UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Software Engineering Curriculum

Oliver-Erik Suik

Generalising health events by using frequent itemset mining

Master's Thesis (30 ECTS)

Supervisor: Sulev Reisberg, PhD

Tartu 2024

Generalising health events by using frequent itemset mining

Abstract:

The digitisation of health data has enabled us to conduct studies that enhance healthcare practices and make clinical processes more efficient. However, the diverse types of health data and their sparse nature create challenges in understanding a patient's health status and utilising it in data mining tasks and analytics techniques. The primary objective of this research is to use frequent itemset mining to generalise similar health events into a higher-level event and assess its practicality and limitations. The study involves extracting health event transactions from Estonian healthcare data using a sliding window technique and applying the FP-Max algorithm to identify frequent itemsets of health concepts. These itemsets are clustered into higher-level events, providing a generalised representation of a patient's health event timeline. Experimenting with different parameters resulted in clusters with varying levels of generalisation which ultimately helped to describe a patient's health status by reducing elements and giving them generalised labels.

Keywords: frequent itemset mining, electronic healthcare records, data summarisation

CERCS: P170 Computer science, numerical analysis, systems, control

Terviseandmete üldistamine sagedaste andmehulkade abil

Lühikokkuvõte:

Terviseandmete digiteerimine on võimaldanud viia läbi uuringuid, mis on parandanud tervishoiupraktikaid ja teinud kliinilisi protsesse tõhusamaks. Siiski on tervishoiu infot patsiendi terviseseisundi hindamisel ja andmepõhistel meetoditel raske kasutada, kuna info on oma olemuselt väga hõre ja mitmekesine. Selle uurimistöö peamine eesmärk on kasutada sagedasi andmehulkasid tervisesündmuste üldistamiseks kõrgemal tasemel sündmuseks ning hinnata nende rakendatavust ja piiranguid. Töö käigus leitakse Eesti terviseandmetest FP-Max algoritmi kasutades sagedased terviseandmete hulgad. Need hulgad klasterdatakse kõrgema taseme sündmusteks, mida kasutatakse selleks, et teha patsientide terviseündmuste ajajoon üldisemaks. Erinevate parameetritega eksperimenteerimise tulemusena tekkisid erineva üldistustasemega klastrid, mis aitasid luua üldistatud pildi patsiendi terviseseisundist, vähendades sündmuste arvu.

Võtmesõnad: sagedaste andmehulkade kaeve, elektroonilised terviseandmed, andmete üldistamine

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Table of Contents

1 Introduction

The digitisation of health data has enabled, in addition to managing patients' health-related information, utilisation of health records in clinical research and the enhancement of healthcare practices. The application of data mining and analytics techniques to Electronic Health Records (EHR) has allowed profound studies of patient cohorts, facilitating various clinical and research challenges such as disease prediction, detection, and progression analysis.

However, using health data creates significant challenges due to the diverse array of data types and their associated complexities. It is very sparse, containing several various facts (disease codes, drugs, laboratory measurements, etc.) with record dates for each patient (see Figure 1).

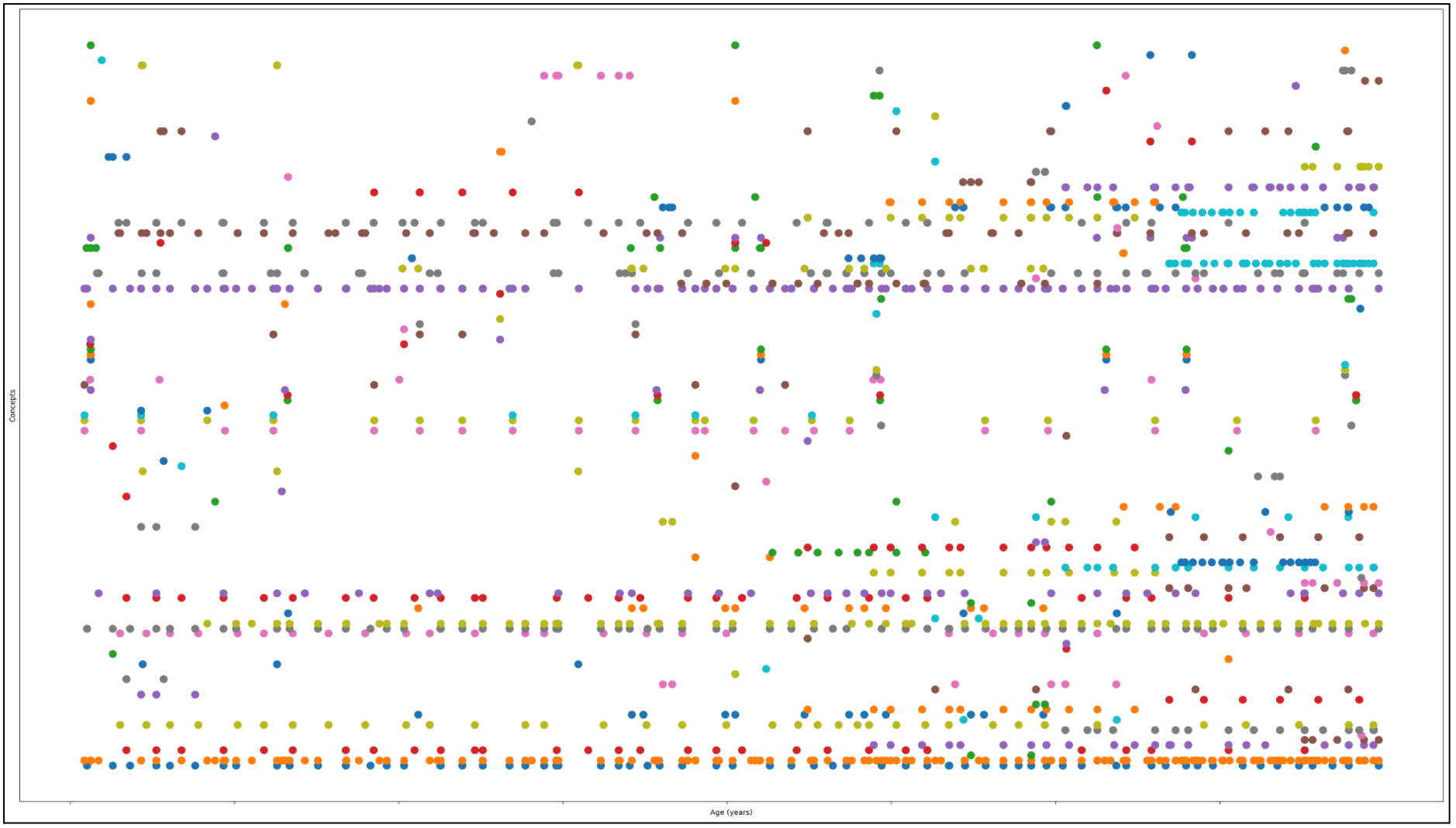

Figure 1. Health event timeline of an elderly person. Each dot represents a fact in the data, dfferent colors indicate different meaning of the facts (e.g., diclofenac, ramipril, surgical biopsy, colonoscopy)

Even the exact condition of the patient (e.g. a particular disease) ends up as a number of facts in the data (e.g. case summary, prescription, bill, etc.) and it is hard to get an overview of the patient's health status, especially when the patient has several concurrent chronic diseases. It also makes it challenging to analyse the data of a set of patients. The individual health records, on their own, cannot succinctly describe a patient's health status. Moreover, the temporal accuracy of these events is often lacking, leading to ambiguity regarding the precise chronological order of occurrences. Some of the inaccuracies can be caused by the policies of general healthcare or organisation (e.g. the case summaries are reported at the end of the disease onset, but bills may be sent out earlier). It has been shown that many variations in the order of the facts make it difficult to apply process mining techniques to get an overview of the underlying processes [6, 7]. Multiple research papers have noted the heterogeneity of the data as a limitation, for example, when applying any pathway, pattern, or process mining methods to health data [5, 6, 7, 9].

The main goal of this research is to generalise similar health events to an upper-level event using frequent itemset mining and to evaluate its applicability. This technique was chosen since it allows finding items from transactional data that appear together frequently and fall into similar concepts. Furthermore, the time of the events in the transaction is ignored, eliminating the problem of the wrong order of health events. This technique is applied to a sample of Estonian healthcare database records comprising 10% of the patients in the database, resulting in clusters of similar events that can be used to generalise health events.

The research questions addressed in this research are as follows:

RQ1: *How can frequent itemset mining be used to generalise Estonian healthcare data?*

RQ2: *What limitations restrict the implication of frequent itemset mining to Estonian healthcare data?*

The thesis begins with an introduction to related work in this field, along with a description of key concepts such as frequent itemset mining and the different algorithms in Chapter 2. Chapter 3 describes the data and the chosen methods for generalising health events. In Chapter 4, the outcomes of applying the methods and validation metrics are presented. Subsequently, the results of applying the methods to health data and future work are discussed in Chapter 5, followed by the conclusion of the research in Chapter 6.

2 Background and related work

This section introduces the core method for achieving the objective of this thesis, gives a summary of the usage of frequent itemset mining in the different fields and ultimately serves to explain the need for generalising health events.

2.1 Frequent itemset mining

Frequent itemset mining is a data mining method aimed at discovering groups of items that frequently co-occur in a dataset. This technique finds applications in transactional databases, market basket analysis, recommendation systems, and other domains where uncovering associations between items holds significance.

The Apriori algorithm [1], proposed by Agrawal and Srikant, is one of the most well-known and widely used for this purpose. Given a list of transactions and a minimum support threshold, it returns frequent itemsets. An itemset is considered "frequent" if its items occur together in at least the specified support percentage of all transactions within the dataset. The algorithm uses a pruning step which is based on the Apriori principle, which states that if an itemset is infrequent, all of its supersets must also be infrequent. Usually, association rules are generated from frequent itemsets as a last step, but this thesis does not focus on this step.

For example, as shown in Table 1, there are 6 transactions where A, B, C, and D are items in the transactions. The minimum support threshold is set to 3.

Table 1. Example transactions.

Transaction ID	**Items**
t1	{A, B, C}
t2	{B, C, D}
t3	{D}
t4	{A, B}
t5	{A, B, C}
t6	{A, B, C, D}

Transactions are scanned to find the support of each item. The result is shown in Table 2.

Table 2. Support of items in transactions.

Item	**Support**
A	4
B	5
C	4
D	3

Since the minimum support threshold is 3, all items are frequent itemsets of size 1. Next, candidate itemsets of size 2 are generated and checked if they do not contain any infrequent itemsets. The support of pruned candidates is found. The result is shown in Table 3.

Table 3. Support of candidate itemsets of size 2.

Itemset	**Support**
{A, B}	4
{A, C}	3
{A, D}	1
{B, C}	4
{B, D}	2
{C, D}	2

Out of these candidates, itemsets {A, B}, {A, C}, and {B, C} are frequent itemsets of size 2. The process is iterated once more and the result is shown in Table 4.

Table 4. Support of candidate itemsets of size 3.

Itemset	**Support**
{A, B, C}	3

Itemset {A, B, C} is also a frequent itemset, hence from the transactions with the minimum support of 3, 8 frequent itemsets were discovered.

Mannila et al. [2] introduced the WinEpi and MinEpi algorithms for mining frequent patterns in sequences. The names of these methods derive from their underlying technique: a sliding window. This window, characterised by a fixed width and movement step, traverses the sequence of events to identify recurring subsequences of events over time. Like the Apriori algorithm, WinEpi and MinEpi employ a frequency threshold and incrementally generate itemsets by constructing candidates. The method used in this thesis is similar since it combines aggregating health events that appear close together using a sliding window and applying an algorithm to these transactions to find frequent itemsets.

In addition to the Apriori algorithm, multiple other algorithms like FP-Growth [3] and FP-Max [4] have emerged, which yield similar outputs. Essentially, they share similarities as they operate on a list of transactions and a minimum support threshold variable. The main difference between the above-mentioned algorithms is the performance on large datasets. Although Apriori may be suitable for smaller datasets or scenarios where simplicity and ease of implementation are prioritised, multiple scans of the dataset are required to compute support counts and generate candidate item sets. FP-Growth constructs a compact data structure called the FP-tree [3], which represents the frequency of itemsets. It requires a single pass over the dataset to build the tree and a second pass to mine frequent itemsets, making this approach attractive for large datasets. FP-Max utilises FP-tree as well, but the output consists only of maximal frequent itemsets. An itemset is maximal if it is frequent and none of its immediate supersets are frequent [4].

Singleton sets and non-maximal frequent itemsets add no value to achieving the goal of this study. Hence, the FP-Max algorithm was selected for the remainder of the study, in addition to its performance, by removing singleton sets from its output.

2.2 Related work

Multiple health data researchers have identified the need for data generalisation, for example when exploring event sequences or applying data mining tasks.

When running a validation study by using health data from the Netherlands on a framework built for detecting clinical event trajectories using Estonian health data, Künnapuu et al. noted that one of the discrepancies resulted from the fact that different medical concepts are used in other cultures or healthcare environments for documenting the same medical condition. They

highlighted that using concepts generalised at a higher level would be considerably more efficient [5].

Salamov explored the applicability of process mining to Estonian health data while analysing the limitations and assessing the feasibility. He highlighted multiple limitations regarding missing or inaccurate time values of the processes' data records, limitations that frequent itemset mining could potentially solve. These limitations could lead to processes appearing in the wrong order in the process model, as seen in one case of a cervical cancer screening event log, where the test result appeared before the test itself. The author brought out the importance of combining concepts into activities, before using them in process mining, due to the complexity of clinical processes [6].

Toth et al. also noted the challenges of applying process mining to health data. One of the challenges they identified is the diversity of the code systems used in healthcare. They proposed aggregating similar medical concepts and creating a hierarchical code system that would allow generating process models on different levels of detail [7].

Campbell et al. introduced a temporal condition pattern mining method to address the sparse utilisation of medical concepts in electronic health record (EHR) data. They applied this method to analyse condition patterns surrounding initial paediatric asthma diagnoses, utilising the SPADE algorithm on datasets with International Classification of Diseases (ICD) codes and expanded diagnostic clusters (EDCs) [8]. The study, conducted on 71,824 patients from the Children's Hospital of Philadelphia, revealed 36 unique diagnoses in the EDC dataset compared to 19 in the ICD dataset. Moreover, temporal trends in condition diagnoses were only identified in the EDC dataset, highlighting the potential of this approach to uncover clinically relevant insights [9].

Several studies have demonstrated how utilising generalised data in some way or form has enhanced the outcomes of their research. The application of frequent itemset mining for generalising health events has not yet been extensively explored in academic research; however, other fields have used similar methods for the same objective.

Gupta et al. proposed a model called "FRI-CL" for summarising biomedical text, which assists researchers in the clinical field with the time-consuming process of extracting precise information from scientific articles and EHR-s. Their model operates similarly to the method used in this thesis. Semantic biomedical concepts were initially extracted from archives and then clustered using an adapted Dirichlet process mixture model (DPMM) clustering [10] to

create sentence representations. Subsequently, sentences were processed using a modified FP-Growth algorithm. Using PageRank [11], highly informative sentences containing frequent concepts were scored to generate summaries. These summaries were evaluated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric [12]. Compared with benchmarks, the model demonstrated improvements in memory utilisation and ROUGE scores [13].

3 Data and methods

This section introduces the data and methodology used for the research.

3.1 Data

In this study, a dataset having the Observational Medical Outcome Partnership (OMOP) common data model (CDM) is used. It is a result of a project by Oja et al. [14], where data from three Estonian national health databases of a 10% sample from the population from the period of 2012-2019 was transformed into this model. In total, the database contained medical information of 149,364 patients. The acquired data includes temporal information about patients' conditions, measurements, procedures and drug exposure with a link to the corresponding medical concept.

The following values from the corresponding data records were necessary to create itemsets:

- *person_id*: Identifier of the patient, identifying the person whose health records are utilised.
- *concept_id*: Identifier of the medical concept.
- *start_date***:** Date of the concept being recorded in the system.

It should be noted that one of the advantages of using frequent itemset mining is that it allows concurrent events and does not necessitate the specification of end times for the events. This is beneficial, particularly considering that the end times of health events are typically either unrecorded or inaccurately documented.

With these database records, examples were constructed of health event timelines for an elderly patient, as shown in Figure 1, and for a child, as depicted in Figure 2. As evident in the figures, these timelines vary regarding the number of concurrent and overall events. Despite Figure 2 having a significantly smaller number of events, it remains challenging to analyse.

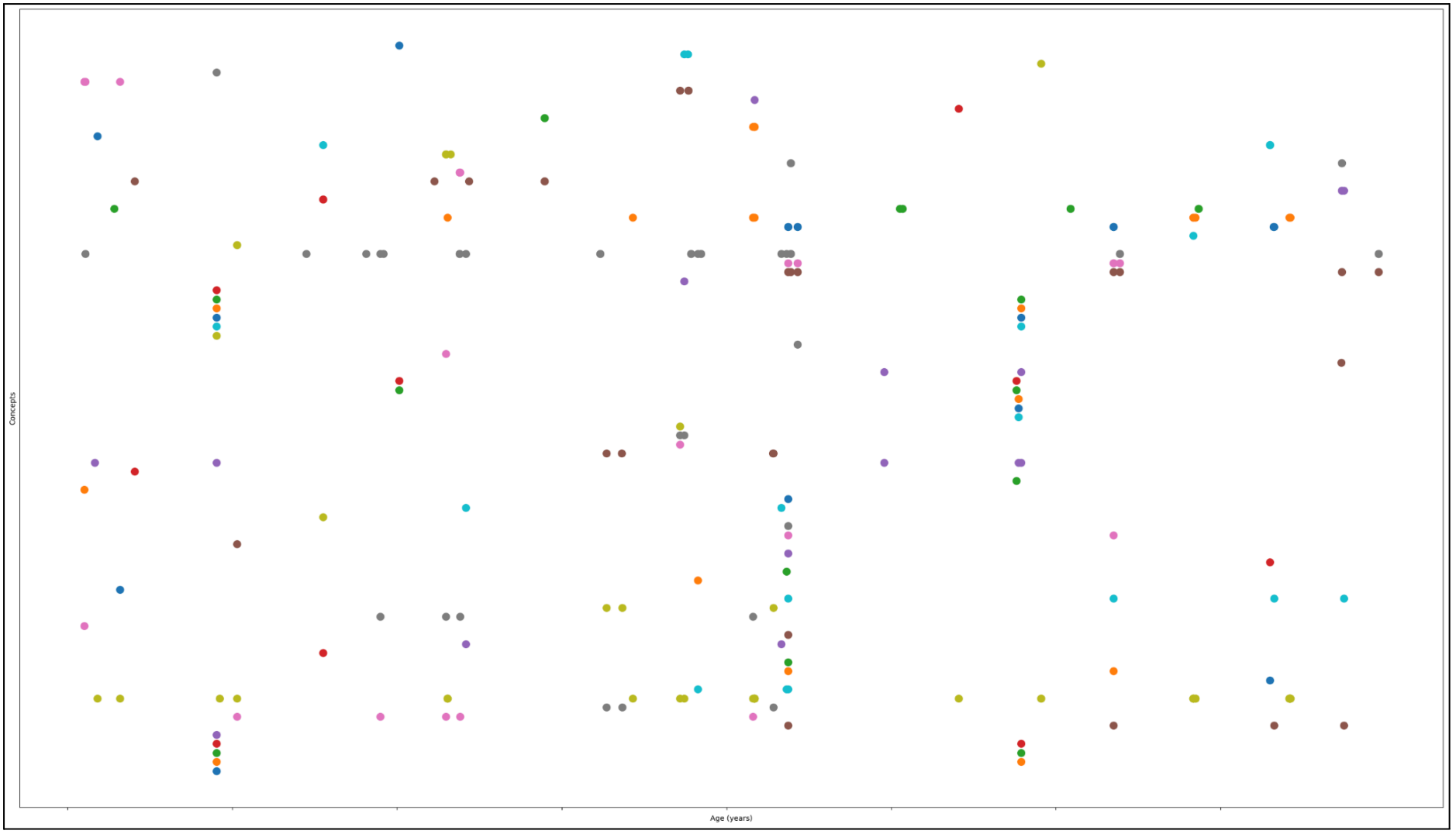

Figure 2. Health event timeline of a child. Each dot represents a fact in the data, dfferent colors indicate different meaning of the facts (e.g., child examination, audiometric test, amoxicillin, poliovirus)

3.2 Methodology

In this study, we are trying to generalise the detailed health events (records in the database) of each patient into higher-level events. For this, we first create itemsets of the records, apply frequent itemset mining, and finally create itemsets clusters. Each cluster of records represents a generalised health event. We try to propose a name for each such event.

The overall steps for this process are illustrated in Figure 3.

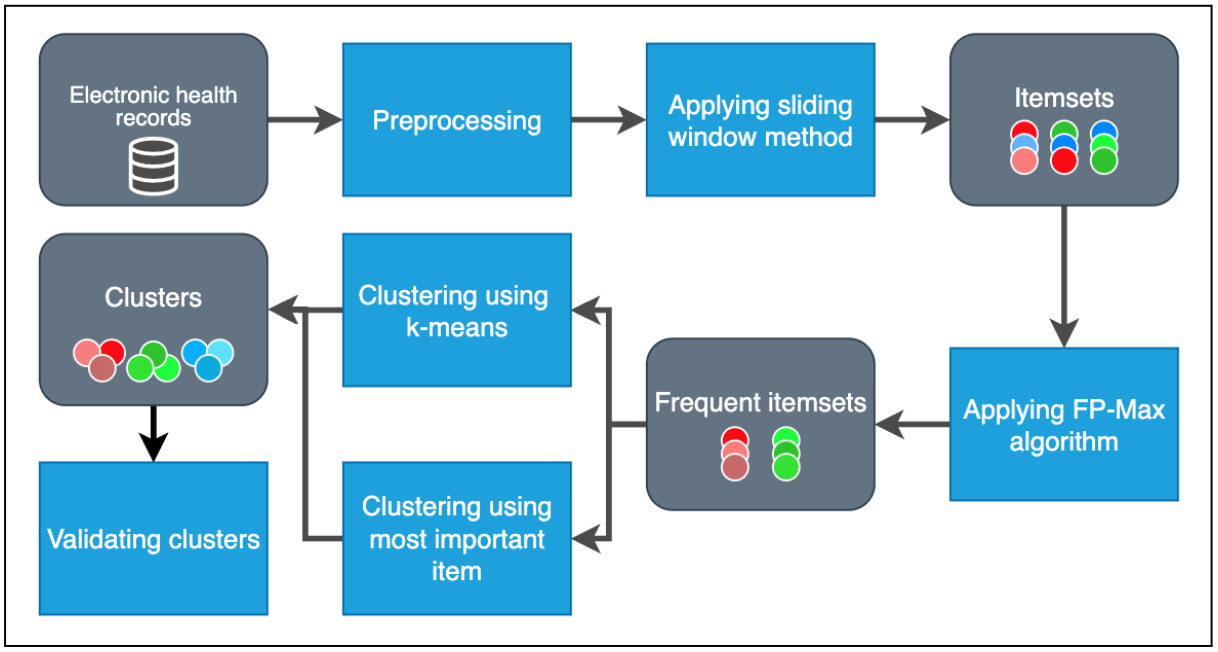

Figure 3. General analysis pipeline of the thesis.

3.2.1 Creating itemsets

A sample size of 10% of all the patients in the database was chosen for this research (n=14,936), as it covers most typical medical concepts observed within the population while maintaining reasonable computational processing times.

For each patient, records are collected from 3 data tables:

condition_occurrence: This table contains records of medical conditions or diseases, including diagnoses, symptoms, or signs. These values are usually inferred from the International Classification of Diseases ICD-10. An example concept name for a clinical finding is "Hyperplasia of the prostate", originating from ICD-code N40.1.

drug_exposure: This table contains records of utilising drugs that are in some way introduced into the patient's body. An example concept name for a drug is "Tamsulosin."

procedure_occurence: This table contains records of procedures and activities that are carried out on the patient by the healthcare professional. An example concept name for a procedure is "Endorectal ultrasonography."

Using our method, all preceding examples would be generalised into a higher-level event related to prostate cancer and treatment.

All records with concepts that appear less than 10 times in the dataset were considered infrequent and filtered out. Consequently, the dates of the events are mapped to integers representing days between the period of 2012-2019 and then converted to a list of events in the format of integer - *concept_id* list pairs. Most records in the database lack precise timestamps, and for the sake of simplicity, the time unit utilised is days.

A sliding window technique is applied to the list of events for creating itemsets. For each patient, the events that appear in the same window within a fixed window length are arranged into itemsets of distinct concepts. Two methods were tested for moving the window along the path of events. The first method uses a fixed-size step equal to the window length, splitting the path into equal-sized sectors (See Figure 4). This would ensure that one event could not belong to multiple itemsets. If no events exist in the window, no itemsets will be created. For the alternative method, every window starts from a new event, including the first event (See Figure 5). This approach is similar to the one used in the Winepi algorithm [2] and allows duplicates.

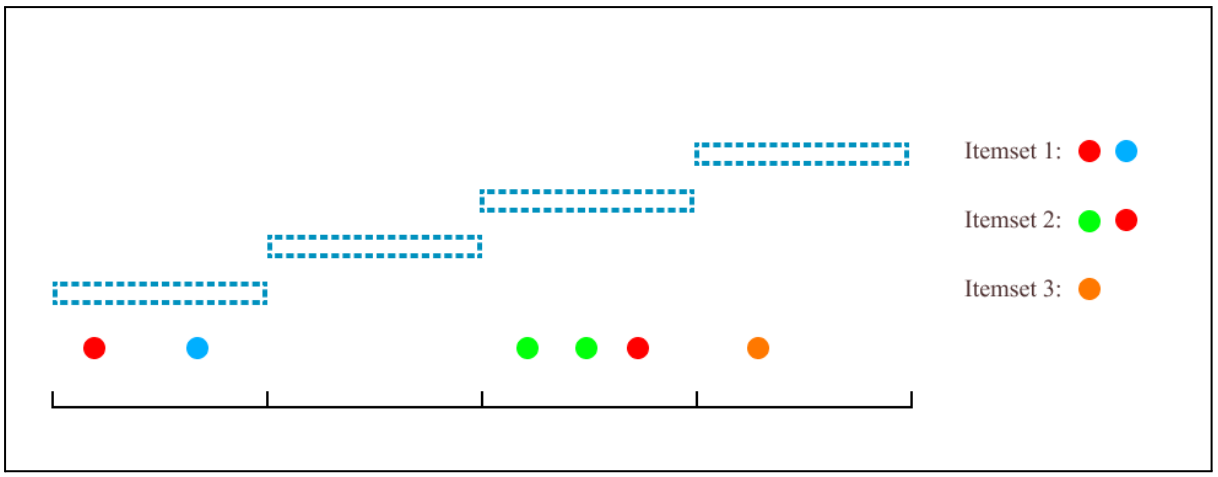

Figure 4. Sliding window method for constructing itemsets. Each dot represents a health event in the patient's timeline.

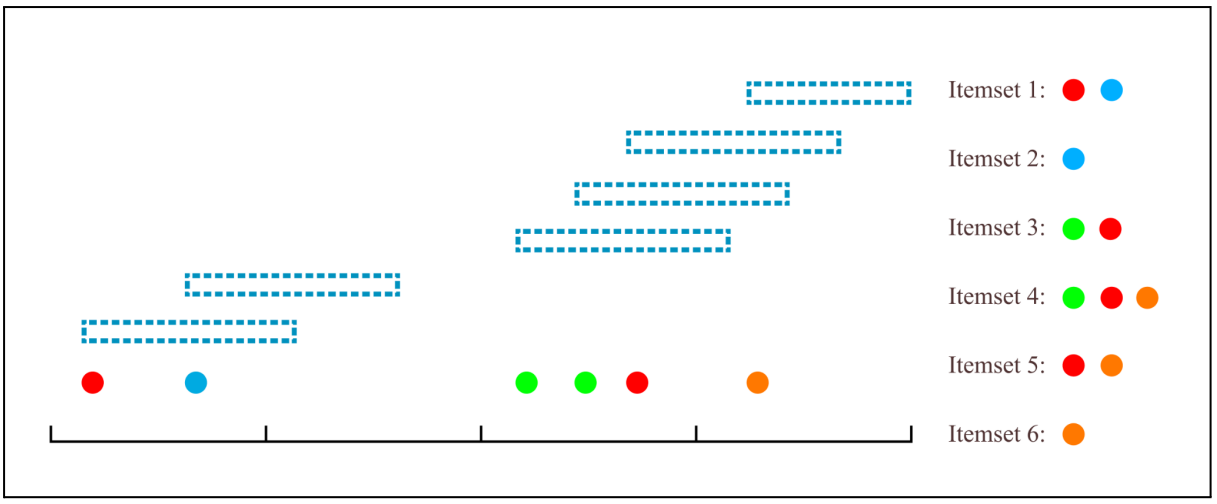

Figure 5. Alternative sliding window method for constructing itemsets. Each dot represents a health event in the patient's timeline.

The following window lengths were used for creating the itemsets: 1, 2, 4, 7, 14, 30.

3.2.2 Frequent itemset mining

The resulting itemsets were first transformed into a one-hot encoded Pandas dataframe [15] which was then inputted into the MLxtend [16] (Machine Learning Extensions) Python library's FP-Max algorithm, along with the minimal threshold support to create frequent itemsets. The itemsets were generated using various minimal supports (0.001, 0.0009, 0.0007, 0.0005, 0.0003), where smaller support led to larger itemsets. Using FP-Max ensured that only frequent super-sets were generated.

3.2.3 Clustering of frequent itemsets

Following the generation of frequent itemsets, clusters that are the base of generalised health events, were created. Two methods were used for clustering.

The first method concatenates all concepts in the itemsets into a string representation. This string representation is then transformed into vectors using the bag-of-words model. Here, each item set resembles a "sentence," with individual items akin to "words." Subsequently, a *k-means* algorithm [17] with a selected number of clusters is applied to the bag-of-words object, enabling the categorisation of distinct clusters.

The following number of clusters were tested in this research: 50, 100, 200, 400.

An alternative method for clustering was also studied. Specifically, within each frequent itemset, the most important concept was identified and the itemsets with the same most

important concept were clustered together. Term frequency-inverse document frequency (TF-IDF) [18] score was used for this, considering each frequent itemset as a document and items as terms.

3.2.4 Proposing a name for each cluster

Chronic diseases often occur frequently alongside less common concepts when dealing with health data, leading to their dispersion across all clusters. To identify the most significant items within the cluster and filter out redundant concepts, TF-IDF score is calculated for each item in the cluster. The cluster is then labelled with the concept name of the item having the highest TF-IDF score. This technique ensures that the labels of the clusters are mostly unique since uninformative concepts are given a lower score.

For the alternative method, clusters that share the same label or the concept with the highest TF-IDF score, are also merged, since multiple frequent itemsets exist for the same concepts.

3.2.5 Validation of the clusters

Ideally, each cluster would exclusively contain concepts that are specific to that cluster. Different metrics that use predefined use cases as validation sets were created to assess the quality and validity of the clusters. All use cases were selected by the author of this thesis and are based on clinical guidelines and discussions with medical doctors. Each use case set comprises concepts that are specific to that particular condition. For example, use case **UC1** contains various concepts related to childbirth (such as "Spontaneous vertex delivery" and "Ultrasound scan") and use case **UC2** contains concepts related to eye diseases (such as "Open-angle glaucoma", "Senile incipient cataract", "Refraction assessment"). Use cases were not created for common chronic diseases like cardiovascular disease and type II diabetes, as their concepts may appear in several clusters, thus not conveying the quality of the clusters. In total, 23 use cases of similar concepts were defined (see Appendix I). All use case concepts are used once only in use cases (no duplicates).

The metrics defined for validating clusters using the predefined use cases:

1. **Average number of clusters per use case.** For each use case, the number of clusters that have matching items are counted. Subsequently, the average of these counts is calculated. The ideal score is 1 (indicating one cluster for each use case). This score helps determine the optimal number of clusters, as ideally, every use case should have one matching cluster.

2. **Standard deviation of clusters per use case.** The number of clusters containing matching items is counted for each use case. Subsequently, the standard deviation is calculated using these counts. This measures the distribution of clusters across use cases. A lower score indicates a better distribution.
3. **Average number of use cases per cluster.** For each cluster, the number of use cases containing matching items is counted. Subsequently, the sum of the counts is divided by the number of clusters with at least one matching item with any of the use cases. This score assists in determining the optimal number of clusters as the goal is to avoid the concepts of several use cases being presented in the same cluster. The ideal value of the metric is the number of use cases divided by the total number of clusters.
4. **Standard deviation of use cases per cluster.** For each cluster, the number of use cases containing matching items is counted. Subsequently, the standard deviation is calculated using these counts. This measures the distribution of use cases across clusters. A lower score indicates a better distribution.

Two more indicative metrics were created for deciding the minimal support threshold:

1. **Use case concepts in frequent itemsets.** The proportion of all use case items included in all items in frequent itemsets. This indicates how many of the predefined use case concepts actually are represented in the frequent itemsets. The ideal score is 1, meaning all the use cases are represented in the itemsets.
2. **Proportion of all concepts included in clusters.** The proportion of all initial items included in clusters. This is calculated by dividing the number of initial items that appear in clusters by the number of items from the initial data. This metric considers the actual frequencies of the items in the original data, better indicating the proportion of the original data covered by the clusters (the larger the better).

Parameters and test values used for validation of the clusters are shown in Table 5. These values were chosen to keep in mind the goal of this thesis and based on how general the resulting events would be.

Table 5. Parameters and values used for testing.

Param	Values
Window length	1, 2, 4, 7, 14, 30
Number of clusters	50, 100, 200, 400
Minimum support	0.001, 0.0009, 0.0007, 0.0005, 0.0003

Each combination of these parameters was iterated three times on different samples of healthcare data to ensure robustness. The average value was calculated from these iterations.

3.2.6 Generalising health events of a single patient

To generalise the health events of a single patient using the clusters of similar concepts, the events on the timeline are first split into itemsets using a sliding window, similar to the previous approach. Subsequently, each item in the set is allocated to a cluster based on where it attains the highest TF-IDF score. Ideally, all events within a window would belong to one cluster. In cases where no clusters contain the event, a "NaN" value is assigned. The result of generalising the timelines of two persons shown in Figure 1 and Figure 2 can be seen in the following section, "Results".

3.3 Location of code

The Python code implemented in this study is available in the GitHub repository[1].

3.4 Ethics approval

This work was approved by the Estonian Bioethics and Human Research Council (EBIN, no. 1.1-12/653).

[1] https://github.com/OliverSuik/fim-health-data

4 Results

This section displays, summarises, and discusses the results of the thesis. It starts with comparing the metric scores using different values for the parameters. Then the results of using an alternative method for creating itemsets and clustering are presented followed by examples of clusters. Lastly, generalised representations of the previous health event timelines in Figure 1 and Figure 2 are shown.

For the 10% sample of the population (n=14,936), the first sliding window method created around 950,000 transactions. From these transactions, 558 frequent itemsets were generated when the window length was set to 1 and with the minimum support is 0.0008. Using the same minimum support, but having a window length of 30, 2187 frequent itemsets were generated.

Table 6 shows the number of unique concepts contained in frequent itemsets using different minimum supports. It can be seen that lowering the minimum support increases the number of unique concepts.

Table 6. Number of concepts in frequent itemsets using different minimum support.

Minimum Support	Number of Unique Concepts
0.015	10
0.01	13
0.005	57
0.001	474
0.0005	743
0.0008	1191
0.0001	1738

4.1 Relationships between parameters and metrics

Relationships between different parameter values and metrics are presented in the following section.

Figure 6 shows that increasing the window length results in higher scores for all metrics, indicating that more concepts appear in clusters but the quality of the clusters declines. The first two metrics begin to increase significantly when the window length is 7, while the other metrics show slower growth.

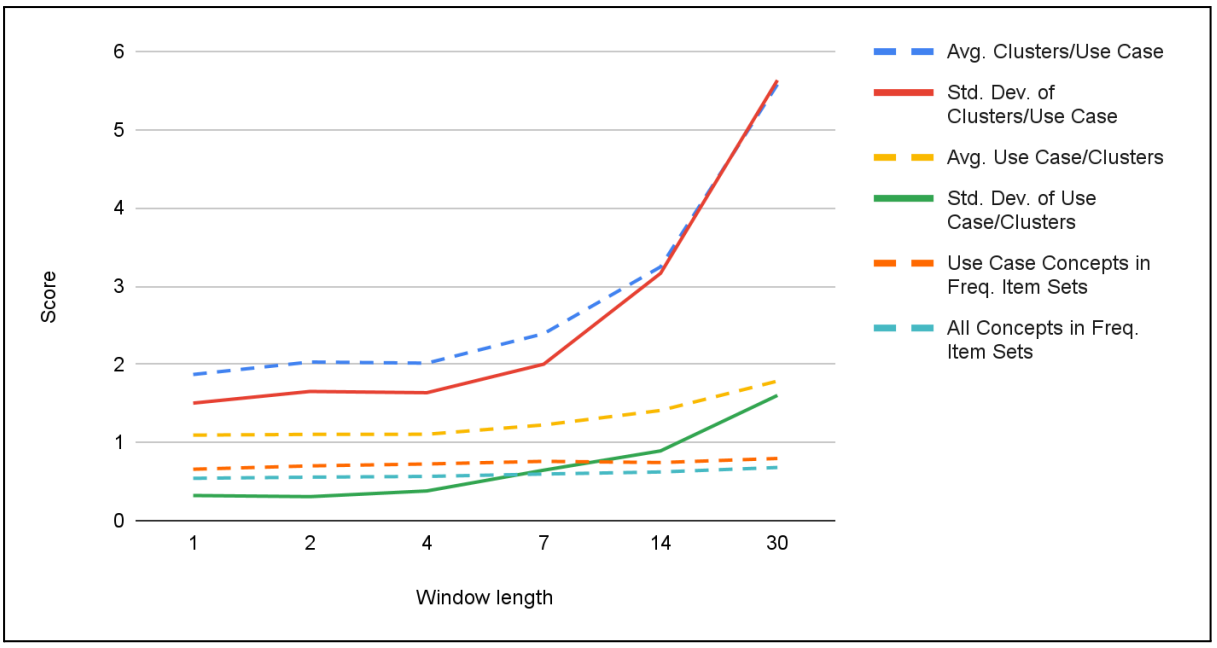

Figure 6. Relationships between window length and metric scores. The number of clusters is fixed at 200 and the minimum support is 0.0008. Dashed lines in the charts represent metrics with an ideal score of 1.

Figure 7 illustrates that increasing the minimum support decreases the values of the last two metrics from 86.9% and 74.2% (min_sup = 0.0003) to 70.65% and 56.7% (min_sup = 0.001), respectively. Conversely, the quality of the clusters improves significantly, with the first and third metrics approaching a value of 1.

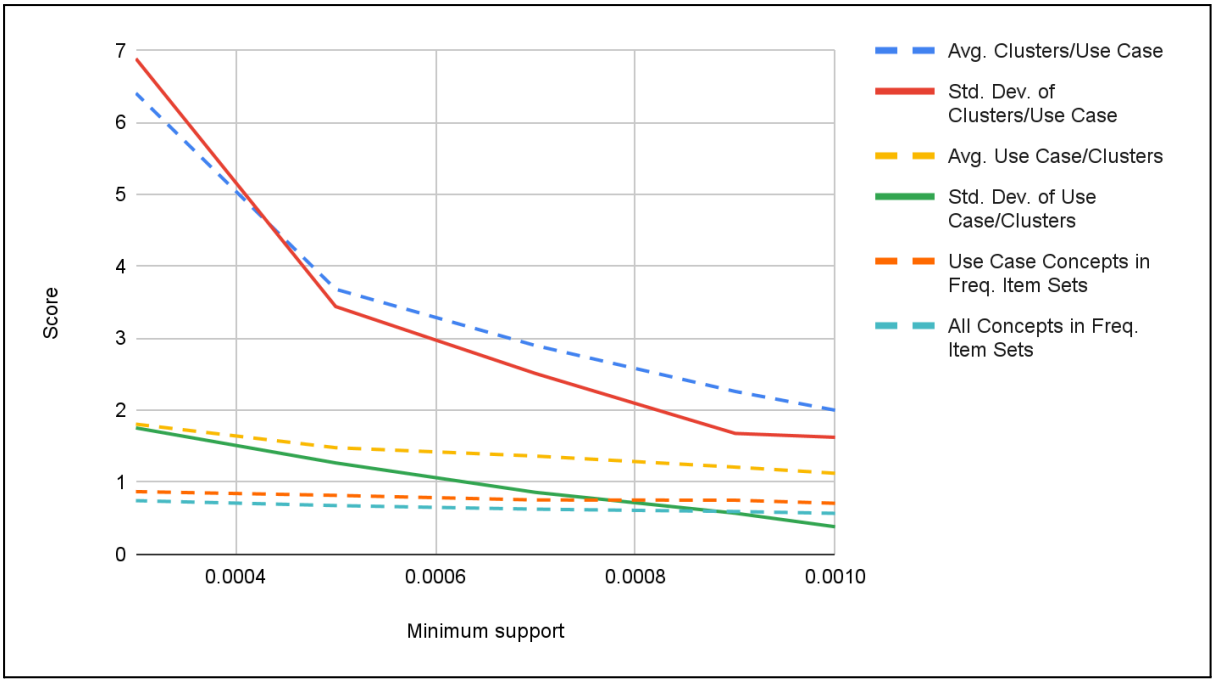

Figure 7. Relationships between minimal support and metric scores. Number of clusters is fixed to 200 and the window length is 7. Dashed lines in the charts represent metrics with an ideal score of 1.

Compared to the previous graphs, Figure 8 demonstrates that increasing the number of clusters affects the first two metrics differently compared to the other two metrics, with the first metrics increasing while the others decreasing. The metrics scores are the most similar when the number of clusters is 75.

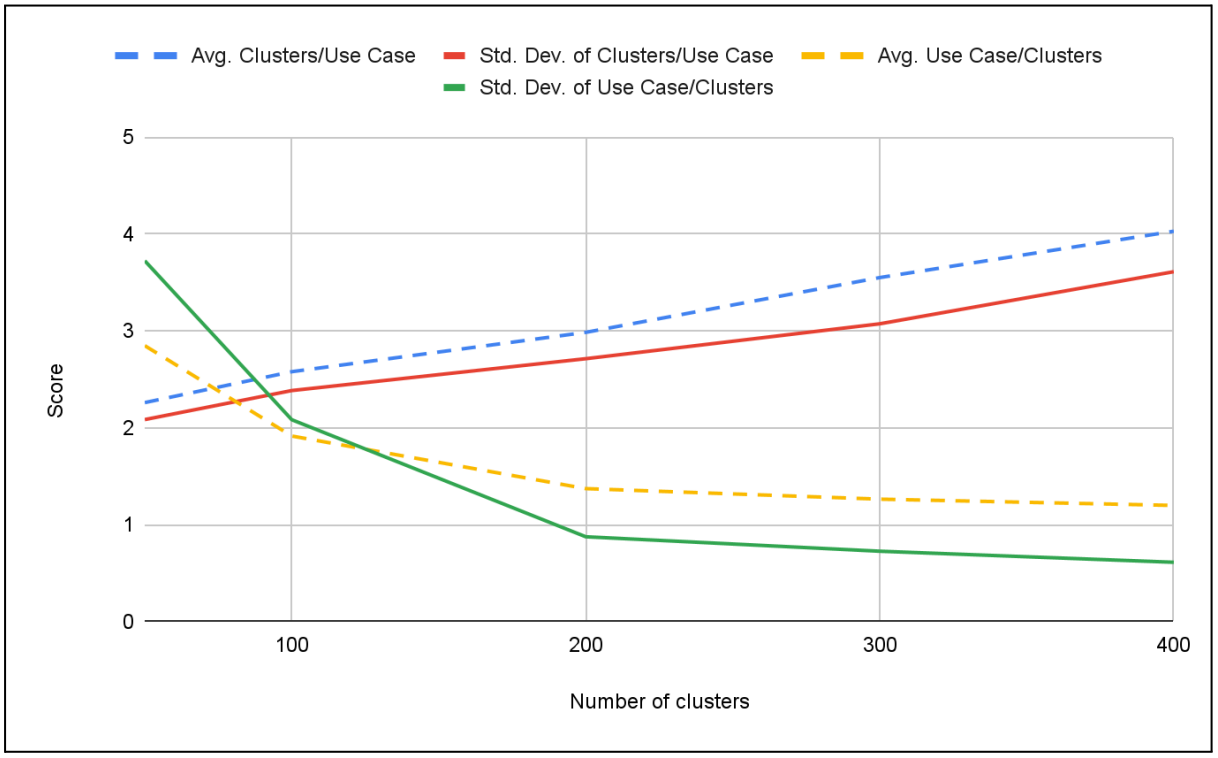

Figure 8. Relationships between the number of clusters and metrics. The minimum support is 0.001 and the window length is 7. Dashed lines in the charts represent metrics with an ideal score of 1.

4.2 Using alternative methods

Using the alternative sliding window method created around 1,100,000 transactions, which is 150,000 more than the first method.

Figure 9 illustrates that employing the alternative sliding window method enhances the scores of the last two metrics by 20% and 15% with every window length, compared to the first method. However, as the window length is increased, all the other scores start to degrade significantly, especially the first two metrics.

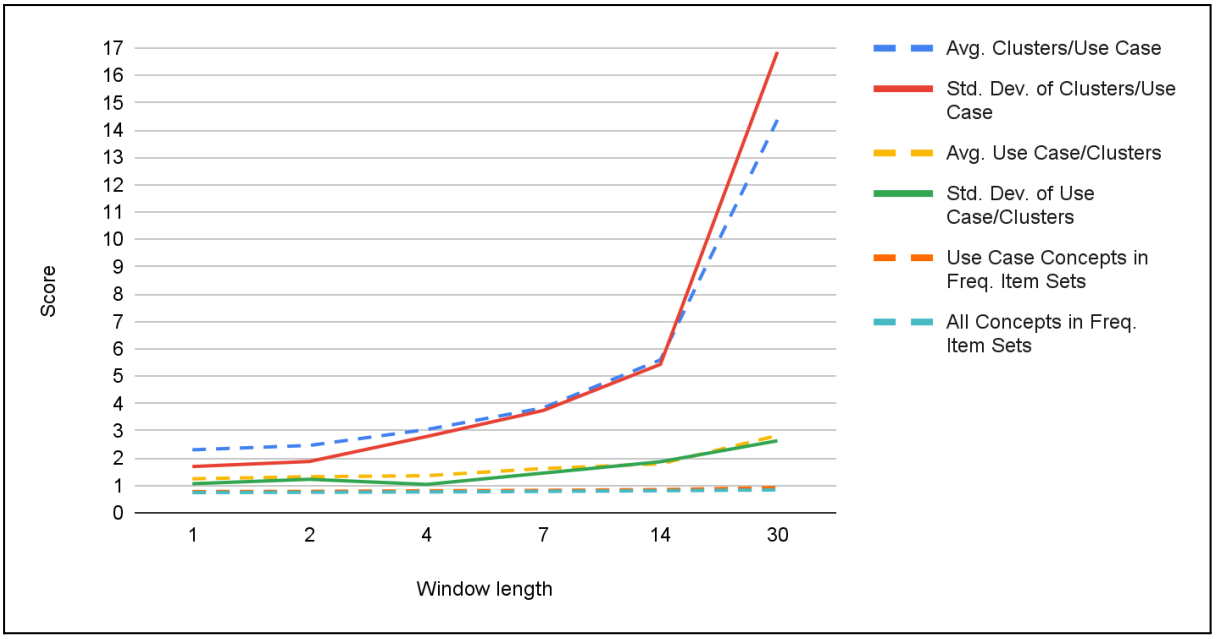

Figure 9. Relationships between window length and metrics using an alternative sliding window method. The number of clusters is fixed at 200 and the minimum support is 0.0008. Dashed lines in the charts represent metrics with an ideal score of 1.

When using the alternative clustering method, out of the initial 558 frequent itemsets (win_length=1), 234 itemsets remained after merging the itemsets containing the same concept with the highest TF-IDF score. An average cluster contained 4 concepts. The metric scores using different window lengths are shown in Figure 10.

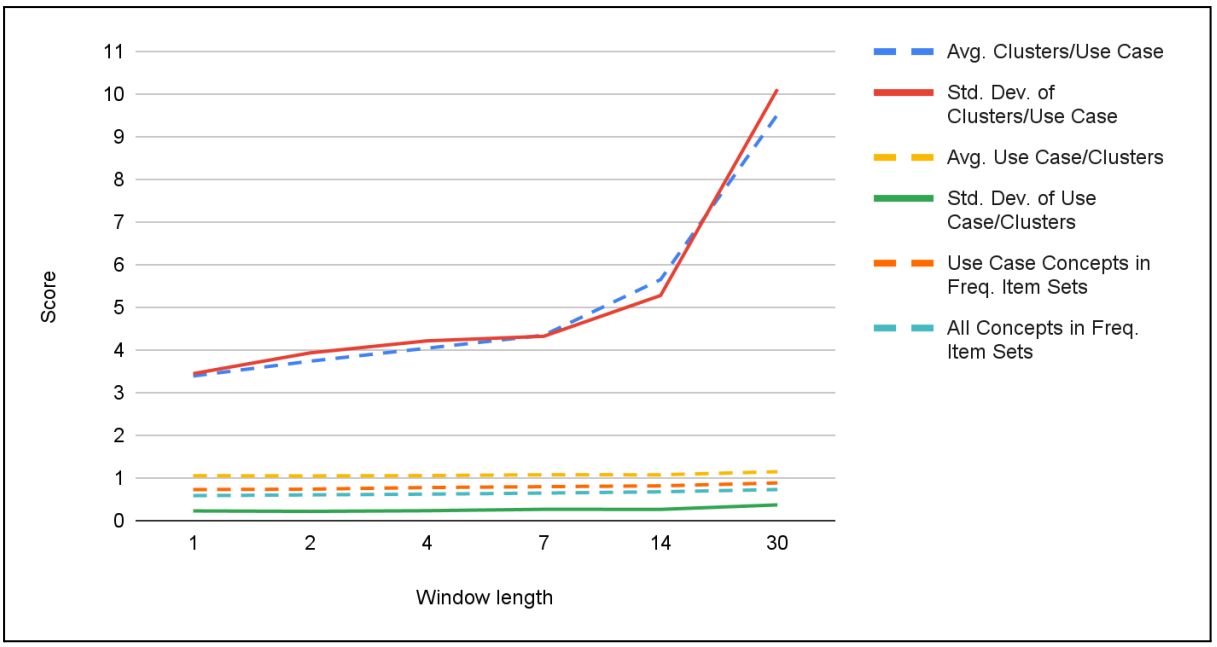

Figure 10. Relationships between window length and metrics using an alternative clustering method. Minimum support is 0.0008. Dashed lines in the charts represent metrics with an ideal score of 1.

4.3 Resulting clusters

Based on the scores obtained using different values for the parameters and the preferences of the thesis author, clusters for closer inspection were created using the parameters *win_length*=4, *num_clusters*=200, and *min_sup*=0.0008. The sizes of the clusters vary from a single concept to 50 concepts, with the average cluster size consisting of 12 concepts. Examples of the resulting clusters, showing the top 7 concepts in each, can be seen in Table 7, Table 8 and Table 9. The given labels for the clusters are at the top of the tables.

Table 7. Cluster containing concepts related to pregnancy and childbirth.

Nr	TF-IDF Score	Concept
1	1.0	**Finding related to pregnancy**
2	0.144	Spontaneous vertex delivery
3	0.144	Infections of the genital tract in pregnancy
4	0.144	Gestational diabetes mellitus
5	0.144	Threatened miscarriage
6	0.144	Uterine scar from previous surgery in pregnancy, childbirth and the puerperium
7	0.125	Ultrasound scan - obstetric

Table 8. Cluster containing concepts related to treatment using antibiotics.

Nr	TF-IDF Score	Concept
1	1.0	**Amoxicillin**
2	0.144	Asymptomatic periapical periodontitis
3	0.077	Amoxicillin 80 MG/ML / Clavulanate 11.4 MG/ML Powder for Oral Solution
4	0.077	Acute pharyngitis

5	0.077	Acute tonsillitis
6	0.077	Acute sinusitis
7	0.059	Acute bronchitis

Table 9. Cluster containing concepts related to chronic diseases.

Nr	TF-IDF Score	Concept
1	0.799	**Nebivolol**
2	0.799	Nebivolol 5 MG Oral Tablet
3	0.062	Primary open angle glaucoma
4	0.054	Type 2 diabetes mellitus without complication
5	0.077	Metformin
6	0.047	Disorder of lipid metabolism
7	0.04	Hypertensive heart disease

4.4 Generalised timelines

Using the same parameter values as in the previous section, generalised versions of timelines similar to those seen in Figure 1 and Figure 2 were created. These can be observed in Figure 11 and Figure 12.

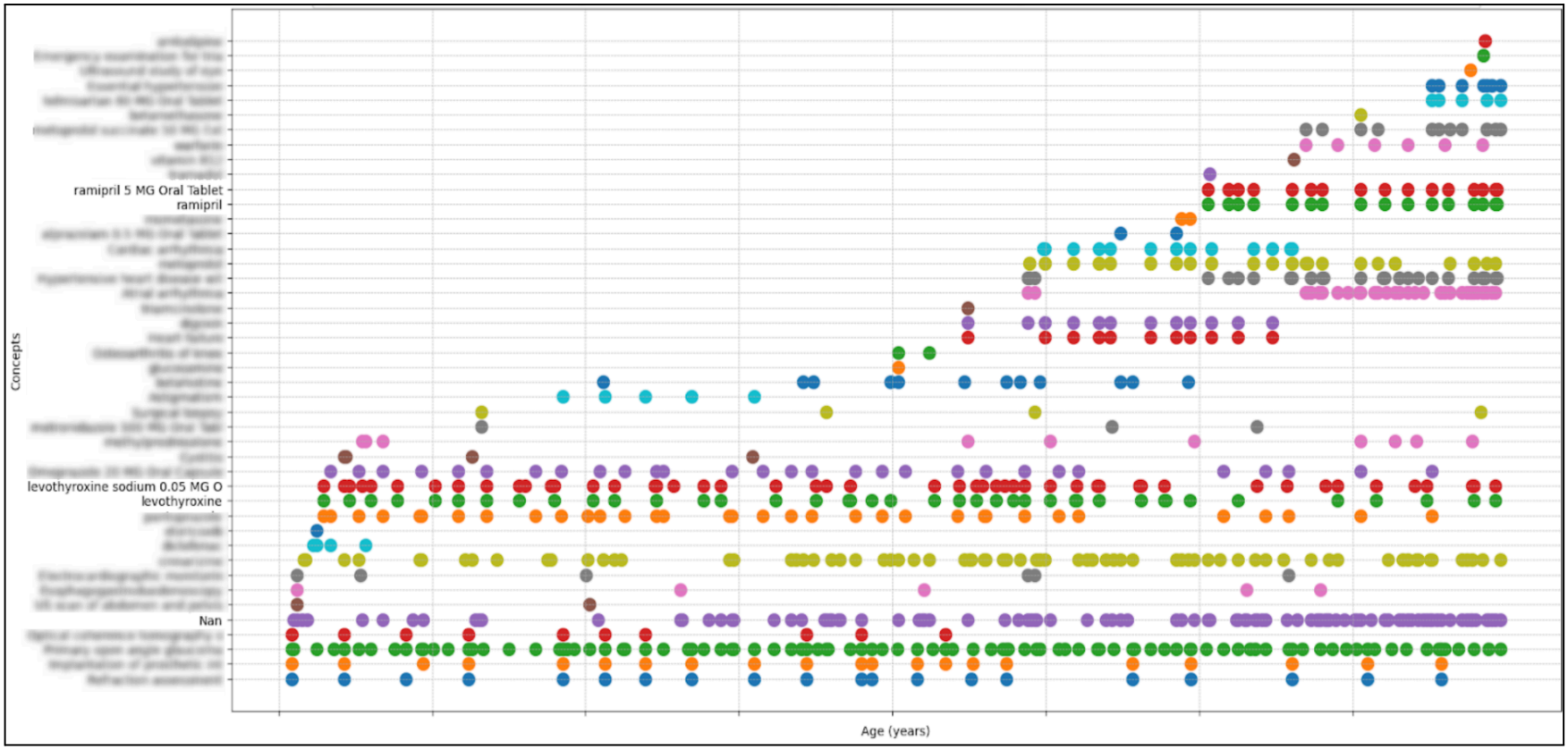

Figure 11. Generalised health event timeline of an elderly person.

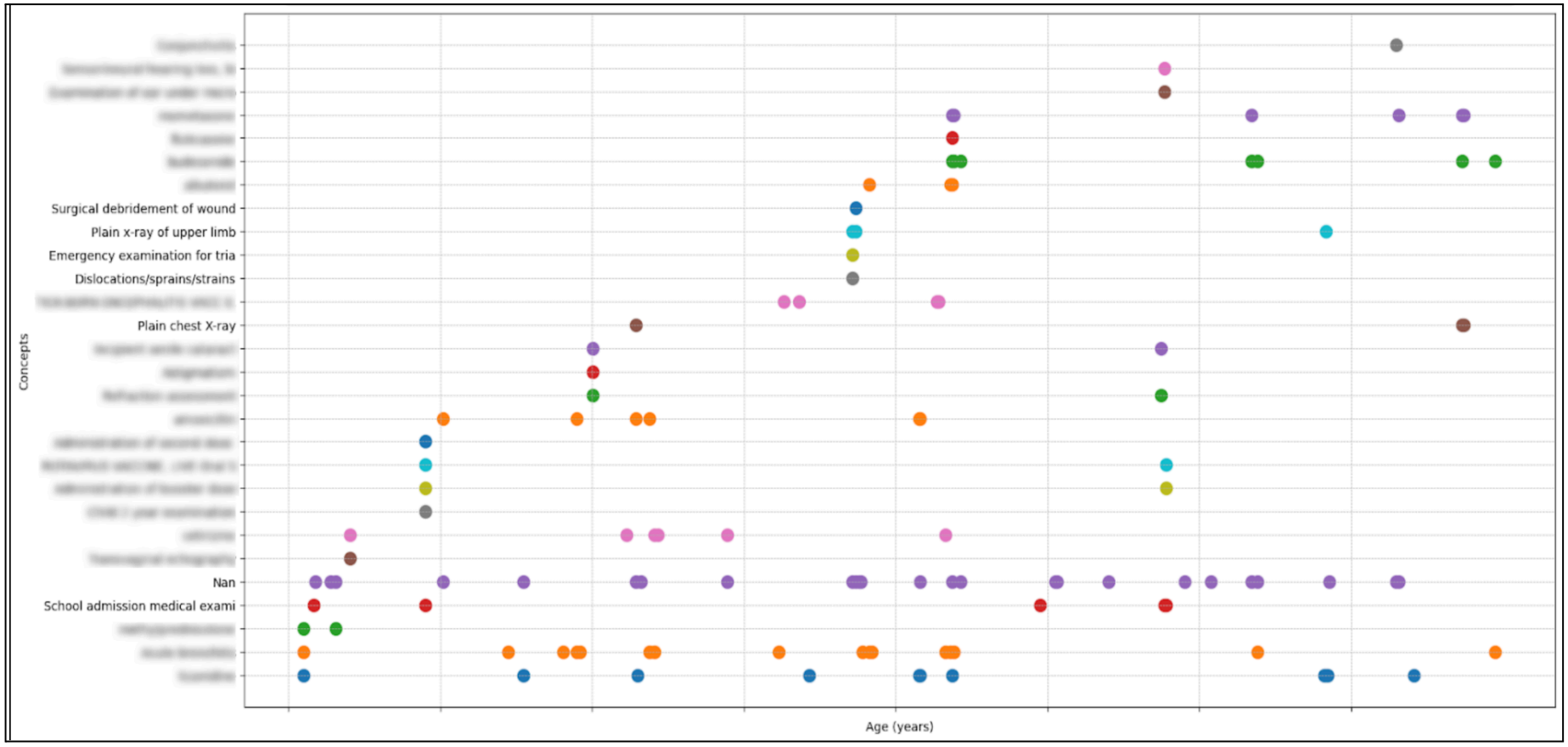

Figure 12. Generalised health event timeline of a child.

5 Discussion

This section analyses the results presented in the previous section, answers research questions, and compares the method proposed in this thesis with similar work. In addition, future work and possible use cases for the method are discussed.

5.1 Interpreting results

5.1.1 Choosing parameter values and alternative methods

Several factors must be considered when selecting parameter values and defining use cases. For example, increasing the window length leads to larger itemset transactions, generating more candidates for the algorithm to combine into frequent itemsets. This results in more concepts within clusters; however, it also causes the clusters to grow in size and become more mixed with concepts from different use cases since the events become more general.

Lowering the minimum threshold support also increases the number of concepts within clusters, exposing rarer events. Conversely, setting the minimum support too high would lead to a lack of frequent itemsets. As seen in Figure 11 and Figure 12, a significant number of concepts are given a "NaN" value, since they are not present in frequent itemsets. In this research, the minimum support value should be set to at least 0.015 (1.5%) to obtain any frequent itemsets (see Table 6).

The number of clusters should be determined based on the number of concepts within frequent itemsets. Setting this value too low may result in noisy and overly generic clusters while setting it too high could create anomalies with clusters containing only a single element.

As seen in Figure 9, using the alternative sliding window method, compared with the first method, resulted in more concepts due to the duplication, which led to larger itemsets. However, this comes with the expense of "Avg. Clusters/Use Case" increasing rapidly, compared to the first method. This could have occurred because the method begins to amplify the occurrence of frequent events in itemsets. If an event occurs clearly separately from others, it is in exactly one itemset. But if there are two other events around this event, then this event already occurs in three itemsets.

The advantages of using the alternative clustering method, which prioritises the most important item, include eliminating the need to choose the number of clusters, which can be

challenging in the first method. Even when the minimal support is set high and a large majority of the concepts are frequent, the resulting clusters remain easily distinguishable compared to the first method. However, the clusters are smaller and more generalised at a lower level. This can be observed in Figure 10, where the "Avg. Clusters/Use Case" score is very high because the concepts in the use cases are distributed across multiple clusters. In contrast, the "Avg. Use Case/Clusters" score is nearly optimal, reflecting the small size of the clusters. To conclude, this approach is suitable when the defined use cases are small, containing up to 4 concepts.

5.1.2 Resulting clusters

It should be considered that many of the resulting clusters represent the same generalised events but with slight differences. These typically include clusters of chronic diseases, as illustrated in Table 9. The number of these clusters can be reduced by adjusting the parameters.

In the cluster in Table 8 and Table 9, it can be seen that multiple variants of a drug can appear in the system. This typically occurs when a drug is reported multiple times: once when prescribed with a general name, and again when purchased from a pharmacy under a specific name. Using the method of this thesis, drugs and their variants are clustered together, however as seen in Figure 11, for clusters labelled "Ramipril" and "Levothyroxine", clusters can also be created for their variants and appear alongside the timeline.

Using the TF-IDF measure to identify the most informative concepts for labelling clusters yields mixed results. In Table 7, the suggested label "Finding related to pregnancy" effectively describes the concepts associated with pregnancy and childbirth in a generalised context. Similarly, the label "Amoxicillin" in Table 8 is appropriate, as the cluster comprises conditions treated with antibiotics, making it sensible to use the name of the drug used as treatment as the label. Some labels are less self-explanatory initially, such as "Plain chest X-ray" in Figure 12. Upon closer inspection, the cluster's name includes concepts related to respiratory system diseases, making the proposed label somewhat accurate since it involves this particular diagnostic procedure. However, some labels are from concepts that indeed are part of the general event but may not be highly significant, for example, "Nebivolol" in Table 9, which is a drug used to treat high blood pressure. This cluster is also a mix of concepts of eye diseases and type II diabetes which appear frequently together but do not contain a concept with a name that describes them all, making it challenging to label.

5.1.3 Generalised timelines

Applying the method to the initial timelines significantly improved the clarity of the pictures. In the child's timeline shown in Figure 12, multiple distinguishable events common to children are visible that were not apparent in the initial figure. For instance, the timeline includes events related to school nurse examinations and vaccinations in line with the national vaccination plan. Concurrent events in the middle of the diagram suggest that the person underwent an emergency procedure (involving concepts such as "Dislocations/sprains/strains," "Emergency examination for triage," "Plain X-ray of upper wound," and "Surgical debridement of wound"). The original figure contained 266 events, which were reduced to 96, with 23 assigned a "NaN" value.

The event timeline of an elderly person exhibits fewer distinguishable events but remains more interpretable than the initial image, revealing occurrences of several chronic diseases and associated procedures and drug prescriptions. The original 1550 events were generalised to 381 events, with 89 classified as "NaN." This illustrates that using our model, frequent items are more likely to be clustered than rarer items, aligning with the objectives of this thesis.

5.2 Reflections and limitations

The primary focus of this research was to explore the potential of frequent itemset mining in generalising Estonian healthcare data. Through developing and implementing a health event generalisation method that combines frequent itemset mining and clustering techniques, this thesis aims to provide insights into how this approach can enhance the representation and analysis of healthcare data. This process addresses **RQ1.** *How can frequent itemset mining be used to generalise Estonian healthcare data?*

In this work, a health event generalisation method was conducted on an Estonian dataset for the first time. The technique used in the thesis can also be applied to health data of other nations, considering that globally, as of 2023, 12% of health data is standardised to the OMOP format [19]. The method reduced the number of events appearing in the same time window and made the timeline of health events easier to interpret. Also, the use cases created to validate the results can be used in other works to generalise health events.

Furthermore, its implementation can notably enhance the efficiency of displaying health data within health information systems. By offering users of the portal a comprehensive overview

of a patient's health status, it significantly enhances the system's utility and effectiveness. A new version of Estonia's health information system is being developed at the time of this research and the data viewer function could benefit from this model [20].

Although the results of this thesis may not be directly applicable in practical contexts, studies requiring the manual creation of generalised health event reference sets can benefit from this work by automating the process.

In Salamov's thesis [6], the author classified concepts related to cervical cancer screening into activities and used them in process mining. When comparing the defined activities with the clusters in this thesis, some show similar generalisations or overlaps. For example, in the activity "Contraception based on ICD10 diagnosis", out of 57 concepts, 4 are present in the cluster labelled "Contraception care management". Among the other 4 concepts in the cluster, 2 are indirectly related ("Transvaginal echography", "DNA analysis"), one is a drug variant ("Ethinyl estradiol 0.02 MG / gestodene …"), and one is the drug "Gestodene," which could belong to the activity as it is used for birth control. This shows that the method could even be used to discover initially left-out relevant concepts.

Some clusters are even identical to the activities, as one of them labelled "Cervical biopsy" contains the exact same concepts ("Cervical biopsy", "Colposcopy") as the activity "CIN or CC related diagnostic procedures".

While frequent itemset mining showed promise in generalising Estonian healthcare data, several limitations emerged during the research process. The following section addresses **RQ2.** *What limitations restrict the implication of frequent itemset mining to Estonian healthcare data?*

The limitations that emerged are related to the initial data, the chosen method, and fundamental problems.

A majority of the concepts in the dataset are rare and do not appear unless the minimal support is set very low. Approximately 5.5% of all concepts in the dataset appear only once. Furthermore, the research was conducted using only Estonian health data, which is limited in scope and lacks comprehensive coverage of various diseases. The dataset also contained records of measurements; however, due to challenges in their incorporation and validation, they were not analysed.

One of the cons of using frequent itemset mining as was demonstrated in this work is that there is no direct way to control the degree of generalisation besides the parameters. It is hard

to tell whether using k-means to cluster itemsets and TF-IDF to label these clusters is the best approach, as there is no validation set to compare the results to and objectively assess if the methods used in this research are sufficient. The metrics developed in this thesis can be applied within predefined use cases; however, there is no metric similar to ROUGE, as used in the research by Gupta et al [13].

In addition, there should have been many more use cases of generalised events created with medical professionals' help. One reason is that determining whether a cluster can be considered generalised is challenging (clusters containing 2 concepts and those containing 20 are both deemed generalised).

5.3 Future work

In terms of future work, there are several directions for advancing this research. First, exploring different clustering methods for grouping frequent itemsets could reveal new insights. Comparative studies using various algorithms like hierarchical clustering [21] or Density-based spatial clustering of applications with noise algorithms (DBSCAN) [22] may reveal different formations of clusters. Additionally, experimenting with new methods for proposing cluster names beyond TF-IDF could enhance the interpretability of results.

In this thesis, health events on the timeline were reduced only to fixed time windows. Grouping these generalised events horizontally, creating a method for determining the start and end dates of these events, and constructing a validation method is another key area for improvement. This would help to eliminate the problem of inaccurate timestamps for events.

6 Conclusion

The main aim of this thesis was to evaluate the use of frequent itemset mining techniques to generalise healthcare data. In addition to creating a model and validation methods for examining the feasibility, the results of using different values for parameters and limitations were discussed.

To explore the applicability of frequent itemset mining, frequent itemsets were generated from Estonia's Electronic Health Records using an FP-Max algorithm, which were then transformed into clusters representing higher-level events. The resulting clusters were utilised to generalise a patient's health event timeline.

Exploring different parameter values led to clusters of varying sizes, levels of generalisation, and numbers of contained concepts. Therefore, depending on the task, one has to find the best balance between the outcome metrics when choosing the parameters for the method. The method presented in this thesis enables investigators to build their own test cases and generate clusters of higher-level events that can be applied in their work or further develop the methods proposed in this thesis. The main limitations identified with this approach included the exclusion of rare events in frequent itemsets and the need for an evaluation metric to assess the results of the chosen methods objectively.

Ultimately, the implemented solution helped to describe a patient's health status by reducing elements and giving them generalised labels, demonstrating the potential applicability of frequent itemset mining in this context.

7 Acknowledgements

This work was supported by the Estonian Research Council (grant numbers PRG1844 and RITA1/02-96).

References

[1] Agrawal, R., Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. *In Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487–499, 1994. https://dl.acm.org/doi/10.5555/645920.672836

[2] Mannila, H., Toivonen, H., Verkamo, I. Discovery of Frequent Episodes in Event Sequences. *In Data Mining and Knowledge Discovery 1*, pp. 259–289, 1997. https://doi.org/10.1023/A:1009748302351

[3] Han, J., Pei, J., Yin, Y. et al. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *In Data Mining and Knowledge Discovery 8*, pp. 53–87, 2004. https://doi.org/10.1023/B:DAMI.0000005258.31418.83

[4] Grahne, G., Zhu, J. Efficiently Using Prefix-trees in Mining Frequent Itemsets. *In Workshop on Frequent Itemset Mining Implementations*, 2003. https://api.semanticscholar.org/CorpusID:58722470 (06.05.2024)

[5] Künnapuu, K., Ioannou, S., Ligi, K., Kolde, R., Laur, S., Vilo, J., Rijnbeek, P., Reisberg, S. Trajectories: a framework for detecting temporal clinical event sequences from health data standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model. In JAMIA Open, Vol 5, No. 1, 2022. https://doi.org/10.1093/jamiaopen/ooac021

[6] Salamov, M. Process mining on Estonian healthcare data. Master's thesis, University of Tartu, 2023. https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=77101 (06.05.2024)

[7] Tóth, K., Machalik, K., Fogarassy, G., Vathy-Fogarassy, A. Applicability of process mining in the exploration of healthcare sequences. *In 2017 IEEE 30th Neumann Colloquium (NC)*, pp. 151-156, 2017. https://doi.org/10.1109/NC.2017.8263273

[8] Bailey LC., Milov DE., Kelleher K., et al. Multi-Institutional Sharing of Electronic Health Record Data to Assess Childhood Obesity. *In PLoS One,* Vol 8, No. 6, 2013. https://doi.org/10.1371/journal.pone.0066192

[9] Campbell, E., Bass, E., Masino, A. Temporal condition pattern mining in large, sparse electronic health record data: A case study in characterizing pediatric asthma. *In journal of the American Medical Informatics Association*, Vol 27, No. 4, pp. 558–566, 2020. https://doi.org/10.1093/jamia/ocaa005

[10] Li, Y., Schofield, E., Gönen, M. A tutorial on Dirichlet process mixture modeling. *In Journal of Mathematical Psychology*, Vol 91, pp. 128-144, 2019. https://doi.org/10.1016/j.jmp.2019.04.004

[11] Brin, S., Page, L. The anatomy of a large-scale hypertextual Web search engine. *In Computer Networks and ISDN Systems*, Vol 30, No. 1, pp. 107-117, 1998. https://doi.org/10.1016/S0169-7552(98)00110-X

[12] Lin, CY. ROUGE: A Package for Automatic Evaluation of Summaries. *In Text summarization branches out (WAS 2004)*, pp. 74–81, 2004. https://aclanthology.org/W04-1013.pdf (06.05.2024)

[13] Gupta, S., Sharaff, A., Nagwani, N.K. Frequent item-set mining and clustering based ranked biomedical text summarization. *In The Journal of Supercomputing*, Vol 79, pp. 1–21, 2022. https://doi.org/10.1007/s11227-022-04578-1

[14] Oja, M., Tamm, S., Mooses, K., Pajusalu, M., Talvik, HA., Ott, A., Laht, M., Malk, M., Lõo, M., Holm, J., Haug, M., Šuvalov, H., Särg, D., Vilo, J., Laur, S., Kolde, R., Reisberg, S. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. *In JAMIA Open*, Vol 6, No. 1, 2023. https://doi.org/10.1093/jamiaopen/ooad100

[15] McKinney, W. Data Structures for Statistical Computing in Python. *In Proceedings of the 9th Python in Science Conference*, pp.56-61, 2010. https://doi.org/10.25080/Majora-92bf1922-00a

[16] Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. The Journal of Open Source Software, Vol 3, No. 24, 2018. https://doi.org/10.21105/joss.00638

[17] MacQueen, J. B. Some Methods for classification and Analysis of Multivariate Observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297, 1967. https://api.semanticscholar.org/CorpusID:6278891 (06.05.2024)

[18] G. Salton, C. Buckley. Term-weighting approaches in automatic text retrieval. *In Information Processing & Management*, Vol. 24, No. 5, pp. 513–523, 1988. https://doi.org/10.1016/0306-4573(88)90021-0

[19] OHDSI. OHDSI. Our Journey. Where the OHDSI Community Has Been and Where We Are Going. 2023. https://www.ohdsi.org/wp-content/uploads/2023/11/OHDSI-Book2023.pdf (06.05.2024)

[20] Health information system. https://www.tehik.ee/en/health-information-system (06.05.2023)

[21] R. Sibson, SLINK: An optimally efficient algorithm for the single-link cluster method. *In The Computer Journal*, Vol. 16, No. 1, pp. 30–34, 1973. https://doi.org/10.1093/comjnl/16.1.30

[22] Ling, R. On the theory and construction of k-clusters. *In The Computer Journal*, Vol. 15, No. 4, pp. 326-332, 1972. https://doi.org/10.1093/comjnl/15.4.326

Appendix

I. Table of use cases

ID	Concept	Relation
1	● 4062484 - Screening for malignant neoplasm of cervix ● 44789520 - Human papilloma virus nucleic acid detection ● 43531329 - Microscopic cytologic examination of smear of specimen from female genital tract prepared using Papanicolaou technique ● 4023405 - Cytologic test ● 4162714 - Negative for intraepithelial lesion or malignancy ● 4191603 - Atypical squamous cells of undetermined significance ● 4161591 - Atypical squamous cells, cannot exclude HSIL ● 196359 - Primary malignant neoplasm of uterine cervix ● 4175471 - Cervical biopsy	Human papillomavirus (HPV)
2	● 4278515 - Biopsy of prostate ● 924566 - Gamsulosin ● 40222572 - Dutasteride 0.5 MG / tamsulosin hydrochloride 0.4 MG Oral Capsule ● 989482 - Dutasteride ● 197032 - Hyperplasia of prostate ● 40166540 - Tamsulosin hydrochloride 0.4 MG Oral Capsule	Prostata cancer
3	● 1167322 - Allopurinol ● 440674 - Gout ● 1167323 - Allopurinol 100 MG Oral Tablet ● 19018787 - Allopurinol 300 MG Oral Tablet	Gout and treatment
4	● 939259 - Budesonide ● 1196677 - Formoterol ● 4110051 - Mixed asthma ● 317009 - Asthma ● 783232 - Budesonide 0.16 MG/ACTUAT / formoterol 0.0045 MG/ACTUAT Inhalation Solution ● 36405113 - Budesonide 0.4 MG Inhalant Powder ● 19063315 - Budesonide 0.064 MG/ACTUAT Nasal Spray	Asthma and treatment

	• 36778958 - 60 ACTUAT Budesonide 0.32 MG/ACTUAT / formoterol 0.009 MG/ACTUAT Inhalant Powder Box of 1 • 1154161 - Montelukast • 19023368 - Montelukast 10 MG Oral Tablet • 1137529 - Salmeterol	
5	• 19044883 - Zopiclone • 374905 - Non-organic sleep disorder • 439708 - Disorders of initiating and maintaining sleep • 744740 - Zolpidem • 40163492 - Zolpidem tartrate 10 MG Oral Tablet • 19044885 - Zopiclone 7.5 MG Oral Tablet	Sleeping disorders and treatment
6	• 1713332 - Amoxicillin • 19073188 - Amoxicillin 500 MG Oral Tablet • 1713412 - Amoxicillin 1000 MG Oral Tablet • 44033419 - Amoxicillin 50 MG/ML Oral Powder • 19115197 - Amoxicillin 875 MG / clavulanate 125 MG Oral Tablet	Antibiotics
7	• 444094 - Finding related to pregnancy • 440457 - Threatened miscarriage • 4205240 - Spontaneous vertex delivery • 4060556 - Uterine scar from previous surgery in pregnancy, childbirth and the puerperium • 4024659 - Gestational diabetes mellitus • 434701 - Anemia in mother complicating pregnancy, childbirth and/or puerperium • 4152021 - Ultrasound scan - obstetric	Birth control
8	• 444094 - Finding related to pregnancy • 440457 - Threatened miscarriage • 4205240 - Spontaneous vertex delivery • 4060556 - Uterine scar from previous surgery in pregnancy, childbirth and the puerperium • 4024659 - Gestational diabetes mellitus • 434701 - Anemia in mother complicating pregnancy, childbirth and/or puerperium	Childbirth
9	• 4152021 - Ultrasound scan - obstetric • 4170107 - Us obstetric doppler	Ultrasound scan
10	• 4307111 - Moderate major depression • 4088609 - Somatic syndrome absent • 4088489 - Somatic syndrome present • 715940 - Escitalopram 10 mg oral tablet • 715939 - Escitalopram • 440383 - Depressive disorder	Depression and anxiety

	• 442077 - Anxiety disorder • 19072934 - Alprazolam 0.5 mg oral tablet • 781039 - Alprazolam • 4077577 - Moderate recurrent major depression	
11	• 381270 - Parkinson's disease • 19123237 - Rasagiline 1 mg oral tablet • 715710 - Rasagiline • 789654 - Benserazide 25 mg / levodopa 100 mg extended release oral capsule	Parkinson's and prevention
12	• 44012547 - Bilastine • 44033226 - Bilastine 20 mg oral tablet • 257007 - Allergic rhinitis • 4320791 - Rhinitis • 4031019 - Allergic contact dermatitis • 133834 - Atopic dermatitis	Allergies
13	• 40484028 - Dental caries extending into dentin • 437589 - Pulpitis • 37397422 - Asymptomatic periapical periodontitis	Dental
14	• 1140643 - Sumatriptan • 318736 - Migraine • 19079711 - Sumatriptan 100 mg oral tablet	Migraine
15	• 197223 - Enterobiasis • 1794280 - Mebendazole • 1794307 - Mebendazole 100 mg oral tablet	Enterobiasis
16	• 4216397 - Nerve root disorder • 4227449 - Spondylosis • 75344 - Intervertebral disc disorder	Back pain
17	• 1000632 - Clotrimazole • 4084966 - Candida infection of genital region • 44097550 - Clotrimazole 100 mg vaginal suppository	Candidiasis
18	• 4237233 - Tympanometry testing • 4334402 - Examination of ear under microscope • 4071702 - Audiometric test	Ear functioning tests
19	• 528323 - Hepatitis b surface antigen vaccine • 4015584 - Administration of second dose of hepatitis b vaccine • 4015017 - Administration of third dose of hepatitis b vaccine • 4015583 - Administration of first dose of hepatitis b vaccine	Hepatitis B vaccine

20	• 4324693 - Mammography • 4178367 - Radiographic imaging procedure	Mammography
21	• 1717704 - Valacyclovir • 444429 - Herpes simplex • 1717708 - Valacyclovir 500 mg oral tablet	Herpes diseases
22	• 4165354 - Administration of measles vaccine • 4179181 - Administration of mumps vaccine • 41111268 - Measles vaccine / mumps vaccine / rubella virus vaccine injectable solution • 4218920 - Administration of rubella vaccine	Measles and mumps vaccines
23	• 46272790 - X-ray of limb • 4094343 - Dislocations/sprains/strains • 4075112 - Emergency examination for triage	Limb fracture

II. Licence

Non-exclusive licence to reproduce the thesis and make the thesis public

I, **Oliver-Erik Suik**,

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

"Generalising health events by using frequent itemset mining,"

supervised by Sulev Reisberg.

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Oliver-Erik Suik
14/05/2024