Exploring Patient Data

Getting insight into treatment processes with data mining techniques

Els Roorda

Vrije Universiteit Amsterdam
Faculty of Sciences
Business Mathematics and Informatics
De Boelelaan 1081a
1081 HV Amsterdam

Oktober 2009

Management Summary

With the current developments in the Dutch healthcare financing system, obtaining knowledge about the treatment processes becomes more important. Hospitals record an enormous amount of medical activities. Analysing these data is often found very complex because of the high dimensionality and complexity of the data.

Every patient is unique, there are thousands of different activities which can be produced at different moments in time (see Chapter 2 for an example). In practice the combination of medical activities done for one patient are seldom the same as the combination of activities for another patient with the same disease.

The Dutch healthcare financing system has changed a lot recently. In the past hospitals received fixed budgets. Since 2005 a part of the Dutch hospital care is financed by a limited supply and demand system.

Standardizing and optimizing the treatment for these products is important to keep the costs low and the quality good. The process of standardizing and optimising requires a good knowledge of the current practice and anomalies. This paper will zoom in on methods and models to gain knowledge about the data.

After a literature research we found that there are no algorithms, methods or tools for exploring and gaining knowledge about these high dimensional hospital data yet. By answering realistic questions, a hospital analyst might have, a number of research field is selected which could be useful for gaining knowledge from hospital treatment data.

The research fields we selected are:

- Visual analytics which combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets.
- Exploratory data analysis which is a philosophy for data analysis that employs a variety of techniques, to maximize insight into a dataset
- Sequence data mining which provides the necessary tools and approaches for unlocking useful knowledge hidden in the mountains of sequence data.
- The Levenshtein edit distance which is a metric for the distance between two strings. The ideas behind this metric could be used for measuring the similarity between trajectories
- Multi dimensional scaling which is a process of mapping objects into a low dimensional space in such a way that the distances or dissimilarities are preserved as good as possible.

Knowledge from a number of these research fields has been put into a tool that has been developed for exploration of hospital treatment data. This tool has the functionality to select the interesting patients, find patterns in the activity sequences, find anomalies and find declaring variables to explain the variation between activity sequences.

These functional demands are covered by the tools properties to give the activities a colour, to hide the activities, to centre the activity sequence on an important moment (for example the surgery), to sort patients on specific variables, to filter interesting patients, to provide extra information on mouse over and to provide zooming functionality. The prototype combines visualisation and sql queries and can be extended with sequence mining in the future. Chapter 5 describes the system more detailed.

At this moment the tool has been shown to more then 20 hospital analysts. Most of those people see a lot of potential in this tool. The costs and effort needed to run the tool are very low. Practice should prove that the tool generates a lot of knowledge.

Contents

1 Introduction

Planning hospital activities is a complex task. Van Vliet [1] states that an average patient for an open-heart surgery in the academic hospital in Utrecht stays for 8 to 10 days in the hospital, visits 5 departments and meets about 100 employees. This demonstrates the large complexity of the planning task in hospitals.

Since 2005 a part of the Dutch hospital care is financed by a limited supply and demand system[1]. Insurance companies buy care from a hospital for a negotiated price. In the negotiation waiting times, quality and price are important factors [2]. To increase efficiency, which can lead to lower costs and lower waiting times, hospitals develop clinical pathways or implement other optimization protocols. Without an efficient planning a hospital can loose a lot of money.

The development of clinical pathways is an approach for improving planning possibilities in hospitals. "A clinical pathway is a collection of methods and tools to guide the members of the multidisciplinary and interdisciplinary team towards patient focussed collaboration. It is a way of identifying and defining the different tasks of the different team members. It is a tool to systematically plan and follow up a patient focussed care programme" [3].
Vanhaecht & Sermeus have developed a step-by-step plan [4] for developing clinical pathways. The plan consists of 30 steps, divided into the phases:

- Plan Phase: Defining population, forming a team, making a first version of the clinical pathway.
- Do Phase: Collecting data about the current practice and about best practice
- Check Phase: Interpreting data from the Do Phase, adjust clinical pathway based on this data
- Act Phase: Implementing and evaluating the clinical pathway.

Vanhaecht & Sermeus state in step 11 (part of the Do Phase), about collecting data about current practice, that, to get insight in current practice, for a number of patients the total trajectory should be analysed. Due to a high diversity between treatments (the sequence of activities which are performed to treat a patient), analysing a larger number of cases lead to more insight.

In practice the co-worker in the hospital who has the responsibility of analyzing the current practice gets an enormous database with patient information and activities that are done for the patient. This data is hard to analyse, because of its high dimensionality and complexity.

[1] This part of the care is the so-called B-segment. Between 2005 and 2008 the B-segment care contained about 10% of the costs of hospital care. In 2008 20% of the hospital care was in the B-segment and in 2009 30%. This percentage is assumed to grow in the future [2].

This high dimensionality and the enormous variability caused by for example interpatient differences and interdoctor differences, make the task of optimizing the processes too ambitious for this paper. Therefore this paper focuses on an important step in the optimization process, namely getting a better insight and understanding about the data that is underlying these processes.

The main goal of this paper is:

> *Provide an overview of existing data mining techniques that could be used for analysis of existing treatment data and to develop a tool for visual exploration of data.*

The papers is organised as follows. The seconds chapter will provide a guiding example of the data used in Dutch hospitals. In the third chapter a number of real life questions are linked with possible methods to answer the questions. The data mining techniques named in chapter 3 are described in detail in chapter 4. Chapter 5 describes the system that is developed to explore the data. Chapter 6 evaluates the system. The last two chapters give a conclusion and some recommendations for further research and development.

2 A guiding example

In our research we used treatment data for about 2000 patients diagnosed with cataract. By a treatment we mean here: "the set of activities that is performed to diagnose and treat the patient". The patient population that was analysed contained about 23000 performed activities containing 300 different activities.

The data which was received and which is commonly used in hospitals, consists of two main tables:

-Diagnosis-Treatment Combinations (DBC's), which describes why the patient visits the hospital
-Activities that are connected to the DBC's. An activity code identifies what was done, a date tells when it was done and an amount tells how many times the activity is done.

To provide the reader with a better understanding of the data, an example of the tables can be seen below. Patient 1 visits the hospital with an eye problem and receives some examination a week later. Patient 2 visits the hospital with an eye problem too, but examination lead to surgery in day-care and some control visits. Patient 3 has less examinations and control visits then the second patient while he had a surgery in day-care too.

The DBC's table contains the trajectory information for each of the patients. All the patients are treated at the ophthalmology department because they are diagnosed with cataract. Patient 1 is treated by Doctor1 and the other two patients are treated by Doctor2. Patient 1 and patient 2 are ensured at Ic1, the third patient is insured at Ic2. The fields treatment cluster and revenue will be explained later in this chapter.
The real DBC's table contains more fields like start date, end date and type of care. They are excluded because we do not use them in our research.

DBC's							
Patient	Dbc id	Specialism	Doctor	Diagnoses	Treatment cluster	Insurance company	Revenue
1	10	Ophthalmology	Doctor1	Cataract	11	Ic1	500
2	11	Ophthalmology	Doctor2	Cataract	32	Ic1	1000
3	12	Ophthalmology	Doctor2	Cataract	32	Ic2	1100

The most important fields in the activities table are: a link to the trajectory, a date, a code describing what was done and a number that says how many times it was done.

Activities			
Dbc id	Date	Activity_code	Number
10	20081117	190011	1
10	20081124	339816	1
10	20081124	411101	1
10	20090706	411101	1
11	20080929	190011	1
11	20080929	339481	2
11	20080929	339819	2
11	20081206	331244	1
11	20081206	190035	1
11	20081207	411101	1
11	20081222	411101	1
11	20090221	411101	1
12	20081002	190011	1
12	20081002	339481	1
12	20081209	331244	1
12	20081209	190035	1
12	20081221	411101	1

Next to the main tables a reference table is used with further information about the activity codes. The information the table contains is description, price per unit and a cluster.

Reference activities			
Activity_code	Description	Costs_per_unit	Cluster code
190011	First patient contact	70	1
190035	Day-care	200	2
411101	Control visit	50	1
339816	Sight examination	85	4
339481	Utrasonography of the eye	20	4
339819	Low vision examination	25	4
331244	Extracaps. extranction implant lens	550	5

In this example patient 1 visits the doctor at November 11th 2008. A week later the patient visits the doctor again for a sight examination and a control visit. Nearly eight month later the patient visits the doctor again for a control visit.

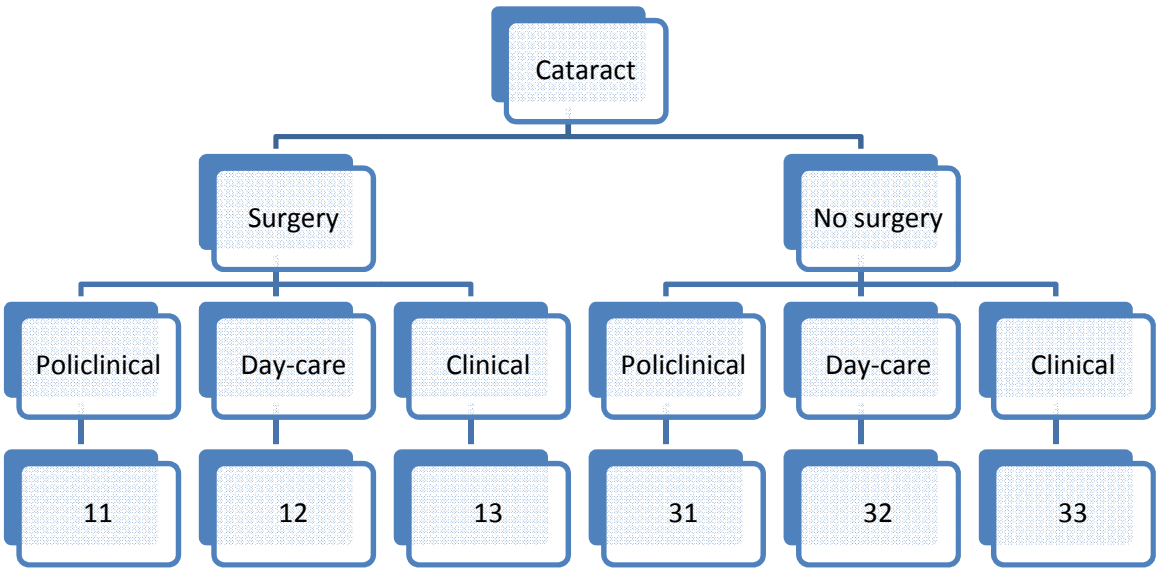
To summarize: A patient has 1 or more DBC's which contain a certain treatment for a certain diagnosis. A treatment consists of one or more activities performed in a hospital. To create a clear image why optimization is important, the next paragraphs will explain how the revenue of a DBC is calculated.

The calculation consists of two steps:
-Derive the treatment cluster of a DBC
-Get the price negotiated with the insurance company of the patient[2]

The treatment cluster is derived with a set of rules made by the government. For cataract there are six important treatment clusters:

Reference treatment clusters	
Treatment cluster	Description
11	Conservative treatment at the policlinic
12	Conservative treatment at the day-care
13	Conservative treatment at the clinic
31	Surgery at the policlinic
32	Surgery while the patient stays in the day-care
33	Surgery while the patient stays in the hospital for one or more nights

The simplified set of rules can be drawn in a diagram:



Patient 1 has no surgery and did not stay at the day-care or in the clinic, which leads to treatment 11. The other two patients did both get surgery in day-care so the DBC's will get treatment cluster 32.

[2] Only for the so-called B segment

For the non-B segment DBC's the profit is defined by the government. For the B segment DBC's the profit is negotiated between each hospital and insurance company. In our cataract example the price table could look as follows:

Reference DBC prices		
Insurance company	Treatment cluster	Price
Ic1	11	500
Ic1	31	1000
Ic2	11	500
Ic2	31	1100

In this case the hospital gets a better price from insurance company 2 then from insurance company 1 for a cataract with treatment cluster 32. The DBC of patient 1 gets a revenue of 500, patient 2 generates 1000 revenue and patient 3 generates 1100 revenue. If we calculate the costs for the hospital the DBC's cost respectively 255, 1060 and 890.

Executing more activities does only lead to more revenue when the treatment cluster changes. In our example only surgery, day-care or clinical treatments can increase the revenue.
In our example the more expensive second patient will lead to less profit than the less expensive third patient. Because of this fact optimizing the process within a treatment cluster is important.

3 Problem formulation

In the previous chapter we noticed that executing more activities do not lead to more revenue automatically. All DBC's with the same treatment cluster and diagnoses (remember that DBC stands for diagnosis treatment combination) lead to one product with one price. For B-segment DBC's this price is negotiated with the insurance company.

Because of the fact that executing more activities does not lead to more profit, optimizing the process within a treatment cluster is important. Optimization in healthcare can't be done without maintaining the quality, because the most important function of a hospital is curing people. At this moment it is very difficult to recognize which activities contributed to the quality of the treatment and which activities are probably inefficient.

Example: Difficulty of recognizing contribution of an activity
Patient 2 in the example from chapter 2 had 3 control visits, while patient 3 had only 1 control visit. Did the hospital deliver a lower quality, because eventual problems are not detected or detected too late? Or was the second patient a more difficult patient because his lens was different, he was older, he has another medical issue? Maybe there are other unknown variables that declare the differences, or did the two control visits not add any added value to the treatment.

To help hospital analyst, we will:

> *Provide an overview of existing data mining techniques that could be used for analysis of existing treatment data and to develop a tool for visual exploration of data.*

To obtain more knowledge about the type of analysis a hospital employee should be able to do a number of realistic example questions have been formulated:
-What percentage of the patients is treated clinically?
-What is the most common set of activities that is performed for the clinical patient population?
-How much does the set of activities vary between patients?
-How much does the time between activities vary between patients (the variation in the process)?
-Which variables cause the variation between the patients?
-Which patient clusters can be made?

After formulating the questions I tried to answer them using methods and techniques described in the literature. Where the literature did not supply a satisfying solution, a self-created technique was used.

3.1 What percentage of the patients is treated clinically

Question background
For a hospital it is very important to identify the setting a patient is treated in, for example to determine the treatment cluster. A patient can be treated on the policlinic when local anaesthesia is used and there is no need to undergo further tests (for example removing the wisdom teeth). Patients who can leave the hospital the same day and need a bed are treated in day-care (for example removing the tonsils). When a patient needs to stay several days in hospital a treatment is called clinical.

Methods to address the question
To find out how many patients are treated in a certain setting a SQL query can be used. A database query can make a count based on self-written specifications, but it is hard to check whether the algorithm detects anomalies and how they are handled.

Relevant sections
Visualisations are often used for the detection of anomalies. Chapter 4.1 provides more information on visualisation. The detection of anomalies is possible with the tool developed for this research. Chapter 0 contains a description of the developed system.

3.2 What is the most common set of activities that is performed for the patient population

Question background
To standardize and improve the planning for a certain patient population it is very important to know what activities are performed, in which order and whether some activities are performed at the same day.

Methods to address the question
The book "Sequence Data Mining" [5] and some other papers [6; 7; 8; 5] on sequence mining led to an algorithm to find frequent subsequences . The implementation of the algorithm was no problem, but interpreting the results was difficult. 17 pages of frequent (frequent defined as in more then 50% of the sequences) subsequences where found in my data set. A result fragment is displayed in the box below:

```
|190011|331244 - 980208|411101| 668 sequences
|190011|411101|411101|411101| 632 sequences
| = transaction separator (activities are performed on different days)
- = item separator (activities are performed on the same day)
```

This result is impossible for an analyst to interpret, unless there is a way to visualize the results.

Relevant sections
Chapter 4.3 provides more details on Sequence Data Mining and the implementation of the frequent subsequences algoritm. A possible way to visualize the frequent

subsequences is showed with the tool developed for this research. Chapter 0 contains a description of the developed system.

3.3 How much does the set of activities vary between patients

Question background
To make an adequate planning the analyst wants to know what variation exists between patients and why the variation exists, because variation can be a sign of inefficiency. In healthcare efficiency and quality do show a positive relation, [9][10] probably because better protocols lead to better quality.

Methods to address the question
Calculating variability and making the variability discussable with doctors is not easy. Trying to solve the problem we investigated the possibilities of using the Levenshtein distance method. This is an algorithm that assigns costs to every transformation that needs to be done to get sequence β out of sequence α. The higher the transformation costs, the bigger the difference between the sequences.

Using this algorithm led to two problems: What costs needs to be assigned to what transformation and when we are able to do this, how can we discuss the variance with the doctors. "Hello doctor Jansen, why is Bertha 102.3 different then John?" will not work in real life.

Relevant sections
Chapter 4.4 provides more information about the Levenshtein distance method. In chapter 4.1 is describe why visual analytics could be an appropriate approach for such a complex task. The developed tool visualises the sequences in such a way that a human could analyze and communicate about variation effectively. Chapter 0 contains a description of the developed system.

3.4 How much does the time between activities vary between patients

Question background
As a hospital analyst you want to know why and how much time there is between certain events. A long time between the last visit to a doctor and surgery can mean waiting times for the OK. Often a patient has to wait very long for a diagnostic activity like MRI or CT so that the diagnostic phase becomes longer, this can cause stress to the patient "Do I have cancer or not" and can cause health risks. The sooner some diseases are detected, the bigger the change is the disease will be cured.

Methods to address the question
As was stated in the last paragraph quantifying variability in an understandable way is not easy.

Relevant sections
The developed tool visualises the sequences in such a way that a human could analyze and communicate about variation effectively. Chapter 0 contains a description of the developed system.

3.5 Which variables cause the variation between the patients

Question background
Certain variables explain the differences between treatments. Often an older patient is kept in the hospital for some extra days. Even doctors can have their own way of treating a patient. Sometimes insurance companies sent more difficult patients to one hospital, so that these patients will have a more expensive treatment. As analyst you want to find what is causing the variability and how variability can be minimalized or how extra costs for difficult patients can be covered.

Methods to address the question
As stated in the last two paragraphs it is difficult too quantify variability, doing a kind of regression is hard when the variation is not quantified.

Relevant sections
The developed tool visualises the sequences in such a way that a human could analyze and communicate about variation effectively. Sorting and filter functions allow the analyst to analyze the influence of the different variables. Chapter 0 contains a description of the developed system.

3.6 Which patient clusters can be made

Question background
To make standardized protocols for the treatment of a patient it is important to recognize patients which are treated "similar" and to cluster them.

Methods to address the question
Clustering is used a lot worldwide to make financing systems for healthcare services. In the Netherlands the DBC system is used, in a lot of other countries the DRG system is used also other systems, for example the ETG methodology classifies patients. [11] The analogy between all systems lies in the fact that humans made the cluster decision rules.

The first question when it comes to clustering is where to cluster on. The Dutch DOT system, currently worked on, states that clinical homogeneity, financial homogeneity and internationally recognizable clustering is important. [12]
To create these clusters teams of analyst, with cooperation of doctors, create decision trees manually.

We applied multi dimensional scaling on the data to cluster automatically. The most important problem we found is the inability to measure the dissimilarity between to sequences.

Relevant sections
Chapter 4.5 provides more information about multi dimensional scaling. This chapter also contains more details about the multi dimensional scaling experiment we did.

4 Overview of techniques

In the previous chapter we tried to answer some questions a hospital employee might have. One of our conclusions is that there are several interesting techniques for analysing the existing sequence data. The other conclusion from the last chapter is that most methods are more valuable when the analyst can visualize and demonstrate the outcomes of an analysis.

This chapter has two purposes:
-Describing the techniques used in the last chapter in more detail
-Describing techniques that can be used to produce a model for visualising the data

4.1 Information visualization and visual analytics

Probably the most accepted definition for Information Visualization is one that comes from Card, Mackinlay, and Shneiderman and that actually is their definition for "visualization". They describe visualization as "the use of computer-supported, interactive visual representations of data to amplify cognition." [13]

They also propose six ways in which visualizations can amplify cognition:
-by increasing the memory and processing resources available to users (by using the visual system or by using the brain as working memory)
-by reducing the search for data (by showing related data in the visualisation)
-by using visual representations to enhance the detection of patterns
-by enabling perceptual inference operations, perceptual inference means that problems can be obvious when they are visualized.
-by using perceptual attention mechanisms for monitoring, large numbers of events can be monitored, when they are organised to do so.
-by encoding information in a manipulable medium, an interactive medium allows better exploration
[13]

Information visualization researchers often have problems to prove the value of a visualisation because it is hard to measure the amplification of cognition. Jean-Daniel Fekete, Jarke J. van Wijk, John T. Stasko, and Chris North state in their paper [10] that an economic approach might help in quantifying the value.
Their conclusion is that a great visualization method is used by many people, who use it routinely to obtain highly valuable knowledge, while having to spend little time and money on hardware, software, and effort.

Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets. [14] Information visualization can be seen as a part of visual analytics. Visual analytics can be seen as an integral approach to decision-making, combining

visualization, human factors and data analysis. The challenge is to identify the best automated algorithm for the analysis task at hand, identify its limits which can not be further automated, and then develop a tightly integrated solution with adequately integrates the best automated analysis algorithms with appropriate visualization and interaction techniques. The philosophy behind visual analytics is combining the strong characteristics of computers and humans. [14]

According to Daniel Keim et all, the fields combined by visual analytics are:
-Visualization which researches the presentation of data
-Data management which researches methods to store data
-Data mining which researches methods to automatically extract valuable information from raw data by means of automatic analysis algorithms.
-Perception and cognition, which focuses on user-centered analysis and modeling
-Human computer interaction which studies the interaction between people and computers

4.2 Exploratory Data Analysis

Exploratory data analysis can be defined, as the examination of data with minimal preconceptions about its structure through which it is hoped that relationships and patterns, at least some of which are unanticipated, will be uncovered. [15]

Exploratory data analysis (EDA) can be seen as a philosophy for data analysis that employs a variety of techniques, which are mostly graphical, to maximize insight into a dataset, uncover underlying structure, extract important variables, detects outliers and anomalies, test underlying assumptions. [16]
In contrast, confirmatory data analysis requires advance decisions about data structures. It examines the precision with which the data reflect these structures and the confidence with which it can be asserted that they will reappear in further samples from the same population. [15]
The difference lies in the approach, for classical analysis, the sequence is
Problem => Data => Model => Analysis => Conclusions
In the EDA approach, analysis is done before the model is made. When EDA is used there will be no assumptions on normality, linearity etc. [16]

Looking at the relation between information visualisation and EDA we can conclude that a visualization fits in the EDA philosophy of analyzing data without assumptions.

4.3 Sequence Data Mining

Sequences are an important type of data which occur frequently in many scientific, medical, security, business and other applications. Examples of sequences are DNA, the items a customer purchases in a store and event sequences.

Sequence data mining provides the necessary tools and approaches for unlocking useful knowledge hidden in the mountains of sequence data. Applications of sequence mining are for example detection of frequent subsequences, classification of sequences and clustering of sequences. [5]

Sequence data has several characteristics:

-length: sequences can be very long and therefore highly dimensional. Sequences can also have different sizes (one patient visits the hospital more often then another patient)
-position: in some sequences absolute position has a meaning, in other sequences the absolute position has no meaning. In our data, a position of March 3rd 2009 does not have a meaning, in DNA does the absolute position say something about the elements meaning.
-ordering: The relative ordering between elements in sequences is often important. The fact that an element occurs before or after or on the same moment as another element is most times important. The distance between two elements is sometimes important too. [5]

The model for a sequence can be formulized as follows:
-A sequence S_i has one ore more transactions s_j
-There are i sequences
-Sequence S_i has j transactions
-A transaction s_j has one or more items x_n
-There are n different items
[5]

Translation of this model to hospital data:
- i is the number of patients
- j is the number of days patient i has activities
- n is the number of different activities (3000 is our case)

Given the example from Chapter 2 filled with the sequence number, the transaction number within the sequence and the item numbers

Activities						
Dbc id	Date	Activity_code	Number	Sequence (i)	Transaction (j)	Item (x)
10	20081117	190011	1	1	1	1
10	20081124	339816	1	1	2	2
10	20081124	411101	1	1	2	3
10	20090706	411101	1	1	3	3
11	20080929	190011	1	2	1	1
11	20080929	339481	2	2	1	4
11	20080929	339819	2	2	1	5
11	20081206	331244	1	2	2	6
11	20081206	190035	1	2	2	7

11	20081207	411101	1	2	3	3
11	20081222	411101	1	2	4	3
11	20090221	411101	1	2	5	3
12	20081002	190011	1	3	1	1
12	20081002	339481	1	3	1	4
12	20081209	331244	1	3	2	6
12	20081209	190035	1	3	2	7
12	20081221	411101	1	3	3	3

The population will be modelled as follows:
There are three sequences S_1, S_2 and S_3
The first sequence had 3 transactions (days with activities), the second sequence has 5 transactions, the third sequence has 3 transactions.
There are 7 different items.

The property that is not present in the model is the distance between the transactions. The model can be completed with a transaction time stamp s_t which contains the number of days from the first activity day.

A lot has been written about finding frequent subsequences, subsequences are also called episodes. [6; 7; 8; 5] A sequence $\alpha = \alpha_1\alpha_2 \cdots \alpha_n$ is called a subsequence of another sequence $\beta = \beta_1\beta_2 \cdots \beta_m$ and β super-sequence of α, if there exist integers $1 \leq j_1 < j_2 \cdots < j_n < m$ such that $\alpha_1 \in \beta_{j_1}, \alpha_2 \in \beta_{j_2} \cdots \alpha_n \in \beta_{j_n}$ [5]
A subsequence is called frequent when the percentage of sequences with the specific subsequence is bigger than p.

The subsequences with a p-value of 1 (the subsequence must be in every sequence) are frequent in our sample:

Frequent subsequences	**Length**
190011	1
411101	1
190011 – 411101	2

The most straightforward way for finding frequent subsequences is collecting all the 1-element sequences and testing all the combinations. When there are three different items {A,B,C}, the following subsequences will be tested to find the length 1 and length 2 frequent sequences: A, B, C, (AB), (AC), (BC), AA, AB, AC, BA, BB, BC, CA, CB, CC, where the items in brackets occur in the same transaction, for 3000 different items

$$Number\ of\ tests = n \cdot n + \frac{n \cdot (n-1)}{2} = 3000 \cdot 3000 + \frac{3000 \cdot 2999}{2} \approx 13.5\ million$$

when we want to search for longer frequent subsequences the number of sequences to test grow exponentially.
A more efficient method is the method from Heiki Manilla et all [8] where a set of frequent length 1 subsequences C_1 is made first. The frequent subsequences with length i, C_i are found in the following way:

-building phase: finding all the candidate episodes using the frequent episodes from C_{i-1} and the frequent building blocks from C_1
-recognition phase: compute the frequencies and add the frequent sequences to C_i
This process is repeated until C_i is empty.

Both methods were implemented in C++. Complexity analysis indicates that the complexity of the first algorithm is O(n^k) where k is the maximum length of the subsequence, the complexity of the second algorithm is in the worst case also O(n^k), but in practice it is of order O(kn), as the the only expensive operation is the computation of frequencies which is done k times and each time it requires a complete scan of the whole data set.
The test with the straightforward algorithm was stopped after 2 hours, the method from Heiki Manilla et all was ready in 3 minutes.

Many other methods can be found in literature, but these were not relevant to describe here, because the last method suits our purpose of finding the frequent subsequences in an acceptable time.

There are a few papers about sequence mining on hospital data but some are not on activity level [17; 18; 19] and some do only provide methods for transactions/episodes and not for the whole sequence [20], so all the papers I found were not very applicable.

4.4 Edit Distance

The Levenshtein edit distance is a metric for the distance between two strings. The distance between two strings is defined as the minimum number of deletions, insertions and reversals which are needed to get one string from another. [21] A reversal is a replacement of one character by another.
For example the words darm and dramt need 2 reversals and 1 insertion, so the distance is between the strings is 3.

In most cases the edit distance is calculated with dynamic programming. The value of the edit distance between string A with length m and a string B with length n can be computed by filling a matrix D with size (m + 1) x (n + 1). The following recurrence is used to fill the matrix [22]:

$$D[i,0] = i$$
$$D[0,j] = j$$
$$D[i,j] = \begin{cases} D[i-1.j-1], if A[i] = B[j] \\ 1 + min(D[i-1,j-1], D[i-1,j], D[i,j-1]), if A[i] \neq B[j] \end{cases}$$

D[m+1,n+1] contains the minimum number of transitions that are needed.
For the words “darm” and “dramt” the matrix looks as follows:

0	1(D)	2(A)	3(R)	4(M)
1(D)	0	1	2	3
2(R)	1	1	1	2
3(A)	2	1	2	2
4(M)	3	2	2	2
5(T)	4	3	3	3

The Damerau Edit Distance is the same metric as the Levenshtein edit distance except it is allowed to transpose two characters. [22] Our example does only cost 2 transactions this time, because the a and the r can be exchanged.
This time the recurrence to fill the matrix is as follows [22]:

$$D[i,0] = i$$
$$D[0,j] = j$$
$$D[i,j] = \begin{cases} D[i-1,j-1], if A[i] = B[j] \\ D[i-1,j-1], if A[i-1] = B[j] \; and \; A[i] = B[j-1] \; and \; D[i-1,j-1] > D[i-2,j-2] \\ 1 + min(D[i-1,j-1], D[i-1,j], D[i,j-1]), otherwise \end{cases}$$

0	1(D)	2(A)	3(R)	4(M)
1(D)	0	1	2	3
2(R)	1	1	1	2
3(A)	2	1	1	2
4(M)	3	2	2	1
5(T)	4	3	3	2

In the second table can be seen that transposing the “a” and “r” leads to less distance.
Both algorithms have a runtime of O(nm) because every cell of the matrix should be filled with a value.

To make these metrics useful for our purpose we should give penalty costs based on the activities that are edited. Exchanging a lung transplantation with a natrium test is very different from exchanging a kalium test with a natrium test. Another problem which should be solved to use an edit distance algorithm is the occurrence of multiple activities on one day and days between activities during which nothing happens.

Because of the differences between treatments and strings, finding one rule to measure dissimilarities between treatments is not easy. A cost matrix should be filled with the edit costs between each pair of activities. Also a rule to define costs for a gap between days with activities should be constructed. Next to these rules a rule should be made for more then one activity on a day.

Therefore finding an edit distance measure requires a lot of research. It is also doubtful whether there is one rule that is applicable for multiple cases. If an analyst researches the use of the MRI in a certain population the distance should be defined differently then when someone researches how long a patient stays in the hospital after a specific surgery.

4.5 Multi Dimensional Scaling

A definition of multidimensional scaling is the search for a low dimensional space in which points in space (not necessarily Euclidean space) represent the objects, one point representing one object in such a way that the distances between the n points in space $\hat{d}_{ij}$ match as well as possible the original dissimilarities δ_{ij}. [23] Where d_{ij} are the interpoint distances and t is the dimensionality. In other words multi dimensional scaling is a process of mapping objects into a low dimensional space in such a way that the distances or dissimilarities are preserved as good as possible.

In ordinary cases the Euclidean (or Pythagorean) distance is used. The distance between x_i and x_j, is given by $d_{ij} = \left[\sum_{l=1}^{t}(x_{il} - x_{jl})^2\right]^{0.5}$

The goodness of fit can be measured with Kruskal's Stress Formula by

$S = \sqrt{\frac{S^*}{T^*}} = \sqrt{\frac{\sum(d_{ij}-\hat{d}_{ij})^2}{\sum d_{ij}^2}}$ [24]

It's widely accepted that when Kruskal's stress is used that less then 20% is a poor fit, between 20% and 10% is a fair fit, between 10% and 5% is a good fit and less then 5% is an excellent fit [25].

The computational problem is to find the best-fitting configuration in t-dimensional space for a fixed value of t. For this optimization problem the method of steepest descent can be used. Previously the configuration was written as n points, but we can consider this as a single point: $(x_{11},\cdots,x_{1t},\cdots,x_{n1},\cdots,x_{nt})$
For given values of δ_{ij} a stress value can be calculated for any point by
$S = S(x_{11},\cdots,x_{1t},\cdots,x_{n1},\cdots,x_{nt})$
Start with an arbitrary point in configuration space and find the direction where S is decreasing most quickly. This can be done by evaluating the negative gradient of the partial derivatives of the function S:
$\left(-\frac{\partial S}{\partial x_{11}},\cdots,-\frac{\partial S}{\partial x_{1t}},\cdots,-\frac{\partial S}{\partial x_{nt}}\right)$
And x_{ij} by (γ is the learning factor)

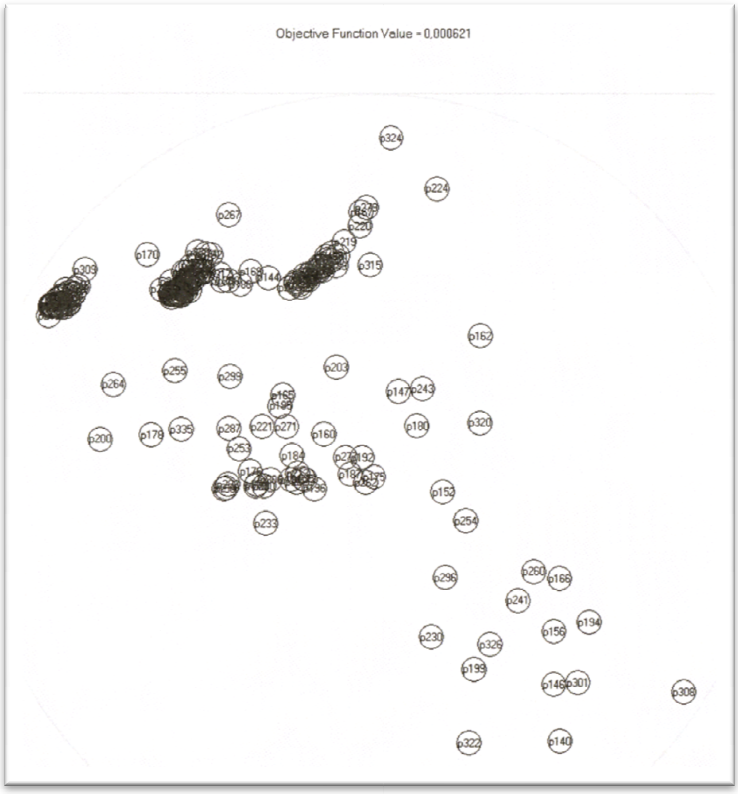
$\Delta x_{ij} = -\gamma\frac{\partial S}{\partial x_{ij}}\ for\ i = 1,\cdots,n\ and\ j = 1,\cdots,t$ [26]

in the direction of the steepest descent. This process is repeated until a minimum has been found. This methods has the risk of finding a local minimum instead of the overall minimum. The probability of finding the overall minimum is bigger when we start from different initial configurations.

With help of MDS we can visualize the patients in a 1 or 2 dimensional graph. The problem with this algorithm is that a dissimilarity measure is needed. How similar the patients are depends on the question the analyst has. If an analyst researches the diagnostic part of the treatment he needs a different dissimilarity measure than an analyst who wants to analyze the number of days a patient have to visit the hospital.

To gain more insight in the possibilities of MDS on treatment data we did an experiment. For the interpatient distance we used the following algorithm:

The results can be seen in the image below, where every dot represents a single patient.



The image shows some clear clusters and outliers, but interpreting the visualisation is very hard. To get more feeling for the value of the graph, the activities of the patients should be made visible.

5 Description of the system

In this chapter we will describe the visualisation tool we developed for the exploration of treatment data. First we will describe what choices we made and what the functional demands are for the tool and which knowledge from the research fields described in chapter 4 we used to make the decisions. Then we will explain the functionality of the tool with the help of screenshots. Because of privacy reasons the tool that is used is filled with fictitious data.

The first functional demand for the tool came from the exploratory data analysis philosophy: *A user with minimal preconceptions about the data should be able to use the tool.*
A number of functional demands came from the field of visual analytics:
-The visual representation should enhance the detection of patterns
-The visual representation should enable perceptual inference
-The tool should be interactive, for a better exploration
-Information related to the trajectories or activities should be added to reduce the search for data

The fields of exploratory data analysis and visual analytics helped us to deduce what our tool should do. The theory from the sequence mining field led us to the right way to visualize the data.
Figure 1 shows the chosen visualization:
-Every line represents a sequence
-The transactions are represented by pies. Because the x-axis contains a time scale (in days) the order and distance between the transactions are visualised.
-The pieces of the pie represent the items. Because there are potentially thousands of different activities, the activity cluster determines the colour. (Figure 6 shows that an exception can be made)

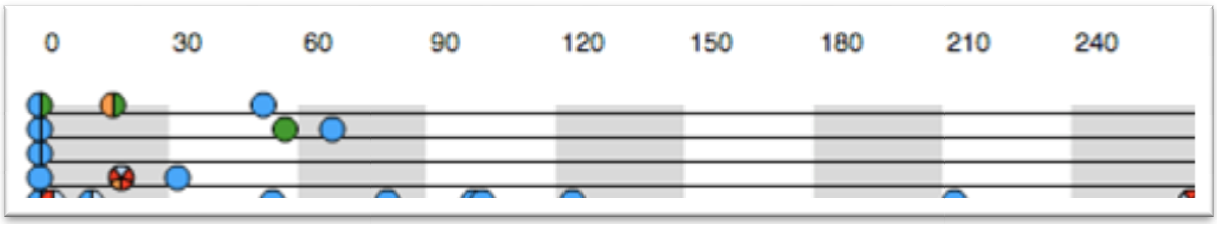


Figure 1: Trajectories, transactions and activities are visualized

Figure 2 shows the first visualization. From this image it is very hard to derive patterns and anomalies. After applying a sort algorithm on trajectory costs, the first patterns appear (Figure 3). Figure 4 and Figure 5 show that aligning on the first day in the clinic reveal the patterns clearly.

Next to the sort and align functionality, a filter functionality has been build (Figure 7) to select a set of trajectories. Possible criteria to select on are age, doctor, insurance company, revenue, costs, number of activities, number of activities of a certain kind and more.

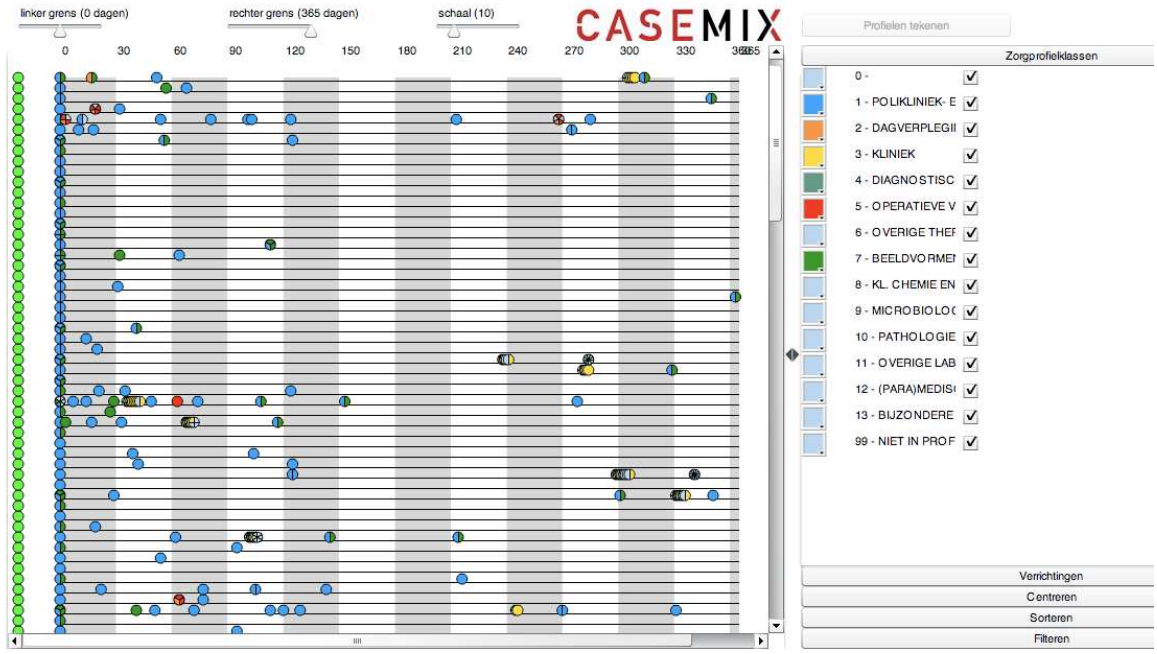


Figure 2: The first visualization

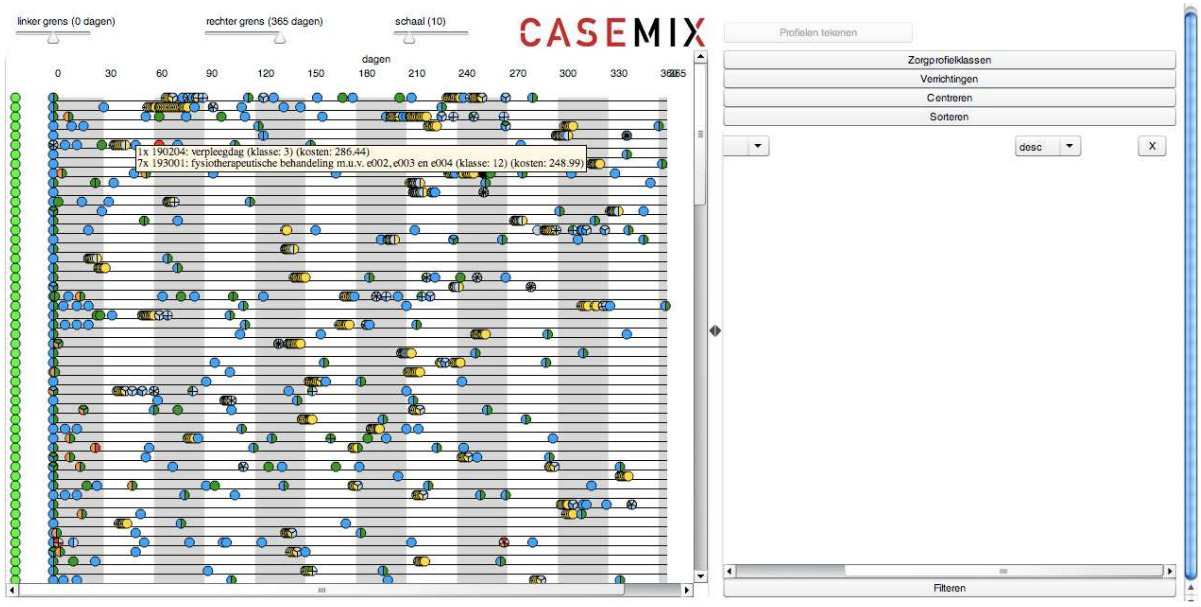


Figure 3: Trajectories are sorted on costs, activity information on mouse over

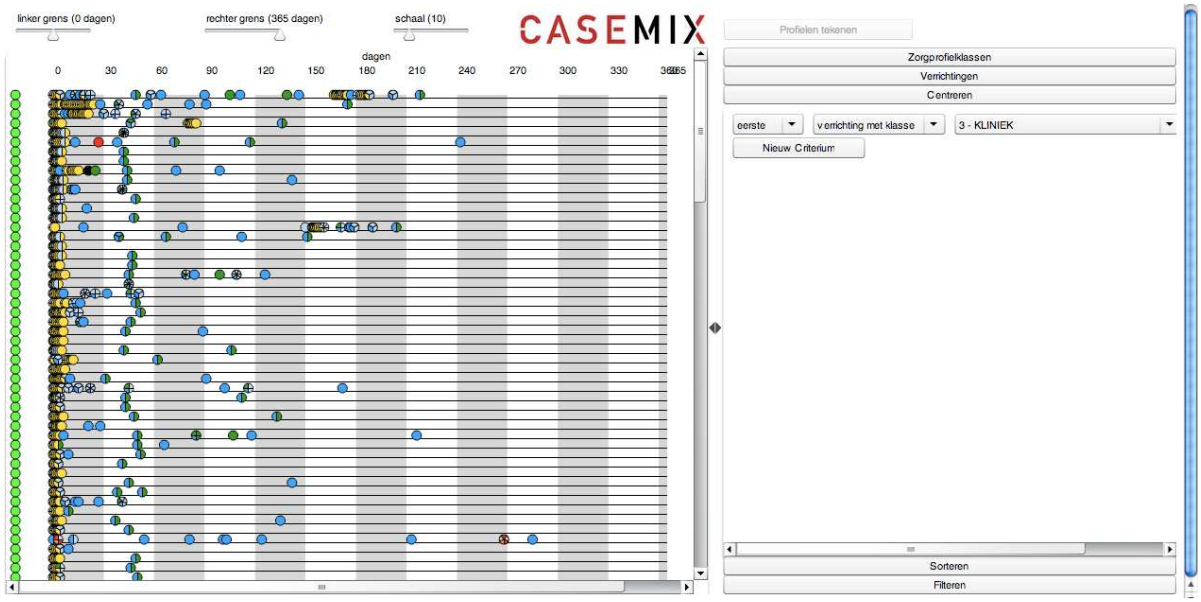


Figure 4: Centering on the first day in clinic reveals the patterns

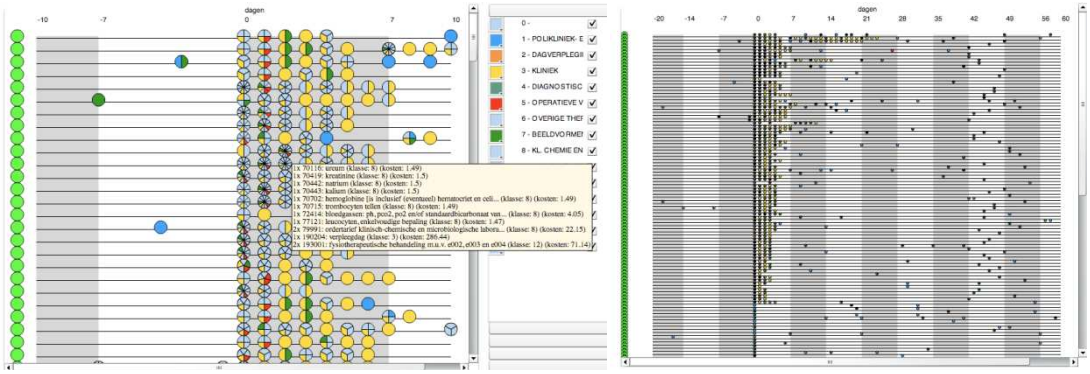


Figure 5: Zooming functionality for more details or better pattern recognition

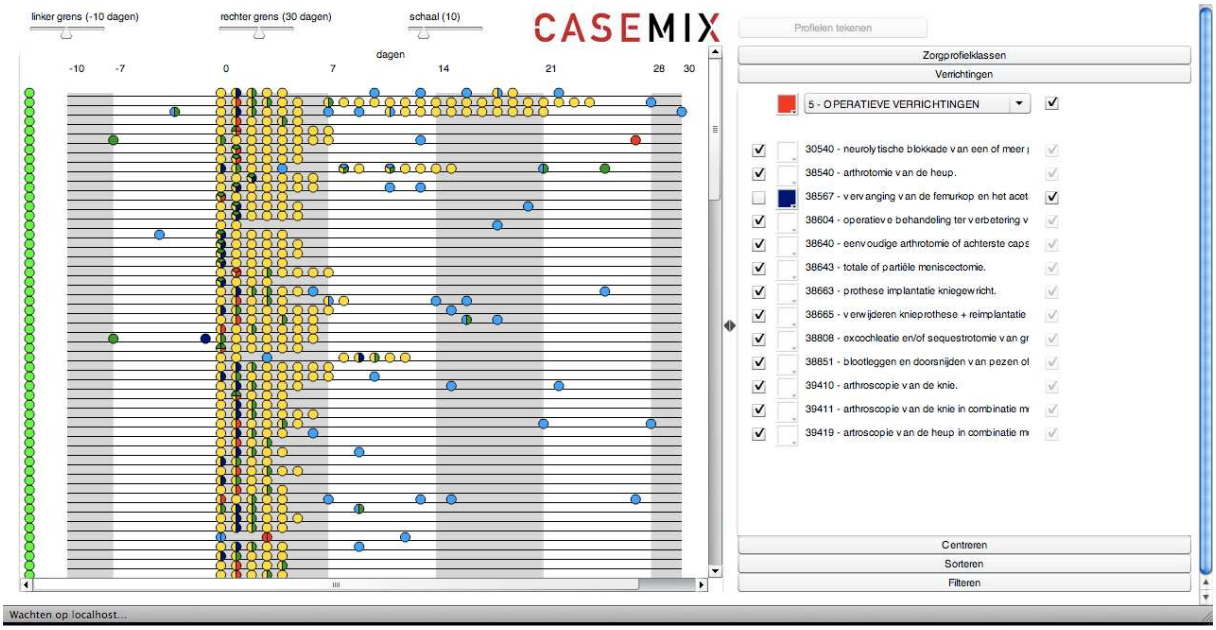


Figure 6: Make an exception for a specific activity

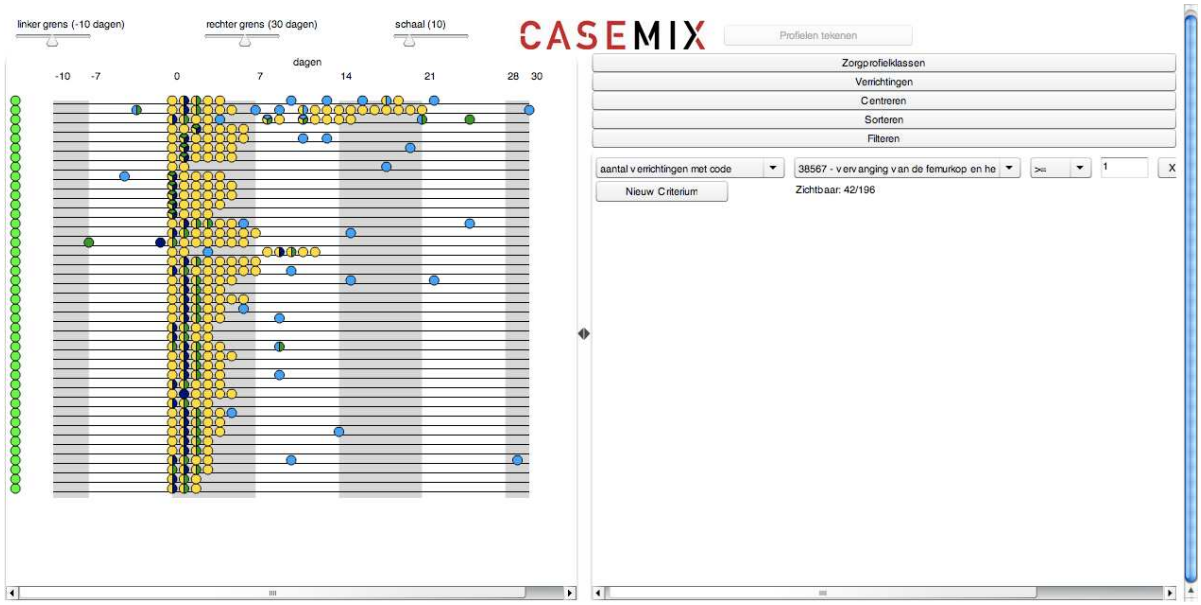


Figure 7: The filter functionality enables trajectory selection

The current tool is build in Flash with Actionscript 3. The hardware and software which is needed for the tool is a computer that runs a web server and a client computer, which can be the same as the web server, that has a browser with a flash plug-in. Nearly all computers in a hospital fit these requirements, so the start up costs are low.

6 System evaluation

In Chapter 4.1 we learned that a visualization is the use of computer-supported, interactive visual representations of data to amplify cognition. Our tool is amplifies cognition in three ways by:
-Making the detection of patterns easier by using visual presentations.
-Adding mouse over information and filter and sorting functions on declaring variables like age of a patient, doctor, origin of a patient and profit to reduce the search for data
-Adding interactivity to the medium for better exploration.

In the previous chapter we formulated a number of functional demands:
- A user with minimal preconceptions about the data should be able to use the tool.
-The visual representation should enhance the detection of patterns
-The visual representation should enable perceptual inference
-The tool should be interactive, for a better exploration
-Information related to the trajectories or activities should be added to reduce the search for data

We have seen in the previous chapter that the interactive tool can detect patterns. The tool does also give related information (Figure 3 + Figure 8). As someone with minimal preconceptions I used the tool to analyze a specific patient group. After a few minutes I knew how long a patient stays at the clinic and whether older patients stay longer.

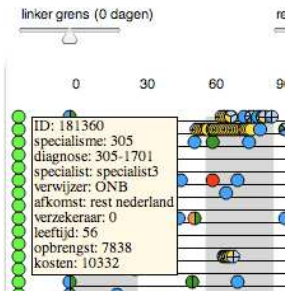


Figure 8: Trajectory information

To test the perceptual inference I tried to find why a certain patient had a lot of costs. To determine why the patient was expensive was a simple task. After a week he got surgery again, probably because of an infection.

In the previous paragraph we showed that the tool meets al the requirements. But that does not make the tool valuable by default. As stated in Chapter 4.1 information visualization researchers often have problems to prove the value of a visualisation because it is hard to measure the amplification of cognition [10]. Because we experience this problem as well we use the economic approach to quantify the value.

According to the economic approach the tool is a great visualization method when it is used by many people, who use it routinely to obtain highly valuable knowledge, while having to spend little time and money on hardware, software, and effort.

This tool is not in use yet so the usage part cannot be assessed yet. The prototype has been shown to more then twenty people from several hospitals working at several departments and several layers of the organisation. Every person it was shown to was enthusiastic and suggested a number of tasks it could be used for. The tasks proposed were: exploring current practice, recognizing severity, searching for outliers to discuss with doctors and making clinical pathways.

The hardware and software requirements are very low as described in Chapter 5. Using the tool requires a little explanation, probably one page of reading material, so the effort to start analyzing is not very big.

The conclusion that can be drawn is that many people see potential in this tool and the costs and effort needed to run the tool are very low. If practice proves that the tool generates a lot of knowledge in a short time by many people, the tool is highly valuable according to the economic approach.

7 Conclusion

Optimizing healthcare processes becomes more important every day in Dutch hospitals. To optimize a process a thorough understanding of the current practice is needed. Due to high variability of treatment sequences, the trajectories of many (or all) patients in the population should be analyzed. This complex task can be supported by the use of data mining techniques.

The goals of this research were:

> *Provide an overview of existing data mining techniques that could be used for analysis of existing treatment data and to develop a tool for visual exploration of data.*

In our research we found five useful research areas for our problem:

- Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets.
- Exploratory data analysis is a philosophy for data analysis that employs a variety of techniques, to maximize insight into a dataset
- Sequence data mining provides the necessary tools and approaches for unlocking useful knowledge hidden in the mountains of sequence data.
- The Levenshtein edit distance is a metric for the distance between two strings. The ideas behind this metric could be used for measuring the similarity between trajectories
- Multi dimensional scaling is a process of mapping objects into a low dimensional space in such a way that the distances or dissimilarities are preserved as good as possible.

The ideas behind visual analytics, the theory of sequence data mining and the philosophy of exploratory data analysis are combined in a visualisation tool that supports the analyst to perform the difficult task of creating a thorough understanding of the current practice.

8 Recommendations

The tool can be extended to a real visual analytics tool. At the moment the tool works for exploring data, but does not contain data mining and other algorithms. At first the frequent subsequence algorithms that was implemented in C++ could be build into the tool, so it could be used by an analyst. One of the choices that has to be made is on which level sequences should be tested, on activity class or on activity code. One solution for this problem could be to take over the colour settings from the activity panel. The activities that have an exception are apparently so important that they should be taken separately, for the other activities a group indication is detailed enough.

At this stage the analyst is able to search for patterns and anomalies. Extending the tool with cluster functionality could be the next step.

9 References

[1] *De invulling van kwalificatieniveau 5.* **van Vliet, J.** 17, 1998, TVZ, pp. 506-507.

[2] **Nederlandse Zorgautoriteit.** *Monitor ziekenhuiszorg 2008* .

[3] *Wat zijn klinische paden?* **Sermeus, W and Vanhaecht, K.** 3, 2002, ACTA HOSPITALIA, pp. 5-11.

[4] *Draaiboek voor de ontwikkeling, implementatie en evaluatie van een klinisch pad. 30 stappenplan van het Netwerk Klinische Paden.* **Vanhaecht, K and Sermeus, W.** 3, 2002, Acta Hospitalia, pp. 13-27.

[5] **Dong, Guozhu and Pei, Jian.** *Sequence Data Mining.* s.l. : Springer, 2007.

[6] **Srikant, Ramakrishnan and Agrawal, Rakesh.** Mining Sequential Patterns: Generalizations and Performance Improvements. [book auth.] Peter Apers, Georges Gardarin and Mokrane Bouzeghoub. *Advances in Database Technology EDBT '96.* 1996.

[7] *Efficient mining of frequent episodes from complex sequences.* **Huang, Kuo-Yu and Chang, Chia-Hui.** s.l. : Elsevier Science Ltd., 2008, Vol. 33. 03064379.

[8] *Discovery of Frequent Episodes in Event Sequences.* **Mannila, Heikki, Toivonen, Hannu and Verkamo, Inkeri.** 3, 1997, Data Mining and Knowledge Discovery, Vol. 1, pp. 259-289.

[9] **Ludwig, Martijn.** *Efficiency of Dutch Hospitals.* Leiden : Datawyse, 2008.

[10] **Fekete, Jean-Daniel, et al.** The Value of Information Visualization. *Information Visualization.* Berlin / Heidelberg : Springer, 2008, pp. 1-18.

[11] *Episode Treatment Groups (ETGs): a patient classification system for measuring outcomes performance by episode of illness.* **MT, Forthman, HG, Dove and LD, Wooster.** 2, nov 2000, Top Health Inf Manage., Vol. 21.

[12] **DBC Onderhoud.** *DBC's op weg naar transparantie Deel III.* 2007.

[13] **Card, Stuart K., Mackinlay, Jock D. and Shneiderman, Ben.** *Readings in Information Visualization.* s.l. : Academic Press, 1999.

[14] **Keim, Daniel, et al.** Visual Analytics: Definition, Process, and Challenges . *Information Visualization.* Berlin / Heidelberg : Springer, 2008, pp. 154-175.

[15] **Tukey, JW.** *Exploratory Data Analysis.* s.l. : Addison-Wesley, 1977.

[16] **Filliben, James J.** NIST/SEMATECH e-Handbook of Statistical Methods. [Online] [Cited: 05 06 2009.] http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm.

[17] **Viveros, Marisa, Nearhos, John and Rothman, Micheal.** Applying Data Mining Techniques to a Health Insurance Information System. *Proceedings of the 22th International Conference on Very Large Data Bases.* 1996.

[18] *An event set approach to sequence discovery in medical data.* **Ramirez, Jorge, et al.** 6, 2000, Vol. 4. 1088-467X (Print) 1571-4128 (Online).

[19] *A Rule Discovery Support System for Sequential Medical Data, in the Case Study of a Chronic Hepatitis Dataset.* **Ohsaki, M, et al.** 2003.

[20] **Semenova, Tatiana, et al.** Conceptual Mining of Large Administrative Health Data. *Advances in Knowledge Discovery and Data Mining.* 2004.

[21] *binary codes capable of correcting deletions, insertions, and reversals.* **Levenshtein, V.I.** 1966, Soviet Physics Doklady, p. 707.

[22] *A Bit-Vector Algorithm for Computing Levenshtein and Damerau Edit Distances.* **Hyyrö, Heikki.** 2003, Nordic Journal of Computing, pp. 29-39.

[23] **Cox, Trevor F. and Cox, Micheal A. A.** *Multidimensional Scaling.* s.l. : Chapman & Hall, 2000.

[24] *Nonmetric multidimensional scaling: A numerical method.* **Kruskal, J.B.** juni 1964, Psychometrika, Vols. 29, no 2, pp. 115-129.

[25] *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.* **Kruskal, J. B.** Maart 1964, Psychometrika, Vols. 29, no 1, pp. 1-27.

[26] **Rojas, Raúl.** *Neural Networks: A Systematic Introduction.* 1996.