principally according to the number of unique features. Hence, especially the FAMD-based k-means clustering distinguishes users who sign out immediately after signing in. In addition, a group of super users who spend more than 10 minutes on the portal using several features is found in the FAMD-based k-means clustering. The other strategy is built on dividing the data according to the used device. This strategy is seen in cosine-based hierarchical clustering and UMAP-based k-means clustering. In these clustering results, most mobile app users sign in within a day after receiving the invite and use only a few features during their initial session.

Furthermore, age and sex distribution are checked for each clustering result, although these user characteristics were not part of the pre-processed dataset used for the clustering task. Inspecting the typical member of each cluster in the clustering results in Tables 5, 6, 7 and 8, it can be seen that most clusters' typical member is a woman between 54 and 63 years old. Clear exceptions are cluster number four in the FAMD-based k-means clustering result, for which the typical member is a 72 years old man and cluster number 5 in the cosine-based hierarchical clustering and in the UMAP-based k-means clustering, for which the typical member is a 67 years old man.

Table 5: Typical member of the Gower's distance-based hierarchical clusters

Cluster	n	Device	Login time(h)	Characteristic features	Unique features	Events	Session length	Invitation to registration	Age	Sex
1	192	Mobile browser	14	home page	1	5	1 min 11 sec	8 days	61	Woman
2	56	Mobile browser	13	additional sign in, home page, profile editing, notification enabling	4	11	4 min 45 sec	8 days	55	Woman
3	13	App	8	additional sign in, home page, profile editing, notification enabling, receive notification, view results, read message, send message	8	23	7 min 43 sec	45 days	57	Man
4	95	Desktop	12	home page, view results	4	9	2 min 22 sec	11 days	62	Woman
5	88	App	10	home page, profile editing, notification enabling, view results, read message	6	16	4 min 41 sec	9 days	55	Woman

Table 6: Typical member of the cosine distance-based hierarchical clusters

Cluster	*n*	Device	Login time (h)	Characteristic features	Unique features	Events	Session length	Invitation to registration	Age	Sex
1	172	Mobile browser	13	home page	2	7	2 min 0 sec	8 days	55	Women
2	104	Mobile app	13	home page, profile editing, notification enabling, view results	3	10	2 min 11 sec	1 day	57	Woman
3	16	App/ Desktop	10	home page, notification enabling, read message, send message	6	14	5 min 50 sec	43 days	60	Woman/ Man
4	27	Desktop	12	additional sign in, home page, profile editing, notification enabling, view results, view documents, read message	6	20	6 min 27 sec	45 days	58	Woman
5	125	Desktop	13	home page	3	7	2 min 3 sec	10 days	67	Man

Table 7: Typical member of the FAMD-based k-means clusters

Cluster	n	Device	Login time(h)	Characteristic features	Unique features	Events	Session length	Invitation to registration	Age	Sex
1	73	Desktop	10	home page, profile editing, notification enabling, view results, read message	6	17	5 min 17 sec	37 days	56	Woman
2	181	Mobile browser	14	home page	1	5	1 min 19 sec	8 days	60	Woman
3	161	Desktop	13	home page, profile editing, notification enabling, view results	4	10	3 min 2 sec	7 days	59	Woman
4	18	Mobile browser	15	-	0	3	0 min 0 sec	10 days	71	Man
5	11	Desktop	8	additional sign in, home page, profile editing, notification enabling, receive notification, view results, read message, send message	8	23	11 min 58 sec	52 days	54	Man

Table 8: Typical member of the UMAP-based k-means clusters

Cluster	n	Device	Login time(h)	Characteristic features	Unique features	Events	Session length	Invitation to registration	Age	Sex
1	56	Mobile browser	13	additional sign in, home page, profile editing, notification enabling, view results	5	12	4 min 20 sec	9 days	54	Woman
2	108	Mobile browser	13	home page	1	4	1 min 5 sec	8 days	56	Woman
3	91	Desktop	12	home page, profile editing, notification enabling, view results, read message	6	17	5 min 17 sec	45 days	60	Woman
4	90	App	13	home page, notification enabling, view results	3	9	1 min 57 sec	0 days	56	Woman
5	99	Desktop	13	home page	2	6	1 min 32 sec	11 days	67	Man

A summary of identified initial session user groups is made to create business-relevant outcomes and make the validated results useful for Kaiku Health and future product development. The validity of all clustering results is confirmed using the decision trees, statistical summations, and domain expert knowledge. The clustering results of the FAMD-based k-means clustering are used for defining and describing the groups found in the dataset. Moreover, the FAMD-based k-means clustering is chosen as it is determined to produce the best clustering result based on expert domain knowledge and the context in which Kaiku Health is used. Therefore, the FAMD-based k-means clustering results are translated into textual descriptions of five user groups based on the centroids in Table 7, the corresponding decision three, and statistical summations. The five initial session user groups are fast mobile browser users and their subgroup no-feature users, notification enablers and result viewers, message-using multi-feature users, and late registered super users. The groups are characterized as follows:

- **Fast mobile browser users** mainly visit only the home page using the mobile browser version of the patient portal. Their initial session takes between 1 and 2 minutes, and they visit one or two features. The initial sessions take place a week after the invitation at 2 p.m. The typical member of this group is a 60-year-old woman. This group is found in all clusterings.
  - **No-feature users** visit no features, and their initial sessions consist of only sign-in. Consequently, their initial session is 0 minutes long, often on a mobile browser around 3 p.m. one and a half weeks after the invitation. The typical member of this group is a 71-year-old man. This group is only found in the FAMD-based k-means clustering.
- **Notification enablers and result viewers** mainly visit the home page, enable notifications, and view results, hence visiting 3-4 different features overall during the initial session. The initial session lasts 2-3 minutes and occurs around 1 p.m. a week after invitation. The typical member of this group is a woman of 59 years. This group consists of app, mobile browser, and desktop users, while the other clustering algorithms resulted in correspondingly separated app and desktop clusters.
- **Message-using multi-feature users** visit the home page, enable notification, view results, edit their profile, and read or send a message during their initial session. Their initial session length is between 4 and 5 minutes during which the user uses 5-6 different features. The initial session takes place a month after the invitation around 10 a.m. on either app, mobile browser, or desktop. The typical member of this group is a woman of 56 years. This group is also found in all clusterings.
- **Late registered super users** visit the home page, enable notifications, view results, edit their profile, read and/or send a message, receive a notification, and/or do an additional login, using 6-8 features in total during their initial

session. Therefore, the initial session lasts 6-12 minutes, and the initial session occurs one and a half months after the invitation at 8 a.m. The typical member of this group is a 54-year-old man. This group is found in all clusterings and consists of initial sessions on the app, mobile browser, and desktop.

There are four principal user patterns in the clusters: patients who visit the home page in their initial session, patients who enable notification and view the result section, patients who use several features, including the messaging feature, and patients who register late and use a great number of features, including those used by the other groups. The user groups are confirmed by domain experts at Kaiku Health. These user groups correspond to the resulting groups of the FAMD-based k-means clustering. In addition, a similar cluster is found in the other clusterings. The best matching corresponding cluster for each clustering algorithm can be seen in Table 9.

Table 9: The corresponding clusters of the defined user groups in the FAMD- and UMAP-based k-means clusterings and Gower's and cosine-based hierarchical clusterings

User group	FAMD	UMAP	Gower's	Cosine
No-feature users	4			
Fast mobile browser users	2	2	1	1
Notification enablers and result viewers	3	4,5	2,4	2,5
Message-using multi-feature users	1	1	5	3
Late registered super users	5	3	3	4

The number of users in each defined user group is seen in Table 10. The number of users in Table 10 corresponds to the number of individuals in the FAMD-based k-means resulting clusters. The largest group is the fast mobile browser users group, consisting of 181 users. Together with the no-feature users group, the fast mobile browser users and the no-feature users constitute 44.8 % of all users. The notification enablers and result viewers are the second largest group, consisting of 161 users or 36.2 %. Thus, only 19.0 % is made up of message-using multi-feature users and late registered super users. Hence, one can conclude that only a fifth of all initial sessions are longer sessions containing several different used features.

Table 10: The number of individuals in each defined user group

User group	Number of users
No-feature users	18
Fast mobile browser users	181
Notification enablers and result viewers	161
Message-using multi-feature users	73
Late registered super users	11
All users	444

5 Discussion

5.1 Findings and their implications

Five different initial session user groups were found: fast mobile browser users, no-feature users, notification enablers and result viewers, message-using multi-feature users, and late registered super users. The fast mobile users visit only the homepage using a mobile browser, while the no features users only complete the sign-in. The notification enablers and result viewers sign in to enable notifications and take a look at their results. Message-using multi-feature users use several features in their initial session, including the message function. The late registered super users do their initial sign-in one and a half months after the invitation and use numerous features during their initial session.

Common for all found groups in the four clusterings results are that the number of unique features follows the total number of events. Hence, the pattern of the total number of events estimates the pattern of the number of unique features. As the usage of some features causes several different events to be logged for the same feature, the total number of events is expected to be a bit bigger than the number of unique features. However, the size difference between the parameters should be only a few numbers, depending on the tracked features. If a majority of the users get lost in the portal or have difficulty navigating between different features, this would be seen as a massive difference in the total number of tracked events and the number of unique features in the clusters. Therefore, the correlation between these parameters indicates that the patient portal is being used qualitatively already in the initial session and that the patient portal is intuitive to use for the patient.

The used features answer the research question regarding what patient users do in their initial sessions. The home page is found among the characteristic features in all user groups except no feature users. However, it is expected to be the most visited "feature" among the defined unique features, as the home page is the first page the patient users encounter after signing in tracked as a log data event. Furthermore, notification enabling, result viewing, and profile editing are found among the characteristic features in the user groups notification enablers and result viewers, message-using multi-feature users, and late registered super users. A possible explanation for this finding is that patient portal users want to enable the notifications and edit their profiles during the initial sessions to facilitate later use. This explanation makes these features expected to be frequently used in the initial sessions. The frequency of the result viewing feature was less expected, but it can be interpreted as the patients being interested in their results. In addition, result viewing is reported to be the most frequently used feature in other studies [10] and also the most used feature among cancer patients [6, 25]. The most frequently used features can also be understood through the structure of the home page [13]. The home page of the Kaiku Health patient portal has links to result viewing, messages profile, and settings, which are consistent with the most frequently used features in the clusters. Thus, the importance of the layout and structure of the home page can be concluded, at least for the initial impression.

The fast mobile browser uses the smallest number of different functions. A possible reason for this is that people might do their initial sign-in and immediately decide to download the app or switch to a desktop. Another explanation could be that people log in to do one single task and then leave the portal as soon as the task is done, as found in [28]. Using a mobile browser or app makes it easy to sign in whenever one needs to use a patient portal feature. Therefore, the answer to the research question regarding device distribution is that fewer unique features users use the mobile browser. Similarly, several unique features users use the mobile app and the desktop.

In general, most initial sessions contain only a few unique features. The user groups fast mobile browser users and notification enablers and result viewers use between one and four unique features in their initial sessions. Together, these user groups count for 342 of 444 or 77 % of all the users according to Table 10. However, the results show that the majority use some features already during the initial sessions. This result is expected as similar findings have been achieved in other studies, showing that median patient portal users visit one feature [13, 28].

Furthermore, initial sessions occurring within a week after the invite are also affected by the time between the invite and the registration. A longer time between the invite and the registration increases the probability of a message being sent to the patient or results being uploaded before the initial session, thus leading to an extended initial session. Therefore, the user group of late registered super users is expected, as late registration is associated with longer initial sessions containing more unique features. This finding also answers the research question regarding how long patients stay on the portal the first time. The session length mainly depends on the time between the invite and the registration.

The initial session occurring during office time is reasonable as people tend to be awake and active during these hours. In addition, they might also be able to contact their hospital regarding the patient portal if needed during office hours. The very similar age and sex of the typical members in the clusters are expected, as the user characteristics are not part of the clustering process, and the data points in the dataset are unevenly distributed regarding sex and age. Instead, the majority of the users in the dataset are women, and the biggest age group is users between 40-65. A pattern seen in the clustering results of the Gower's- and cosine-based hierarchical clustering results and the UMAP-based k-means clustering result is that older users tend to prefer the desktop. Similar patterns are reported in another patient portal study [7], making this result expected.

The no-feature users' user group is found only in the FAMD clustering result and can be understood and explained in numerous ways. The group is not unexpected as similar zero-use patient portal user groups have been identified in another study [13]. One possible explanation for the group is lower computer literacy and possible attitudes toward new digital tools. Older users over 65 are reported to have lower computer literacy compared to younger users [11], and the typical members of the no-feature group are significantly older than the typical members of the other user groups. Nevertheless, other possible explanations are the data preprocessing and tracking of log data.

5.2 Limitations

This clustering analysis has several limitations. The dataset used in the cluster analysis consists of only 444 users, making the dataset a small dataset. Therefore trends in the multidimensional data are more challenging to spot. Also worth noticing is that the portion of users registered later than a month after the invitation is expected to be smaller than the potion registered sooner after the invitation. The potential later registered users invited in August are not included in this data set, as the data is collected between the first of June and the last of August 2023.

The dynamical definition of the initial session affects the further analysis, as non-dynamical restrictions are applied to all data. In this analysis, a single session is not restricted to a single sign-in, thus potentially containing an additional sign-in. This decision can result in longer sessions. However, it is determined to be necessary for catching the actual first session, as users can be required to sign in several times during a session, e.g., because of password change. In addition, similar decisions have also been made in another patient portal study [14]. Defining the end of a session is challenging when no sign-out is encountered. No session is identical to another, and some time limit is often needed [9]. Moreover, the end of a session definition impacts the balance between the proportion of longer sessions and no-feature short sessions. In this analysis, the absolute end of a session is determined to be 30 minutes after the initial sign-in, and the time passed between any two following events is allowed to be at maximum 5 minutes to be counted to the same session, as seen in Figure 2. Hence, longer first sessions consisting of several shorter sessions are avoided, causing some first sessions to be considerably short.

Tracking log data events may cause incorrect data at times, consequently affecting the data pre-processing and the final result of the data analysis [28]. Some events might be logged at incorrect timestamps due to errors or delays in the technical login process. These types of possible delays and errors in log data events are also a potential explanation for finding no-feature initial sessions, making the time between the initial sign-in event and a possible following event longer than 5 minutes.

The binary tracking of unique features in the feature selection part of the clustering process affects and limits the results of the clustering algorithms. Tracking whether a specific feature is encountered among the logged data events or not in an initial session leaves out information about the order of the encountered features in the initial session. Nor is the frequency of a specific feature occurrence in an initial session captured in the binary tracking of unique features. Yet the correlation between the number of unique features and the number of tracked log events implies that quantitative unique feature tracking would have added only a little or no extra valuable information. In addition, the distribution of user patterns and routes in patient portals is reported to be very diffuse [13]. Therefore, the binary tracking of unique features in the initial sessions is chosen to generate valuable results and answer what happens in the initial sessions.

Other clustering process limitations are the dimension reduction in the k-means clustering-based methods usage and the order of the data points in the hierarchical clustering-based methods. Dimension reduction always results in data loss. In hierarchical clustering, the order of the data points in a dataset affects the results. Hence, another order of the data points could have generated different results. However, all clustering approaches resulted in sensible clustering results.

Limitations beyond the data and the chosen data analysis method are the lack of other analysis methods to verify the results generated in the data analysis. No interviews with the users were conducted, which could have been used to validate the conclusions regarding the results. Furthermore, as the Kaiku Health patient portal is used mainly at private hospitals and clinics, the socioeconomic advantage of the patients might have affected the results. The hospital and clinics perhaps also have their practices regarding giving patient portal instructions to the patients, which could have affected the registrations, the content of the initial sessions, and the general attitude towards the patient portal. These are aspects that cannot be verified by analyzing the real-time generated objective log data.

The study's cross-sectional nature makes it impossible to draw wider conclusions regarding the users and the user groups in a larger context over time. No conclusions on patient behavior, portal usage, and the pattern of usage development during an extended period can be made, as no longitudinal data is available in this cluster analysis. Hence, it also remains unclear whether the found initial session user groups have significance in the long run. Still, it can be concluded that clear user groups and patterns are found in the initial sessions.

5.3 Further research

At Elekta Kaiku, further research on this topic could involve a follow-up longitudinal analysis of the same data set to reveal the possible impact of the initial session on further use of the Kaiku Health patient portal. Questions to answer could be whether the individual user patterns from the initial login will repeat in potential later sessions. Other questions are how frequently the later sessions (excluding the initial session) will occur and whether the frequency is impacted by the type of the initial session. One could also investigate which features the following sessions contain and whether this content is impacted by the user pattern of the initial session.

One of the user groups found in the cluster analysis is the no-feature users. As the no-feature users' user group can have several possible explanations, further research is needed on the underlying causes. Additional research would also create a better understanding of the Kaiku Health patient portal users. The result could be used in the Kaiku Health patient portal development, and make the portal suit the needs of the patient users better. Another interesting aspect would be to investigate the user characteristics and possible association with different patient portal initial session user groups, both in a cross-sectional study and in a study over time.

As a conclusion, potential further research on the topic is interesting as clear user groups are found among the initial sessions in this thesis. The user groups use different features and different amounts of unique features during their initial session, depending on which device they use and how soon after the invitation they register and sign in.

6 Conclusion

Five different user groups were found in the initial sessions: fast mobile browser users, no-feature users, notification enablers and result viewers, message-using multi-feature users, and late registered super users. The user groups are based on the number of unique features used in the initial session: the no-feature users only sign-in, and fast mobile browser users sign-in and visit the home page. Notification enablers and result viewers enable notification and view results in addition to visiting the home page, and message-using multi-feature users use all features already mentioned in addition to the messaging feature. Late registered super users do their initial sign-in considerably later after being invited than the other user groups, using the most different features of all users.

The final clustering results were achieved by identifying the initial sessions from the log data and performing feature selection, followed by trying four approaches based on two different clustering methods: Hierarchical clustering based on first Gower's distance and then cosine distance as well as k-means clustering based on first FAMD dimension reduction and then UMAP dimension reduction. The resulting clusters from the four approaches were evaluated using decision trees, statistical summation, visual inspection, and domain knowledge. The most distinct clusters were produced by the UMAP-based k-means clustering when using internal validation measurements. Still, the found user groups were mapped in the other clustering results, implying that all used clustering approaches were appropriate for this clustering task.

Returning to the research questions, which are what happens during the patients' initial sessions in the patient portal, how long the initial sessions are, and what the distribution of the used devices for the initial sessions are, answers have been provided through the found distinct user groups. Hence, the goal of identifying user groups based on initial session characteristics is achieved by utilizing different clustering approaches on a dataset consisting of log data events from patients using the Kaiku Health patient portal.

The identified user groups show differences in the activity among cancer patients when registering and using the Kaiku Health patient portal for the first time. In a broader perspective, the results might be significant. Understanding patient portal users' first impression is important, as it often impacts further use [14]. Therefore, to improve the usability of patient portals and cancer care, these results should be used in future product development and system design.

As a conclusion of this work, the following findings and points should be taken into account in further patient portal development and research:

- Different types of initial session behavior and patterns exist, but no signs of unfocused and non-strategic use were found in the data of this work

- The importance of the patient portal home page layout was identified, as the layout impacts the features used in the initial session

- Longitudinal research on the impact of the initial session user types on longtime patient portal use is needed

- The group using no features in the initial session should be further analyzed to reveal possible reasons and explanations behind their low activity

References

[1] Thomas Hofmarcher et al. "The cost of cancer in Europe 2018". In: *European Journal of Cancer* 129 (2020), pp. 41–49.

[2] Rebecca L Siegel et al. "Cancer statistics, 2023". In: *Ca Cancer J Clin* 73.1 (2023), pp. 17–48.

[3] Lucia Ferrara et al. "How to improve efficiency in cancer care: Dimensions, methods, and areas of evaluation". In: *Journal of Cancer Policy* (2022), p. 100355.

[4] Jessica S Ancker et al. "Use of an electronic patient portal among disadvantaged populations". In: *Journal of general internal medicine* 26 (2011), pp. 1117–1123.

[5] James Brian Jones et al. "The wired patient: patterns of electronic patient portal use among patients with cardiac disease or diabetes". In: *Journal of medical Internet research* 17.2 (2015), e3157.

[6] Rebecca P Luoh et al. "Patterns and predictors of cancer-specific patient health portal usage among patients with cancer: results from the UWCCC Survivorship Program". In: *Cancer Medicine* 10.20 (2021), pp. 7373–7382.

[7] Kelvin K Pho et al. "Mobile device applications for electronic patient portals in oncology". In: *JCO clinical cancer informatics* 3 (2019), pp. 1–8.

[8] Sanna Iivanainen et al. "ePRO symptom follow-up of colorectal cancer patients receiving oxaliplatin-based adjuvant chemotherapy is feasible and enhances the quality of patient care: a prospective multicenter study". In: *Journal of Cancer Research and Clinical Oncology* (2023), pp. 1–8.

[9] Jeong Yeob Han. "Transaction logfile analysis in health communication research: challenges and opportunities". In: *Patient education and counseling* 82.3 (2011), pp. 307–312.

[10] Raymond Tsai et al. "How patients use a patient portal: an institutional case study of demographics and usage patterns". In: *Applied Clinical Informatics* 10.01 (2019), pp. 096–102.

[11] Mita Sanghavi Goel et al. "Disparities in enrollment and use of an electronic patient portal". In: *Journal of general internal medicine* 26 (2011), pp. 1112–1116.

[12] Kelvin Pho et al. "Characteristics of patients using patient portals in oncology". In: *JAMA oncology* 4.3 (2018), pp. 416–418.

[13] Floor Sieverink et al. "The added value of log file analyses of the use of a personal health record for patients with type 2 diabetes mellitus: preliminary results". In: *Journal of diabetes science and technology* 8.2 (2014), pp. 247–255.

[14] Floor Sieverink et al. "Opening the black box of electronic health: collecting, analyzing, and interpreting log data". In: *JMIR research protocols* 6.8 (2017), e6452.

[15] National Cancer Institute. *Understanding cancer*. https://www.cancer.gov/about-cancer/understanding. (Visited 2023-08-30). (2021).

[16] National Cancer Institute. *What is cancer?* https://www.cancer.gov/about-cancer/understanding/what-is-cancer. (Visited 2023-08-30). (2021).

[17] World Health Organisation. *Cancer*. https://www.who.int/news-room/fact-sheets/detail/cancer. (Visited 2023-08-30). (2022).

[18] Jacques Ferlay et al. "Cancer statistics for the year 2020: An overview". In: *International journal of cancer* 149.4 (2021), pp. 778–789.

[19] National Cancer Institute. *Cancer Staging*. https://www.cancer.gov/about-cancer/diagnosis-staging/staging. (Visited 2023-08-30). (2021).

[20] World Health Organisation. *Types of Cancer*. https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies. (Visited 2023-11-3). (2022).

[21] HealthIT.gov. *What is a patient portal?* https://www.HealthIT.gov.faq/what-patient-portal. (Visited 2023-09-29). (2017).

[22] Ashley Griffin et al. "Patient portals". In: *Applied clinical informatics* 7.02 (2016), pp. 489–501.

[23] Elettra Carini et al. "The impact of digital patient portals on health outcomes, system efficiency, and patient attitudes: updated systematic literature review". In: *Journal of Medical Internet Research* 23.9 (2021), e26189.

[24] Andrew L Laccetti et al. "Increase in cancer center staff effort related to electronic patient portal use". In: *Journal of oncology practice* 12.12 (2016), e981–e990.

[25] David E Gerber et al. "Predictors and intensity of online access to electronic medical records among patients with cancer". In: *Journal of oncology practice* 10.5 (2014), e307–e312.

[26] Steven S Coughlin, Lee Caplan, and Lufei Young. "A review of web portal use by oncology patients". In: *Journal of cancer treatment & diagnosis* 2.6 (2018).

[27] Christian Johnson, Vaishali Patel, and Melinda Krakow. "Access and Use of Electronic Health Information by Individuals with Cancer: 2017-2018". In: *2020 Virtual Annual Research Meeting*. AcademyHealth. 2020.

[28] Gennaro Di Tosto et al. "Metrics for outpatient portal use based on log file analysis: algorithm development". In: *Journal of medical Internet research* 22.6 (2020), e16849.

[29] Nicole A Maher et al. "Passive data collection and use in healthcare: A systematic review of ethical issues". In: *International journal of medical informatics* 129 (2019), pp. 242–247.

[30] Talayeh Aledavood. "Temporal patterns of human behavior". English. Doctoral thesis. School of Science, 2017, 83 + app. 42. ISBN: 978-952-60-7724-6 (electronic), 978-952-60-7723-9 (printed). URL: http://urn.fi/URN:ISBN:978-952-60-7724-6.

[31] Talayeh Aledavood et al. "Daily rhythms in mobile telephone communication". In: *PloS one* 10.9 (2015), e0138098.

[32] Talayeh Aledavood, Sune Lehmann, and Jari Saramäki. "Digital daily cycles of individuals". In: *Frontiers in Physics* 3 (2015), p. 73.

[33] Rainer Kujala, Talayeh Aledavood, and Jari Saramäki. "Estimation and monitoring of city-to-city travel times using call detail records". In: *EPJ Data Science* 5 (2016), pp. 1–16.

[34] Stephen M Mattingly et al. "The effects of seasons and weather on sleep patterns measured through longitudinal multimodal sensing". In: *NPJ digital medicine* 4.1 (2021), p. 76.

[35] Brian Everitt et al. *Cluster analysis*. eng. 5th ed. Wiley series in probability and statistics. Chichester, West Sussex, U.K: Wiley, 2011. ISBN: 0-470-97844-9.

[36] Amir Ahmad and Shehroz S Khan. "Survey of state-of-the-art mixed data clustering algorithms". In: *Ieee Access* 7 (2019), pp. 31883–31902.

[37] Michel Van de Velden, Alfonso Iodice D'Enza, and Angelos Markos. "Distance-based clustering of mixed data". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 11.3 (2019), e1456.

[38] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. "On clustering validation techniques". In: *Journal of intelligent information systems* 17 (2001), pp. 107–145.

[39] Rui Xu and Donald Wunsch. "Survey of clustering algorithms". In: *IEEE Transactions on neural networks* 16.3 (2005), pp. 645–678.

[40] Glenn W Milligan and Martha C Cooper. "Methodology review: Clustering methods". In: *Applied psychological measurement* 11.4 (1987), pp. 329–354.

[41] Tsang-Hsiang Cheng, Chih-Ping Wei, and Vincent S Tseng. "Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches". In: *19th IEEE symposium on computer-based medical systems (CBMS'06)*. IEEE. 2006, pp. 165–170.

[42] The pandas development team. *pandas-dev/pandas: Pandas*. Version 1.5.3. Dec. 2023. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[43] Charles R. Harris et al. *Array programming with NumPy*. Version 1.20.3. Sept. 2020. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[44] Frank Nielsen and Frank Nielsen. “Hierarchical clustering”. In: *Introduction to HPC with MPI for Data Science* (2016), pp. 195–211.

[45] Fionn Murtagh and Pedro Contreras. “Algorithms for hierarchical clustering: an overview, II”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7.6 (2017), e1219.

[46] John C Gower. “A general coefficient of similarity and some of its properties”. In: *Biometrics* (1971), pp. 857–871.

[47] Alexander H Foss, Marianthi Markatou, and Bonnie Ray. “Distance metrics and clustering methods for mixed-type data”. In: *International Statistical Review* 87.1 (2019), pp. 80–109.

[48] Michael Yan. *Gower*. Version 0.1.2. Dec. 2023. URL: https://github.com/wwwjk366/gower.

[49] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[50] Sirine Sayadi et al. “Secure distribution of Factor Analysis of Mixed Data (FAMD) and its application to personalized medicine of transplanted patients”. In: *Advanced Information Networking and Applications: Proceedings of the 35th International Conference on Advanced Information Networking and Applications (AINA-2021), Volume 1 35*. Springer. 2021, pp. 507–518.

[51] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[52] Caterina Labrín and Francisco Urdinez. “Principal component analysis”. In: *R for Political Data Science*. Chapman and Hall/CRC, 2020, pp. 375–393.

[53] Jérôme Pagès. *Multiple Factor Analysis by Example Using R*. eng. 1st edition. Chapman and Hall/CRC, 2014.

[54] Max Halford. *Prince*. Version 0.13.0. Dec. 2023. URL: https://github.com/MaxHalford/prince.

[55] Leland McInnes. *umap-learn*. Version 0.5.5. Dec. 2023. URL: https://github.com/lmcinnes/umap.

[56] Stuart Lloyd. “Least squares quantization in PCM”. In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[57] Douglas Steinley. “K-means clustering: a half-century synthesis”. In: *British Journal of Mathematical and Statistical Psychology* 59.1 (2006), pp. 1–34.

[58] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[59] Christian Hennig. “What are the true clusters?” In: *Pattern Recognition Letters* 64 (2015), pp. 53–62.

[60] Yanchi Liu et al. "Understanding of internal clustering validation measures". In: *2010 IEEE international conference on data mining*. IEEE. 2010, pp. 911–916.

[61] Jannica Savander et al. "Clustering of radiotherapy patients based on the usage patterns in the Kaiku Health application and electronic patient-reported outcomes". In: (2023).

[62] Olivier Parisot, Mohammad Ghoniem, and Benoît Otjacques. "Decision Trees and Data Preprocessing to Help Clustering Interpretation." In: (2014), pp. 48–55.

[63] Sayan Bandyapadhyay et al. "How to find a good explanation for clustering?" In: *Artificial Intelligence* (2023), p. 103948.

[64] Yan-Yan Song and LU Ying. "Decision tree methods: applications for classification and prediction". In: *Shanghai archives of psychiatry* 27.2 (2015), p. 130.

[65] Ian H Witten et al. "Practical machine learning tools and techniques". In: *Data mining*. Vol. 2. 4. Elsevier Amsterdam, The Netherlands. 2005, pp. 403–413.

[66] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: 96.34 (1996), pp. 226–231.