UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Markus Lippus

Predicting Illness and Type of Treatment from Digital Health Records

Master's thesis (30 ECTS)

Supervisors: Sven Laur, PhD
Anna Leontjeva, MSc

Tartu 2017

Predicting Illness and Type of Treatment from Digital Health Records

Abstract: The rising costs of healthcare and decreasing size of the working population is a dire problem in most of the developed world. While it is inevitable that new methods are costly, it is possible to reduce avoidable expenses by better planning and prevention. Most hospitals keep digital records of everything that happens to a patient during their treatment and in Estonia all medical bills are also presented to the Estonian Health Insurance Fund (EHIF) for reimbursement. In this work the data from EHIF is used to build a model that as the first step uncovers the different clinical pathways followed for the treatment of patients with an illness. As a second step the model is used to predict the number of patients that will be provided the uncovered treatments in the future. The output of such a model could be a valuable asset for planning resource allocation and preventative health care.

Keywords:
Process mining, health care, predictive modeling, model based clustering
CERCS: P170

Haiguse ja ravitüübi ennustamine kasutades digitaalseid raviarveid

Lühikokkuvõte: Kasvavad kulud tervishoius ning samaaegne töötava populatsiooni kahanemine on kriitline probleem kõikjal arenenud maailmas. Ühest küljest on paratamatu, et uued ravimid ja meetodid on kallid, on teisest küljest võimalik vähendada välditavaid kulutusi parema plaanimise ja ennetustööga. Enamik haiglad salvestavad digitaalselt kõik, mis patsiendiga ravi jooksul toimub ja Eestis esitatakse kõik raviarved ka Eesti Haigekassale (HK) hüvitamiseks. Käesolevas töös kasutatakse HK andmeid ehitamaks mudelit, mille abil on võimalik tuletada erinevad raviprotsessid, mida patsientide ravimisel kasutatakse ning samuti ka ennustada patsientide hulka, kes tulevikus vastavat ravi vajavad. Selline mudel võiks olla kasulik suunamaks otsuseid vahendite jaotamisel ja enntustöö suunamisel.

Võtmesõnad: protsessi kaevandamine, ennustav modelleerimine, tervishoid, mudelipõhine klasterdamine

CERCS: P170

Table of Contents

Used abbreviations

EHIF – Estonian Health Insurance Fund

HMM – Hidden Markov Model

ICD-10 – International Classification of Diseases, revision 10

NMF – Nonnegative Matrix Factorization

LDA – Latent Dirichlet Allocation

COPD – Chronic Obstructive Pulmonary Disease

SVD – Singular Value Decomposition

1 Introduction

Although in general decreasing costs accompany growth in efficiency as technology advances, this does not hold true for health care (Kumar, 2011). The rising costs in health care are a dire and growing problem for all developed nations regardless of how the system in the respective country is set up. This puts remarkable strain on the societies as a whole and therefore solutions that mitigate this problem are greatly needed.

Most hospitals today use some sort of information systems to collect and store information about their patients in the form of numbers, text and images, with reports saying that in the US alone the amount of data has reached zettabyte scale. This data holds immense potential, but is rarely thoroughly analyzed to aid practitioners in their work or to bring about an increase in the quality of health care in general.

As in Estonia the whole population is covered by state provided health care, the system is set up so that all medical bills pass through a central organization called Estonian Health Insurance Fund (EHIF). As a result the data collected by the EHIF contains every health care service provided to every person in Estonia. The implications of this for data analysis compared to a more decentralized system where data is stored in various information systems employed at each hospital are astounding. The notable benefits include the unified naming and pricing scheme for services and analyses and a single data format. Although as a drawback the granularity of the data is lacking in several aspects - EHIF operates an insurance scheme and they collect data that's relevant to fulfilling this function. As a result the data is lacking in granularity, meaning more detailed information, such as drug dosages and test results are not available.

This places Estonia in a favorable position regarding health care data analysis as at least the high level data about the treatment of any illness is available in a unified format for the whole population. This enables, through careful analysis, both discovering high level models for the whole population and pointing to areas where more specific data could help in generating more useful insights.

1.1 Objective

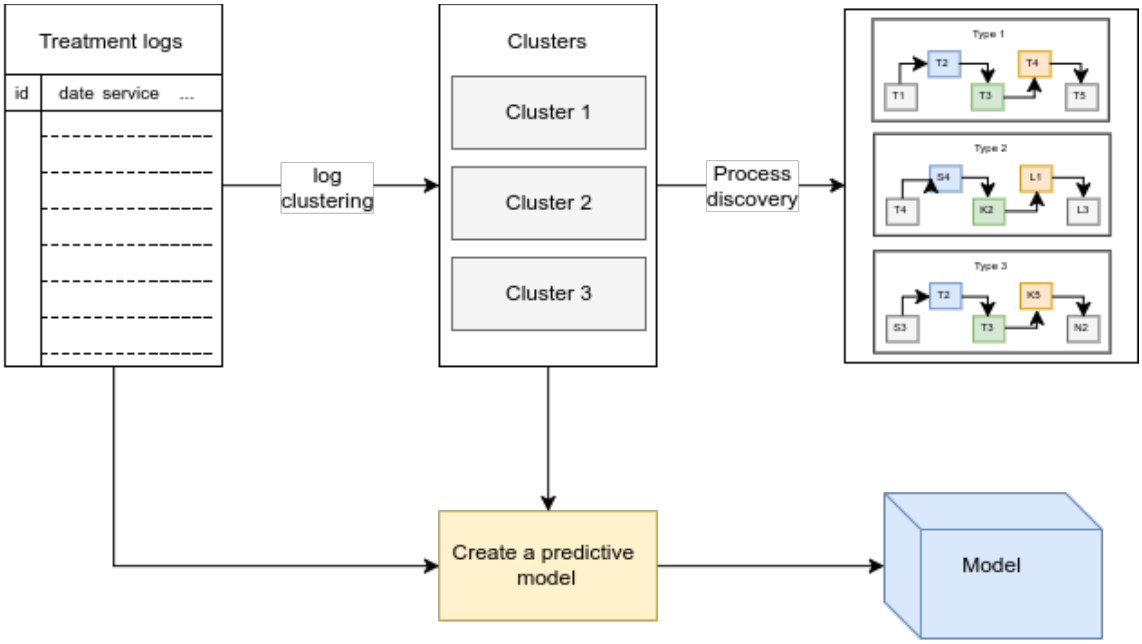

Figure 1: The planned path and objective of the work. The treatment logs will be used to find different types of treatment for an illness. From these filtered logs process models will be mined that give an overview of the procedure used on patients in a group.

Using the historical information about a patient a predictive model can be built, using both the whole database of logs and filtered logs, that can predict the likelihood of a patient getting a disease and then the type of treatment they will receive.

The aim of this work create a framework for population based prediction of costs and other parameters related to the treatment of illnesses. The approach proposed here consists of multiple steps illustrated on Figure 1:

1. Finding different groups of treatments based on the events taking place over the course of the treatment;
2. Using the patients' history and other relevant and accessible data to create a prediction model for the likelihood of a person getting the illness;
3. Using background information about the patients with the illness to create a prediction model for the likelihood of a person going to be assigned to any of said groups;
4. Infer a descriptive and easily readable model of the treatment course in said groups.

This work would result in a workflow, shown on Figure 2, that would enable one to predict the prevalence of some disease in the population in the future and also obtain information about the likely future treatment of these patients.

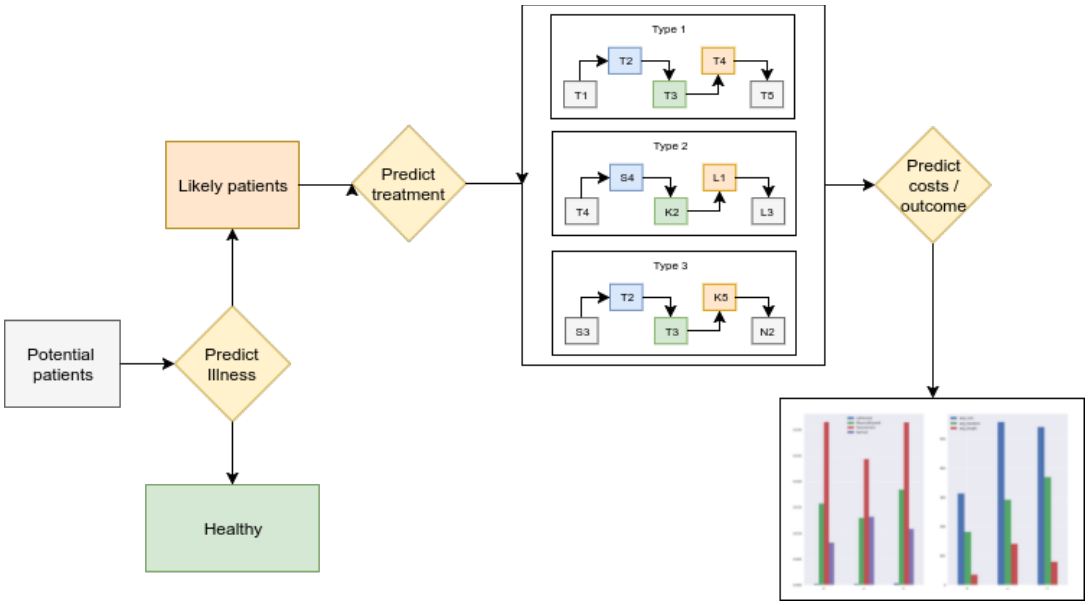

Figure 2: The future use of this work. From the patient history it would be possible to predict the likelihood of a patient getting a specific disease. Fro the patients classified as likely to get the illness, it would be possible to predict the type of treatment they would receive.

This would allow making better predictions about the future cost and outcome of such diseases. It is difficult to make accurate predictions about health care as it is a rapidly changing field with advances in preventive work and treatment happening at a fast pace. The advantage of the approach used in this work is that knowing the kind of treatment a patient is likely to receive provides additional information about the illness such as severity or type. This information could aid in making informed decisions about the possible corrections in the predicted distribution as increase in preventive efforts may offset some type of the illness while not effecting the others.

1.2 Previous work

The variability of the data in health care allows for a multitude of approaches to prediction and modeling. A lot of work has been done using only clinical data such as analysis results and genetic information about patients (Marinov, Mosa, Yoo, & Boren, 2011; Palaniappan & Awang, 2008). These works have shown considerable success in using methods from data mining to identify important factors in the development of specific illnesses and predict the risk of developing illness.

In the process mining side a lot of work has been done regarding process discovery (Lakshmanan, Rozsnyai, & Wang, 2013; Lang, Bürkle, Laumann, & Prokosch, 2008; R. S. S. Mans et al., 2009; R. Mans & Schonenberg, 2008). The aim of these works has mainly been to discover the underlying processes in treatment and administration. Knowledge of such workflows would allow for better planning as it enables calculating realistic timelines and better understand the frequency of occurrence of certain workflows, but measuring conformance could also provide better insight into guideline planning. Sadly it does appear, that current approaches do not satisfy all the requirements necessary for effective process discovery from the treatment log data of patients (Yang & Su, 2014).

The studies that concern predicting the costs in health care vary widely in both the data and methods used for this purpose. There have been fairly successful reports predicting high cost patients as a proxy of the actual cost (Moturu, Johnson, & Liu, 2007) and even the specific monetary value of future health care requirements (Sushmita et al., 2015). The latter is also relevant to this work in terms of the underlying data as one of the datasets used consists of clinical claims.

Previous work on similar datasets has been done in Estonia by the National Institute for Health Development and Raul Kiivet (Tonsiver et al., 2014). Also Estonian Genome Center includes EHIF data in the analysis of genome data and EHIF itself regularly analyses the data they collect (Estonian Health Insurance Fund & Group, 2015).

1.3 Outline

Materials and methods gives an overview of the data and the methods that are used. Puts the methodological background into the context of this work.

Results explains the results obtained using the methods previously described.

Discussion and future work provides summary of the results, discusses the shortcomings, implications and possible further advances of various aspects of the work.

Conclusion summarizes the results and includes the final notes and takeways of the work.

Bibliography lists articles, books and software used in this work for reference is specifics are required.

2 Materials and methods

2.1 Data

The data used in this work is a part of a larger dataset that originates from the Estonian Health Insurance Fund containing all visits to all doctors excluding dentists and general practitioners, that occurred during the period of 2010-2016. In this work two groups of illnesses were investigated: malignant neoplasm of the breast or C50 by International Classification of Diseases, revision 10 (ICD-10) nomenclature (WHO, 1992) and J44 or chronic obstructive pulmonary disease (COPD). The features available for each service are shortly described in Table 1.

The choice of the diagnosis was based several factors such as frequency of occurrence, severity of the illness and duration of treatment. Both diagnoses can bring about a number of difficult complications, such as metastasis on the case of breast cancer and osteoporosis and heart failure in the cases of COPD. This makes this illnesses more likely candidates for having multiple treatment procedures. The treatment of chosen illnesses is also short and the cases frequent enough to have a reasonable amount of cases start and end during the time period for which data is available – an important factor for characterizing the treatment.

The data used in this work was extracted as follows:

1. all bills with the diagnosis code and its subtypes as the principal or secondary diagnosis were queried from the database;
2. for all patients, all medical bills preceding the initial diagnosis of the illness under investigation were retrieved (historical bills);
3. for all patients, all illnesses diagnosed with them preceding the initial diagnosis of the illness under investigation were retrieved.

In the database there were 10,420 patient with the diagnosis C50 and 30,162 patients with the diagnosis J44.

As the treatment of these illnesses lasts over a considerable time, most of the cases do not start or end in the period for which data is available. For purposes explained later in the work, cases that have a beginning and an end in the scope of the data were separated.

Table 1: The available features for every row in the services table.

year	The financial year to which the bill is attributed
bill id	Unique identifier for the medical bill
patient id	Unique identifier for the patient
age	The age of the patient at the time of start of the medical bill
date of birth	The date of birth of the patient
sex	The sex of the patient
place of residence	county of residence of the patient
service provider location	parish of the service provider
service provider code	unique identifier of the service provider
service provider name	name of the service provider
service provider type	type of service provider (e.g. hospital, private clinic)
service provider type code	numerical identifier for the previous feature
doctor specialty code	Specialty of the doctor providing the service
doctor specialty name	Specialty of the doctor providing the service
unavoidable health care	Boolean showing if heath care was avoidable or not
start date	Start date of the bill
end date	End date of the bill
end code	Reason for ending the bill
specialty code	Domain code for the treatment
specialty name	Domain name for the treatment
continued bill	Shows whether the bill was made as continuation to a previous one
DRG proportion	the amount of the bill that is payable using the disease related group (DRG) funding method
total of the service	The total cost of the service
amount of the service	the amount the service was provided
net total of the service	cost multiplied by the amount
principal diagnosis code	the principal diagnosis at the time of providing the service
principal diagnosis name	the principal diagnosis at the time of providing the service
amount	The number of times the service was provided
coefficient1	The coefficient NHIF used to pay for this service
multiple	times the service was provided
time of provision of the serv	the date at which the service was provided
service type code	a group to which the service belongs to
service type name	a group to which the service belongs to
days	how many days the service was provided for
paid sum	how much NHIF paid for the service
service code	code of the service provided
service name	name of the service provided
type of treatment	type of the treatment (ambulatory, stationary...)

There are a total on 16 different reasons for closing a bill used in the dataset. Here we used four of these. The reason all others were dismissed here is that they indicate either that the patient was either asked to return at a later date or directed to another doctor –

both of which imply continuation of the treatment. In this work the end of a case is defined as having one of the following reasons for closing the bill:

- patient left on their own volition against the doctor's recommendations;
- death;
- other reasons;
- improvement or convalescence.

2.1.1 Processing

For every patient a a set of background information was produced. This was formulated as a vector, comprising of the attributes described below.

1. Patient age in days at the time of the initial diagnosis.
2. The sex of the patient;
3. Previous diagnoses – for that purpose the previously extracted set of diagnoses was used.
4. The exact diagnosis code given under ICD-10 C50 or J44.
5. The specialty codes of all doctors previously encountered.
6. Occurrences of frequent sets of services.

For identifying the frequent sets of services to use as features the services were extracted from all the historical bills of all the patients. These were treated as transactions and from these transactions frequent item sets were found using the FPgrowth algorithm (Han, Pei, & Yin, 2000). These item sets or sets of services were then used as features in the background vector as either having occurred in the patients medical history or not. The thinking behind this is that although some services are very informative on their own, such as previous surgery for example, but most of the services, such as various blood tests, might only be informative in sets of services which together indicate some nuances in the previous treatment of the patient.

2.2 Process mining

Information systems log enormous amounts of data about the activities they handle in various processes in various kinds of industries. These logs hold information about the process that is being carried through and also how well a model characterizes real life.

Process mining is a field concerned with extracting knowledge from these event logs with the aim to improve efficiency and better understand the underlying processes.

Most of the work in the field has been focused on logs from manufacturing systems and customer support (Greco, Guzzo, Pontieri, & Saccà, 2006; Pospíšil, Mates, Hruška, & Bartík, 2013), but recently more focus has been given to medical field and clinical process discovery specifically (Dalianis, Hassel, Henriksson, & Skeppstedt, 2012; Lang et al., 2008; Yang & Su, 2014).

The field is usually categorized into three types by objective (W. M. P. van der Aalst, 2011), which are

Discovery – finding a process from the event logs without any prior knowledge of it.

Conformance - testing if the information in the event logs corresponds to the model under investigation.

Enhancement – Similar to discovery, but in this case the investigator has prior knowledge in the form of a process model which they aim to improve.

In this work we are interested in process discovery as our interest is finding the different processes underlying the treatment logs in our disposal.

2.2.1 Event logs

In the context of process mining, event logs, an example of which is shown on Table Error: Reference source not found, are a type of logs that contain information about the execution of a process over many instances of such execution. This means the events in the log are recorded so that (W. van der Aalst, Weijters, & Maruster, 2004):

1. each event refers to a task in the process;
2. each event refers to an instance of the workflow or case;
3. events are totally ordered.

pat_id	duration	type	treatment_profile	diagnosis	price	service	kpv
*	5	A	A38	C50.3	176.72	3002	*
*	5	A	A38	C50.3	193.64	6074	*
*	5	A	A38	C50.3	134.42	7903	*
*	5	A	A38	C50.3	115.62	66707	*
*	5	A	A38	C50.3	140.06	3004	*
*	0	A	A38	C50.3	11.42	3002	*
*	2	A	A38	C50.3	11.42	3002	*
*	2	A	A38	C50.3	12.51	6074	*
*	2	A	A38	C50.3	8.66	7903	*
*	2	A	A38	C50.3	7.46	66707	*
*	2	A	A38	C50.3	9.04	3004	*

Figure 3: An example of an event log using the data used in this work. Here the service column refers to the task in the process, pat_id – patient id, refers to an instance of the workflow and the events are totally orderable using the column kpv – date. In this example the dates and patient id-s are removed for privacy concerns.

In the data used in this work it would seem obvious to use the bill id as a case identifier. This would result in a distorted image of the logs as new bills are sometimes started in the middle of the treatment cycle for various causes like getting assigned to another doctor or accounting reasons. For that reason patient id is used as the case identifier so a case is formed of services that are provided to a single patient in the scope of the diagnosis under investigation and is ordered by the dates at which the services were provided.

The medical process is inherently very varied and the number of different procedures in the logs very large. It would be of great use if it was possible to disregard the services that are irrelevant to the current diagnosis or if a hierarchy existed using which the services could be merged. Alas no such hierarchy exists for the services used by EHIF and defining less relevant services is non trivial.

2.2.2 Discovery

In his work we are most interested in the discovery aspect of process mining as we would like to infer the process that has generated the logs under investigation.

While some sources of logs lend themselves well to process discovery algorithms and produce consistent, well defined process models, some sources contain more variation and using naïve algorithms on such logs often result in spaghetti-like models, shown on

Figure 4. This may be the result of either an inherently variable process, different models producing the logs or both. In the case of health care processes the latter is arguably most likely. Not only is every patient different, which brings about some variation, but also the severity and type of the same illness varies and thus different treatment procedures are required.

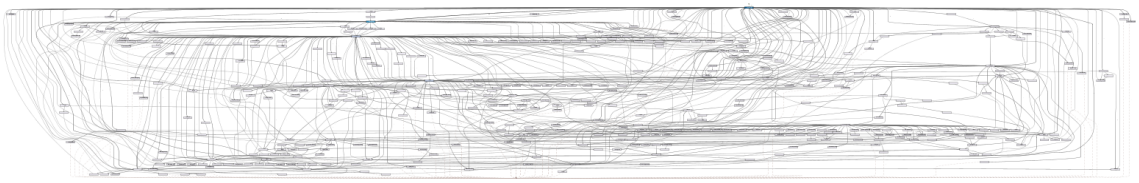

Figure 4: An example of a "spaghetti model" uncovered from a heterogeneous process.

This problem can be partially mitigated by using clever algorithms such as Fuzzy Miner (Günther & Van Der Aalst, 2007), but these methods may also remove important, but rarely occurring relations and events such as a specific kind of surgery. It also does nothing to uncover the various models that may have generated the logs under investigation, but tries to explain all the logs with a single model.

For such cases trace clustering has been attempted, which involves grouping traces similar by some metric into groups to get more easily readable models (Bose & van der Aalst, 2010; Delias, Doumpos, Grigoroudis, Manolitzas, & Matsatsinis, 2015).

2.3 Cluster analysis

Although broadly the diagnosis in either group of the patients in the dataset used in this work is the same, there are very likely important differences between specific cases. A part of this difference is of course captured in the specific diagnosis code assigned to the patient (C50 and J44 would be called a “non-specific diagnosis” in the ICD-10 hierarchy), but these codes do not always specify the severity of the illness or the specifics of the patient that may cause the case to unfold in a different manner.

Such differences are likely important in determining the type of the treatment a person will receive and also strongly influence both the outcome and the total cost of the treatment. It is likely that such information can be inferred from the treatment procedure

itself and that using this information it could be possible to find groups of patients that required different type of treatment arising from differences in disease type and the patient's background.

Cluster analysis is a method for grouping objects under investigation into groups based on some similarity measure where the objective is to assign group labels to clusters so that intra-group similarity is maximized while the similarity between groups is minimized. This kind of analysis enables inferring the most natural structure of the data at hand (if there is one) with minimal knowledge about it beforehand. In the context of this work, the clustering of the cases would serve three purposes:

1. finding clusters based on treatment type would enable us to characterize the clusters such as duration, end result and total expenses of the treatment;
2. it would allow us to infer a process model for different treatment processes of the illness to better explain what happens to the patient over the course of the treatment;
3. it would allow for predicting into which cluster a person is likely to fall. This would enable us to predict the possible outcomes and the likely cost of the illness in the future. The clusters here are important as these also describe the illness itself and so make it possible to take into account the changes to the treatment in the future.

2.3.1 Clustering

A multitude of clustering methods have been introduced with their own benefits and drawbacks and while some work better with some type of data than others, it is often up to the investigator to choose the best method for the data they have. Following (Han & Kamber, 2006) these can be broadly classified into partitioning, hierarchical, grid-based and model-based clustering. Often these methods are also combined in an effort to infer structure in the data under investigation.

Most clustering methods treat objects as points in space and assume that there is a defined similarity measure to assess the objects being analyzed. In many cases this is the case and a simple metric such as Euclidean or Manhattan distance ca be used. There are cases though, where defining a usable distance metric is either very difficult or impossible. As previously stated the data used in this work are sequences of medical

procedures – sequences of discrete values, for which finding a suitable distance metric is not a simple matter.

2.3.2 Model based clustering

As most data can be represented in a vector form it is easy to think of them as points in n-dimensional space and use an appropriate distance metric to separate these into groups.
The problem with this approach is two-fold:

1. some information may get lost if certain types of data (sequences, sound) are handled in such a way;
2. different groups of data may have different parameters and as such spatial closeness may not imply same origin.

These problems can be solved by using model-based clustering. This type of clustering does not try to group objects into k clusters by similarity, but find the models that are most likely to have generated the data. Besides being a better fit for certain types of data the model also provides a better description for the groups that it finds from the data.
The most common models for this purpose are Gaussian mixtures and multinomial models, but for some more complex data such as time series, Markov chains and Hidden Markov Models (HMMs) have been used widely (Bicego, Murino, & Figueiredo A.T., 2003; Panuccio, Bicego, & Murino, 2002; Smyth, 1997). As HMMs are inherently very suitable for capturing the sequential nature of medical records, these are used in this work.

Hidden Markov Models

A Hidden Markov Model is a method for modeling sequences and discovering the underlying properties of the process that generates the observable sequences. In the scope of this work the observable sequence would be the services provided to a patient and the underlying model would represent the conditions of the illness the patient is suffering.

An HMM consists of a number of hidden states H, a transition matrix A and an emission matrix B. At any time point t the model is in a single hidden state. The emission matrix specifies for each hidden state the likelihood of generating an observed value while the

transition matrix specifies the likelihood of moving from state n_i to state n_j. A formal definition can be found in Table 2 and an illustrative image on Figure 5.

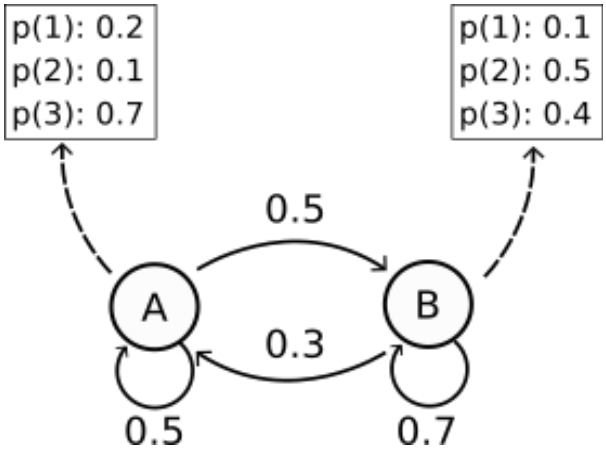

Figure 5: An illustrative figure of a HMM. The states are A and Band possible emitted values are 1, 2 and 3. On the edges are the transition probabilities and in the boxes the emission probabilities of the indicated hidden states.

Table 2: The formal definition of an HMM.

Model	$\lambda = (A,B)$
Number of states	I
Number of observations	T
A set of states	$N = \{n_1, n_2, \ldots n_i\}$
Transition matrix A	$A = (a_{ij})$
Emission matrix B	$B = (b_{ij})$
Sequence of observations	$Y = (y_1, y_2, \ldots y_t)$
Sequence of hidden states	$\Pi = (\pi_1, \pi_2, \ldots \pi_n)$

The observed values here are the units of the sequence and the hidden states model some underlying labeling of the process. In this work the observed values could be the services provided to the patient and the hidden states are the specific issues that lead to providing these services. Rabiner defines three fundamental problems that characterize HMMs: (Rabiner, 1989)

1. likelihood – determining the likelihood of an observed sequence using a HMM $P(Y|\lambda)$;

2. decoding – discovering the most likely sequence of hidden states, given a sequence of observations and a HMM

   $\Pi = \underset{\pi}{argmax}\, P(\pi|Y,\lambda)$;

3. learning – find the transition matrix *A* and emission matrix *B*, given a set of sates *N* and an observation sequence *Y*

   $A,B = \underset{A,B}{argmax}\, P(A,B|N,Y)$.

For the purposes of using HMMs for clustering, points 1 and 3 are most important, as here we aim to both find the most likely model for each trace and also learn the most likely model given the traces assigned to it. Finding the likelihood is done using the forward algorithm, while fitting the model is done using the forward-backward algorithm.

Forward algorithm

To find the likelihood of a sequence of observations given a HMM, one needs to compute the probability for all possible sequences of hidden states for having generated the observed sequence. As the number of hidden states and possible values increases, calculating this directly quickly turns infeasible

The forward algorithm makes this possible by using dynamic programming. This is done by computing a dynamic programming matrix *f* where the rows are hidden states and the columns are elements of the sequence. An element $f_k(i)$ of the matrix *f* is the probability of being in state *i* after the first *k* observations:

$$f_k(i) = P(x_{1,} x_2, \ldots, x_i, \pi_i = k)$$

where $\pi_i = k$ is the path through the hidden states and x_i is the probability of the *i*-th element of the observed sequence having been generated by the model.

Using dynamic programming this can be calculated recursively using the probabilities already calculated for the previous time point:

$$f_k(i) = b_{ij} \sum_{j=1}^{N} f_j(i-1)\, a_{jk}$$

where b_{ij} is the probability of emitting element *i* at state *j*, and a_{jk} is the probability of transitioning from state *j* to state *k*.

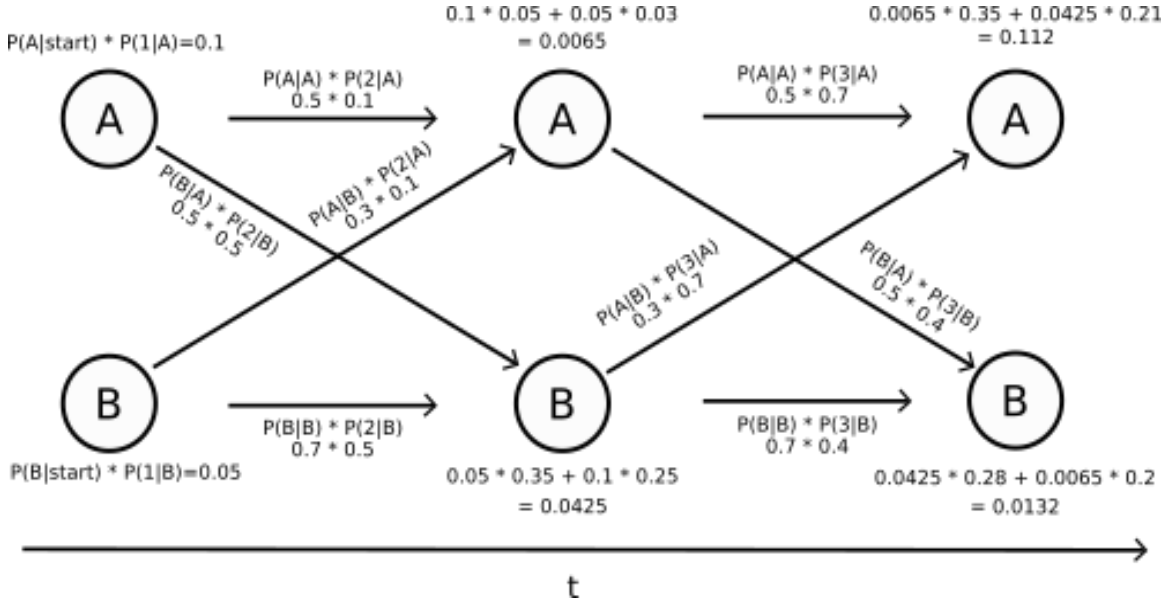

Figure 6: An illustration of the forward algorithm. The example has been done for calculating the probability of a sequence '123' using the model seen on Figure 5. On the nodes are the probabilities of being at that node at the corresponding time step. At every state the probability of getting there and emitting the required value is found and stored. The stored values from the previous time step can be used at the next step reducing the amount of computation necessary.

Forward-backward algorithm

The forward-backward algorithm, or the Baum-Welch algorithm (Rabiner, 1989) is an iterative algorithm, which trains both the emission and transition probabilities by iterating over cycles of computing estimates for both probabilities.

As our aim was to use HMMs to find optimal clusters for the models we used the following hard-clustering algorithm:

1. k HMMs were initiated randomly – the values in both the transition matrix and emission matrix were generated randomly;
2. for every trace the probability of having been generated by all of these models was found;
3. every trace was assigned to the model that most likely generated it;
4. all models were trained with the traces assigned to it.

5. Repeat steps 2 to 4 until less than 1% of traces change model assignment.

As with any type of clustering there is the issue of choosing the right number of clusters K, but as HMMs are the models we also must choose the number of hidden states *m*.
Usually for choosing the right number of clusters methods like comparing intra-cluster variance against extra-cluster variance or the silhouette method could be used, but these methods assume a distance metric of some sort. As the data used in this work consist of sequences of discrete values, defining a distance metric suitable for this purpose is a non trivial task. For this reason Bayesian information criterion (BIC) (Schwarz, 1978) and cross validation are used here to aid in choosing the number of clusters.
BIC is a method for model selection and in essence chooses the model that provides the most benefit without introducing too much complexity. It is formally defined as:

$$BIC = k\ln(n) - 2\ln(L)$$

where *k* is the number of free parameters, *n* is the number of data points and L is likelihood of the model. In the literature Akaike information criterion (AIC) (Akaike, 1974) has also been used for similar purposes, but the choice here was made in favor of BIC as it prefers simpler models.
The same approach was used to find the optimal number of hidden states *m*.

2.4 Topic modeling

The nature of the data used in this work makes clustering in this way a bit less trivial. In the data used in this work the precision with which the time a service was provided can be pinpointed is one day and the sequence of events during one day is unknown. In very rare cases only one service was provided in a day so the total ordering of the services cannot be done. This poses two problems:

1. the sequence of events in a day may be important and this information is lost
2. the number of services provided in one day is not equal and this may cause anomalies.

The first problem can be mitigated by sorting the services in a uniform manner so that if the same set of services is provided during a day these would be considered equally. This

does not mitigate the problem of different number of services though and it is possible that this produces artifacts of its own. To address the problem of varying number of tasks topic modeling was used as preprocessing.

Topic modeling is a method for extracting the main themes or topics from a collection of documents, usually a collection of texts. The algorithms that do this are probabilistic and analyze the frequency of the words in the texts to assign a topic or a mixture of topics to every text. Although mainly used for modeling texts, many other applications have been found for these methods such as pattern finding in images and social networks (Blei, 2012).

In the context of current work this could be used in a number of ways. The simplest approach would be to assign a single, most likely topic to each day as the treatment phase that generates the set of services provided, and try to model the sequence of these principal topics. The problem with such an approach is that the sequence of events would have to be mostly described by a single topic or a lot of information about the day would get discarded while disregarding the less prevalent topics of the day.

A little more complex model could be treating the topics as sets of services, as can be easily done in the case of nonnegative matrix factorization (NMF). Each day could be considered a fixed size set of topics such as *{A, B, C}* and the model would attempt to describe each day as such a combination. The model describes the data better, but is more complex than the first option and requires defining a specific loss function.

As topic modeling algorithms output a fixed size mixture for each sample, it would also be reasonable to attempt to describe each day as the specific mixture of the available components. This is by far the most computationally difficult task of the three and requires the most data, but would also most accurately describe the data.

From the discussed options the first and the last were attempted with different topic modeling algorithms.

2.4.1 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a probabilistic generative method that attempts to model every text as a collection of topics while a topic itself is a distribution over words (Blei et al., 2003).

While all topics are distributions over the same fixed vocabulary different probabilities the distributions are different. For example in this work the vocabulary is the list of services provided to the patients at a hospital and a topic could be for example “admission” or “surgery”. In the admission topic services pertaining to reception and some initial tests like bloodwork would be assigned higher probabilities, while radiation therapy would be less likely to occur. LDA assumes that these topics are known beforehand and all the documents are generated from these using the following process:

1. a distribution over topics is chosen randomly;
2. for every word in the document:
    1. a topic is chosen using the distribution produced in the first step;
    2. a word is chosen randomly from the distribution of the topic;

The method assumes that all documents have been generated in this manner and are thus a mixture of these topics. As in actuality the topic structure is hidden and the texts are observed, the objective is to reverse the generative process and use the documents as evidence to find this hidden structure.

In the case of LDA there is the matter of choosing a good value of topics. As perplexity is a measure often used to assess how well a number of topics describe the text, we used this as the likelihood in BIC estimation. To do this we split the cases by day and treated the set of procedures in one day as a document. When topics were generated we used these to label each day with a topic. LDA assigns a mixture of topics to each day so the most influential topic was chosen for each day. If there was no clearly dominant topic for the day a “no-topic” label was assigned to the day.

2.4.2 Nonnegative matrix factorization

NMF has been used both for topic modelling and dimensionality reduction [CITATION]. As shown on Figure 7 it is a matrix factorization technique, meaning it tries to construct a factorization of the form $V = WH$, while minimizing the reconstruction error. As the matrices have sizes $n \times k$ and $k \times m$ and the number k is up to the user to choose, this method can be used as a compression method when choosing a k smaller than n or m. (Seung & Lee, 1999). As computing exact NMF is NP-hard (Vavasis, 2010) in this work an implementation described in (Lin, 2007) is used.

In the current work the rows of the matrix *V* are the visits made by the patients and the columns are all the services provided to them. Each row is a vector of services provided to a patient during one day. An example of the resulting matrix can be seen in Table 3.

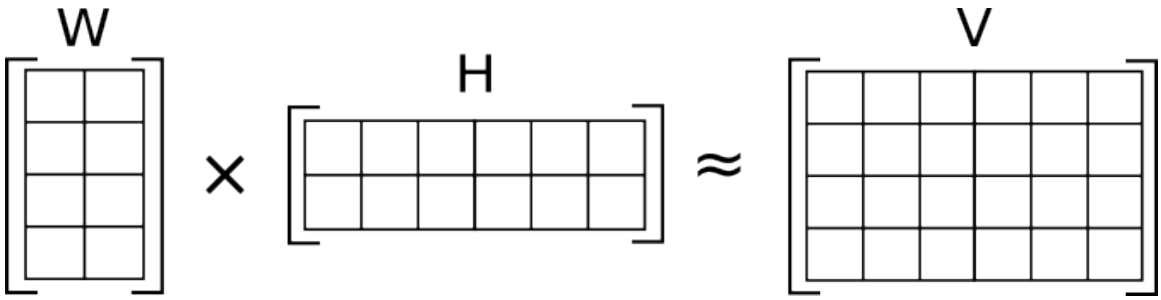

Figure 7: An illustration of nonnegative matrix factorization. V *is the* $n \times m$ *matrix being factorized into smaller matrices* W *and* H *of sizes* $n \times k$ *and* $k \times m$ *respectively (Qwertyus, 2013).*

As the name implies the method poses a constraint that all elements in matrices *W* and *H* must be nonnegative and thus NMF will not find results that exclude factors from the original matrix. As all the combinations are additive, the result can intuitively be thought of as separating the features into k groups of features. This makes NMF very suitable for solving the aforementioned problems, as using it makes it possible find the latent groups of services and at the same time provides, for each day, a composition of these *k* groups.

In this work *k* was chosen so that it would be possible to discover groups of approximately 5-10 features. The rationale here is that the groups of services provided at hospitals would most likely be at around this size and this also corresponds well to the average number of services provided in a day.

Table 3: Table showing a sample of the matrix being factorized. There is a row for each day a patient was provided medical services. The columns correspond to the services provided and for each row the number of times a service was provided to the patient during that day is stored.

kpv	**5**	**7**	**...**	**923O**	**999O**
*	0	0	...	0	0
*	0	1	...	0	0
*	2	0	...	0	0
*	0	0	...	0	0
*	0	0	...	0	3
*	0	0	...	1	0
*	0	0	...	0	0

2.5 Classification

As the aim of this work is to create a system that would require as little fine tuning as possible, the classification algorithms were chosen to be as robust as possible, with a reputation of requiring little parameter optimization and being easy to use.

2.5.1 Random forest

Random forests were first introduced by Breiman (Breiman, 2001) and the algorithm is an ensemble method that can be used for both classification and regression tasks.

As many ensemble methods, random forest combines the predictions of multiple weak learners to build a single good classifier. In the case of random forest these classifiers are decision trees. Each tree is trained with approximately 2/3 of the data available while the rest of the samples are considered out-of-bag data and used to evaluate both the error of each tree and importance of each variable (Breiman, 2001).

The last point makes random forests very useful for our purposes and in health care data analysis in general. If a system is supposed to give answers to help humans make better decisions in a critical field such as health care, it is of utmost importance that the system is also able to explain its suggestions as well as possible. From a random forest model it is possible to extract the relevance of each parameter, which lets the user better understand why it made the predictions it did. This is useful as both reassurance to the user and it may also point out good predictors, which could otherwise be overlooked by humans.

2.5.2 Gradient Boosted Trees

Gradient boosted trees (Friedman, 2001) is similar to random forest as it is also an ensemble of decision trees, but the principles behind fitting the trees is slightly different. In a random forest the trees are fitted in parallel with them being independent of each other. Gradient boosted trees on the other hand is a greedy algorithm that trains each next tree to better classify the samples with which the previous ones had trouble with.

The algorithm can be described in the following steps:

1. train a weak learner on the data;

2. calculate the loss and reweigh the examples by giving extra weight to examples that the current ensemble has trouble with;
3. train a new learner on the newly weighed examples;
4. add the tree to the ensemble and repeat from step 2 until instructed.

The algorithm is prone to over fitting so it is important to limit the individual trees to a very small size. It is not uncommon for the individual trees to have two leaves.

As a gradient boosting method this algorithm introduces learning rate as an extra parameter, but it is still comparatively simple to use and serves as a comparison to the effectiveness of random forest.

3 Results

3.1 Topic modeling

To process the data with both LDA and NMF it was transformed to a matrix where each row corresponds to a day for a patient and every column is a service. There are as many columns as there are services provided under the diagnosis, which are 1020 for C50 and 1628 for J44. An example is shown in Table 4. The values in every cell correspond to the number of times the service was provided during this day to this specific patient.

Table 4: Table showing an example of how the data was formatted for NMF and LDA.

service_code	pat_id	kpv	5	7	...	923O	999O
0	*	*	0	0 ...		0	0
1	*	*	0	0 ...		0	0
2	*	*	0	0 ...		0	0
3	*	*	0	0 ...		0	0
4	*	*	0	0 ...		0	0

For finding the number of topics suitable for LDA perplexity was used with cross validation as a measure of goodness. For this the dataset was separated into two while keeping the test set size at 10% of the whole data. We tested various numbers of topics and recorded the perplexity measures, shown on Figure 8. The results of the experiment were similar for both the patients with the diagnosis C50 and J44 and 10 topics were chosen as the optimal number.

After fitting the topics, a distribution over these topics was assigned to every visit by every patient. For most of the visits the best topic could be easily chosen, but when no clearly dominant topic could be found a label of “no-topic” was assigned to the visit.

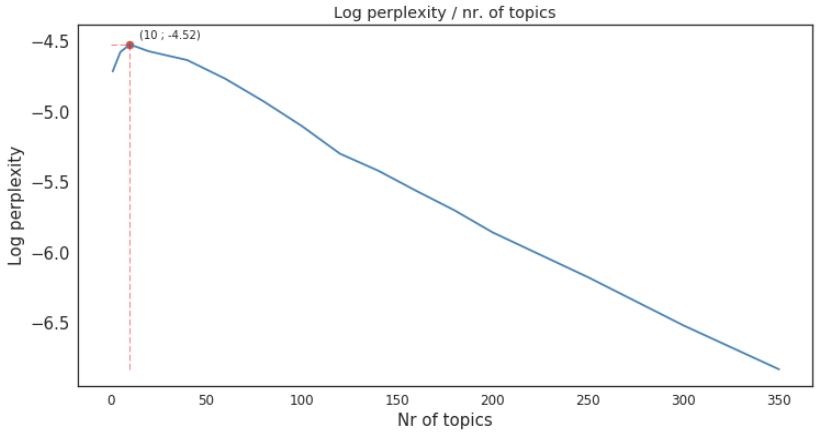

Figure 8: The perplexity values at various n topics. The graph indicates that 10 topics provides the most optimal results.

```
7 9 9
8 8 8 8 3 8 3
7 7 7 7 7 9 4 4 4 4 4 4 2 8 4 6 4 4 8 6 8
8 2 6 10 3 9 8 9 10 9 5 9
5 0 1 4 4 2 2 6
```

Figure 9: A random selection of five traces over topics from the patients with diagnosis J44. For each day a patient visited the doctor a topic is recorded. "10" corresponds to the day where no single topic could be selected.

For NMF an intuitive number of 200 components was chosen, the rationale being that every component would most likely contain 5-10 services and 200 components would satisfy that condition for both datasets.

3.2 Clustering

As two different preprocessing methods are used on the data, the structure of the HMMs must correspondingly be different to accommodate that. In the case of LDA it is fairly straightforward as every state in the HMM simply emits the main topic of the day or the "no-topic" label if there is none. As NMF produces are 200 values for each day, a a multivariate gaussian HMM was used with 200 covariates.

To find the optimal number of hidden states H and number of models K, multiple values for both were tested and BIC was calculated based on the log likelihood of the model.

3.2.1 Choosing the parameters

The data sets in all cases were split 80/20 to training and test sets, model was trained on the training set and then the log probability of the test set was found. This log-probability was used to compute the BIC values shown on Figure 10. From the figures it is evident that no conclusions about a suitable number of hidden states can be drawn from the BIC values. The reason for the apparent linear growth here is that the complexity penalty component of BIC completely overpowers the increase in the likelihood of the model.

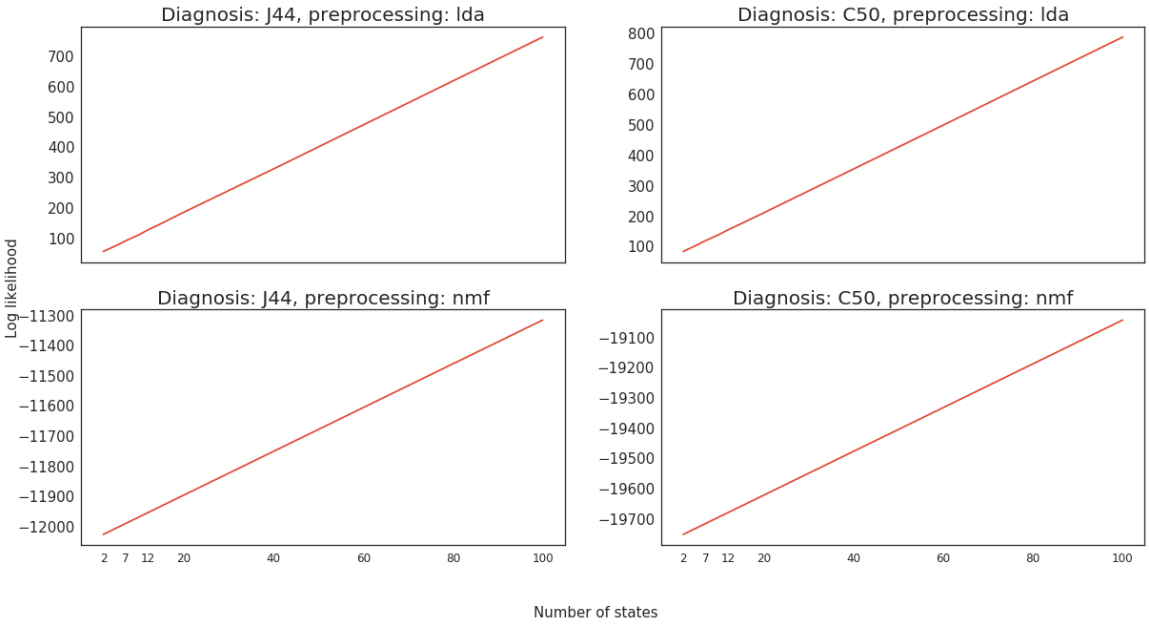

Figure 10: The BIC values of the clusterings of the two datasets using two different preprocessing methods. The penalty of BIC completely overpowers the increase in the model likelihood and no conclusions can be made about the hidden state selection.

This is not uncommon when using BIC as the penalty term is dependent on the number of any chosen parameters. The increase in the likelihood of the model may be proportional to the increase in complexity induced by the larger number of parameters , but the constant in front of the penalty term may cause it to completely drown it.

As such, the log likelihoods themselves were investigated and the corresponding plots can be seen on Figure 11. As expected, the increase in log-likelihoods is very small indeed and completely nonexistent in the cases where NMF was used a preprocessing method (There is a difference in the sixth decimal). The result from NMF is very

unexpected and the most prudent conclusion to draw from these results is that our approach using NMF is not valid here and we should not proceed with this method. As such, going forward only the results from LDA based clustering is reported. 40 and 20 hidden states were chosen for models for J44 and C50 respectively.

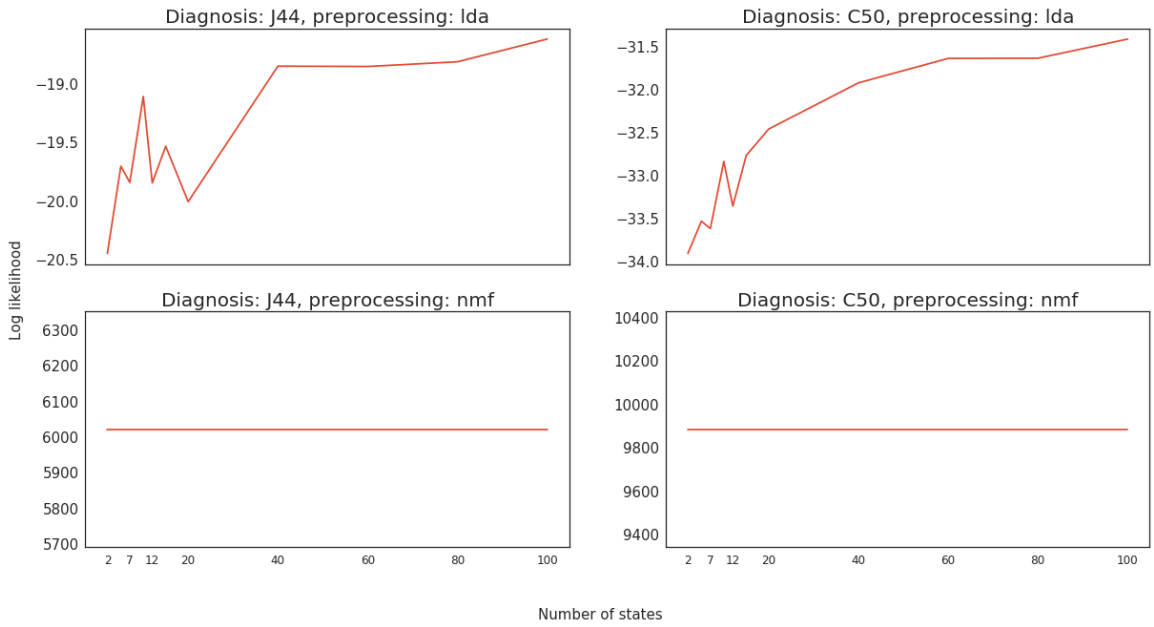

Figure 11: The log likelihoods of the models at different numbers of hidden states.

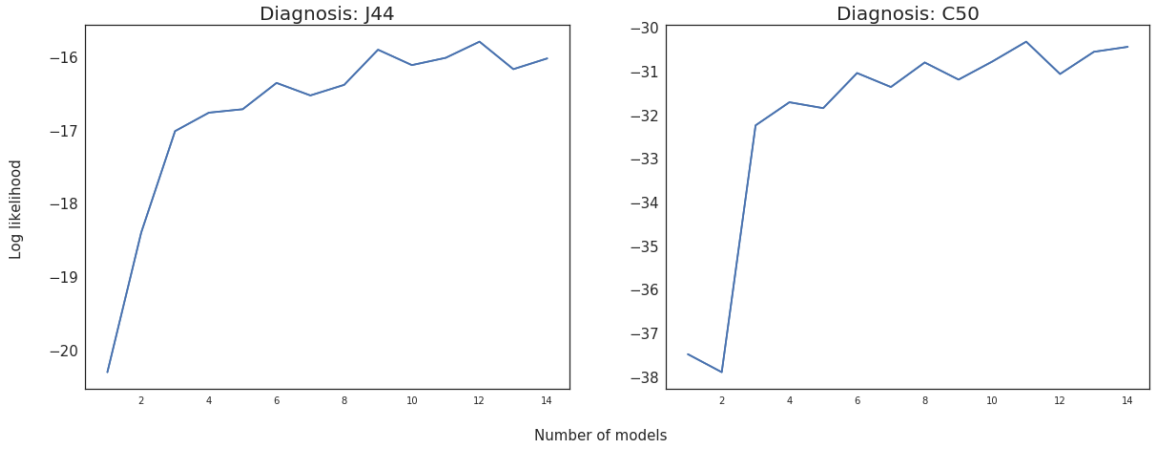

Figure 12: The log likelihoods of the different numbers of models.

The number of models *k* was chosen in a similar manner to the number of states. The experiments were with the traces from LDA using various values of *k*. The results can be seen on Figure 12. Judging from the plots 3 models were used for both diagnoses.

3.2.2 Visualizing the clusters

To visualize the clusters a number of parameters of interest were chosen by which to compare the clusterings. A number of diagnosis end codes were chosen, shown in Table 5. There are a total of 16 different end codes and the basis of this selection was that all other codes concern directing the patient to another doctor or setting the next visit and are thus not indicative of the result of the case.

Besides the codes also the length of the treatment, the number of services provided to a patient and the average cost of the whole treatment were selected as attributes for comparison. The described attributes make the assumption that the case has already found its end so only cases that have a beginning and an end in the time frame the data is available for were used for computing these attributes.

Table 5: Table showing the reasons for closing a treatment bill and the corresponding codes.

Code	Description
9	Left on own volition
10	Death
11	Other reasons
15	Convalescence

The resulting plots for both COPD and breast cancer can be seen on Figures 13 and 14 respectively, the cluster sizes for both diagnoses are shown in Table 6. The clusters are fairly distinctive in regards to the measured parameters and strikingly the clustering for both diagnoses look very similar with regards to the cost/duration/length scale.

For J44 there is a clear separation of difficult and simple cases as there are almost no cases resulting in death in the first cluster, the cases are far shorter, cheaper and require less services on average. In contrast the proportion of cases resulting in death in relation

Table 6: Table showing the number of traces in each cluster for both diagnoses. The columns marked "total" display the total number of available traces assigned to each cluster. The marking "full traces" denotes the number of traces assigned to the specified cluster that both begin and end in the time frame for which data is available.

Diagnosis	J44 total	C50 total	J44 full traces	C50 full traces
Cluster 1	16 529	4 090	5 232	307
Cluster 2	2 872	3 544	246	156
Cluster 3	10 763	2 657	1 508	168

to convalescence changes drastically for the other two clusters. The other two clusters could be described as the difficult with the second one, while far smaller than the other two clusters, contains the most fatal cases with the ratio of convalescence to death being turned on its head in relation to the other two clusters and the costs skyrocketing to several times that of the cases in other clusters.

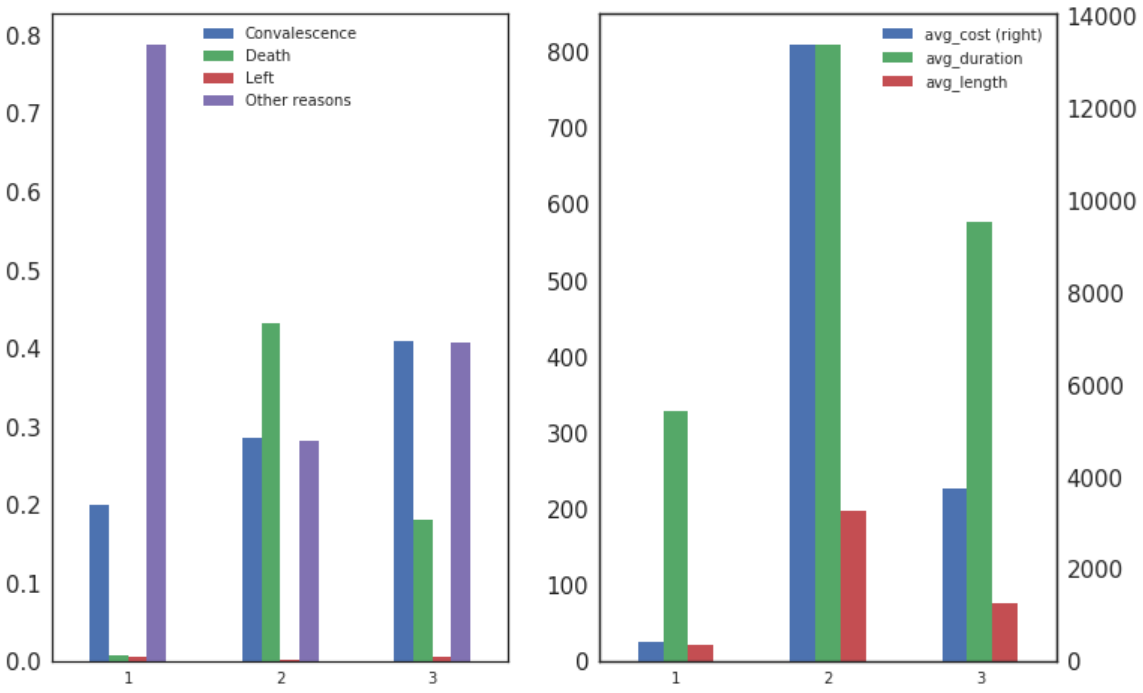

Figure 13: Figure visualizing the clustering of traces with diagnosis "J44".

The left hand plot shows the distribution over the possible end codes with the y-scale showing the proportion of all cases.

The right hand plot shows the distribution over the chosen parameters. The left y-scale shows the average duration in days and average length in number of services while the right y-scale shows the average cost in euros.

The clusters are similar in the case of C50 in the sense that cluster 1 contains the “simpler cases” with proportionally fewer cases resulting with death and cases being cheaper overall and requiring less services on average, while the other two clusters represent the more difficult scenarios. There is a significant difference though, in the two clusters representing the more difficult cases. The most obvious is that all the clusters are of similar size so that if in the case of J44 the worst case was more of a rare situation, in the case of C50, the number of cases in cluster 2 even exceed the number of cases in the less sever cluster 3.

For both diagnoses there is a high number of cases ending for “other reasons”. For a person outside the domain, this result is difficult to interpret, but could provide some useful insight to a person with more knowledge about the actual treatment procedure of the illness as for some illnesses there may be a few more common reasons for ending treatment with this specific code.

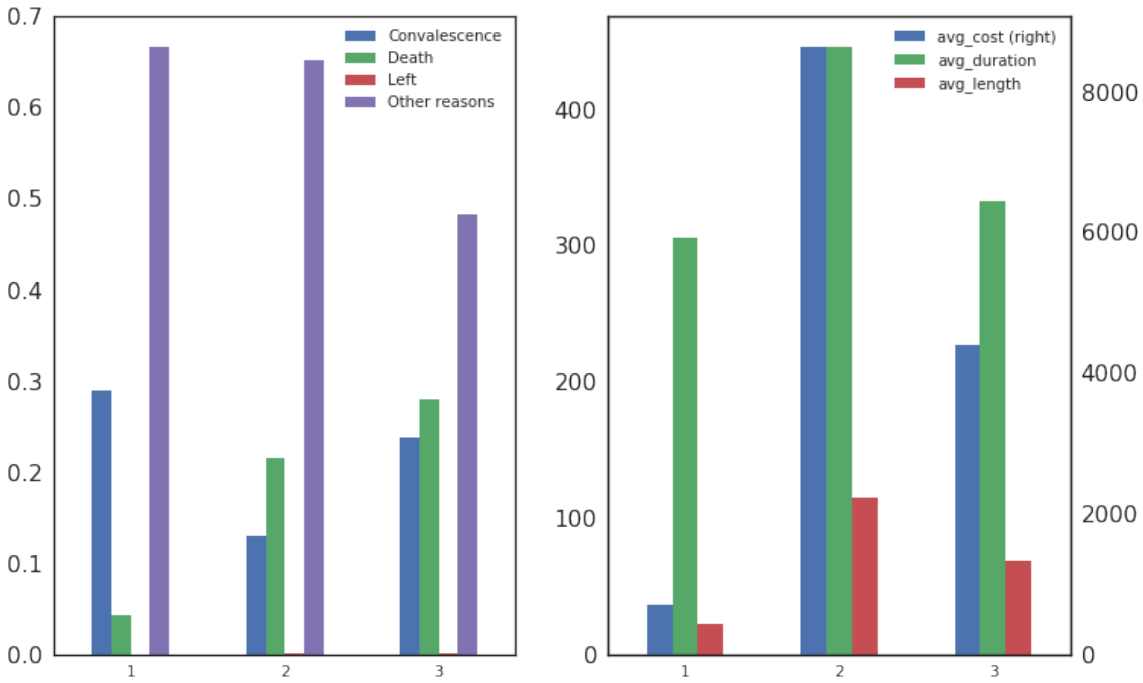

Figure 14: Figure visualizing the clustering of traces with diagnosis “C50”.

The lefthand plot shows the distribution over the possible end codes with the y-scale showing the proportion of all cases.

The righthand plot shows the distribution over the chosen parameters. The left y-scale shows the average duration in days and average length in number of services while the right y-scale shows the average cost in euros.

3.3 Discovering the clinical pathways

Measuring the parameters as in the previous section gives us an idea of the differences between the clusters, but does little to explain the cause of the differences. Here process discovery is used to discover the underlying clinical pathways that correspond to each previously found cluster.

Even though clustering the traces drastically reduces the number of unique services that would be used to build the models, which makes the models a lot easier to understand, there is room for improvement in reducing the number of services. This is especially true in the more difficult cases as the reason why these cases take so much more time and money to treat and have a higher chance of not ending with recovery, are the complications accompanying the main diagnosis.

The complications are myriad and range from infections to cancer. These additional conditions require a host of services to treat and increase the complexity of the pathway. Simplifying this is a difficult matter as the aim is to produce a pathway that is informative, but at the same time not overwhelming.

Attempts were made at filtering the services using rules, combining services based on similarity and using the Fuzzy Miner algorithm to automatically abstract away clusters of services.

3.3.1 Fuzzy Miner

At first, attempts were made at using Fuzzy Miner on the logs as they were. For testing purposes we chose cluster 3 from the J44 cases as this had the largest number of distinct events – 422.

There are two principal software packages used for process mining tasks that have Fuzzy Miner implemented: ProM (Verbeek, Buijs, van Dongen, & van der Aalst, 2010) and Disco (Fluxicon, 2017). Here Disco has been chosen. Disco allows setting parameters that effect the work of the Fuzzy Miner algorithm, such as the percentage of events and connecting edges to display. When displaying all the events the resulting model is unreadable, as shown on Figure 15. The reasonable level at which to show the model is very dependent on the logs and thus on the diagnosis and cluster. This makes this part of the process hard to automate. A reasonably sized model for obtaining an overview of the process is found at 2.6% of the traces, shown on Figure 16 . This looks more

comprehensible, but this most likely removes important data, as the most significant edges that lead to closing a case start from triage and “Eriarsti korduv vastuvõtt” - repeat reception by a specialist doctor.

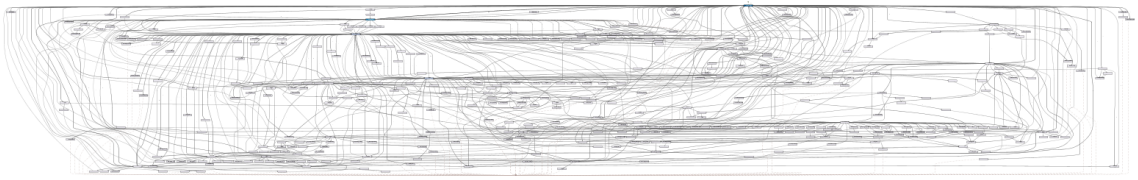

Figure 15: Figure showing the process model used in cluster 3 of diagnosis J44 with all events and edges displayed. This is what is called a “spaghetti model” in the field of process modeling.

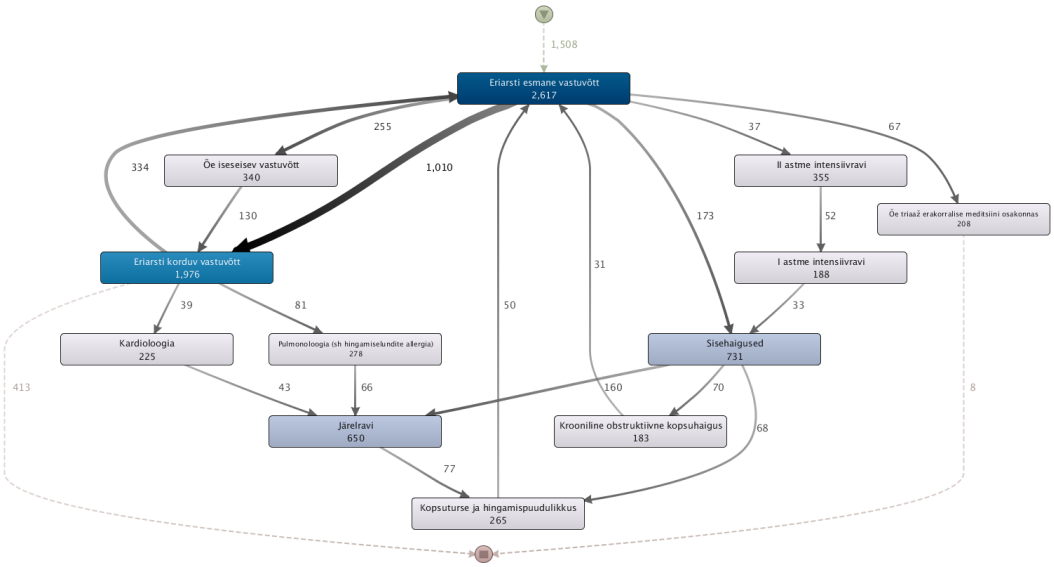

Figure 16: A process model found using Disco. Nodes limited to 2.6% edges to just the most significant ones.

To alleviate this problem a set of rules were devised, shown in Table 7, by looking at the set of 422 services to merge services where the nuances may not be that important for a general understanding of the clinical pathway underlying the care given to patients in a cluster. Such merging is not ideal as it requires human input, but as the hierarchy for the services that is available in the data is very limited this was explored as an option. As a more easily automatable steps we also removed all events pertaining to transportation and

reception and events that occur in less than 1% of the cases as these are most likely not representable of the cluster under investigation.

The resulting model discovered after such simplification is shown on Figure 19.

Although not evaluated by and expert in the field the model looks fairly comprehensible and informative about what happens to patients over the course of the treatment such as complications that develop over the course of the treatment, likely reasons for admission and the events prior to the end of the case.

Using this method models were found for all the clusters for both diagnoses. The pathways for patients with diagnosis J44 are shown on Figures 17, 18 and 19 and the pathways for patients with diagnosis C50 are shown on Figures 20, 21, 22 and 23.

As the number of services that were merged or pruned differed with the clusters, the parameters for the Fuzzy Miner algorithm were separately tuned for each graph generated.

Table 7: Table showing the rules generated by hand to merge similar services in the treatment logs.

Source	**Target**
Contains “Anesteesia”	ANESTEESIA
Anything with service category is “VERI JA VERETOOTED”	LAB
Contains “Recovery”	RECOVERY
Contains “Intensiivravi”	INTENSIVE CARE
Contains “taastusravi”	REHABILITATION
Contains “kemoteraapiakuur”	CHEMO
Contains “kasvaja” and “operatsioon”	TUMOR SURGERY
Contains “triaaž”	TRIAGE
Contains “infektsioon” or “nakkus”	INFECTION

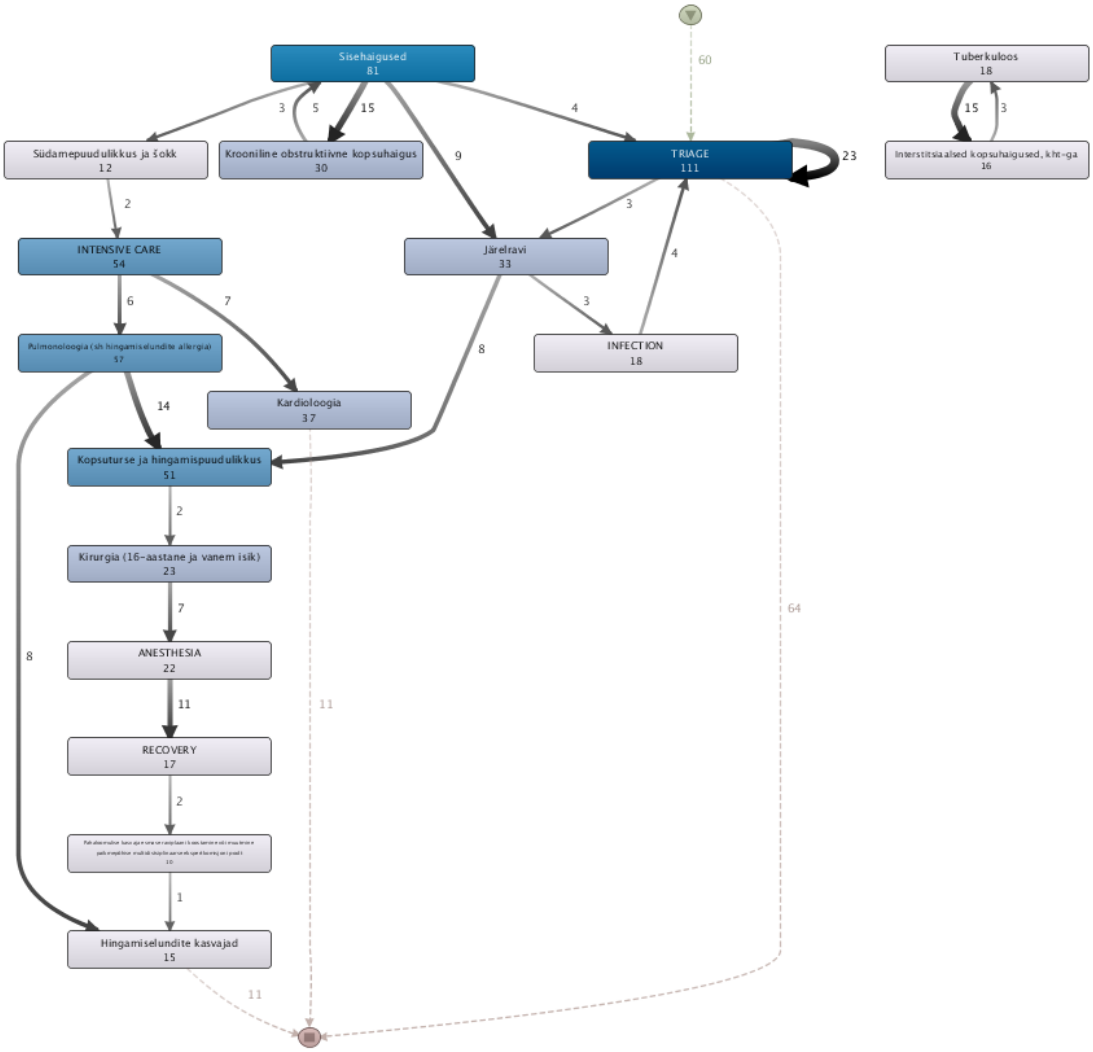

Figure 17: Clinical pathway for cluster 1 of diagnosis J44 on logs filtered with manually defined rules. There is a seemingly separated segment consisting of tuberculosis and a lung disease on the right side of the graph. This is in fact connected to the rest of the process, but fuzzy miner deemed the connecting edges insignificant and these are thus removed from the graph.

In this pathway the main pathways seem to be centered on cardiology and cancer besides COPD itself.

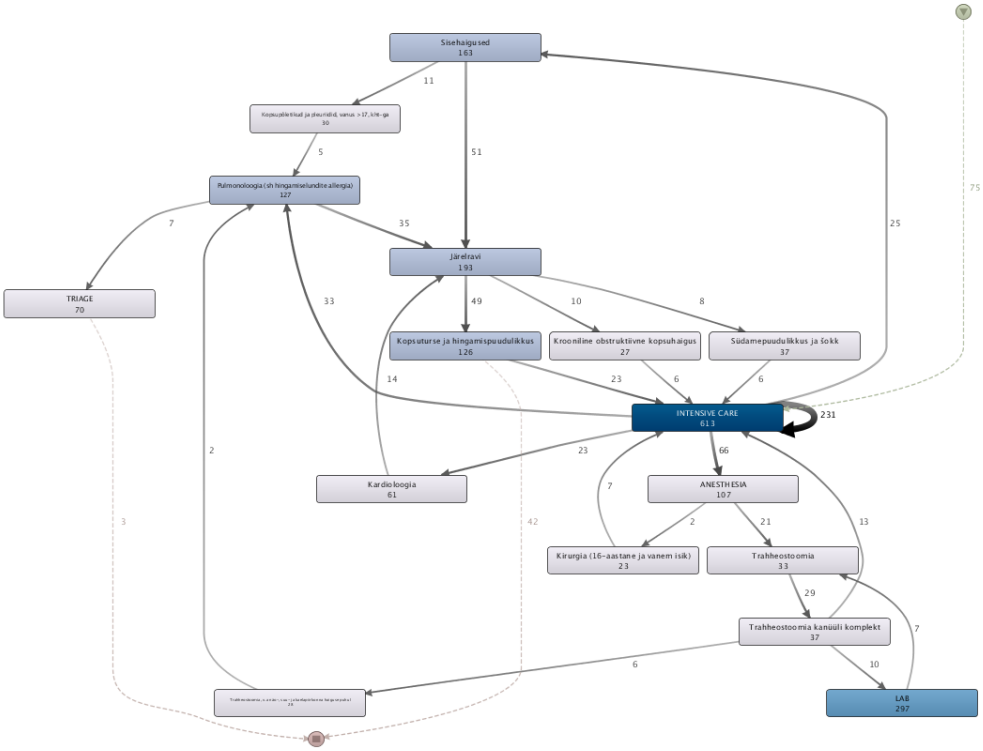

Figure 18: Clinical pathway for cluster 2 of diagnosis J44 on logs filtered with manually defined rules.

The pathways for this cluster are noticeably more complex, with more complications related to respiratory organs and heart. Interestingly, cancer seems to have much less importance in this cluster.

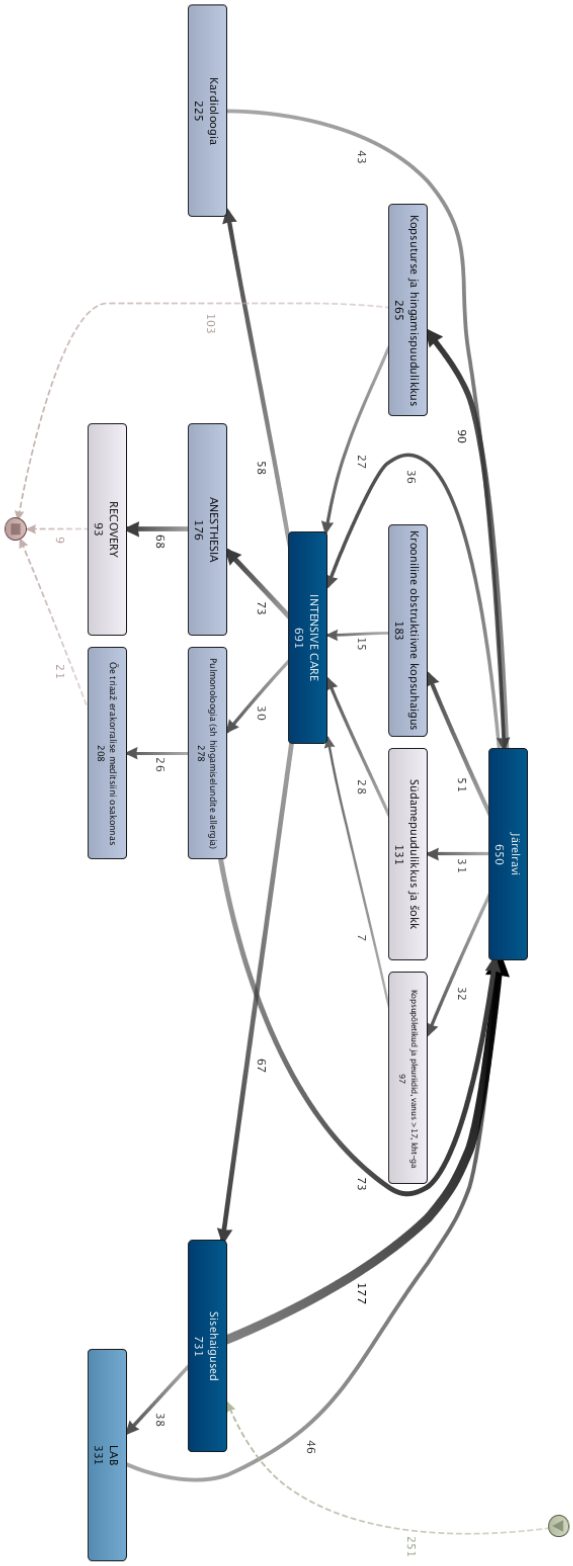

Figure 19: Clinical pathway for cluster 3 of diagnosis J44 on logs filtered with manually defined rules.

Principal complications seem to be the same as in cluster 2.

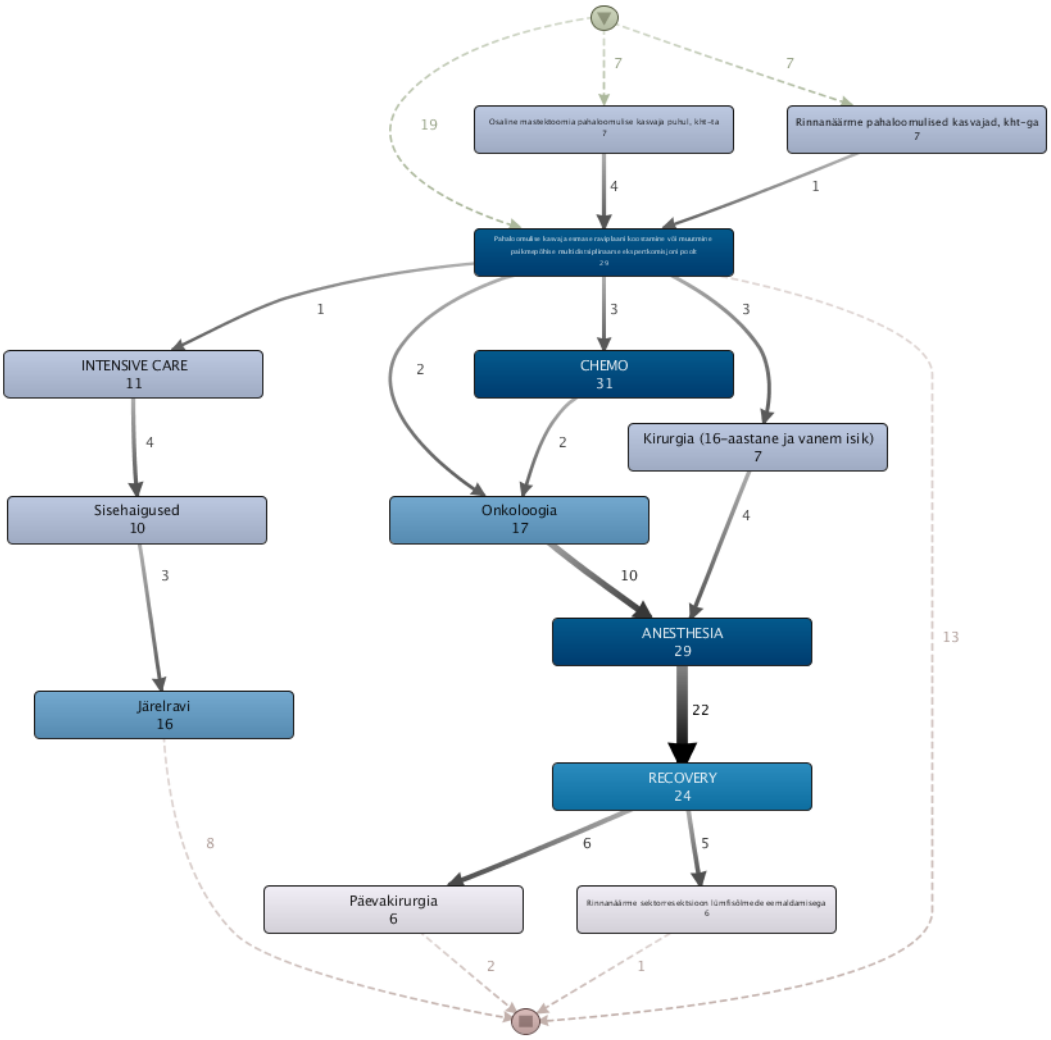

Figure 20: Clinical pathway for cluster 1 of diagnosis C50 on logs filtered with manually defined rules. This cluster has the proportionally highest survivability and we can see from the model that the paths don't lead back meaning that recurrence of the cancer is rare among these patients.

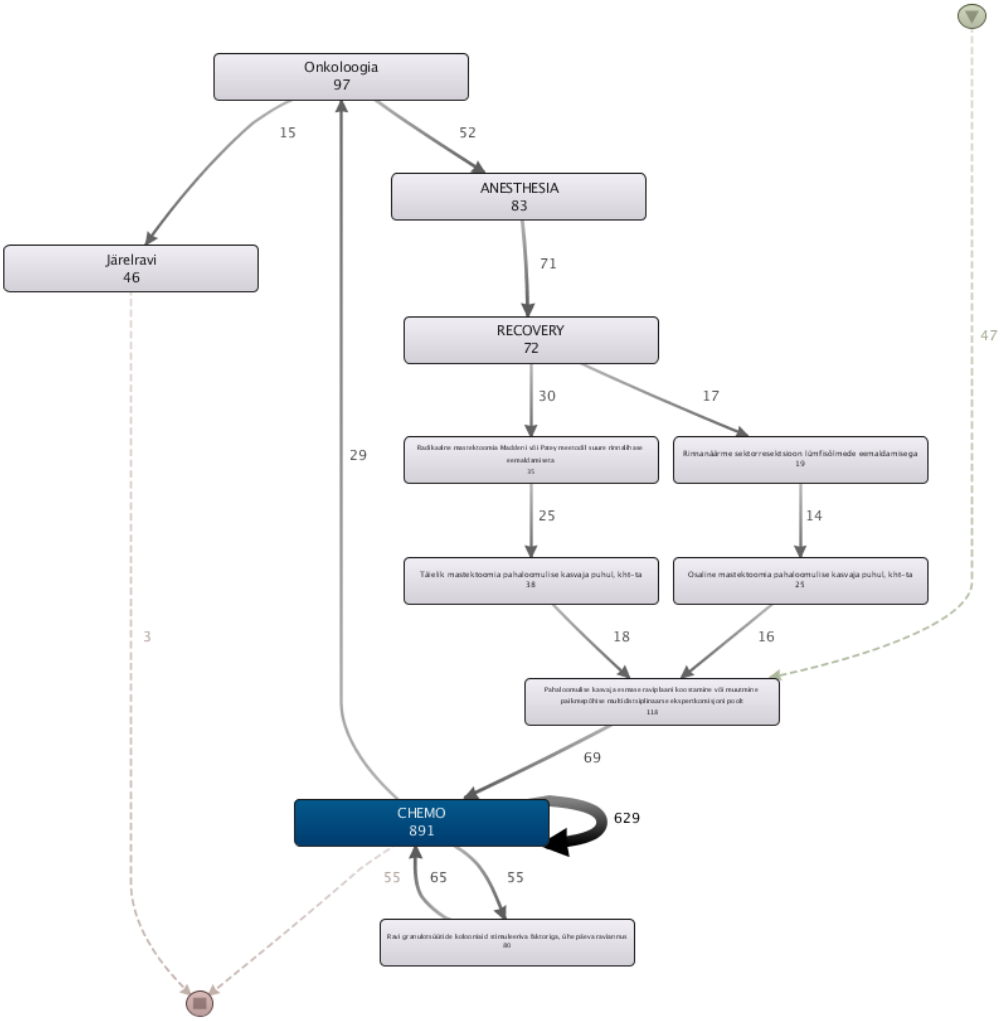

Figure 21: Clinical pathway for cluster 2 of diagnosis C50 on logs filtered with manually defined rules. A closeup of the events where the text is too small to read is shown on figure 23.

Here a rather surprisingly clear pathway can be seen as the patients go through surgery then chemotherapy and may or may not go through such repeated cycles depending on the possible recurrence of the cancer.

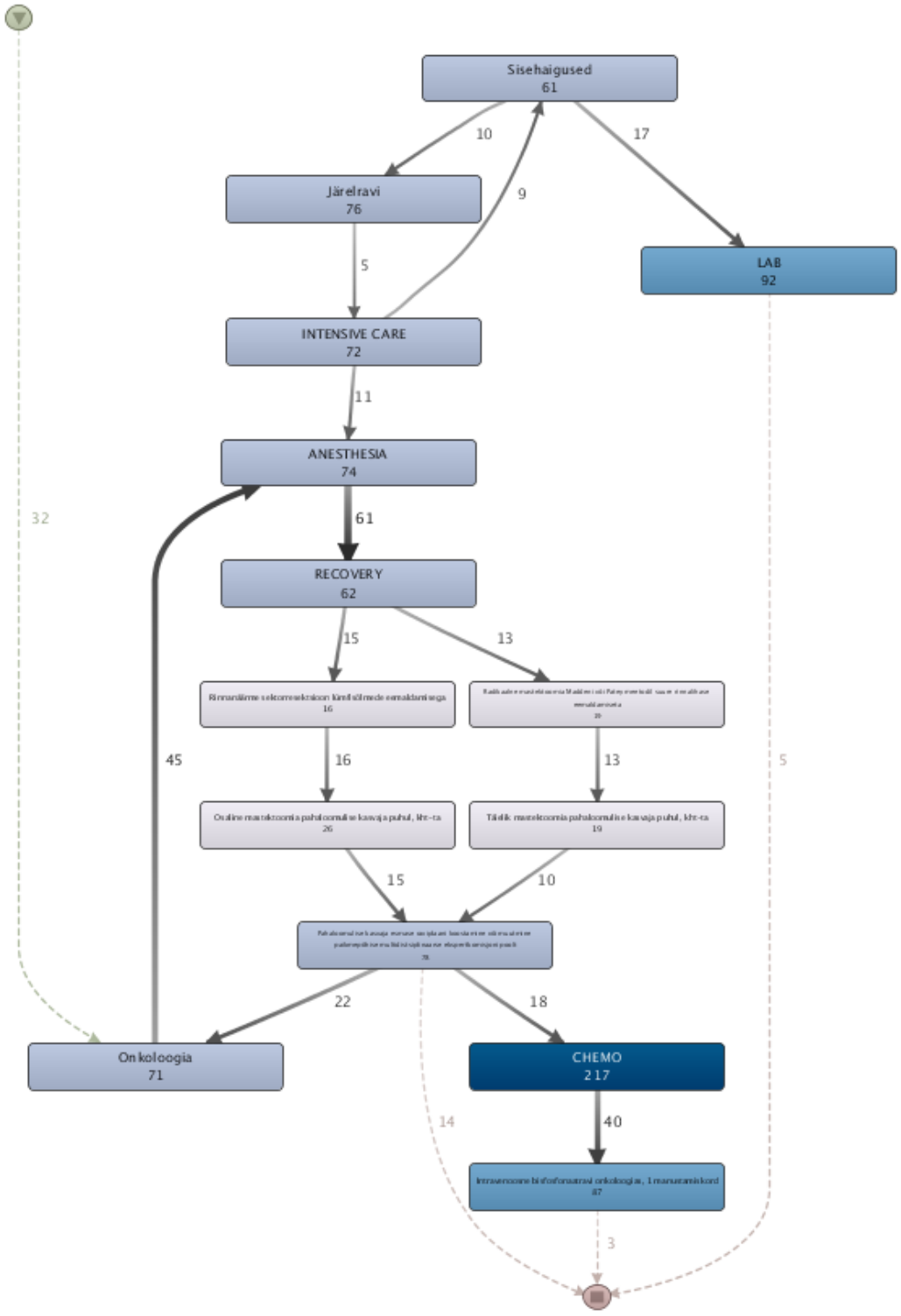

Figure 22: Clinical pathway for cluster 3 of diagnosis C50 on logs filtered with manually defined rules. A closeup of the events where the text is too small to read is shown on figure 23.

Similar in nature to cluster 2 with the additional unspecified complications in the field of internal medicine.

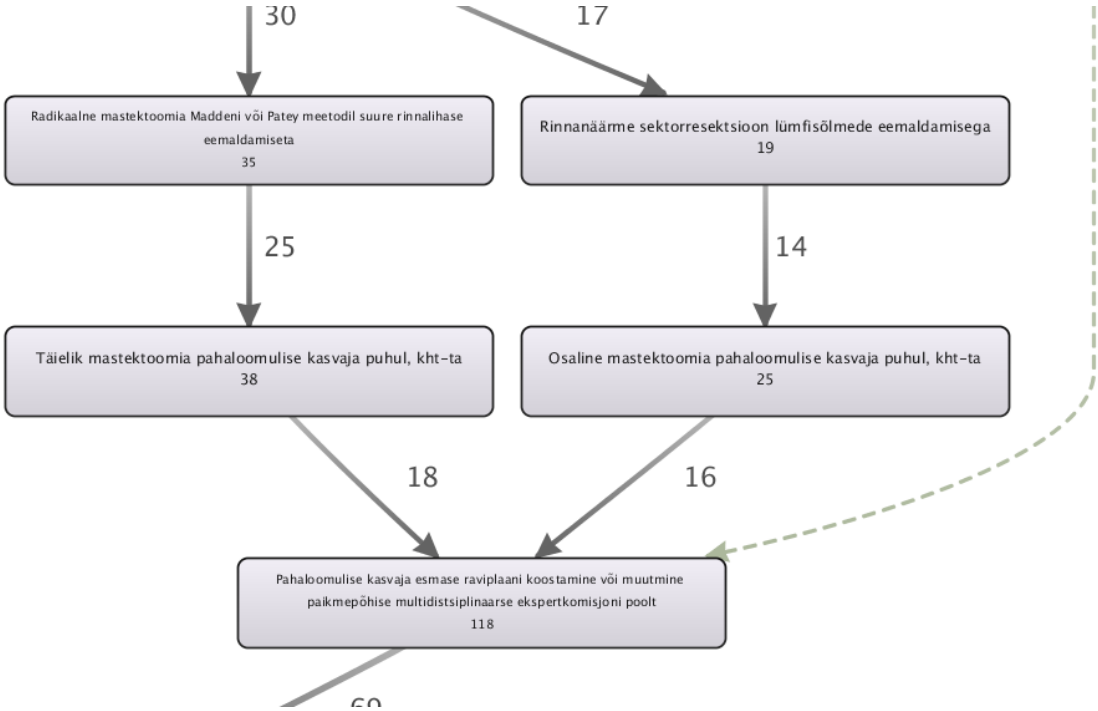

Figure 23: A closeup of a bundle of events from the clinical pathway graph for clusters 2 and 3 of diagnosis C50.

The closeup is a set of services concerning partial or total mastectomy.

3.4 Predicting illness and type of treatment

The predictions are made in two steps:

1. Predict the number of people getting the illness.
2. Predict the the treatment these patients are going to receive

If these two steps are reasonably accurate, the results could be used to predict the likely number of future patients and estimate the burden they will place on the health care system through costs and treatment requirements.

For the people that already have the illness, defining their treatment history is fairly straightforward: everything preceding the initial diagnosis of interest is treatment history. This is not possible for people who never got ill during the time range available for this research. To create a reasonable treatment history for these patients, a limit was imposed at 1st January 2015 and everything preceding that date was considered treatment history. The whole known treatment history could not be used for these patients, as they might have fallen ill instantly after the point where data is available. So the assumption was made that if they have not fallen ill after at 1st January 2015, then they can be considered healthy people for the purposes of predicting the illnesses. The treatment histories found using this method were combined with the histories of the verified ill patients found as described previously.

To build the feature vector for each patient all previous diagnoses including secondary diagnoses, specialty codes of the doctors and provided procedures were extracted in addition to the age and sex of the patient. These were then processed in various ways.

As the diagnoses are fairly specific under ICD-10, the number of different diagnoses was too large for practical purposes. As ICD-10 is hierarchical structure it is possible to group the diagnoses, but to control loss of information singular value decomposition (SVD) (Lange, 2010) was used as dimensionality reduction instead. For that a vector of all possible diagnoses was composed for every patient, TF-IDF was performed on these vectors, the resulting matrix was normalized and finally SVD applied to this matrix. On Figure 24 the cumulative explained variance of these components is shown. Judging from these images the first 200 components were used as features.

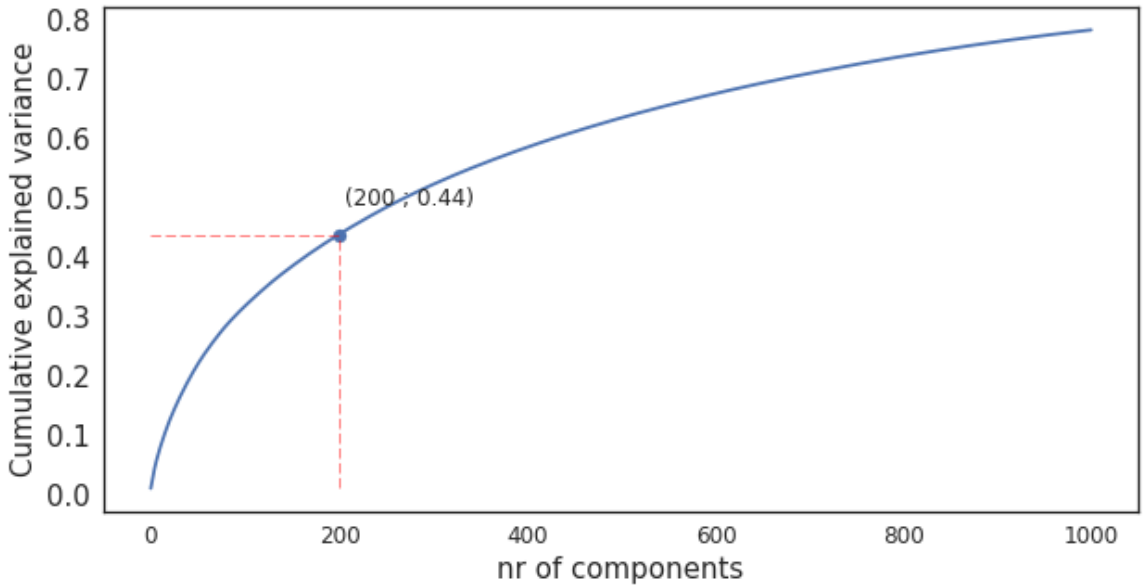

Figure 24: Cumulative explained variance of the components of the SVD performed on the diagnoses. 200 component point is marked as the one chosen.

All services provided to a patient during one day were considered a transaction and frequent item sets were mined from these transactions. For that FPgrowth algorithm (Srikant & Agrawal, 1996) was used with parameters shown in Table 8.

Table 8: FPgrowth parameters. Only sets if items containing at least two items were considered and were considered frequent only if the item set was in at least 2% of all transactions being mined.

minimum number of items	2
minimum relative support	2%

Then for each patient the frequencies of all mined item sets were found and these frequencies were used as features.

The specialty codes of the doctors were counted and used “as is” without any further processing.

3.4.1 Predicting the number of people getting the illness

Predicting the people who will get ill from the whole population is a complicated classification problem as the classes are very uneven. There are in total more than 1.3 million patients so the number of actual patients is two orders of magnitude smaller in

case of both diagnoses. To mitigate this issue the classes were weighed while training the random forest classifier.

The results of the classification for both diagnoses using both random forest classifier and gradient boosted trees are shown on Figures 25, 26, 27 and 28 and the metrics from the classification are shown in Table 9. As the classes are highly imbalanced, the accuracy score is rather meaningless, but the confusion matrices and ROC curves are more informative and indicate that the classifier is far above random.

We see that for both diagnoses the model classifies roughly half of the people who get ill as healthy, which is an issue, but it never wrongly classifies a healthy person as ill. For he purposes of this work, this means that the classifier will strongly underestimate the number of people that will get sick in the future.

What is very surprising is the fact that both algorithms are more accurate when predicting breast cancer. With little specialist knowledge in the medical field, one would assume that predicting breast cancer would be more difficult than COPD and this does hold true when using random forest classifiers, but the difference evaporates when using gradient boosted trees.

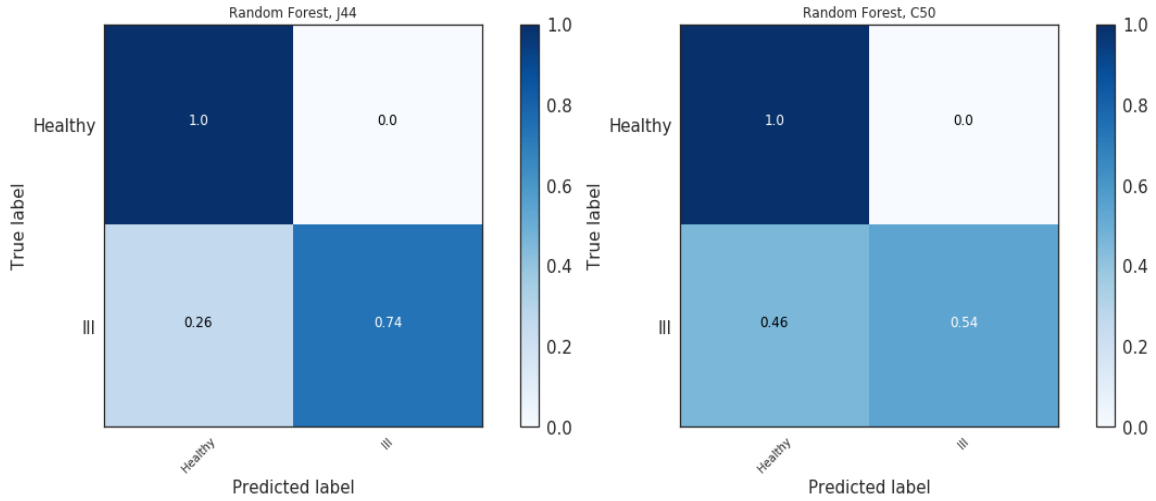

Figure 25: Confusion matrices of predicting the likelihood of people getting ill using random forest classifier. Graphs for both diagnoses J44 and C50 are shown. The matrices are normalized over rows.

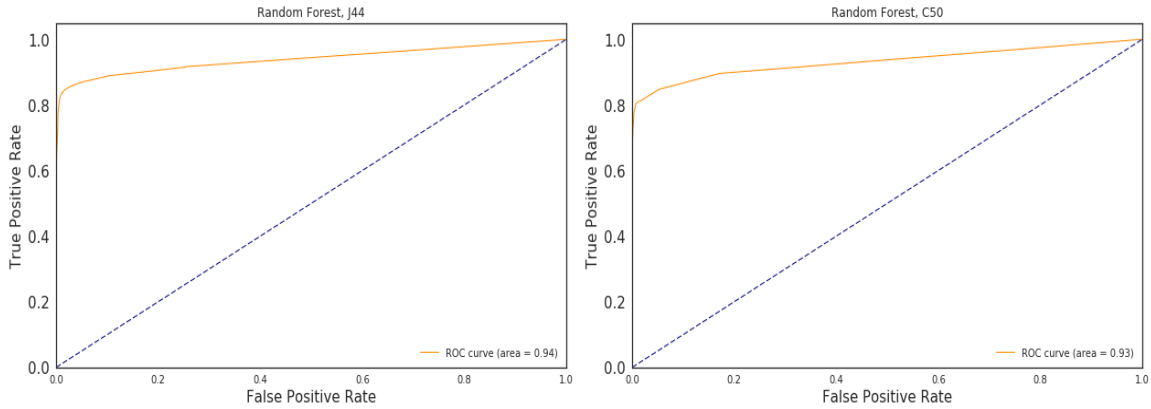

Figure 26: ROC curves for the prediction of likelihood of people getting ill using random forest classifier. Graphs for both diagnoses J44 and C50 are shown.

Table 9: Metrics of the classification of the likelihood of people getting ill. As the classes are highly unbalanced the accuracy metric is not very informative and the others are slightly skewed as well.

	J44 - RFC	J44 - XGB	C50 - RFC	C50 - XGB
Accuracy	99.07112%	99.29104%	99.61377%	99.75317%
Recall	89.04961%	91.29143%	84.2341%	90.72961%
F1	86.667%	87.21789%	77.23235%	85.78549%

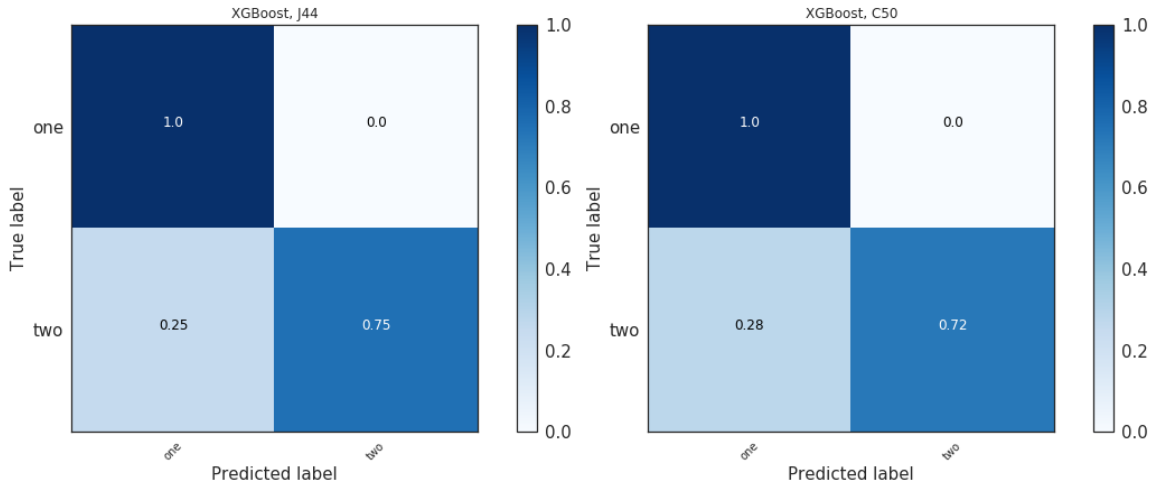

Figure 27: Confusion matrices of predicting the likelihood of people getting ill using random forest classifier. Graphs for both diagnoses J44 and C50 are shown. The matrices are normalized over rows.

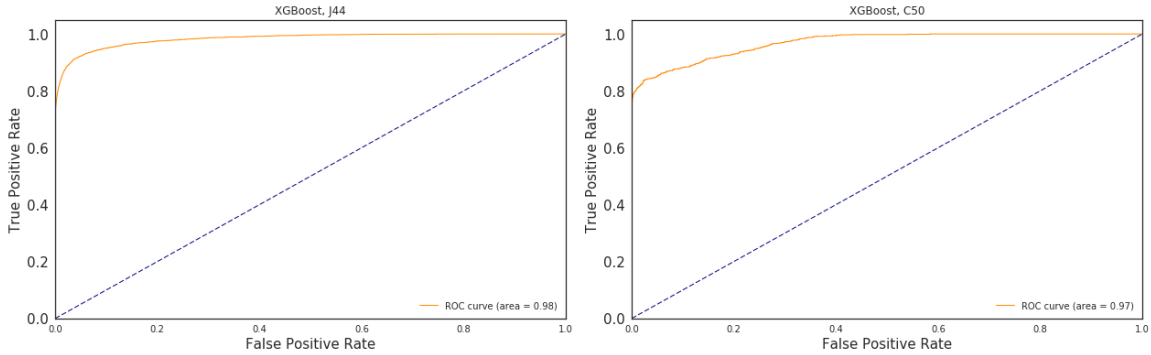

Figure 28: ROC curves for the prediction of likelihood of people getting ill using gradient boosted trees. Graphs for both diagnoses J44 and C50 are shown.

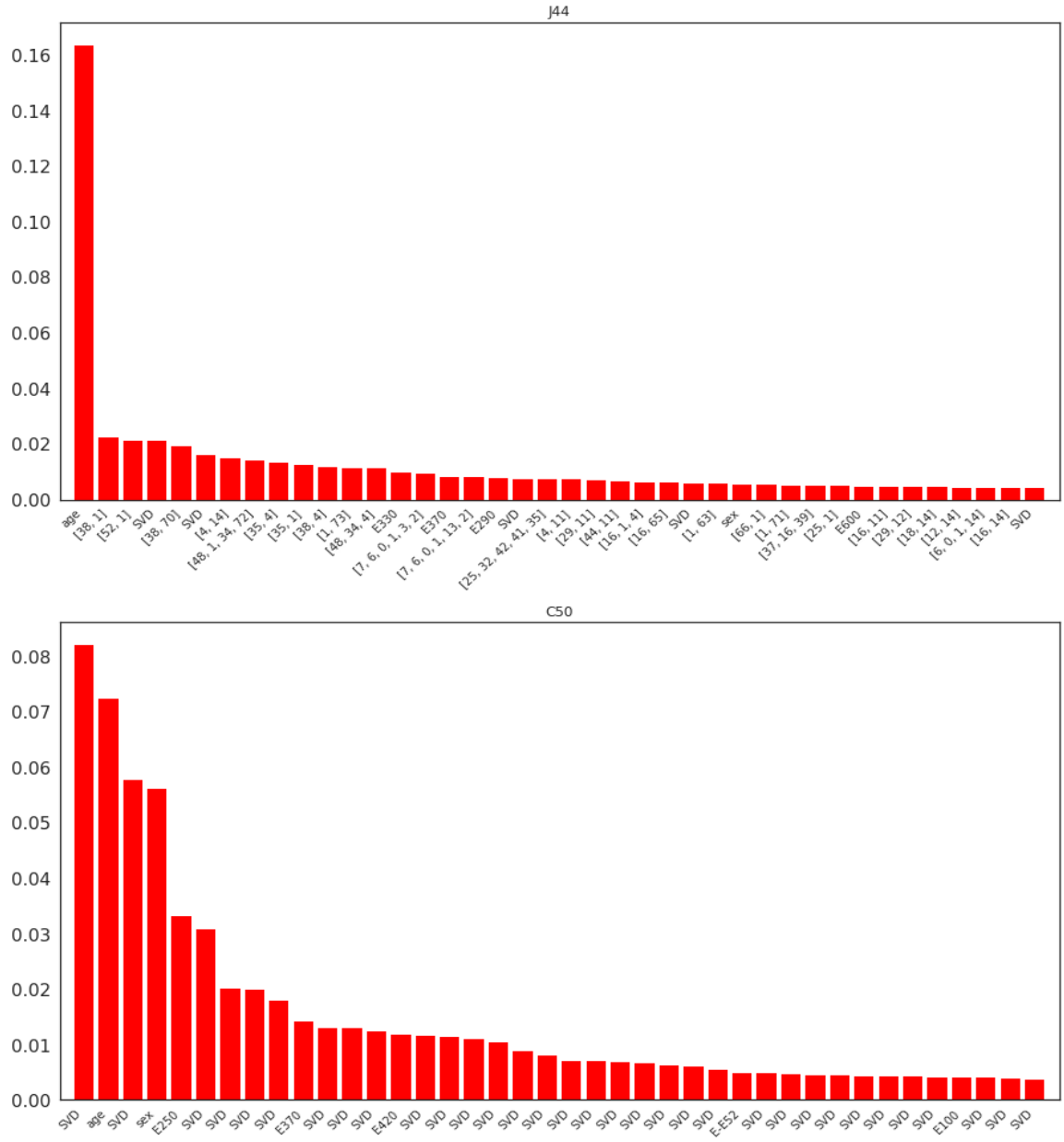

Figure 29: Graphs showing the feature importances for the initial classification task of categorizing how many people will get sick overall. The services marked with "SVD" are the features that were found when performing LSA on the diagnosis vectors built from the treatment histories of the patients. The legend for mapping the service codes to service names is shown in Table 10

We also looked at the feature importances for both diagnoses, which are shown on Figure 29 while the corresponding legend for the service names are shown in Table 10. It can be observed that age plays a far more important role in predicting J44. Also interesting to note is that services provided pay an important part only for predicting J44 while these

are relatively irrelevant for predicting C50. The specialty of the doctor only shows up in the most relevant 40 features for predicting C50.

Table 10: Table for mapping service codes to service names. Accompanies Figure 29.

0	Kreatiniin, uurea, kusihape*
1	Eriarsti esmane vastuvõtt
2	Ensüümid: ALP, ASAT, ALAT, LDH, CK, GGT, CK-Mba, alfa-amülaas*
3	Aneemia-, südame-, kasvajamarkerite määramine, haigustekitajate uuringud, antikehade, vitamiinide ja ensüümide määramine immuunmeetodil*
4	Eriarsti korduv vastuvõtt
6	C-reaktiivne valk
7	Naatrium, kaalium, kaltsium*
11	Elektrokardiograafia koos kompuuteranalüüsiga
12	Kolesterooli fraktsioonid: HDL, LDL*
13	Bilirubiin, konjugeeritud bilirubiin*
14	Röntgeniülesvõte rindkere piirkonnast (üks ülesvõte)
14	Röntgeniülesvõte rindkere piirkonnast (üks ülesvõte)
16	Bioloogilise materjali aeroobne külv põhisöötme(te)le
18	Uriini sademe mikroskoopiline uuring
25	Kompuutertomograafia natiivis (iga järgmine piirkond)
29	Täismahus ehhokardiograafia
32	Kompuutertomograafia kontrastaine 10 ml
34	Silmapõhja uuring kolmepeeglilääätse või Volke luubiga
35	Kompuutertomograafia natiivis
35	Kompuutertomograafia natiivis
37	Mikroorganismi samastamine üksikute biokeemiliste või immunoloogiliste reaktsioonide abil
38	Spirograafia
39	Ravimitundlikkuse määramine diskdifusiooni meetodil kuni kuue preparaadi suhtes
41	Kompuutertomograafia kontrastainega
42	Kompuutertomograafia kontrastainega (iga järgmine piirkond)
44	Järelravi
48	Silmade refraktsiooni uurimine autorefraktomeetri abil
52	Röntgeniülesvõte alajäsemetest (kaks ülesvõtet)
63	Diagnostilisel või ravi eesmärgil organi/õõne punktsioon
65	Bioloogilise materjali aeroobne külv lisasöötme(te)le
66	Röntgeniülesvõte lülisamba piirkonnast (kaks ülesvõtet)
70	Bronhodilataatortest
71	Otomikroskoopia

The feature importance graph also shows us that while previous diagnoses play little part in predicting J44, most of the services seem to be tests and other procedures related to assessment of the patient's health. This might indicate that while the diagnoses themselves play a smaller part, the guesses of the medical practitioner, reflected in what they test for, are more important.

3.4.2 Predicting type of treatment

The same features were used to predict the distribution of patients into clusters. As with the previous classification task the classes are again unbalanced, but the situation is a lot less severe than was in the case of predicting the number of people who would get sick at all. The largest difference in size is between clusters 1 and 2 of J44, with the number of patients in either being 16529 and 2872 respectively. The results from the classification are shown on Figures 30 and 31, the classifier metrics can be found in Table . Classes were again weighed when using random forest classifier, but the prediction accuracy for the second cluster of J44 was still lacking.

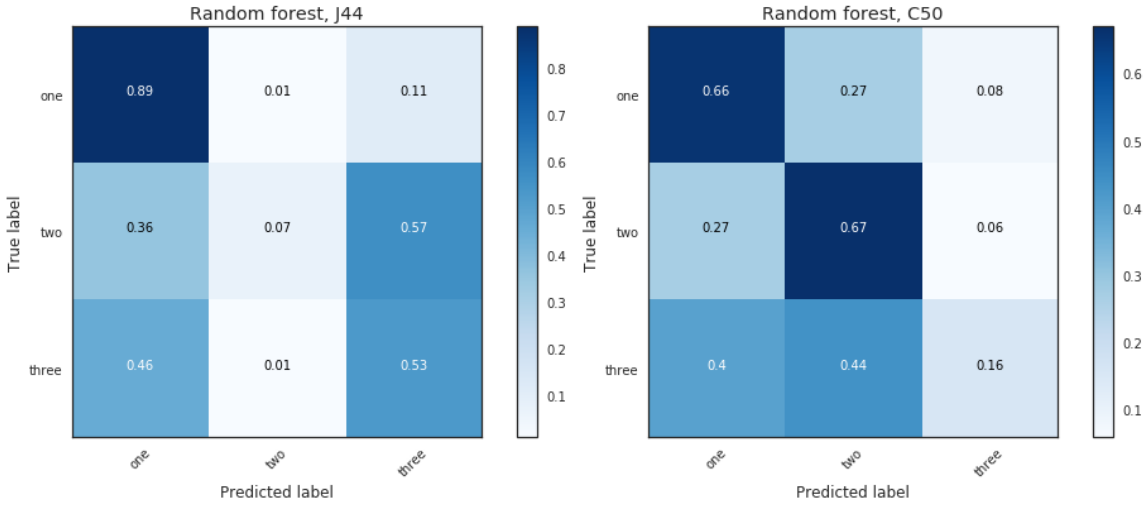

Figure 30: Classification results using random forest classifier. The second and smallest cluster is the most difficult to predict in the case of J44.

In the case of C50 classifying patients in the third cluster seems to be very difficult which corresponds well to the clinical pathways shown earlier, where the differences between the first and second cluster were rather clear, but the differences between the third and fourth were hard to see.

The fact that the classifier seems to prefer clustering the cases between two clusters rather than the three we have, may indicate that our choice of clusters may not have been appropriate. Especially as for both illnesses one of the clusters It should be kept in mind that in this work very high level data is used – no lab results, other measurements or treatment outcomes are used. It seems reasonable to hypothesize that given more precise data about the treatment the classification accuracy could be improved.

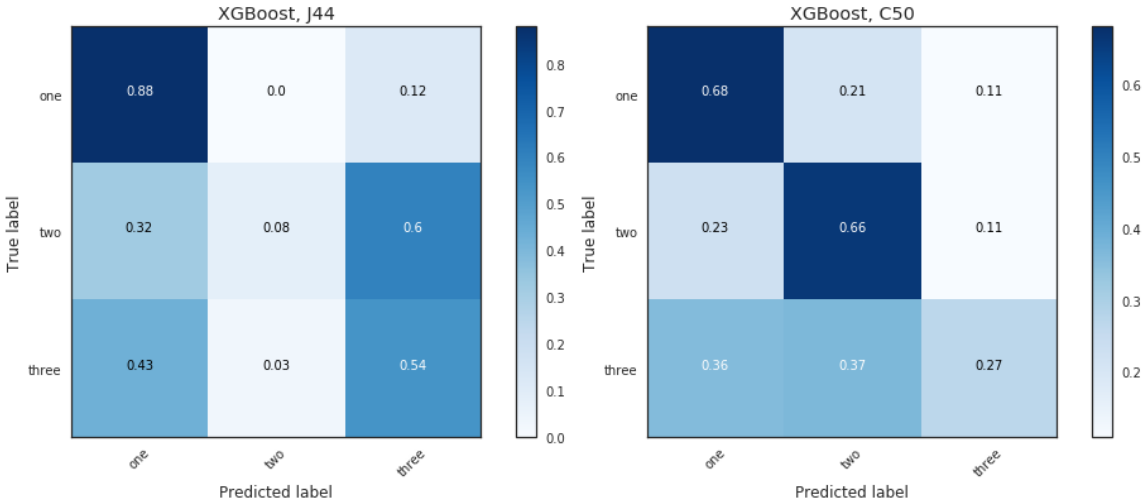

Figure 31: Classification results using gradient boosted trees. The results are better than those from random forest classifier, but not very drastically and the weaknesses are the same on both algorithms.

Table 11: Metrics from both classifiers and diagnoses. RFC stands for random forest classifier and XGB stands for gradient boosted trees. XGB gets slightly better results on both diagnoses, especially at recall.

	J44 - RFC	J44 - XGB	C50 - RFC	C50 - XGB
Accuracy	67.78256%	68.54491%	52.71845%	55.92233%
Recall	48.87862%	50.18704%	46.83285%	52.83515%
F1	48.95388%	50.04882%	50.02193%	53.8613%

The feature importances were also extracted from the random forest classifier and the most important 40 features are shown on Figure 32. The sets of services were coded as the names are too long to reasonably graph and a legend for both diagnoses is presented in Table 12.

For both diagnoses the most influential predictor by far is age, which is to be expected. This is especially true for breast cancer where also surprisingly no feature related to previous diagnoses even makes it onto the figure. Also most of the more important sets of services previously provided to the patients, contain a mammography, and x-ray of the chest or other related services.

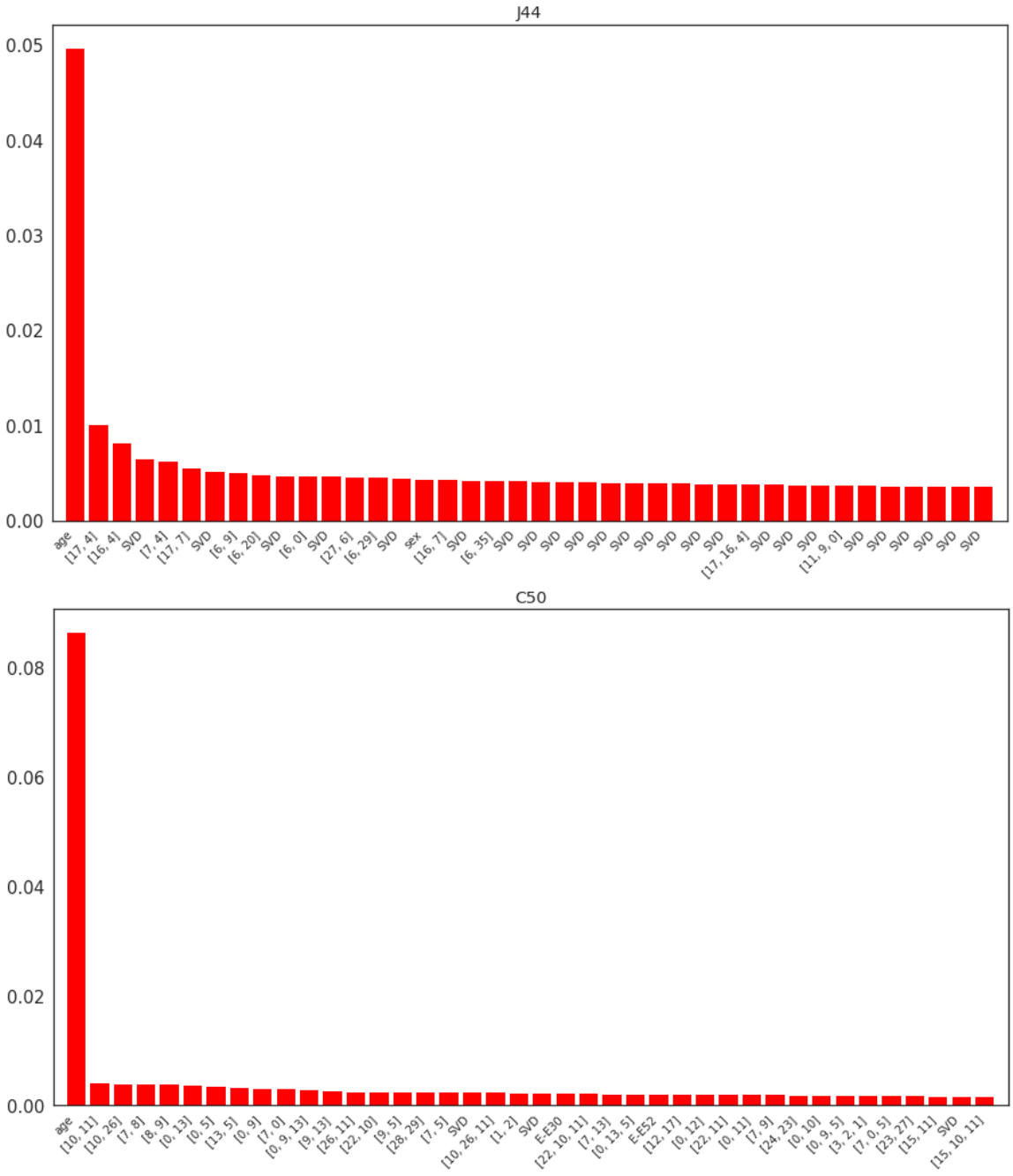

Figure 32: Feature importance plots when classifying patients into their respective treatment clusters. The features marked with numbers are coded sets of services. The legend for decoding the sets is shown in Table 12.

The services marked with "SVD" are the features that were found when performing LSA on the diagnosis vectors built from the treatment histories of the patients.

For COPD the other features have a bit more of significance and also most of the features in the top40, shown on the figure are related to previous diagnoses. Here the more relevant sets of services also contain a visit to the doctor and some service that was performed during this visit. It serves as a nice sanity check that the sex of the patient is

represented among the features of COPD while it is absent from the important features for breast cancer.

Table 12: Legend to use with the feature importance plots for both diagnoses shown on Figure 32.

#	C50	J44
0	Eriarsti esmane vastuvõtt	Aneemia-, südame-, kasvajamarkerite määramine, haigustekitajate uuringud, antikehade, vitamiinide ja ensüümide määramine immuunmeetodil*
1	Ensüümid: ALP, ASAT, ALAT, LDH, CK, GGT, CK-Mba, alfa-amülaas*	Kreatiniin, uurea, kusihape*
2	Kreatiniin, uurea, kusihape*	C-reaktiivne valk
3	Naatrium, kaalium, kaltsium*	Ensüümid: ALP, ASAT, ALAT, LDH, CK, GGT, CK-Mba, alfa-amülaas*
4	C-reaktiivne valk	Eriarsti esmane vastuvõtt
5	Eriarsti korduv vastuvõtt	Naatrium, kaalium, kaltsium*
6	Bilirubiin, konjugeeritud bilirubiin*	Elektrokardiograafia koos kompuuteranalüüsiga
7	Aneemia-, südame-, kasvajamarkerite määramine, haigustekitajate uuringud, antikehade, vitamiinide ja ensüümide määramine immuunmeetodil*	Eriarsti korduv vastuvõtt
8	Elektrokardiograafia koos kompuuteranalüüsiga	Bilirubiin, konjugeeritud bilirubiin*
9	Röntgeniülesvõte rindkere piirkonnast (üks ülesvõte)	Röntgeniülesvõte rindkere piirkonnast (üks ülesvõte)
10	Rinnanäärme ultraheliuuring (üks rind)	Sõeluuringud, hormoonuuringud, haigustekitajate uuringud immuunmeetodil*
11	Jämenõelabiopsia või punktsioon ultraheli või röntgeni kontrolli all	Harvaesinevad ja kinnitavad uuringud, erakorralised analüüsid immuunmeetodil*
12	AB0-veregrupi ja Rh(D) kinnitav määramine (AB0-grupp määratud nii otsese kui ka pöördreaktsiooniga)	Hüübimisjada sõeluuringud: PT, APTT*
13	Mammograafia, üks rinnanääre kahes sihis	Fibriini laguproduktide uuringud: fibriini D-dimeerid, fibriini monomeerid*
14	AB0-veregrupi määramine patsiendi identifitseerimisel või erütrokomponentide kontrollil	Kolesterooli fraktsioonid: HDL, LDL*
15	Pehmete kudede ultraheliuuring (üks piirkond)	Kolesterool, triglütseriidid*
16	Vaginaalne ultraheliuuring	Spirograafia
17	Erütrotsütaarsete antikehade sõeluuring kahe erütrotsüüdiga	Bronhodilataatortest
18	Sõeluuringud, hormoonuuringud, haigustekitajate uuringud immuunmeetodil*	Kompuutertomograafia kontrastaine 10 ml
19	Kolesterool, triglütseriidid*	Kompuutertomograafia kontrastainega (iga järgmine piirkond)
20	Silmapõhja uuring kolmepeegliläätse või Volke luubiga	Bioloogilise materjali aeroobne külv põhisöötme(te)le
21	Papanicolaou meetodil tehtud ja skriinija hinnatud günekotsütoloogiline uuring	Kompuutertomograafia natiivis (iga järgmine piirkond)
22	Iga järgnev jämenõelabiopsia	Kompuutertomograafia kontrastainega
23	Kompuutertomograafia kontrastainega	Glükohemoglobiin
24	Kompuutertomograafia kontrastainega (iga järgmine piirkond)	Silmapõhja uuring kolmepeegliläätse või Volke luubiga
25	Silmade refraktsiooni uurimine autorefraktomeetri abil	Silmade refraktsiooni uurimine autorefraktomeetri abil
26	Hematoksüliin-eosiin värvinguga pahaloomulisuse diferentseeringuga biopsiamaterjali uuring (1 blokk)	Kompuutertomograafia natiivis
27	Kompuutertomograafia kontrastaine 10 ml	Uriini sademe mikroskoopiline uuring
28	Kompuutertomograafia natiivis	Raud, magneesium, fosfaat*
29	Kompuutertomograafia natiivis (iga järgmine piirkond)	Sisehaigused
30	-	Erütrotsütaarsete antikehade sõeluuring kahe erütrotsüüdiga
31	-	AB0-veregrupi ja Rh(D) kinnitav määramine (AB0-grupp määratud nii otsese kui ka pöördreaktsiooniga)
32	-	Mikroorganismi samastamine üksikute biokeemiliste või immunoloogiliste reaktsioonide abil
33	-	Kõhu- ja vaagnapiirkonna ultraheliuuring

4 Discussion and future work

4.1 Preprocessing the data

In an effort to create a framework for population based prediction of costs and other parameters related to the treatment a number of approaches were used for data processing, clustering, process discovery and classification.

Due to the specifics of the data at hand some level of preprocessing was required and LDA and NMF were chosen for this purpose. Different approaches were used for both preprocessing methods and from the results it is observable that of these two only LDA approach seems to give meaningful results. It should be noted though, that not all the possible uses of NMF, discussed in the methods section, were attempted. It is possible that some of these approaches, such as treating each day as a fixed size combination of sets of services, would give better results. It could also be beneficial to look at other dimensionality reduction or embedding methods used for example in text analysis.

4.2 Clustering

HMMs were used as the models for clustering with the rationale of taking into account the sequential information about the treatment. This approach seems to work fairly well at uncovering the underlying treatment procedures as is suggested by the characterization of the clusters by our chosen parameters on Figures 13 and 14. This also supported by the reasonable predictability of the clusters and the discovered process models which are clearly differentiable.

But it is clear that input from a domain expert would be of great help here. An experienced practitioner would most likely be acquainted with the typical cases for an illness and this knowledge could be used either as a prior to the number of treatment groups to be inferred from the data or as a validation method for the clustering method.

4.3 Process discovery

The expert knowledge would also be of great use in generating and describing the process models. On a high level these can be to some extent understood without expert

knowledge, but as shown on the figures in the results section, this requires abstracting away some details both with rules and the Fuzzy Miner algorithm.

An expert could be of great use here in two ways. Firstly, their input would enable creation of better rules or even a hierarchy for merging the events to a manageable level. Secondly they would better understand the nuances of the treatment shown on a process model and could give a better description to it. Better understanding of the predicted models is critical if they are to provide a support for decision making with regards to health care spending such as directing resources to prevention.

4.4 Predicting

The section of this work dealing with predictions is separated into to parts: predicting falling ill and predicting the type of treatment received. The results from both sections are notable, especially considering the type of data used in this work: high level billing data with no information about test results or other more detailed parameters.

The engineered features in the form of components from performing SVD on the diagnoses and theitem sets mined from the provided services, worked reasonably well. From the results it can be seen that previously provided services are a much better predictor of J44 than for C50 for which most of the predictive power lies in previous diagnoses and the specialties of the doctors visited. Also of notice is that C50 can be predicted almost the same accuracy as J44, which is surprising as one would expect that cancer is a more difficult to predict illness compared to a chronic disease. This may be the case for only this type of cancer as breast cancer is often screened for and information about the screening could make prediction of breast cancer much easier than others.

The prediction accuracy was not as high when predicting the treatment type of the patients. This may result from multiple factors such as less than ideal clustering, inherently difficult classification problem and too high level data. The latter issue could be improved by more thorough feature engineering and investigating what other attributes could be gained from the data, but at one point there would be a limit to the accuracy we can get from billing data. The possible advantages of more detailed data can be seen from the fact that a lot of the item sets that have high predictive power include a type of screening or testing. It seems likely that including the result of these screenings would improve the results.

It could also be of interest to make an attempt at unpacking the SVD components to get a better understanding into what diagnoses in the medical history could be important in assessing the risk of a patient.

Conclusion

The objective of this work was to create a framework for population based prediction of costs and other parameters related to the treatment. An important problem in the context of constantly rising health care costs and diminishing working population. For that purpose it was necessary to develop a method for both discovering the various treatment types from the data and for predicting the number of people who are likely to become ill and be treated according to each of these treatment types. The work was done using billing data from EHIF for the period of 2010-2017 and two diagnoses: C50 and J44, were used as example diagnoses.

The results show that creating such a framework from this type of data is feasible. It is possible to cluster these treatment processes and discover the likely underlying clinical procedures. Although this process would most likely benefit from input from a domain expert.

The results also show that it is possible to predict, with limited accuracy, the number of people likely to fall ill based on previous treatment history and little background information about the patient. The accuracy of these predictions is of course dependent on the illness as some illnesses are more predictable from previous treatment history. There are indications that given more granular data it would be possible to increase this accuracy in a meaningful way.

Using such a framework could serve useful to planning resource allocation in health care as it would provide information about the number of people receiving a certain kind of treatment, what the treatment costs and how long it lasts on average and what services are provided to the patients during their treatment. This could help in estimating the future costs in health care and indicate the optimal prevention methods to which to allocate resources.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. http://doi.org/10.1109/TAC.1974.1100705

Bicego, M., Murino, V., & Figueiredo A.T., M. (2003). Similarity-based clustering of sequences using Hidden Markov Models. *Third International Conference on Machine Learning and Data Mining in Pattern Recognition*, 86–95. http://doi.org/10.1007/3-540-45065-3_8

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. http://doi.org/10.1145/2133806.2133826

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bose, R. P. J. C., & van der Aalst, W. M. P. (2010). Trace Clustering Based on Conserved Patterns: Towards Achieving Better Process Models, 170–181. http://doi.org/10.1007/978-3-642-12186-9_16

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

Dalianis, H., Hassel, M., Henriksson, A., & Skeppstedt, M. (2012). Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. *Swedish Language Technology Conference*.

Delias, P., Doumpos, M., Grigoroudis, E., Manolitzas, P., & Matsatsinis, N. (2015). Supporting healthcare management decisions via robust clustering of event logs. *Knowledge-Based Systems*. http://doi.org/10.1016/j.knosys.2015.04.012

Estonian Health Insurance Fund, & Group, W. B. (2015). *Ravi terviklik käsitlus ja osapoolte koostöö Eesti tervishoiusüsteemis*. Retrieved from https://www.haigekassa.ee/sites/default/files/Maailmapanga-uuring/veeb_est_summary_report_hk_2015.pdf

Fluxicon. (2017). Disco. Retrieved from https://fluxicon.com/disco/

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. http://doi.org/DOI 10.1214/aos/1013203451

Greco, G., Guzzo, A., Pontieri, L., & Saccà, D. (2006). Discovering expressive process models by clustering log traces. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1010–1027. http://doi.org/10.1109/TKDE.2006.123

Günther, C. W., & Van Der Aalst, W. M. P. (2007). Fuzzy Mining – Adaptive Process Simplification Based on Multi-Perspective Metrics. *Business Process Management, 5th International Conference (BPM 2007)*.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. *Annals of Physics* (Vol. 54). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://doi.org/10.5860/CHOICE.49-3305

Han, J., Pei, J., & Yin, Y. (2000). Mining Frequent Patterns Without Candidate Generation. *SIGMOD Rec.*, *29*(2), 1–12. http://doi.org/10.1145/335191.335372

Kumar, R. K. (2011). Technology and healthcare costs. *Annals of Pediatric Cardiology*, *4*(1), 84–86. http://doi.org/10.4103/0974-2069.79634

Lakshmanan, G. T., Rozsnyai, S., & Wang, F. (2013). Investigating Clinical Care Pathways Correlated with Outcomes (pp. 323–338). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-40176-3_27

Lang, M., Bürkle, T., Laumann, S., & Prokosch, H.-U. U. (2008). Process mining for clinical workflows: challenges and current limitations. *Studies in Health Technology and Informatics*, *136*, 229–34. http://doi.org/10.1007/978-3-642-19345-3

Lange, K. (2010). Singular Value Decomposition. In *Numerical Analysis for Statisticians* (pp. 129–142). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4419-5945-4_9

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, *19*, 2756–2779. http://doi.org/10.1162/neco.2007.19.10.2756

Mans, R. S. S., Schonenberg, M. H. H., Song, M. S., Van Der Aalst, W. M. P., Bakker, P. J. M. J. M., Aalst, W. M. P. van der, … Bakker, P. J. M. J. M. (2009). Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital. *Proceedings of BIOSTEC 2008, 25*, 425–438. http://doi.org/10.1007/978-3-540-92219-3_32

Mans, R., & Schonenberg, H. (2008). Process mining techniques: an application to stroke care. *Studies in Health …*, *136*, 573–578. http://doi.org/10.3233/978-1-58603-864-9-573

Marinov, M., Mosa, A. S. M., Yoo, I., & Boren, S. A. (2011). Data-mining technologies for diabetes: a systematic review. *Journal of Diabetes Science and Technology*, *5*(6), 1549–56. http://doi.org/10.1177/193229681100500631

Moturu, S. T., Johnson, W. G., & Liu, H. (2007). Predicting Future High-Cost Patients: A Real-World Risk Modeling Application. In *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)* (pp. 202–208). http://doi.org/10.1109/BIBM.2007.54

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS International Conference on Computer Systems and Applications* (pp. 108–115). IEEE. http://doi.org/10.1109/AICCSA.2008.4493524

Panuccio, A., Bicego, M., & Murino, V. (2002). A Hidden Markov Model-based approach to sequential data clustering. *Structural Syntactic and Statistical Pattern Recognition*. http://doi.org/10.1007/3-540-70659-3_77

Pospíšil, M., Mates, V., Hruška, T., & Bartík, V. (2013). Process Mining in a Manufacturing Company for Predictions and Planning. *International Journal on Advances in Software*, *6*(3 & 4), 283–297. http://doi.org/10.1.1.672.4578

Qwertyus. (2013). Illustration of approximate non-negative matrix factorization (NMF). May also serve as an illustration of other matrix decomposition methods. Retrieved from https://upload.wikimedia.org/wikipedia/commons/f/f9/NMF.png

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*(2), 257–286. http://doi.org/10.1109/5.18626

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. http://doi.org/10.1214/aos/1176344136

Smyth, P. (1997). Clustering sequences with hidden Markov models. *Advances in Neural Information Processing Systems*, *9*, 648–654. http://doi.org/10.1017/CBO9781107415324.004

Srikant, R., & Agrawal, E. (1996). Mining Sequential Patterns: Generalization and Performance Improvements. *5th International Conference on Extending Database Technology (EDBT ’96)*, 3–17. http://doi.org/10.1109/ICDE.1995.380415

Sushmita, S., Newman, S., Marquardt, J., Ram, P., Prasad, V., Cock, M. De, & Teredesai, A. (2015). Population Cost Prediction on Public Healthcare Datasets. *Proceedings of the 5th International Conference on Digital Health 2015 - DH ’15*, 87–94. http://doi.org/10.1145/2750511.2750521

Tonsiver, T., Ehrenberg, A., Ringmets, I., Lepik, K., Saare, K., & Kiivet, R.-A. (2014). Kehaväline viljastamine Eestis: efektiivsus ja kulud. *Eesti Arst*, *93*(3), 143–150. Retrieved from http://ojs.utlib.ee/index.php/EA/article/viewFile/11694/6878

van der Aalst, W. M. P. (2011). Introduction. In *Process Mining*. Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-19345-3_1

van der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow mining: discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142. http://doi.org/10.1109/TKDE.2004.47

Vavasis, S. A. (2010). On the Complexity of Nonnegative Matrix Factorization. *SIAM Journal on Optimization*, *20*(3), 1364. http://doi.org/10.1137/070709967

Verbeek, H. M. W., Buijs, J. C. A. M., van Dongen, B. F., & van der Aalst, W. M. P. (2010). XES, XESame, and ProM 6. In P. Soffer & E. Proper (Eds.), *Information Systems Evolution - CAiSE Forum 2010, Hammamet, Tunisia, June 7-9, 2010, Selected Extended Papers* (Vol. 72, pp. 60–75). Springer. http://doi.org/10.1007/978-3-642-17722-4_5

WHO. (1992). *ICD-10 Classification of Mental and Behavioural Disorders; Diagnostic Criteria for Research*. World Health Organization. Retrieved from https://books.google.ee/books?id=HlnzVSbec18C

Yang, W., & Su, Q. (2014). Process Mining for Clinical Pathway Literature Review and Future Directions. *Service Systems and Service Management (ICSSSM), 2014 11th International Conference*, 1–5. http://doi.org/10.1109/ICSSSM.2014.6943412

Non-exclusive licence to reproduce thesis and make thesis public

I, Markus Lippus

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

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Predicting Illness and Type of Treatment from Digital Health Records

supervised by Sven Laur and Anna Leontjeva,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 18.05.2017