Aalborg University Copenhagen

Semester: 4th

Aalborg University Copenhagen
A.C. Meyers Vænge 15
2450 København SV
Semester Coordinator:
Henning Olesen
Secretary: Maiken Keller

Title: Personalized Medicine based on patient journals and family medical history records

Project Period: Fall 2013

Semester Theme:

Master Thesis

Supervisor(s):

Henning Olesen

Allan Hamershoj

Members:

Stavris Solo (20120549)

Copies: 4

Pages: 89

Finished: 10/01/2014

Abstract:

The aim of this work was to develop a procedure for analysing and processing large amount of medical data in order to help physicians to make better decisions during the diagnosis process and to forecast high risk patients for developing a specific disease.

The suggested solution is a **procedure** that combined different techniques that were applied to pre-process, analyse and visualize patient medical record data in order to foresee the patient high risk of developing a specific disease. This is achieved through the analysis of inter, intra-family relationships among patients and their similarities in terms of diagnosed diseases. The above procedure led to the development of a system prototype that will help the physician to make better decisions during the diagnosis process.

The methods followed during the research process in order to reach to the proposed solution consists of primary research, such as an interview with the physician and secondary research such as related work, journal articles, papers, internet sources.

The collaborative filtering technique was used to explore, not only the patients' intra-family relation, but also, the inter-family relations, in order to define a similarity score among patients. The patients are grouped based on their similarity score.

The Latent Semantic Index clustering technique was used to analyse the group of similar patients in order to foresee the risk of developing a specific disease.

Cluster visualization was used to visualize the results. In this way, the physician can further investigate and understand the results i.e. why a patient has a high risk of developing a specific disease.

Acknowledgements

The last two years of this master program have been an amazing experience. There are some persons I would like to thank for making possible to complete these studies and specifically this thesis.

First I would like to thank my supervisors about their advices and discussions and their insightful comments during this thesis period. Also many thanks to the CMI the department and the teachers for providing a stimulating environment.

Second, I would like to express special thanks to Kostas and Arisa for the inspiration, support and help from the first day i started this journey.

Furthermore, I would like to thank Antonela for being patient and the everyday support especially during the last three months. Last but not least many thanks to my family for their warm support.

Contents

List of Figures

List of tables

Reading Guide

Chapter 1 – Introduction

This section introduces the area and the purpose of this study. Moreover, discusses the challenges and the problems that are faced while dealing with medical records. Furthermore, presents the research question and defines the delimitation of this project.

Chapter 2 – Understanding domain problem

Presents the dataset used for this project in order to have an overview of the amount and the complexity of the data contained. Moreover, introduces two illustrative scenarios that describes the as-is and to-be process.

Chapter 3 – State of the art

This chapter presents and discusses the state-of-the-art techniques and technologies used to analyse and manage different types of big data. Presenting the most known and used techniques and technologies. Furthermore, introduces the related to the research question academic and industrial work.

Chapter 4 – Methodology

Chapter 4 presents the overall research approach followed in order to answer the research question. It introduces all the steps followed during this research approach such as the primary research, the secondary research, the experimental approach and the development software method followed for the development of system prototype.

Chapter 5 – Problem and data analysis

This chapter discusses different area of the problem domain. It presents also the physicians suggestions regarding the problem domain. Furthermore, the analysis of the dataset through pre-processing is presented. Moreover, a detailed discussion take place into this chapter in order to define the appropriate techniques that can be applied in order to answer the research question.

Chapter 6 – Problem Solution

In this chapter a proposed solution in order to address the research question is presented. Furthermore, the development software method used to develop the system prototype is described. Moreover, different experiments and investigation of the results are discussed into this chapter.

Chapter 7 – Conclusions and Future Work

The conclusion chapter presents the findings and the answer to the research question and emphasizes the contributions of my work. Moreover, presents improvements that can be done in the near future.

1. Introduction

During the last years the amount of information all over the world has increased more than ever. According to [1] the amount of digital information is growing ten times as much every 5 years. The effect of this growth is being felt everywhere from business to science, from government to arts [1].

The rise of new technology which facilitated the gathering, storing, processing and analysing of data made possible this explosion. Companies and organizations are using powerful tools to collect a tremendous amount of data to support their decision-making. Moreover, people now more than ever interact with information [1] in new and novel ways.

"We are at a different period because of so much information" [1].

According to [2] data have become a torrent flowing into every area of the global economy. The numbers that MGI research shows are incredible. Facebook collects 30 billion of content shared on their network every month, 5 billion of smartphones devices are in the market, Google had more than 3 billion queries about the H1V1 flu, and there are more than 340 million tweets per day having a 40% of projected growth in global data generated per year [2]. This new phenomenon has been called as the industrial revolution of data, but scientists and computer engineers have given to this phenomenon a new term "Big Data" [1]. W*hat is Big Data exactly?* There are many definitions and explanations about the term of big data. According to [2] Big Data:

> *"Refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse."*

A more detailed definition about what exactly is considered to be Big Data according to [3] is:

> *"Big Data defines a situation in which data sets have grown to such* ***enormous sizes*** *that conventional information technologies can* ***no longer effectively handle*** *either size of the data set or the scale and growth of the data set."*

Big data is creating unprecedented opportunities for businesses, governments and organizations to achieve deeper, faster insights that can strengthen decision-making, improve the customer experience, and accelerate the pace of innovation [3].

According to [3] the Big Data paradoxically is not that new. Even though, the most massive data set are created during the last years, big data have roots in medical communities where analysis on medical datasets were made in order to derive value from them. Collecting various types of medical data about patients that consist of medical history, monitoring sensors, patient treatment and drug information allows hospitals to share this information through different information systems and different personnel within it. This information is available to physicians in different departments, but are they able to process, combine and understand all this flow of information? Do they have the time to go through all that information about a specific or a large number of patients in order to come up with some valuable results?

This project will investigate and process medical data records gathered in hospitals in order to extract valuable information that can be used from physicians during the diagnosis process. Furthermore, the system will allow the physician to predict if a patient is in a high risk to develop a specific disease.

1.1 Motivation & Background

According to [2] the recent development and maturation of technology have made possible to address different problems faced in the past such as data storage and data capturing. Amounts of medical and health data have been collected at hospitals and medical organizations at extraordinary speed. Those data are a very valuable resource for improving health delivery, health care, decision making and better risk analysis and diagnosis [2].

Research indicates that analyzing large datasets which include patient characteristics, cost and outcomes of treatments can help to identify the most clinically effective and cost-effective treatments to apply [2]. Furthermore, new studies have shown that working with medical data, it's possible to analyze the data gathered from monitoring the patient in order to

improve future drug costs and treatment options. Based on recent research [2], the use of data from remote monitoring systems can reduce the time that a patient stays in-hospital bed, make better nursing home care and improve outpatient physician appointments, as well as reduce long-term health complications. Moreover, current studies demonstrate that there are still many opportunities to derive value from the big data medical area. Such opportunities are predictive modeling, the use of simulations and modeling on a given clinical dataset in order to predict clinical outcomes [2].

Statistical tools and algorithms can be used in order to improve clinical trial design. Furthermore, the analyses of these clinical trials can be valuable in order to identify additional indications and discover adverse effects. Personalized medicine is another lever, where the use of medical data can derive value by analyzing a large dataset of medical records. The objective of the personalized medicine is to examine the relationships among genetic variation, predisposition for specific diseases, and specific drug responses and then to account for the genetic variability of individuals in the drug development process [2].

Motivated by the research done in this area, the need for innovation and the opportunities in medical big data I decided to look further in the field of medical data. Specifically, I will investigate the use of patient medical data such as medical history, family relation, and patient diseases similarities in order to extract valuable information. Analyzing and processing the medical data offers a lot of possibilities as described above. My goal will be to use this information in order to make it possible to predict the risk of a patient affected by a specific disease without having developed full disease's symptoms.

1.2 Problem Formulation

1.2.1 Future of medical data records

As mentioned above, medical data records were the stepping stone in the new era of Big Data. Medical Data records were used in the beginning in order to extract value for drug development. Throughout the years this dataset has been grown intensively because almost every person has a medical record. The patient medical records is a complex dataset and full

of different types of information such as historical health information, medicines, symptoms, treatments, vaccination, patient medical record and clinical test.

According to [2] analysing disease patterns in big data medical records creates a value lever and trends to model future demand and costs and make strategic investment and decisions. Another promising big data innovation in medical records that could produce value is the analysis of emerging large datasets (e.g., genome data) to improve research and development productivity and develop personalized medicine [2]. The principles of personalized medicine consist of early prediction and prognosis before the patient shows diseases symptoms, more effective therapies, and adjustment of the dosage according to the patient profile [2].

More or less every person has a medical record with some information about his health history, treatment received and symptoms developed. We expect in the future this information will grow since every time a person visits a hospital, new information about him is stored. The fact that today we are able to collect, store and share medical data, it is really important. However, more effort should be put on how to leverage all this vast amount of information .

1.2.2 Challenges and problems faced

Although technology has made unprecedented advances in the medical area, it is still very difficult for the doctor to make prognosis before the patient has developed symptoms of a specific disease. Even in the case where the patient has developed symptoms, the physician needs to wait for the test results in order to make his decision. Moreover, it is difficult to identify the inheritance of specific diseases, making it more challenging to identify the high risk patients that could be affected by the disease.

In addition, there are cases when patients have inherited a specific disease but it is difficult for the doctor to predict when it will affect him or if it will not. It is not possible for the physician to make this prediction because he does not have all the necessary information regarding the patient family historical records and symptoms. Even if this information was available, it is still time consuming for the physician to go through the whole family tree and try to find patterns that can help him to assess the disease risk.

As [4] claims it's very difficult almost impossible for a physician to completely analyze the complex indicators of risk (in real time) for every disease at the time of patient interaction. Even nowadays that physicians in hospitals are taking detailed history record, making physical examinations and running laboratory test to assess the risk for future diseases, these methods are in general limited for specific diseases and depends in the knowledge of the individual physician [4]. There is a need for innovations in this specific area in order to facilitate the working process of the physicians.

1.2.3 Research question

With this study I aim to capitalize on the previous studies on personalized medicine and develop a system prototype that helps physicians to assess the disease risk based on historical information and patient similarities. In particular, the purpose of this research is to derive value from patient medical record datasets by analysing and visualising different hidden patterns, based on the principles of the personalized medicine.

In this way, a physician will be able to foresee and estimate the risk a patient has in terms of historical and genetic diseases. This could be possible by taking in consideration not only the medical history of one individual patient but also the information we have from his family related persons and by other similar patients. Based on the above hypothesis, we have the possibility to discover and investigate the patient risk of developing a disease [4].

There are many techniques that can be applied and used in order to analyse a dataset. During the last years different techniques were developed in order to capture value from big data. Existing techniques [2] that were used for analyzing smaller datasets were effectively adapted so that they could be applicable to big data. In sections 3.1 and 3.2, I analyse and discuss different techniques and technologies used to analyse medical data.

In a broad view, I am interested in how data analytics and data visualization of medical records based on patient health history, family tree, patient symptoms and patient similarities will help the physician to predict possible future risk diseases that can affect

the patient. Furthermore, this project will investigate the appropriate techniques and technologies to achieve the above goal.

More specifically, this research will focus on the analysis of the relationship between patients' ancestors, relatives, medical records, symptoms and patient similarities in order to suggest to the physician the risk of developing a specific disease for a specific patient.

Moreover this research will address the following sub questions:

- *Which are the proper technique(s) that would make possible to identify patient similarities based on patient relations and according to their diagnosed diseases?*
- *Which are the appropriate technique(s) that will make easy to identify patients who are at high risk of developing a specific disease (e.g., diabetes) and would benefit from a preventive care program?*
- *How to visualize the above information to the physician in order to allow him to investigate the reasons of this prediction?*

In order to address the above questions I will build a system prototype that will facilitate the physician during the diagnosis process by providing specific information based on preliminary analysis of the medical dataset, suggesting the risk of a specific disease for a patient and visualize this information inspired from an explore and play approach.

The purpose of this prototype is to make possible for the physician to perform a number of activities. First the system will allow the physician to analyse the medical dataset chosen. Second, the physician will be able to classify the data set according to patients' medical history, patients' similarities and family relations. The third step consists of analyzing the results from steps 1 and 2 through Latent Semantic Analysis in order to achieve the goal of suggesting to the physician the high risk patients for a specific disease.

I will investigate if there is any already similar solution and use this information to build my prototype. The system prototype reduces the amount of information from the medical records making visible to the physician relevant information. Only information that is important and needed to support diagnosis process will be presented to the

physician. Every step described above will be visible to the physician in order to make the visualization part a two way interaction between the prototype and the physician.

Extracting patterns and deriving value from medical record datasets is not an easy process, but it can be valuable not only to the physicians for diagnostic purposes, but also for hospitals and governments for statistical purposes.

1.3 Project delimitations

Big Data has many challenges such as data acquisition, data storage, data analytics, data sharing, data visualization, and data searching but it is not the purpose of this project to focus on all of them.

The focus of this project is not to address big data acquisition, data searching and data sharing challenges. Diverse data storage technologies will be described and discussed in order to have an overall picture of the big data challenges, but it is not the scope of this project to investigate which is the appropriate technology to be used to store the data.

In addition, it is not the scope of this project to describe and investigate the business value and different stakeholders included into this concept.

2. Understanding the problem domain

In order to understand the problem domain there is a need to gain an overview of the medical record dataset. Medical datasets are usually very complex and large. For this reason, first, I believe it is necessary to provide a synopsis of what a medical record dataset is and how the dataset that will be analysed in this project looks like.

Furthermore, to better understand the system in context I will present two illustrative scenarios that will provide a guide to as-is and to-be processes respectively. A scenario is a story or a thinking experiment, which helped me to investigate and understand how the system I am building can be used by users [11]. The use of scenarios was also important to me in order to understand and describe the concept of the system being used [11]. Specifically, the first scenario allowed me to understand the as-is process, and helped me discover and define the challenges and problems faced during this process. The second scenario describes the to-be process. In this scenario potential solutions are assumed based on my understanding and thinking in order to address the problems and challenges faced on the first scenario.

2.1 Dataset - Electronic Medical Record

The dataset selected for use on this project is an electronic medical record (EMR) from a Danish hospital and all information presented is according to the Danish medical system and laws. Moreover, all the sensitive information represented such as CPR number and patient ID, was de-identified, in order to unlink the information with the real person. The dataset were used during recent research and acquired from K.P who was in charge for the research [5]. Furthermore, a confidential contract was signed between the counter parts in order to restrict the use of the data for academic purposes only.

The EMR during the last years has been grown rapidly because of the amount of information stored on it [6]. Almost every person has a medical record in a hospital or a clinic and if he does not most probably he will have in the future. The research [6] has shown that in almost all

medical institutions, the EMR contains a large volume of historical patient data, such as personal information, clinical test, medication, diagnosis, and drug treatment.

The type of information stored in the selected dataset consists of two data types, structured and unstructured data. The structured data, in general, consists of observational data such as diagnosis code, medications and clinical test results while unstructured data include free medical text such as the physician's notes.

The structured data are easy to read and to process, since these data are numerical and nominal. The unstructured data such as free medical text can be difficult to read and could present challenges in processing them [7].

In appendix 1 [figure 11, 12] examples of the structured and unstructured data, as shown in the tables of the dataset under investigation, are available. The selected dataset contains even more complex information that the information mentioned above. The information presented on the selected dataset specifically, consists of semi-automated coding of diagnosis, family relation code, drug record, laboratory record, area of data information and client vaccination information. The semi-automated diagnosis code is based on classification systems like ICPC, ICD9 / ICD10 according to the World Health Organization (WHO) [8]. The drug record information is stored according to the WHO classification system with defined daily doses (ATC/DDD). The family relation information between patients is stored according to the degree of their relations. Data area information contains information about the different type of hospitals among which the data are shared.

Moreover, the given dataset has information about a patient's medical records. This information grows and is updated every time the patient visits the hospital. The patient medical record contains information about the patient such as the patient ID, the diagnosed diseases, summary of patient diagnosis and the free medical text that the physicians has wrote for the specific patient. The tables below presents an instance of the patient medical record information and other type of information stored in the given dataset.

Table 1. An example of the patient medical information

MRID	MRT	ICD10ID	ICPCID	MS	MRLID	DA
00134	Lidt højt fast-BS på 6,4. Der er for flere år siden fundet en let abnorm. Vil prøve at tabe et par kg og stramme diæten op.	i109	K86	Ukompliceret hypertension	4265	at

MRID = Medical Record ID, MRT = Medical Record Text,
ICD10ID = World Health Organization Disease Code,
ICPCID = International Classification of Primary Care code based on the ICPCID-2 –DK,
MS = Medical Summary, MRLID = Medical Record Line ID,
DA= Data Area of the hospital that the data are retrieved,

Table 2 Client Diagnosis Information

DD	DC	MRID	DA	ICD10ID	ICPCID	MRLID
9/25/2006	4	00000142	lsk	S608	S19	00344596
4/11/2003	2	00006980	lkc	f101	P15	00112991
2/19/2004	2	00009173	skk	r298	L17	00221017

DD = Diagnosis Date, DD = Diagnosis Category
MRID = Medical Record ID, DA= Data Area of the hospital that the data are retrieved,
ICD10ID = World Health Organization Disease Code,
ICPCID = International Classification of Primary Care code based on the ICPCID-2 –DK,
MRLID = Medical Record Line ID,

Further screenshots are presented in appendix 1 [figure 11] which provide a better overview of the patient medical record stored in the selected EMR.

The EMR used in hospitals contains different types of information according to the physicians' specialties. It is possible to share data and information based on these specialty needs. The physicians use this information to have an overview of the patient's history medical records and to follow the patient treatment. Every time that a new patient visits a hospital the physician creates a new clinical record and during the diagnosis process tries to find relations between him and family related patient's in order to define the type of the disease.

A clinical entry recorded by a physician beside the usual information may contain other important information stored as free-text such as the thoughts of the physician about the diagnosis, the treatment information and not only.

This information is valuable but is also difficult to be processed, since it contains complex language, different abbreviations and a wide range of terminology. Moreover, is time consuming for the physician to go through every record related to the patient under the diagnosis process and read the available information [9].

In the future, all the information presented in different EMRs can be really useful to the physician during the diagnosis process, not only the patient's family historical records but other similar patients as well. Due to the amount of this information and the complexity of this task there is a need to define which are the most important and related features of the available patient medical information.

> *"Nowadays, EMRs are not only used for supporting the care process, but are often reused in observational epidemiological studies, e.g. to investigate the association between drugs and possible adverse effects"* [7].

Defining exactly what data are available and mapping out the most important features from the data features is a key part of the process in order to provide with extra help the physician [9]. So the very first step was to investigate the EMR dataset in order to define the most valuable information for use.

While it seems that EMR contains enough and detailed patient information, there are also other challenges to deal with. Usually, such datasets are incomplete (sometimes important parameter values are missing). These types of datasets are sparse, not all the representable information for the patient is available. Moreover, in such datasets the research has shown that systematic or random noise is always present, making it very difficult to further process and work with it to achieve reliable results [10].

According to [10] there are some steps to follow before one proceeds and works with a medical dataset in order to derive patterns and gain knowledge from it. The steps consist of understanding the domain problem, forming a new dataset or cleaning the existing one and finding different hidden regularities.

The overall goal of the above process is to find patterns or rules in order to achieve expected results. The last step of this process is also known as Data Mining.

What data mining process achieves is that allows us to extract information from a selected dataset and to manipulate it into a new dataset that it would be more comprehensible and useful for further use [13]. Section 5.1 presents a more detailed description of the steps followed to pre-process the selected dataset in order to extract the most important features of each table.

Moreover, recent research [9] shows that different techniques have been used in order to analyse the information available in order to classify patients with various conditions, patients in need of a specific therapy and for categorizing clinical entries based on the EMR records.

After having presented the dataset selected and shown how it is currently used in hospitals, in the next sections two illustrative scenarios follow to demonstrate the as-is and to-be process.

2.2 As-is Process

Tom has been diagnosed with two diseases, diabetes and hypertension during the last year. In the medical record database there are records about his diagnoses and the treatment he is receiving and a lot of other information as well.

One day he is not feeling very well, and he decides to visit the nearby hospital. During his visit the doctor discovers some suspicious symptoms but isn't very sure about them. The doctor asks Tom about his family members if they have had similar symptoms, but Tom doesn't remember anything. The doctor tries to look up in the system Tom's history and his ancestor and relatives' history so he can find some similar useful symptoms in order to make his decision about Tom's diagnosis. But there is no previous record of Tom and he cannot find any persons related to Tom in the database.

The doctor's decision is to run some blood test and wait for the results. After two days, Tom is again at the hospital to get the test results which are ok. The doctor says that there is nothing wrong about him and they will wait to see how things will move on. Tom goes home relieved that nothing was wrong with him.

After 6 months Tom has the same symptoms but this time he is feeling much worse. He went at the hospitals and follows the same procedure as 6 months before but this time he has been diagnosed with a heart disease. The doctor explained him that was not possible for him to diagnose the disease the first time Tom was at the hospital, because the symptoms were not enough and he did not have enough information about Tom's family medical records nor could he compare Tom's symptoms to other similar patients that could have similar symptoms.

From the scenario described it can be easily understood that the doctor is facing problems with his system in order to easily access the information he needs. Maybe this information is missing because the patient family history records are missing. Furthermore, the information even if it is available could be difficult and time consuming for the doctor to be accessed analysed and used.

2.3 To-be Process

Tom goes at the hospital because he is not feeling well. Tom has been diagnosed with diabetes and hypertension some months before. During the visit in the hospital he explains to the doctor his symptoms. The doctor decides to run some blood test because he isn't sure about his diagnosis. After two days Tom gets his test results which according to the doctor does not show much. Tom is concerned about his health, so is his doctor also. The doctor decides to get some help about Tom symptoms based on similar symptoms that other patients have had as well.

He goes to the computer and starts the PreMed system prototype. The prototype will allow the doctor based on Tom patient id to search in the EMR database and discover first, family related patients and second, other patients that are similar to Tom in terms of his diagnosed diseases. The system prototype shows to the doctor only the patients that have been diagnosed with the same diseases as Tom. The system asks the doctor if he wants to proceed with the analysis of the free text of those medical records. The doctor presses the continue button and waits for the result. The processing of the medical record free text is based on Latent Semantic Analysis.

The system visualises to the doctor in a two dimensional space that according to the calculation Tom is a very high risk person for developing a heart disease. The doctor's thoughts are now clear, he also investigates the reasons of this result by seeing in the screen that Tom is are very close to other

persons that had the same diseases as Tom and they were diagnosed with heart disease. So the doctor decides that Tom should enter in a preventive care program regarding this disease.

The system will allow the doctor to understand why this person is a high risk person. The doctor can always follow the flow of events and understand and elaborate the results of the system by investigating intra and inter relationship, doing so he is getting a second opinion or a suggestion before making a decision about his patient.

The difference between the two scenarios is the type of information used to make the diagnosis and the way this information is processed. In the first scenario the doctor has not enough information about Tom's family historical records and it is difficult for him to find other similar patients to Tom. The challenge on this scenario is that Tom's diagnosis is based on the doctor's knowledge alone. In the second scenario, even if Tom's family medical records are not available, the doctor will rely on information from other similar-to-Tom patients in order to extract important information regarding Tom's diagnosis.

The following section describes the state-of-the-art techniques and technologies used in the medical area in order to analyse and process the EMR patient information.

3. State of the art

The new information era and the rise of technologies have enabled organizations, companies and governments to gather and store a huge amount of data. Nowadays, data have been developed into an important factor of production in every sector [2]. Especially in the medical area, data have been used for different purposes. Today, with all information gathered about patients' medical records and new techniques and methods discovered, it is easier to analyse medical record data in order to uncover patterns and capture the value hidden in them. Having presented the dataset selected for this project in section 2.3, this chapter presents the state of the art techniques and technologies. Thus, I define and select the appropriate ones, which can be used in order to analyse and process the dataset.

This chapter is organized as followed; sections 3.1 and 3.2 describe and discuss different types of Big Data techniques and technologies available for medical data. Section 3.3 presents the academic and industrial related work made during the years, which are specifically related to the research questions.

3.1 Big Data Techniques

As information technology moves forward a range of techniques and technologies have been developed and adapted to analyze, manipulate and visualize big data. New techniques and technologies are still being developed in order to solve new problems and capture value from big data [2].

While I acknowledge that there are many techniques for general purpose used to analyze big data, I have only chosen to describe techniques relevant to my research questions. Therefore, the scope of this section is to describe, discuss and reflect on techniques that are adapted and used in the medical area.

According to [2] classification, association rule learning, clustering, data mining, predictive modeling, genetic algorithms, pattern Recognition, neural networks, collaborative filtering and machine learning are different techniques used to analyse large amount of data.

Basically, not only the above mentioned techniques, but also others (i.e. statistics, regression and spatial analysis) can be applied to this project. Below, I present the state of the art in big data techniques which are used to analyze large datasets.

Machine learning is a subcategory of the field called artificial intelligence that allows computers to discover and extract patterns and support the decision made, based on the data [12]. This process allows the machine to generalize because almost every non-random dataset contains patterns that makes possible the generalization [2]. In particular machine learning techniques could be very useful while analysing a dataset, because of the variety of the algorithms this category supports.

According to [12] machine learning has two major categories: supervised and unsupervised learning. The idea behind supervised learning is that the algorithm predicts based on a target value of the given dataset. Unsupervised learning is the opposite of supervised learning. There is no target value given for the dataset and it is not clear what we are looking for. In this project, we used clustering. Clustering is a technique used in unsupervised learning [12] that automatically forms clusters of similar things.

Machine learning can be categorized in supervised and unsupervised learning as shown in the table below [12]. Following the example presented in [12] it can be said that classification techniques are examples of supervised learning and clustering techniques are examples of unsupervised learning.

Table 3 Different algorithms for Supervised and Unsupervised learning [12].

Supervised learning	**Unsupervised learning**
Classification techniques	Clustering techniques

Machine learning techniques are used with medical datasets and they have shown great results [20].

> *"Current machine learning algorithms provide tools that can significantly help medical practitioners to reveal interesting relationships in their data."* [20]

However, which type of machine learning technique should be applied depends on the type of data (i.e. numerical, nominal or both). Section 3.4.1 provides a better explanation and examples of different techniques used in medical area, and specifically for predicting diseases.

Data Mining is a set of techniques uses in order to derive patterns from large datasets [2]. In order to achieve this a combination of different methods from statistics, machine learning and database management is needed. This combination requires at least one of the following techniques: association rule learning, cluster analysis, classification, and regression [2].

Predictive are those techniques which deal with building a model that will predict different results according to the type of data [14]. In this project, we employ visualization which is a descriptive technique. Further, I could utilize classification and regression that are predictive techniques in my dataset [14]. The use of data mining techniques applied to the given dataset can make possible to extract only important data features from the selected dataset and to derive value from them.

According [18] **Collaborative Filtering** is a technique that selects a set of users' data and compares it with other users' data in order to find a smaller set that has a similar taste. The main idea behind this technique is to find a way to determine the user similarities.

One way of achieving this is to compare one person with another and to calculate their similarity. The similarity score can be calculated with different methods, such as vector and Jaccard similarities, Pearson and Euclidean distance, or Cosine Similarity. However, in order to select the appropriate method to measure similarities between users, depends on the type of application one will develop and the type of data [18].

The above technique is used often in recommendation systems in different areas with successful results [4] and could be applied to our dataset in order to find patient similarities according to patient diagnosed diseases.

Classification is a set of techniques that makes it possible to categorize data by predicting the category that data will fall into [12]. The main idea behind classification is to build or train a classifier and to feed it with data that we know their category beforehand. This set of data used to feed the classifier is called training set. This technique could be very useful

when one wants to classify data according to the already known classes [12]. In the selected dataset, this technique is applied to classify patients according to their diagnosed diseases.

Clustering is a technique known for grouping items into different groups and splits different groups into smaller groups of similar items, whose characteristics of similarity are not known in advance [2]. The items are grouped according to the distance they have between them. The items that are in the same group should have smaller distance from each other; items that are in different groups should have larger distance between them [16].

Clustering as a technique is a general task to be solved, which uses different algorithms to achieve it. Clustering is an unsupervised learning technique [12].

In the selected dataset, clustering could be used for many purposes. One could be to cluster patients according to responses they have for a specific treatment. Another purpose could be that patients could be clustered according to their diagnosed diseases. Clustering technique could be applied to the selected dataset during the investigation of the patient symptoms and allow us to understand and explore why the specific patient is closer to one of the clusters.

Visualization Techniques are other techniques that make possible to visualize the results of big data analysis in a way that can be more understandable and comprehensive to readers [2]. Potential techniques are Tags, Clustergram, History flow, Spatial Information flow [2].

Tags visualization is a technique that differentiates a larger body text word that appears frequently from other words that does not appear so frequently, according to a weighted visual list. The words with higher number of appearances are larger in size than other words, making it easier to the reader to distinguish the concept of a word that it is more significant in a large body text set [2].

This technique could be very helpful to physicians in order to distinguish different symptoms of specific diseases, which is important to them. Further according to WHO [8], a specific disease has specific type of symptoms in order to be diagnosed. Therefore, it would be really helpful to a physician to have a visualization of those symptoms and distinguish the ones that are more frequent.

Clustergram is another visualization technique used to visualize how a specific item of the selected dataset could be assigned to the potential clusters presented. As such, it allows readers to better understand how the flow changes with the increasing number of clusters [2]. In our case, this technique allows a physician to understand how the flow of patients will change if the numbers of clusters grows during their classification according to their diagnosed diseases.

All the above techniques are frequently used in big data analytics depending on what one plans to achieve while working with different dataset. They all could be applied to our dataset. However, the goal of this project is to select the most appropriate technique in order to address the research questions. According to [12] there are some rules and steps to follow, which help in selecting the appropriate technique. Those rules and steps depend on the goal one wants to achieve and in the type of data contained on the dataset [12]. In section 5.3, I present and describe the reasons and the steps followed in order to select the appropriate techniques.

3.2 Big Data Technologies

The last years the number of technologies used to store and analyse big data is growing as the data collected are growing with an exponential rate according to Moore's law [3].

During the past decade many companies such as IBM, HP, and Oracle were offering data warehousing technologies. However, now more than ever, there is a need to store different data format such as structured, unstructured and semi-structured data.

The variety of the data stored has brought up different requirements about their storage and management. Traditional data systems face difficulties to handle the diversity of data because they were not designed to manage such a variety of data [3].

According to [3] because of this need, technology has moved to a parallel processing across multiple servers in order to manage the store and management of big data. Below in this section a description and discussion of data technologies, that can be used and could became potential solution to store and manage the selected dataset, is presented. To understand how

those technologies have been evolved we must look back at 2004 where Google presented the Big Table data storage system built on Google File System [17].

Big table architecture provides a column and row oriented structure that could be expanded into a large number of nodes. Everything depends on a master node from which all the other nodes are coordinated. This is the greatest limitation of the technology since every other nodes need to be coordinated from the master node and if the master node goes down then the whole systems is useless [3].

Hadoop is an open source distributed non-relational database which was based on the Google Big Table model [3]. Yahoo was one of the first companies to adopt and make use of it with great success [2, 3]. Since Hadoop is open source solution, this technology is being used in different companies and organization. For instance, hospitals could use this platform in the near future to process and storage the medical data.

Cassandra is an open source database management system which makes possible the storage and management of vast amount of data on a distributed system. Cassandra can store up to 2.000.000 columns in one row. It was build according to the Dynamo Amazon architecture and in contrast with Big Table it is not depended on the master node of the system. Every node can accept data from the whole system and can answer queries independently where they are coming from [2, 3].

MapReduce is another technology used in order to process vast dataset on distributed systems. This framework was introduced by Google Inc. and because of its success, it was implemented in Hadoop technology as well [2]. MapReduce technique can be useful while working on the specific data set since I can use the MapReduce concept in order to pre-process the selected dataset and only extract information that is valuable for the following development process.

NoSQL is another technology that restores and manages large dataset and characterizes a new category of non-relational, distributed, open-source and horizontally scalable databases [18].

According to [18] there are different categories of NoSQL technologies. One of them is **Graph Database.** It can deal with the complexity of different datasets. Since in medical data is

complex, because of the nature and the variety of information stored, the graph database could be a potential solution to store this information because any type of information can be represented in an accessible way. Graph databases stores such information in a single data structure called graph that contains a node, edge and property [18].

Relational Database Management Systems (RDMS) are a set of related tables where the data are stored in rows and columns. Usually the data stored in RDMS are structured and unstructured data [2]. One common language, which makes possible the management of different abilities such as insert, update and delete the data, is SQL. Microsoft SQL Server is one of the above technologies used to store and retrieve information according to users' requests [2]. The selected dataset is stored in an MS-SQL server since that was the technology used from the source hospital where the data were collected.

All the above technologies can be used in order to store and manage different types of dataset including the selecting one, but the selection of the proper technology depends on many factors. One important factor is to get known with the dataset you want to store and analyze. Other important factors are flexibility in consistency, scalability and performance of those technologies [3].

In the section 5.5 all these factors will be discussed and evaluated in order to select which of those technologies can be a potential solution.

3.3 Related Work

This section elaborates the related work done in the medical area. According to [3], medicine was the primary area where analysis and investigation of different big data were made in order to extract value. During the literature study I discovered that the amount of research made in this area is quite large. There are different approaches and different techniques used in the medical area and as the research [10] claims.

> *"Machine learning methods have been applied to a variety of medical domains in order to improve medical decision making".*

However, my focus is to predict the risk of future diseases for a specific patient and help the physicians make better decisions during the diagnosis procedure using an electronic medical records (EMR) database.

3.3.1 Academic Related Work

Data mining and machine learning techniques has been used very often in order to make prediction in medical area.

A recent research [19] describes different techniques used in order to predict and diagnose sub types of diseases, especially in the genomic medicine. Large dataset of DNA microarrays that contains huge number of genes for every patient where analyzed in order to attain a prognosis which can be used for personalized patient therapeutic decisions. Such techniques are the decision tree, neural networks, naive bayesian classifier, bayesian networks, support vector machines (SVM) and k-nearest neighbors [19].

Moreover, the paper [20] describes the state of the art and a historical overview of predictive techniques used in medical area. Below, I will present some of the techniques that are used in order to improve the diagnostic procedure for ishaemic heart disease derived from [20].

The techniques used for this purpose are: Decision Tree algorithms such as Assistant-R, Neural Networks, Naïve Bayes, Semi Naïve Bayes, and K-NN. All these techniques had a very successful prediction in the context used. Author of this the research [20] presents also the appropriateness of those various algorithms according to their performance, transparency, explanations, reduction, and missing data handling. One of those techniques can be a potential solution to my dataset. However, all these techniques have been applied in only one specific disease according to the trained dataset for every disease [20]. In my case the system will suggest the high risk patients for every disease chosen, and not only for one singular disease such as [20].

During the literature review, I discovered an interesting paper [21] which uses a different predictive approach in medical area using the collaborative filtering technique. The

collaborative filtering method is widely used in other areas, specifically in recommended systems. The main idea of collaborative filtering was presented in section 3.1.

This approach is completely diverse from other research made before in the medical area because they present a predicting system based on all types of diseases not focusing only in one specific disease. The result is a list of potential future diseases that can affect patients. Moreover, the prediction of patients is based to a set of similar patients that have been diagnosed with the same diseases.

The above research [21] draws an analogy between the marketing use of the collaborative filtering such as recommendation systems and medical predictions. The analogy made is that patients are acting the role of users and the diseases the role of items. Moreover, this research introduces the binary rating instead of the normal rating. A patient is diagnosed with a specific disease (1) or not diagnosed with it (0), hence there are no set of rating to deal with [21].

Furthermore, the work above [21] is based in the International Classification of Diseases code (ICD-9) and the patient visits at the hospital. This code allows defining the type of the disease(s) that the patient has been diagnosed. Specifically, [21] used vector similarities collaborative filtering method in order to predict the list of potential future diseases.

This academic work is closer to my research question where I want to investigate patient medical records in order to predict the risk of developing a specific disease based on his family historical records and other patient similarities.

According to the described related work [21] and their discussion, the framework was developed in order to assist a physician during decision making and inform patients for the risks they have on developing specific diseases. The results presented were impressive and their approach [21] is innovative. However, the question is how close is this framework in order to be used by the physicians?

According to [20] there are some guidelines in order such a system to be accepted and understood by physicians. The system must be able to have good performance, to deal with missing and noisy data, and to explain the decisions and recommendations made. This explanation should be transparent to the physician and he must be able to analyse and

understand the generated knowledge from the system. The system has to explain the reasons of his decisions and predictions otherwise the physician will not accept them [20].

Therefore, I use the most important concept described such as patient similarities, ICD-10 codes, and research [21]. Specifically I will use the concept of collaborative filtering only to find similar patients according to their diagnosed diseases. In addition to this technique I will use a latent semantic analysis technique to the set of similar patient extracted from the collaborative filtering technique in order to predict the disease that can possibly affect the specific patient according to his symptoms and his medical history. In addition to my work I give more detailed information about patients' similarity by involving the family relation between patients.

Furthermore, the visualization part of the system will allow the physician to investigate the clusters and to understand why the selected patient is on high risk to develop the specific disease based on the patient family relation and other patient similarities.

This work is inspired by the research [21] and addresses the research questions and provides physicians with a tool that can be easy used and help him make better decisions. Specifically, my approach finds similar patient. Further it predicts the high risk of the patient on developing a specific disease, and addresses most the challenges pointed by [20] making possible for the physicians to understand the decisions made by the system at every step.

3.3.2 Industrial Related Work

Electronic Medical Record IT solutions used in the medical area are a more general category that contains the EMR information. Moreover, they contain integrated information of hospital information systems, patient administrative system, a large number of administrative tools for or billing and financial management and a patient portal [22].

One of this solutions used in Denmark is the My Clinic IT Solution that combines all the above information. According to their website the solution is used in a large number of hospitals and their biggest installation contains electronic health record information about 800.000 persons [22]. The solution provides a large number of functionalities such as medical record and electronic health record, medication and treatment, calendar, time booking,

examinations electronic communication, clinical management information, bookkeeping and financial-management in order to make as easy as possible to collect all the necessary information regarding Electronic Health Record (EHR). The My Clinic solution is also able to filter and share data between different departments and support semi-automated coding diagnosed according to the world health organization classification such as ICD-9 or ICD-10 and IPCPC [22].

What it seems to be missing is an additional feature or tool that can be applied upon this solution in order to help physicians not only access and share the available patient information but also elaborate and analyse that, in order to facilitate the decision making during the diagnosis process. My system prototype is a potential solution that can be applied to this type of solution. Such an extra tool will allow physicians not only to access, elaborate and analyse the patient electronic records but also to visualize the results of the analysis in a way that the physician can investigate the reasons of these findings.

As mentioned before the amount of gathered and available medical data stored in that type of solution such EHR is growing very fast and the large part of the collected data are in unstructured text and based on natural language. Recent research [23] has shown that those data are doubled every 5 years. It is difficult for a physician to deal with large amounts of data.

This is the reason why the technology and science is looking for solutions in order to address these challenges. Watson IBM is an innovative technology solution based on the natural language processing that tries to solve and address those challenges. The idea lying behind this technology is the use of natural language capabilities, hypothesis generation, and evidence-based learning to support physicians as they make decisions during diagnosis process [23].

The process followed is described below. The physician makes a query about a patient describing the symptoms and other potential related factors. Then, the system selects key component of the query based on the incorporated system medical terminology. Next, the system digs into important patient medical data in order to discover family related data, medical treatment and other information. The combination of all this information and the information of clinical test, doctor and nurses notes and journal articles is used in order to

extract different hypothesis and test them. After those steps the system provides the physician a list of potential diagnosis according to the score of each hypothesis made [23].

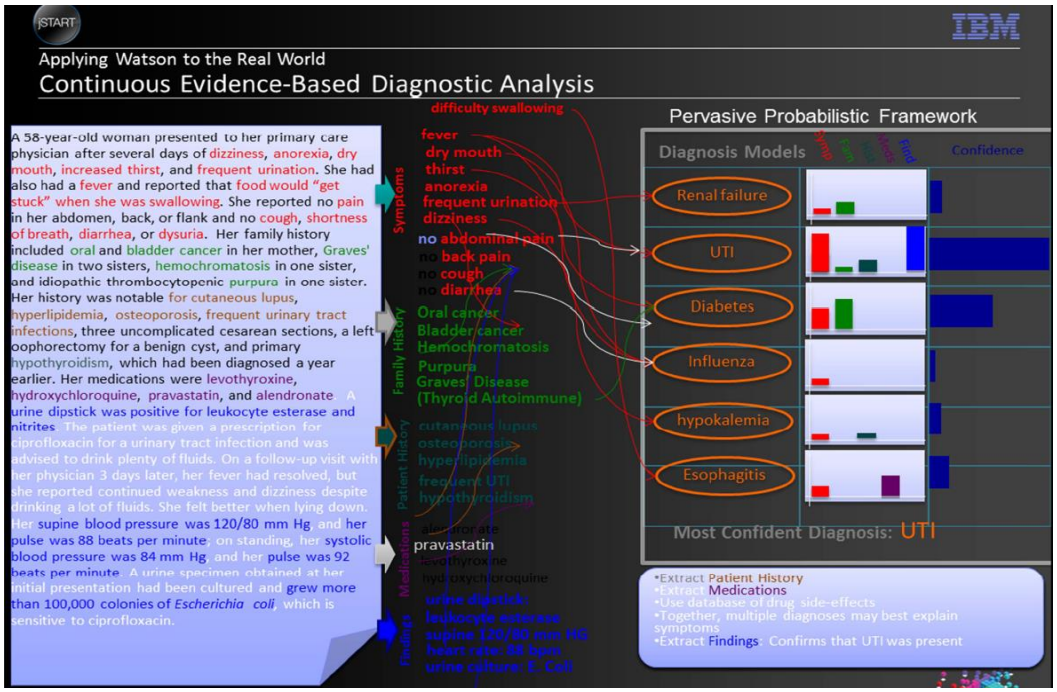


Figure 1 Watson at work [24]

Figure 1 presents an example of the steps made in order to predict a patient diagnose based on the Watson system and according to the processing of the unstructured medical text based on natural language processing (NLP) [24].

I was fascinated by the ability of the Watson system that takes the context as an important component while performing all the described tasks in order to help the physician make better and more accurate decisions. The visualization of how the decisions were made inspired and gave me ideas about how to visualize the result of my system.

However, research [25] has shown that the use of Watson somehow limits the physician's participation during the diagnosis process. Furthermore, according to [25] physicians are in charge for training of Watson, and they have to take care of it by explaining to the machine which is right and wrong. In addition not all the unstructured data are used for diagnosis, that's why during the analysis of the unstructured data Watson removes a large part of them because Watson is based on rule learning according to different disease's symptoms

and does not take into account other unstructured data that could contain valuable information [25].

According to [25] analysing all the unstructured data gives mores possibilities for latent semantic analogies.

> *"This is especially useful when looking for indirect causes for a disease, like for example residence or eating habits"* [25].

The system under development will only suggest the diagnosis to physicians and will help them understand the reasons of this prediction but it is up to him to take the final decision according to his opinion and intuition.

To summarize, in this chapter the state-of-the-art Big Data techniques and technologies were described in order to familiarize the reader, as well as myself, of potential components of my solution or answer to the research question. Last but not least, I presented related work, done in the medical area, so as to provide the reader with an overview of the techniques that are currently used to deal with similar problems.

4. Methodology

This section describes the method followed to answer the research question. Specifically, I provide a thorough account of the empirical and theoretical data that was collected during this project.

After the problem formulation, I started to design an appropriate research approach in order to address the research question. First, I started looking into the literature in order to better understand the concept of big data and the use of analytics and visualisation in the medical area. Through the literature I gained insights on the use of big data analytics in previous studies in the medical area, as well as on the techniques used to analyse and process datasets. This knowledge helped me in identifying the criteria in order to select the appropriate techniques for my problem area. Second, in order to understand better the medical dataset, I conducted an interview with a doctor which has experience on the use of such systems in hospital settings in order to understand what information is valuable to doctors from the entire dataset. Finally, the knowledge obtained from the semi-structured interview and the literature study allowed me to explain and argue about the assumptions, choices and the decisions made during this work process.

Moreover, I started to plan the developing process of such a system and I chose the software development method used in order to address the research question. The figure below describe the overall process followed from the beginning of this project until the end of it.

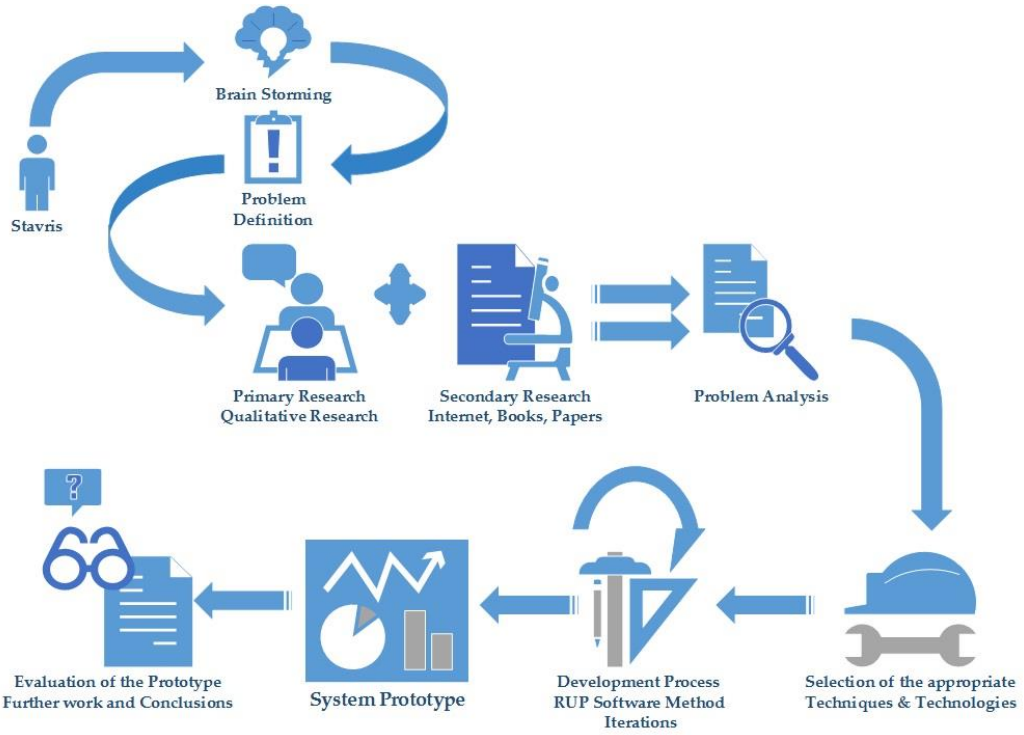


Figure 2 Design research approach

The structure of this remaining chapter is divided in different subsections such as primary research, secondary research, experimental approach, development process. In the below sections, I describe how conducted each of these processes.

4.1 Primary research: Interview

The collection of empirical data was very important because it allowed me to gather more specific information about the problems faced by the physicians while dealing with patients' electronic medical data. Though, the first step was to conduct a semi-structured interview with the doctor and try to understand the process that they follow during patient visits. I conducted a semi-structured interview with a doctor which has a good knowledge about the medical data records that hospitals store about patients. The purpose of this interview was to have a better understanding of the most important data that are presented in the patient medical records and to better understand the doctor needs and expectations from the analysis and the visualization of these data.

The next second step was to combine all the information gathered by the qualitative research, in order to understand and define the problems and the challenges the physicians are facing today while trying to process the EMR patient medical data.

In the semi-structured interview, I presented some hypothetical questions regarding the prototype system under development. The idea was to gather more information about the potential solution that will help me to address and answer the research question.

4.2 Secondary Research: Literature Study

Secondary sources are important resources for collecting data related to the topic of the research question. I used different sources for the research study such as the Internet, books, and papers. During the search process, I used different key words, such as medical data, big data techniques and technologies, medical records, machine learning in medicine, and visualization methods for big data. These keywords were used in different combinations to find relevant literature.

The use of Internet, books and articles helped me to identify literature related to the medical data record and how those data have been processed and analyzed for different purposes. After having a general understanding of what exactly medical data records are and what type of information they contain, a deeper and more specific investigation followed. Specifically, I investigated the area of medical data and searched for studies related to medical records, patient treatment, symptoms, journal and relation between patients such as ancestors, relatives and descendants.

Moreover, the literature search helped me to identify the different big data techniques and technologies that are used to analyze and visualize the information contained in medical records in order to extract valuable information.

The purpose of the secondary research was to facilitate and to get a better understanding of the techniques technologies used in big data analytics and the visualization process with a focus on the area of medical record data in order to select the appropriate combination of techniques and technologies to answer the research question.

4.3 Experimental Approach

After collecting all the necessary information from the primary and the secondary research regarding the problem area, the next step was to analyse the given dataset. The problem that this project is trying to address is quite complex and as a consequence needs an experimental approach where I experiment with the dataset in different ways. Due to the complexity of the problem I decided that the experimental approach process should to be divided into 4 tasks: pre-processing the data, intra and inter-family relations' investigation and suggesting the high risk of developing a disease.

The **pre-processing** of the given dataset was an important task, during this task, was possible to reduce the information flow of the medical record data. This process made possible to filter the medical records features, only to features that are important to the physician during the diagnosis process. To achieve the above goal a combination of SQL queries was used following the Map Reduce programming model. According to [26] the concept of Map Reduce actually refers to two separate and distinct tasks that are performed. The map reduce procedure allowed me to work with the chosen dataset and to convert it into other different dataset in order to define and finalize a dataset that contains only the patient medical records and data features we are interesting in. A further explanation about the MapReduce process will be given in section 5.2.

The **intra-family** investigation allowed me to understand and identify the patient family relations based on the Family Relations table stored on the selected dataset. This task was necessary to define the relations of a patient with his ancestors, descendants and relatives because according to the physicians suggestions is important to know the family medical history of the patient. During this process was possible to group the patient medical data of the patients according to their family tree having ordered one family at the time.

The **inter-family** investigation allowed me to explore and define the relations that a specific patient could have with other patient and not only with his family members according to the diagnosed diseases. During the inter-family investigation process a technique was chosen in order to define patients relations based on the similarity of the patients according to

the collaborative filtering technique. Collaborative Filtering technique is based on the similarities between users or items. According to [21] it can be an analogy between the use of collaborative filtering in marketing and in medical prediction. In this project this analogy is used in order to create groups of the patient according to their similarities based on their diagnosed diseases.

The method used for grouping patients according to their similarities was Jaccard Coefficient. This method was chosen because every patient record is seen as a binary vector. Every patient has been diagnosed (1) or not been diagnosed (0) with a specific disease. According to [27] for data that are binary, it is more common to look at each vector and try to find how the entries of one vector match to the other. More discussion about this method is described in the section 6.2.2.

After the pre-process of the dataset, the exploration of the inter and intra family relations between patients an analysis of the patients free medical text was necessary in order to **define the high risk** of the selected patient based on the patients similarity method defined during the previous task. During last task a Latent Semantic Analysis technique was selected to make possible to analyse the medical record text of the specific patients according to other similar patients and extract valuable information. This information in combination with cluster visualization technique was used in order to predict for a given patient the risk of being affected from a specific disease.

4.4 Development of the prototype

In order to develop a system that address and describes the above challenges and problems, a software development method was needed. I used the principles of the Rational Unified Process (RUP) method for software development [28]. The RUP method is used only during the development of the prototype not for the whole project.

The use of RUP can be divided in four phases: Inception, Elaboration, Construction and Transition. The inception phase is the first phase of the iteration and during this phase it is possible to define the delimitations of each task. The second phase, the elaboration phase, makes it possible to analyse the problem domain and the requirements. The third phase is the

construction phase, where the application is developed. The transition phase is about launching the product to the market [28].

Hence, the system under development is a proof of concept the transition phase is not included and described in this project. As the RUP method makes it possible to have different iterations during the development process, it allowed me to test and investigate the output of each iteration.

During the development 3 main iterations were carried out. The first iteration allowed me to make the patients mapping according to their diagnosed diseases. The choice of the diseases, for the purposes of this project, was made according to the physician's suggestions. During this iteration, it was possible to present the patient diseases information as a binary vector with 0 and 1 values based on recent research [21] and as explained above in section 3.3.1. The second iteration of the development was to implement the selected techniques in order to find similar patients. Specifically the collaborative filtering technique and the Jaccard Coefficient method were implemented during this iteration.

The third iteration was the implementation of the latent semantic analysis technique used in order to analyse the free medical text of similar patients in order to define the risk of developing a specific disease for a specific patient. Moreover, during this iteration the visualization of the results are presented in a way that will allow the doctor to understand and investigate more effectively the reasons why this patient has a high risk of developing a specific disease.

The visualization of the results will be presented in a two dimensional form and different colors were used in order to understand better the results. The patient with a high risk will be shown in red color. Moreover, after the final iteration different experiments took place to test the cluster efficiency and investigate the results. More information about the testing and the experiments made is presented in the section 6.3 and in the problem solution chapter.

5. Problem and Dataset Analysis

In order to collect all the information needed to answer the research question a detailed investigation of the different challenges during the problem analysis was conducted. First, I conducted a semi-structured interview in order to understand the problem faced and to define the physician's needs during diagnosis procedure. Second, the use of as-is process and to be-process as described and explained in the section 2.1 and 2.2 helped me to understand better those needs and gave me ideas how to address the challenges faced. Third, different techniques and technologies were investigated to assure the selection of the appropriate one in order to pre-process, analyse, and visualize the selected dataset and derive value. Finally, having all this information available made it possible to make the right decisions and define the right sequence of steps to follow in order to work with the selected data and discover value that can be useful to the physician during the diagnosis process? In the following sections, I describe the different steps followed during the problem analysis.

5.1 Physicians overview and suggestions

The purpose of the interview was to better understand the process physicians follow during the diagnosis process, define his needs and listen to his suggestions about the importance of medical history patient data and specifically the family relations.

During the discussion with the physician an interesting fact came up. The problem of family history records which plays an important role to the physician decision during the diagnosis process. As the interviewee explained, she had experience using EMR in other countries and in Denmark as well and she could see big differences.

> *"When you take medical history of a patient in my home country you cannot make a relation between family members. You just have to fill a box with the information about the family members if the patient explain that a family member had the same disease. Here in Denmark it is possible to define the relation of a patient according to the relation degree (ancestor, descendant, relative). This is important information for us in order to keep family history records related to the symptoms of patient in order to see similarities" [Doctor Interview].*

It was clear that one of the most important problems to deal was the amount of information available in the EMR system and the need to analyse it. As mentioned in section 1.2.2, even nowadays when the hospitals are taking detailed information about the patient, it is difficult and time consuming for the doctor to analyse the amount of information.

> *"It is time consuming when you need to elaborate the family historical records because of the free text written. First you have to find the person related to the patient then you have to read the free text or the diagnosis made" [Doctor Interview].*

According to the physician there is also a need to better understand the patient needs and symptoms in order to prevent different inherited diseases. However, this is a complex process and creates a number of difficulties that slows down and makes difficult the diagnosis process. One of the difficulties is the complexity of the information presented in the EMR. As explained in section 1.2.2, the use of this complex information depends also in the knowledge and experience of each physician.

> *"Until now in order to make a diagnose we have to wait for the clinical test. But what if the test does not show anything we possible may use some help by analyzing the patient family historical records of other similar patient. Patients are becoming more and more demanding about their health care" [Doctor Interview].*

As the physician explains above and according to research the diagnosis process nowadays is based on the **disease – centered** model [4]. Recent research [4] has also presented also another model in healthcare called **patient – centered**. The difference with the old **disease – centered** model where the physician decisions were based on clinical expertise and medical evidence of different tests, is that now the patient contributes and receives services based on his individual needs and preferences [4]. More focus is been given to the patient in order to prevent and manage the different diseases that can affect him. The challenge is to make possible, to develop and use such a system that will achieve the goal to prevent and manage potential patient diseases before developing the specific disease.

5.2 Pre-processing the Dataset

The selected dataset has around 281.000 patient medical records. As described in section 2.1, the EMR dataset contains complex information regarding a patient medical record. In the selected EMR database there are many tables related to each other that contains information about a patient such as Personal Information, Patient Records, Patient Drug Table, and Patient Family Relation, Medical Record, Medical Record Line, ICD10ID Table, Medical Clinical Data Table. In Appendix 2, an overview of the structure of the database and the relation between the tables is available.

However, not all the information contained in these tables is necessary and valuable information in relation to my research question. As a result, there is a need to filter a large part of the above information. Specifically, the dataset was pre-processed in order to gather and analyse from the chosen dataset only the information that is most significant and important for the future work. The most important information was identified with the help of the doctor. As the doctor clearly explained during the interview:

> *"It is very important to know the medical history of the patient, the medical history of the family related person but also of other similar patient." [Doctor Interview]*

According to the doctor suggestions and the investigation of the dataset, I decided to work on a subset of this dataset in order to gather the necessary information for the further development process. Below, I present a description list of the tables selected in order to extract the necessary information about a patient and the reasons of those choices.

- **Client** table contains information about client personal data such as Name, Surname, Address, CPR number, year of birth.
- **Family Relation** table comprises of information regarding the family relation type and the degree that the patients may have between each other (Ancestors, Descendants, relatives).
- **Medical Record** table is about the medical record ID of each patient based on their CPR number. Contains also information about the patient diagnosis based on the

ICD10ID (World Health Organization disease code), a summary of the patient diagnoses and patient treatment.

- **Medical Record Line** table consist of information regarding the medical record text, patient diagnosis based on the ICD10ID (World Health Organization disease code), treatment summary, patient ID, medical record line ID and medical record ID.

The picture below presents an overview of the information contained in the above tables and the relationship between them.

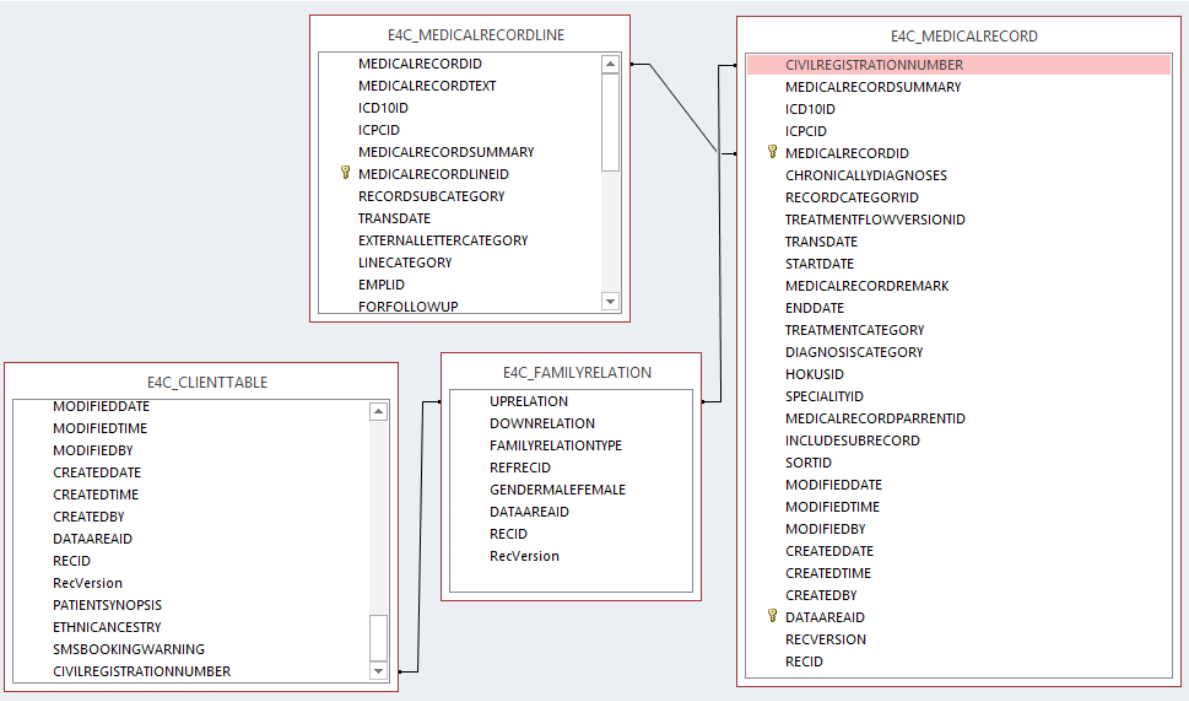


Figure 3 Information data contained on the selected tables

Furthermore, all the above tables also contain other data as well e.g. about the exact date and time of when the above data was recorded and information of the person who inserted them. A better overview for each of the presented tables and the data enclosed into these tables can be found in Appendix 1 [Figure 11, 12, 13, 14].

According to the research question, I want to investigate the relationship between diseases, patient diagnosis, and intra and inter-family relations in order to identify high risk patients for a specific disease. The tables described above provided me with all the necessary information to address the research question. However, the information, presented on the

selected tables, still needs to be analysed and processed in order to accomplish the research goal.

Since all the selected tables displayed above in [Figure 3] contain complex and different type of information there is a need to filter and extract exactly the piece of information which is most important according to the doctor suggestions and the analysis made in these tables. Specifically, the main idea was to reduce the amount of information contained in the above tables in order to process only information that will be useful during the diagnosis process.

The procedure followed to reduce the flow of information was based on SQL query analysis according to the MapReduce programming model. The MapReduce method is widely used while working with big data in distributed systems. The main idea behinds this programming model, lays behind the two procedures that the model has implemented. The map procedure consists of filtering, sorting and extracting something one cares from the dataset. While, the reduce procedure consist of aggregation, summary, filtering, or transformation of those results [26].

The procedure followed was to take the datasets extracted through the map procedure as input and combines them into a smaller dataset in the reduce procedure according to the needs of the development [26]. Specifically, I selected from the chosen dataset tables that contained information about the patient medical record. Every time I extracted information from the selected tables a new table with combined patient data was created. Later each of the created tables was used in combination with the already existing tables in the database in order to achieve the goal and finalize the dataset.

Figure 4 shows new tables created through every step made during pre-processing phase. Furthermore, this figure demonstrates how the information flow changed during the pre-process task. Moreover, it shows how the information flow of the selected tables was reduced through the MapReduce model.

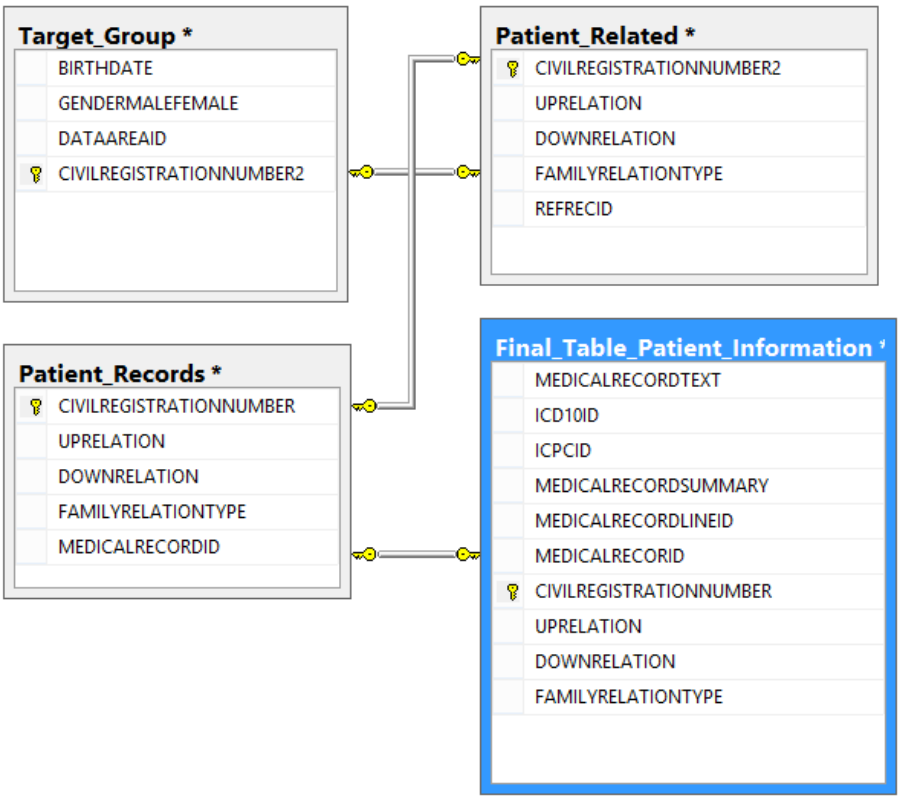


Figure 4. The tables created during the pre-process task

Table 4 summarizes the steps followed during the pre-processing of the dataset. The purpose is to finalize the dataset according to the investigation made it and based on the physician's suggestions.

Table 4. Step description followed during the preprocess

Step followed	**Description**
Identifying the patient target group.	Selecting patient born from 1945 until 2013.
Investigating the Inter-Family relation of the selected target group.	Find the family relation between the selected patients and ordering them according to the up relation. Building in this way the patient family tree if this information was available.
Matching Patients ID from the previous step to Medical Record ID.	Selecting patient information from Medical Record table based on the condition *Patient ID = Medical Record ID*
Creation of the final table with all the necessary information about the patient.	Combining all the previous gathered information with the patient medical record line table in order to finalize the dataset.

The subsections below present a detailed description of the SQL queries developed and the decisions made in order to make possible the processing of the dataset according to the programming model of MapReduce.

5.2.1 Defining the patient target group

The first step was to identify the patient target group by selecting only patients born from 1945 until 2013. I have filtered information based on the year of birth of each person from the client table. Only persons that are born from 1945 until today will be considered for further investigation. The information extracted above was inserted into a new table in database named Target Group Table.

```
SELECT [BIRTHDATE], [GENDERMALEFEMALE], [DATAAREAID], [CIVILREGISTRATIONNUMBER2]
Into [Target_Group] FROM [CLIENTTABLE]
Where BIRTHDATE < '2013-04-17' and BIRTHDATE > '1945-01-01' order by BIRTHDATE
```

The purpose of this selection was to maximize the chances that for every patient there are enough medical data about his health problems and the related to the patient ancestors and relatives. The selection query could be made for persons born before 1945 as well. I am aware that making this decision probably I have a loss of information regarding the patients filtered out. However, selecting all the patients from the dataset there is the risk to have patients' with no information about their family medical history.

The table below presents an example of the information extracted from the client table of the database.

Table 5 Example of Patient information in Target Group table

Patient CPR	Gender	Year of birth
1201450033	0	1945
1905621738	1	1962
0602681388	1	1968

5.2.2 Intra-Family relation between the selected patients

The second step of the pre-process analysis was to define and explain the family relations between patients. In order to find the patients' relationship (i.e. ancestor, descendant and relatives) I analyzed the Family Relation Table and in particular the information included in the columns Up Relation, Down Relation and Relation Type. The Up Relation and Down Relation columns show the CPR number of the ancestors, descendants and relatives of a specific patient. The Relation Type column defines the type of relation between the patient and the Ancestor, Descendant and Relatives, as well as the degree of the relation. The table below shows an example of those records and how they are presented in the Family Relation Table.

Table 6 Instance of the Family Relation Table

Up Relation	Down Relation	Relation Type	Gender
xxxx 02 7395	xxxx 73 2993	1	0
xxxx 01 6093	xxxx 93 2622	12	0
xxxx 01 6093	xxxx 99 1101	13	0
xxxx 61 1570	xxxx 57 2905	3	1

According to the discussion with the doctor and the analysis done based on the 5th, 6th and the last digit of the CPR number, it is possible to define those relations between patients. The 5th and the 6th digit of the CPR number declare the year of birth of the patient. The last digit of the CPR defines the gender of the patient male or female. Based on those findings, it is possible to define and describe every relation type in the family relation table. The table below describes the different type of the relationships between patients discovered on the previous step.

Table 7. Examples of relationships type between patients

Family Relation Type	Description of the Relationship
1	Father - Daughter, Father - Son
2	Mother - Son, Mother – Daughter
3	Wife – Husband
10	Brother – Brother

However, there are some special cases to deal with while trying to understand the family relations of the patients. These cases consist of the adopted children, I was not able to identify

a relation that could describe these types of relations. Such cases need further investigation. However, the information available for the adopted persons can be used when trying to find similarities between all the patients and not only between related patients. After understanding and defining the type of relation between patients, the next step was to combine the information from Family Relation table with the Target Group table in order to find the related patients and group by them according to their family relation

```
SELECT [CIVILREGISTRATIONNUMBER2], [UPRELATION], [DOWNRELATION],[FAMILYRELATIONTYPE]
Into [Patient_Related]
FROM [Target_Group] Left join [FAMILYRELATION]
On [Target_Group].[CIVILREGISTRATIONNUMBER2] = [FAMILYRELATION].DOWNRELATION
Order by UPRELATION desc
```

The results of the above query are inserted into a new table named Patient Related table. Selecting left join between those two tables allowed me to keep the information about patient that probably do not have any family related information. The medical information of non-related patients is also important and it can be used to find similarities between all patients. The last line of the SQL query makes it possible to order the patient in such a way that it is possible to view the patients' family tree. The results of the above query are inserted on a new table which contains information about every patient, their family related persons, their gender and the CPR number.

Table 8 Example of the table patient related

Patient CPR	Up Relation	Down Relation	Family Relation	Gender
3012980112	0502722878	3012980112	2	1
0210941447	3012970298	0210941447	12	0
1801922598	3012990915	1801922598	13	1

5.2.3 Matching Patients and Medical Record

The third step is to process and analyse further the dataset by combining the results of the created **Patient Related table** with the **Medical Record table**. The purpose of this combination is to assign to each patient in the Patient Related Table his medical records according to the medical record ID. The result I wanted to achieve was to have information about the patient

such as, CPR number, Up Relation, Down Relation, Family Relation, Gender and Medical Record ID into one table. In order to achieve this a new table called **Patient Records** was created**.** This table was a result of the inner join between the Patient Related and Medical Record table. These two tables can be joined because they both contain the CPR column. I was interested only in the patient medical record ID where the patient CPR number matches in the both tables. Once more in this step there is information loss, because the patient medical record related to the patient not presented in the selected target group are filtered out. But as explained above this information was left out because it does not help a lot since, patient data are missing from these record.

```
SELECT [MEDICALRECORD].[CIVILREGISTRATIONNUMBER],
[MEDICALRECORD].[MEDICALRECORDID],[Patient_Related].[UPRELATION],
[Patient_Related].[DOWNRELATION], [Patient_Related].[FAMILYRELATIONTYPE]Into
Patient_Records FROM [MEDICALRECORD] Inner join [Patient_Related] On
[MEDICALRECORD].CIVILREGISTRATIONNUMBER=[Patient_Related].CIVILREGISTRATIONNUMBER2

Order by [Patient_Related].UPRELATION desc
```

It is important to have the patient medical record ID. This information is valuable and useful in order to extract the free medical record text for each patient according to his medical record ID field that is presented in the medical record line table also. The field medical record ID will allow me to extract information from those two tables on the condition that this field is equal in both tables. An example of the information inserted in the new created table is presented below.

Table 9 Example of Patient Record Table

Patient CPR	Up Relation	Down Relation	Family Relation	Medical Record ID
3012980112	0502722878	3012980112	2	00004468
3012991591	0801692380	3012991591	2	00089064
1801922598	3012990915	1801922598	13	00008891

5.2.4 Finalizing the dataset

Last but not least is the step to finalize the dataset we will use in order to develop the system prototype. This step consists of combining and filtering the information stored into the previous table created and the **Medical Record Line table**. As mentioned before, this table

contains information about every patient during their visits in the hospital. This information consists of free medical record text, ICD10ID (the diagnosed diseases), IPCID (the category of the disease), Medical Record Summary, Line ID, Transaction Date and Medical Record ID. The SQL query that allowed me to combine the information between these two tables is presented below.

```
Select [Patient_Records].[CIVILREGISTRATIONNUMBER],
[Patient_Records].[UPRELATION], [Patient_Records].[DOWNRELATION],
[Patient_Records].[FAMILYRELATIONTYPE],
[Medical_Record_Line_Distinct].[MEDICALRECORDTEXT],
[Medical_Record_Line_Distinct].[ICD10ID],
[Medical_Record_Line].[MEDICALRECORDID],
[Medical_Record_Line].[MEDICALRECORDLINEID], [Medical_Record_Line].[ICPCID],
[Medical_Record_Line].[MEDICALRECORDSUMMARY], [Medical_Record_Line].[DATAAREAID]
Into Final_Table_Patient_Information
From [Medical_Record_Line] Inner join [Patient_Records]
On [Patient_Records].[MEDICALRECORDID]=[Medical_Record].[MEDICALRECORDID]
Order by UPRELATION desc
```

During this selection the rest of information on the table Medical Record Line was filtered out. In the following table, I present the information extracted from the combination of the above SQL queries. Moreover the importance and the utility of this information is summarized for every column in the table below.

Table 10 Data description of the final table

Information	**Description**
Patient CPR	The patient ID that distinguish patient between them
Up Relation	The Ancestor relation of the patient (e.g. mother, father, grandmother, grandfather).
Down Relation	The Descendant relation of the patient (e.g. daughter, son)
Family Relation Type	Type of relation (e.g. mother, son, father daughter)
Medical Record Text	Medical description about the patient symptoms or diseases.
ICD10ID	The universal code for a specific disease based on the WHO classification codes.
ICPCID	ICPC-2 DK codes used in Denmark for the classifications of diagnosis according to WHO ICPC-2.
Medical Record Summary	The summary of the patient medical record,

The most significant and important information regarding the patient medical history which can be used to facilitate the diagnosis process are:

- **The Medical Record Text** describes the symptoms of the patient and physician's notes made during the visit and the diagnosis process regarding patient symptoms or disease. This field will be used during the LSA analysis in order to discover patterns among similar patient based on their free medical record text.
- **ICD10ID** is the code of a specific disease that a patient is diagnosed with according to the WHO standard. Only after the doctor is sure about the patient diagnose he can fill this specific field [2]. This field is important for my further work because according to this field the classification of similar patient will take place. Based on the selected field is possible to define which are the diseases diagnosed for every patient.
- **ICPC-ID** states for International Classification of Primary Care; it is the code used in the Danish health system in order to categorize the diseases according to the WHO standards. According to [29] the use of ICPC is widely used in Denmark from 1998 the Danish version of ICPC has been implemented in the electronic health record systems, and has been renovated in 2008. ICPC system used in Denmark does not only provide a simple translations of ICPC-2 codes from WHO but it also provides a complete mapping for the ICD-10 classification used in hospitals making in this way more coherent the patient record between different health sectors [29].
- **Up Relation** and **Down Relation** define the family tree of each patient. Specifically, these relations find the ancestors and relatives of a specific patient and their medical history. This information is important during the process of finding similarities patterns between similar patients in order to investigate why these patient are similar or have similar symptoms.
- **Patient CPR** allows to distinguish patients between them, it is a unique ID for every patient

After analyzing the original dataset through the above queries (subsections 5.1.1, 5.1.2, 5.1.3, 5.1.4) we have the final table with the most important features of information needed for the development of the system prototype. The following section discusses and describes the

selected techniques for finding patient similarities and the prediction of the high risk of a patient disease.

5.3 Collaborative Filtering

During the last years, research in the medical area has shown that a large number of techniques are developed and used in order to make medical predictions. Most of them are used to make prediction about specific cases or a class of diseases based on some combination of basic data.

According to [4] patients exposed to similar risk factors such as similarities in lifestyle, genetic predispositions and environmental factors may develop similar diseases. Also, during the interview made with the physician I found out that family historical records and other patient similarities play or might play an important role during the diagnosis process.

As described in section 3.1, there are different techniques that can be used in order to classify a group of data. The classification technique could be used to classify patient in different groups according to their diagnosed diseases. However, as discussed in section 3.1, in order to use this technique to classify the data, one needs to already know the target classes. In my case, I want to group patient according to their similarities but I do not have a known target class where a patient can belong to according to his diseases.

Collaborative filtering is a well-known technique used in marketing, and more specifically, in recommendation systems for making recommendations according to users or items' similarities. The technique of collaborative filtering consists of defining similar users based on their similar taste and recommending to them what similar users like. Furthermore, it makes it possible to take a dataset of users' data and compare it with other users and define the similarity between users [16].

The author(s) of [4] describe a collaborative filtering method that has been applied in the medical area but with some small challenges to solve. According to [4], an analogy of the use of collaborative filtering can be drawn from marketing area to the medical area. There were of course some challenges to solve, such as the rating or the similar taste in different items.

When you talk about diseases, you cannot rate them or like them, you can just define that a patient has been diagnosed or not with a specific disease.

Based on [21] it is possible to use the ICD-10 code defined from WHO in order to define the diseases diagnosed for a specific patient. For each patient, it is possible to have the information about his diseases as a binary vector, 1 has been diagnosed with a specific disease and 0 has been not diagnosed with the specific disease [21]. It seems that a collaborative filtering technique can be applied in order to find similarities between the vectors presented in the above table. Inspired and based on the research made by [4] and [21] and I will use the collaborative filtering technique in order to group patients according to their similarities. The section below, describes the type of algorithm chosen to perform the collaborative filtering method.

5.3.1 Memory based vs Model based

The Collaborative filtering technique consists of two categories of algorithms i.e. memory based and model based.

Memory based algorithms make predictions using the entire dataset, by calculating a weight of similarities between the active user and all the other users in the dataset. The prediction for the active user is determined by the average weight of other users' opinions [30]. Furthermore, this type of algorithms maintains a matrix of all users' preferences for all items, so they need to perform some computation across the entire matrix. They work rationally well and can easy be updated. However, memory based algorithms can be expensive in a computational manner. This is because they grow on size and as such their response time could be longer [15].

Model based algorithms make predictions that are generated based on a model of user preferences that is pre-constructed in the dataset [30]. According to [15], the first step is to construct a descriptive model of users and then to compile the user preferences to it and then generate recommendation(s) from that model.

Model based algorithms need less memory than the memory based but adding new data to the model requires full recompilation. In addition, the model constructed, beyond the

prediction, can also derive value by exposing certain correlations between the data. Moreover, in contrast to memory based, the model based algorithms are faster and more scalable. On the other hand, the descriptive model is difficult to build [30].

Table 11 summarizes the advantages and disadvantages of each category in order to better understand which type fits better my problem.

Table 11 Comparing Memory Based and Model Based CF methods

	Advantages	Disadvantages
Memory based	• Easy updated • Simple to use • Good results	• All the dataset must be reused
Model based	• Faster • More scalable	• Difficulties while building the model • Problem while introducing new data

For the purposes of this project a memory based collaborative technique was selected based on the results from Table 11. Even more, research [15] claims that the use of memory based technique it is often more suitable in memory dataset which changes very frequently. The dataset selected for this project changes often. Every time a person visits a hospital changes are made to his medical record. Specifically, the selected algorithm is a user based collaborative filtering, since we are trying to find patients similarities according to their diagnosed diseases. After selecting the type of algorithm to use i.e. a **user based** and **memory based,** the next step is to find the method that can measure the patient similarities according to their diagnosed diseases. There are different methods used to define users' similarities such as Euclidean Distance Score, Pearson correlation and Jaccard similarities. The section below presents and describes the selected method to measure patient similarities.

5.3.2 Measuring similarity method

Based on research made by [15], the similarity method used depends on the dataset that one is working and on the type of application one is building. The dataset that I am working with in this project has some differences from datasets used in collaborative filtering application in marketing. The biggest difference of the given dataset, as explained before in section 3.3.1, is

that there are no user ratings or user taste. This made more challenging to define the method used for finding patient similarities. Hence, the applied binary values 0 and 1 make possible to compare two patients.

According to [15], the most common methods used in collaborative filtering in order to find similarities between users are Euclidean Distance Score, Pearson correlation and Cosine similarity. The first one, the Euclidean Distance Score is a simple method used for finding similarities between two persons according to the ranking that these two persons have given to common items. The formula used to measure the distance between these two users is presented below.

$$d(p,q) = \sqrt{\sum_{i=0}^{n}(p_i + q_i)^2}$$

The smaller it is the Euclidean distance between these users, the more similar they are [15]. As explained before, we are not dealing with user rating in the selecting dataset. In place of ratings we have binary vectors 0 and 1 values.

Another common method used to find similarities is the Pearson correlation method which measures exactly how two set of data or vectors fits on a straight line by showing how similar two vectors are. This method is much more complicated than the Euclidean Distance score but it has an advantage because it is unaffected from the amount of users' ratings [15].

The Pearson correlation is a value between 1 and –1, where 1 indicates that the variables are perfectly correlated, 0 indicates no correlation, and –1 means they are inversely correlated. In addition, to use the mathematical type of this method during the development process some changes need to be made in order to modify the range value results from -1 and 1, to 0 and 1 making the process more difficult [15].

The cosine similarity is also another method used to measure similarities between two vectors [15].

According to [15], the cosine similarity method makes it possible to measure the cosine of the angle between the two vectors selected. If the angle measured between them is 90, then

the similarity between these two vectors is 0. Only if the vectors point in the same direction, then they have a similarity 1 between them.

The Cosine Similarity is comparable to the Jaccard coefficient similarity method due to the fact that both methods compare one type of attribute distributed among all data [31].
All the above methods could be applied to the chosen dataset, however research [27] has shown that if the similarity is binary, then it will be more valuable to look at these vectors and try to match the same attribute type between them. Therefore, in this case, Jaccard similarity or coefficient method was selected as a solution while working with binary vectors to measure the similarity between them. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

According to [31], the Jaccard approach looks at the two vectors sets and finds the matches where both values are equal to 1. The result of this approach will give us how many 1 to 1 matches there are in comparison to the total number of the vectors compared [31]. In my case, I am comparing patients that are presented as vectors with binary values 0 (has not been diagnosed with a disease) and 1 (has been diagnosed) with a specific disease. When comparing two vectors of binary values 0 and 1, I am interested only in the diseases they have in common so only in the value 1 of the vectors in order to measure their similarity. It looks like the Jaccard Coefficient method fits better to the set of data I am comparing.

After defining the collaborative filtering technique, the type of algorithm and the method used to find the similarity between patients, the next step was to use the information of similar patient to predict and investigate the high risk of a specific patient who is similar to this group. The following section presents and describes the selected method used to further analyse the group of similar patients discovered using the collaborative filtering technique and the reasons for this decision.

5.4 Supervised & Unsupervised Machine Learning

Studies [12] describe two major techniques for machine learning categorized as supervised and unsupervised machine learning. Most of the techniques described in section 3.3.1 were supervised learning techniques and were based in a training dataset in order to achieve their predicting goal.

According to the literature [12] in order to select the right machine learning technique, supervised or unsupervised, the first thing is to get to know the data you are working with. Based on [12], I understood that the process of knowing the dataset was very important because it allowed me to understand what type of features the dataset has, nominal or continuous, how to process the data, to classify or to just group them without having any target value, are there any features missing from the dataset? Only after understanding and going through this process was it possible for me to select the appropriate machine learning technique to apply in order to derive value.

Based on the above suggestions and during the dataset exploration, it was clear that the type of the selected data were both nominal and continuous data features. The first step was to apply on the dataset the collaborative filtering method that allowed grouping the patients based on similarities according to their diagnosed diseases. The second step was to elaborate more on the data presented on the group of similar patients. In order to do that, I investigated the field of free medical text, which contains information about the patient medical record and is stored as unstructured data. The reason to investigate this field was to derive hidden value from it and try to define for a specific patient the high risk of developing a specific disease.

Following [12] suggestions, I decided that in order to analyse and investigate this type of information an unsupervised learning machine method is needed. As [12] explains, when there is no target value for the given data, a reasonable technique to apply is unsupervised learning. In my case, I do not know what the target value is for the selected patient that I am trying to define the high risk of being affected by a disease. Therefore by using an unsupervised learning technique, the machine will tell us what it knows about these data [12].

Clustering is well known unsupervised learning technique, which forms automated clusters of similar things. The difference between clustering and classification is that in classification one knows what he is looking for. It is already known that the instance of data will fall into one of the predefined classes. In clustering one does not know what he is looking for, because there are no predefined classes, but the results are almost the same as in classification [12].

Clustering technique also is widely used to discover patterns and derive value from latent concepts in a collection of unstructured text documents. According to [32]

> *"A starting point for applying clustering algorithms to unstructured text data is to create a vector space model for text data. The basic idea is (a) to extract unique content-bearing words from the set of documents and treat these words as features and (b) to represent each document as a vector of certain weighted word frequencies in this feature space."* [32].

The cluster analysis in relation with the family relation and the patient similarities also make possible for the physician to investigate the reasons of the suggestion provided by the system. Having defined the type of the unsupervised machine learning technique, to further process the available data, the following section presents some of the potential clustering methods that can be used and applied to the specific dataset.

5.4.1 Clustering Method

In the finalized dataset there is a field named Medical Records Text, which contains large amount of text that requires analysis in order to discover and extract value form it. It is important to find a method that will allow physicians discover hidden relationship between similar patients analyzing unstructured text presented in this field.

As explained in the above section the first starting point for applying a clustering algorithm to large amount of unstructured data text is to create the vector space. Research [32] has shown that the most important part of the unstructured text analysis is to discover semantic meaning by exploring the hidden relationships between the words and documents. Latent Semantic Indexing (LSI) is a machine learning technique that allows us process a large group of text documents and analyse in order to find the underlying meaning or concepts of those

documents. LSI during the analysis of the text maps both words and documents into a "concept" space and does the comparison of these documents in this space. [33]. Therefore, this method can be applied to the finalized dataset in order to analyse the free medical text field presented there. Every medical record text can be seen as a document and is mapped with the use of LSI into the concept space and compared with other medical records in order to explore and identify different hidden relationships that the patient can have between them.

LSI is not a traditional natural processing language or artificial intelligence program; the main attribute of LSI is the use of singular value decomposition (SVD) which makes possible to filter out some of the noise introduced into the vector space relationship and also to find the smallest set of vector space that spans over the documents [33].

The technique of latent semantic indexing consists of the following steps [33]:

- A matrix (m, n) is constructed in order to map the terms and the text documents as vectors. The matrix created represents the information how many times a term is presented to each document. Every row of the matrix presents the terms that exists into all documents and every column is the document itself.
- According to the appearance of each term into the text documents a weight is applied to the text documents array and the terms array.
- SVD is an automatic mathematical technique used to reduce this matrix as a product of three matrices U, S and V. The diagonal matrix S created from the decomposition where the entries outside the main diagonal are all zero and only on diagonal contains all the singular values that are non-negative and non-zero.
- Since the matrix created can be very large and computationally difficult to be processed, the matrix dimensionality has to be reduced by selecting the *k* largest values on this diagonal, and the corresponding number of columns in the other two matrices. In this way, the beginning vector space dimensionality is reduced by *k*-dimensional vectors.
- The relation that text documents can have between them can be calculated based on how close they are using a cosine, or Euclidian measure.

Although LSI was and is used with success for different types of applications, there are some disadvantages. LSI requires high computational processing power and fast memory. However,

during the last years as the technology has moved on and the cost of RAM has decreased [33]. A potential solution to those disadvantages could be a distributed system solution where the information is shared between a network of computers that can share the amount of the computational processing power and memory. Every computer presented in the network has its own memory, making easier to share the load of large information.

5.5 Data Storage

The technology selected to store big data must be flexible and easy to change in order to meet the growth scale of the data stored in such a system. Furthermore, the system must be secure especially, in the medical area where the data are sensitive because it contains personal information. Latency is another challenge to be faced since in many cases the big data employs a real time component [3].

In order to store big data and process them some of the technologies were described in section 3.2 and could be applied to the selected dataset.

During the investigation of the dataset, it was decided to work with the existing storage system. The dataset selected for the project was stored in a RDMS, precisely was a MS-SQL database.

In section 3.4.2 a description about a company that provides hospitals and different medical institution with EMR solutions was discussed and the storage system provided by this company was an MS-SQL database. The decision to work with the existing storage system seem to have practical dimensions when it comes to applying in real world since different medical institutions are using the same storage system.

Other technology could be used as well but as known in medical area there to many stakeholders involved and the cost of changing the existing storage infrastructure could be enormous.

However, taking in consideration the growth rate of the EMR dataset that is doubled every 5 years [23], a better investigation in order to define the appropriate technology to store and manage the type of information is needed. This investigation could be part of the future work.

6. Problem Solution

The big data analytics as explained before has been used in early years in medical area, and now more than ever, scientist are trying to apply new and existing techniques in the medical area [2]. However, they are facing different challenges and difficulties as described in problem formulation section 1.2.2 and discovered during the semi-structured interview with the doctor. The following section presents a solution that is based on the latest research outcomes and state of the art techniques and technologies in order to address some of these problems. Moreover, I will present a procedure defined during this project work, which will make possible to elaborate and extract value from the selected medical dataset.

6.1 Solution Logic

According to [12], in order to develop a system that will analyse large amount of data and derive value from them, some steps should be followed. These are collecting, pre-processing, analyzing the data, selecting the technique algorithm, testing the technique and using it. Those steps will help me to face and answer the research questions and address the different problems discovered during this work.

One of the most common problems, described in the problem formulation and pointed out from the literature [9] as well, is the complexity of medical data, which makes time consuming for the physicians to deal with them. To address the above problem according to [13] the data need to be transformed into something more accessible, more query able and relatable. The first step, as described in section 5.2 [Table 4], was pre-processing of the selected dataset according to MapReduce concept, which allowed to define the important data features that will be used for the further development process. During this step it was also possible to investigate the inter-family relations between patients according to patients' relations' degree. This step enabled me to group patients' information according to their family relations and showing one family members after the other. In this way I achieved to build the family tree for each patient.

The second step followed was to explore and define the inter-family relations based on the similarities between patients. In order to address the first research question I need to

analyse and determine the similarities that a patient has with ancestor, relatives and then with other similar patients as well. As explained in section 5.1, it is important to the physician to analyse the family related persons in order to find similarities between them. In addition, it will help the physician to make better decisions during the diagnosis process.

Furthermore, patient similarity is important and can be used in order to predict specific diseases that can affect him [21]. The model used to define patient similarities was based on the recent research, and patients who have been diagnosed with the same disease had similar symptoms [4].

According to [21] and as explained in section 5.3, it is possible to map every patients' diseases that they has been diagnosed. This was achieved using the ICD- 10 code that is almost presented at every patient medical record. Based on the 5 selected diseases and the ICD – 10 codes, it could be possible to present all the patients and their diagnosed diseases. Table 12 shows a theoretical example of how this information could be presented to a physician.

Table 12. Presentation of the patients ID and their diagnosed diseases

Patient ID	**Disease Diagnosed**
xxxxxxx 53	Disease2, Disease 4
xxxxxxx 91	Disease 2, Disease 3, Disease 4
xxxxxxx 92	Disease 2, Disease 3, Disease 4

The information presented in the above table, made possible for every patient and the 5 selected diseases to define a patient vector. In this vector, the value 1 defines the diseases that had already affected the patient and the values 0 defines the patient has not been diagnosed with the specific disease.

Table 13 presents theoretically how the mapping of patients' diseases could be represented to a physician.

Table 13 Mapping patients and diseases

Patient ID	Disease 1	Disease 2	Disease 3	Disease 4	Disease 5
xxxxxxx 91	0	1	1	1	0
xxxxxxx 92	0	1	1	1	0
xxxxxxx 53	0	1	0	1	0

The above mapped information allowed me to define patient similarities using a collaborative filtering technique. Specifically, I used the Jaccard coefficient to define patient similarities between patients according to their diseases. The selection of this technique allowed me to address the problem that physicians are facing when they are trying to investigate patient relations and similarities for a specific disease. Furthermore, this information uses the intra-family and inter-family, which are based on patients' diagnosed diseases.

The collaborative filtering allows physicians group and investigate only patient that are similar to the selected patient. The grouped similar patients are presented to physicians according to their family relations and order by the family relation type field. Table 14 presents a theoretical example of how information can be presented.

Table 14 Example of similar patient data

Patient ID	Diseases	Medical Record Text	Family Relation	Similarity Score
xxxxxx xx91	D2, D3, D4	Yes	2	0.67
xxxxxx xx92	D2, D3, D4	Yes	13	0.6
xxxxxx xx53	D2, D4	Yes	Null	0.4

**D=Diseases*

In this way, persons with a family relation will be presented first, since according to the physician the family history records plays important role during the diagnosis problem. Moreover, the information above [Table 14] is presented in such a way that will allow the physician to explore firstly the intra-family similarity of the selected patient. Secondly, the physician can investigate the inter-family similarity based on how similar to other patients is the selected patient. The Jaccard similarity coefficient presented in the last column makes possible to compare the selected patient with related patients using inter and intra-family relation. The third step is to use a machine learning technique in order to predict the selected patient risk according to similarity with the patients group presented above.

The selected technique, as described and discuss in section 5.4.1, was the LSI. Specifically, this technique allowed me to investigate through latent semantic analysis the

unstructured data stored in Medical Record Text. This column contains the risk of the selected patient to develop a specific disease.

Let us assume that the selected patient has been diagnosed with disease 2, 3 and 4 as presented in [Table 13]. This means that we will investigate a cluster that contains all similar patients with disease 2, 3 and 4, because they match with the selected patient. Further, we use the LSI clustering technique to analyse the medical record text from the patients presented in this cluster. Hence, we run LSI for disease 2, 3 and 4, searching for patterns visible from the analysis of structured data [table 13].

Table 15 Investigating medical records for the selected patient and diseases

Patient	D2, unstructured text	D3, unstructured text	D4, unstructured text
xxxxxx xx91	Yes	yes	No
xxxxxx xx92	Yes	yes	Yes
xxxxxx xx53	Yes	No	Yes

Table 15 describes an example of the selected patients and their diseases that were identified in structured data, as shown in table 13. The LSI technique made possible to answer the research question: "identify patients who are at high risk of developing a specific disease such as diabetes or heart problems."

Last but not least, the visualization method used to present the above clusters allows physicians to investigate and understand the prediction made by the system. The selected patient is presented with different colour in order to be visible to the physician. The clusters formed during the cluster analysis are presented with different colours as well. The technique used for the data visualisation of the SVD results is a scatter plots. According to [34] this technique is suitable to visualize diagnostic data. The SVD makes possible to calculate the scater plots based on computation of coordinates using the equation AV=US derived from the main equation of SVD: A= USV. The scatter plot is calculated based on Pearson correlation coefficient between the documents and in my case using the medical record text.

Furthermore, the physician interacts with the cluster visualization by clicking on the selected patient. This interactive feature allow physicians understand and investigate the

reasons of being close to a specific cluster based on family relations and other patients' similarities.

With the above selected techniques and steps followed, it was feasible to accomplish most of the requirements of machine learning pointed out by [13] explained in section 3.4.1 as well.

The reason SVD is useful, is that it finds a reduced dimensional representation of the matrix that emphasizes the strongest relationships and throws away the noise, addressing one of the challenges pointed out from [13] (how to deal with the noisy data).

The presentation and visualization of every step made by the system achieves the transparency requirement of such a system. More information about the overview of the system developed is presented in Appendix 3. The procedure followed allows a physician to have a new point of view for the investigated data and reveal patterns that he could not see explicitly before. The interaction with the system allows the physician to understand the prediction made by the system when the selected patient is diagnosed. Tables 16 summarizes the problems, challenges faced throughout this research and the solution logic followed.

Table 16 Overall overview of the problems faced and the solution logic

Problem faced	Provided solution
Large amount of information and complexity of the medical dataset.	Pre-processing of the selected dataset according ti MapReduce programming model in order to define the most important data features for the development process.
Investigation on patient relation according to their family relation and patient similarities.	Collaborative filtering technique based on the Jaccard similarity method to define patient similarity according to the already diagnosed patients' diseases.
How to define the high risk of developing a specific disease for a patient according to his symptoms.	LSI technique and SVD algorithm will suggest to the physician the possible risk of developing a specific disease based on cluster visualization.
How to visualize the prediction of the system to the physician so it can be transparent.	The similar patient group clusters visualization and the two way interaction will allow the physician to investigate the reasons of this result.

The following section describes and presents the development method and the procedure followed to develop the system prototype.

6.2 Development of the prototype

In order to develop a system prototype that will address all the problems and challenges described in the previous section 6.1 [Table 16], a software development method was needed. As described in methodology section 4.4, the rational unified process (RUP) method was selected.

The reasons of this selection were based on the idea that RUP allows us to do several iterations during development and at the end of each iteration it provides us with a prototype. Having a small prototype for each iteration is important because we can test his functionality and then continue with the next iteration and build upon the previous one [28]. Figure 5 shows the process followed during the development process including the three iterations.

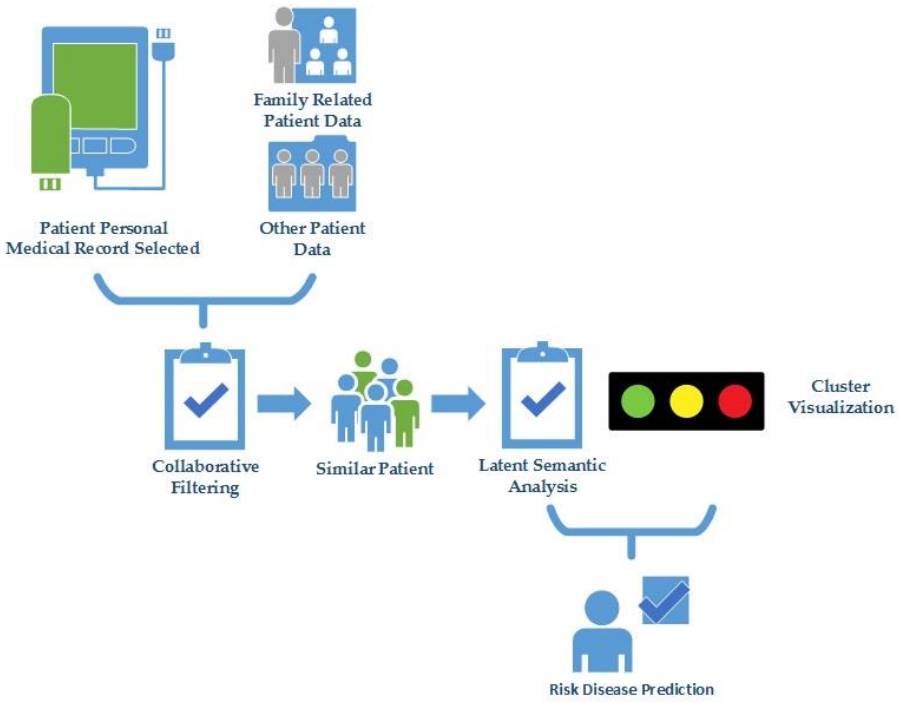


Figure 5. The development process

The first phase of the RUP method the inception phase. This made possible to delimit the scope of the project and to define the primary scenario and use cases. This helped in have a potential design of the system. The second phase was the elaboration phase. This phase allowed me define rapidly the basic practical architecture of the system under development

[28]. The third phase was about the construction and implementation of the system. The next subsections describes the three iterations in detail.

6.2.1 Phase 1: Mapping patient according to their diagnosed diseases

The first iteration consisted of 2 tasks: the selection of available patient diagnosed with the selected diseases according to the doctor's suggestion and their mapping. During this process only 5 diseases are mapped form the overall number, in order to simplify the development process. The selection of the diseases was initially made in collaboration with the doctor. Below is presented an example of the code. For every patient presented in the dictionary Patient.Diseases, checks are made. These checks look at the selected patient and evaluate if the patient is diagnosed with one of 5 selected diseases. If the condition is true then into the value 1 is given, otherwise the value remains 0.

```
if (PatientDiseases.ContainsKey(col1))
     {
     int[] patientCurrentList = PatientDiseases[col1];
            if (col2 == "e10")
                patientCurrentList[0] = 1;
            if (col2 == "e105b")
                patientCurrentList[1] = 1;
            if (col2 == "z489")
               patientCurrentList[2] = 1;
            if (col2 == "i9")
               patientCurrentList[3] = 1;
            if (col2 == "i10")
                patientCurrentList[4] = 1;
            PatientDiseases[col1] = patientCurrentList;
}
```

Figure 6 presents a screenshot from the system developed in first iteration.

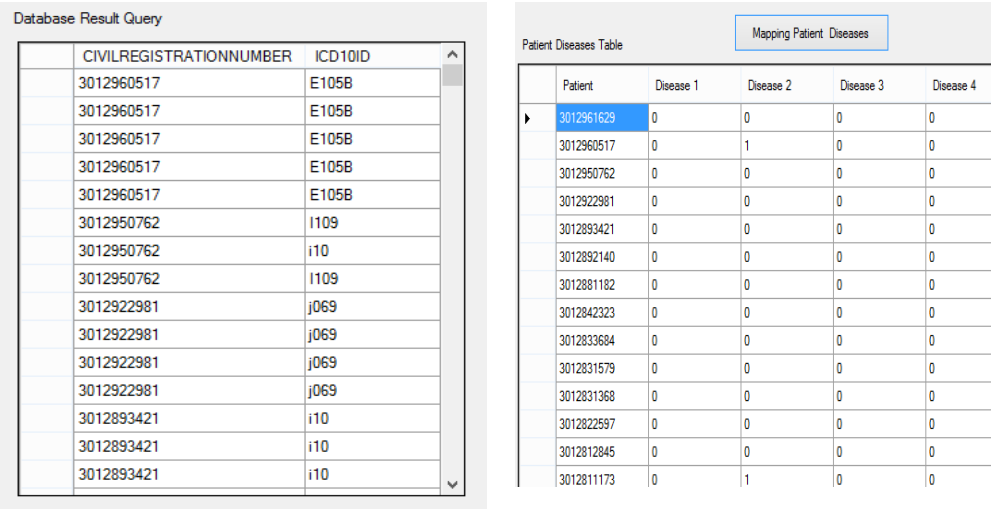


Figure 6. Presentation of the first iteration

During the testing and investigation of the first iteration, I discovered that for the selected diseases, suggested from the doctor, the matrix created was not very dense. The number of patients having two or more diagnosed diseases was small. The small number of patients and diseases will have negative impact on the size of the similar groups that will be used in the next iteration. Therefore, I went one step back on the patient query selection and changed it. The selected patients are based on a combination of the most common diseases found on the dataset and the doctor suggestions.

6.2.2 Phase 2: Collaborative filtering method implementation

During the second iteration the implementation of collaborative filtering method took place. It was based on the Jaccard Coefficient as discussed in section 5.3.2. According to the Jaccard coefficient type in order to measure the similarity between two vectors this information is needed:

J= (number of matching presences) / (number of attributes not involved in 00 matches) [31].

$$J = \frac{N_{11}}{N_{01} + N_{10} + N_{11}}$$

An example of how the Jaccard coefficient could be applied and how to measure the similarity between two vectors according to [31] follows.

$v_x = \{1,1,0,1,0,1\}$ and $v_y = \{1,1,1,0,0,1\}$

N_{01}=1, the number of times where x was 0 and y was 1. N_{10}=1, the number of times where x was 1 and y was 0. N_{00}=1, the number of times where x was 0 and y was 0. N_{11}=3, the number of times where x was 1 and y was 1. According to the Jaccard type those two vectors has a similarity 0.6 with a maximum 1. The code below describes the implementation of the Jaccard coefficient used for this project.

```
public double calculateSimilarity(int[] v1, int[] v2)
        {
            int n11=0;
            int n10=0;
            int n01=0;
            double val;
            for(int i=0;i<5;i++)
            {
                if(v1[i]== v2[i] && v1[i]==1)
                    n11++;
                if (v1[i] == 0 && v2[i]==1)
                    n01++;
                if (v1[i] == 1 && v2[i] == 0)
                    n10++;
            }

            return val=n11/(n10+n01+n11);
        }
```

The physician has to select the ID of the respective patient in order to find other patient similar to him and to proceed their information. After this stage the dat grid will show only the similar patient mapped information. The next step of the second iteration was to show to the physician only the similar patients and their data information available as presented in section 6.1 [Table 14]. The data presented in the new data grid will allow the physician to have a better understanding of those results based on the similarity score and family relation degree that the selected patient has with other patient. The Figure 7 show how this information is presented.

Similarity's Patient Group

Patient	Diseases	Medical Record Text	Up Relation	Down Relation	Family Relation Type	Similarity Score
3012950762	I109	Bellavej OG Korsebøllevej SYGEHUSE Langholtvej AMBUL...	1501632212	3012950762	10	0,6
3012950762	i10	Velbefinde. Medicin gennemgås. Centyl m KCl ændres til Ce...	1501632212	3012950762	10	0,6
3012950762	I109	Bangs OG Lindvej SYGEHUSE Delasvej AMBULATORIUM,...	1501632212	3012950762	10	0,6
3012842323	i10	Hypertension (1995)og ischæmisk hjertelidelse (1995) med sj...				0,6
3012842323	i10	BT 146/74- 147/74- 136/72. P 80. Velb.				0,6
3012842323	i109	Uændret svimmel, men har opdaget at det hjælper at gå tur. ...				0,6
3012842323	i10	Noget svimmel og ang. pect. næsten dagligtBT: 169/98, 17...				0,6
3012842323	i109	Af og til lidt svimmel og utilpas. BT: 164/82, 156/81, 147/78,...				0,6

Figure 7 Example of similar group information data

As the picture above shows all the information is available to the physician. The patient similarity score and the family relation type are important information that allows him to

realize why those patient are similar between them. Furthermore, this type information will be valuable during the cluster visualization while allows the physician to investigate for the selected patient the reasons why this patient has been placed in the specific cluster.

6.2.3 Phase 3: LSI implementation and cluster visualization

As described in section 5.4.1, the selected method suggests to the physician the high risk of a patient, who may develop a specific disease. Further the physician uses the LSI method to investigate why the selected patient is closer to a specific cluster.

Below I present and explain the implementation of the LSI algorithm used during the development process. In order to implement the LSI algorithm and visualize the clusters, I used two additional libraries that helped me create the SVD and the two dimensional graphics.

- ZedGraph is a library written in C#, for drawing 2D Line, Bar, Pie Charts, etc. [35].
- Bluebits.Matrix.Library is a .net library used to perform fast and accurate matrix operations in a .NET application [36].

Below it is presented and described the main method used during the implementation of the cluster technique used in this project. The most important classes used during this process are also described below:

```
public static void createWordList()
```

For every document in the list of similar patients, we call the method **getWordList(),** which extracts all the words from the medical records of similar patients. It inserts them into a new list (wordlist) that is sorted alphabetically in order to make the search more faster. During this process the algorithm also checks the word frequency with the help of the function **getTF()**. This counts how many times the words appear for each medical record, and also removes from the list all the words that occured less then two times.

```
public static void createWordList()
            {
             foreach (string doc in docs)
             {wordlist = getWordList(wordlist, doc);}
             wordlist.Sort(); //sort the wordlist alphabetically
             foreach (var item in sortedList)
             if (item.Value < 2) wordlist.Remove(item.Key);
             double[] queryvector;
             q = new double[wordlist.Count];
             A = new double[wordlist.Count, docs.Count - 1];
             for (int j = 0; j < docs.Count; j++)
                 {
                  queryvector = new double[wordlist.Count];
                   for (int i = 0; i < wordlist.Count; i++)
                       {//calculate Term Frequency
                   double tf = getTF(docs[j], wordlist[i]);
                   if (j == 0) //if Query term then add it to query array
                   {q[i] = tf;}
                    else //if document term then add it to document array
                    {A[i, j - 1] = tf;}
                  }
```

`private static double getTF()`

It calculate how many times a word appears in the list of medical record afer spliting the medical record into single words.

```
      private static double getTF(string document, string term)
              {
                 string[] queryTerms = Regex.Split(document, "\\s");
                  double count = 0;
                  foreach (string t in queryTerms)
                  {
                      if (t == term)
                      {
                         count++;
                      }
                  }
                  return count;
              }
```

`Private static StopWords()`

This function is important while the pre-processing of unstructured data from the medical record. The main idea is to remove from the unstructured data numerical values and stop words. Stop word are these words that does not have any meaning. Filtering these words and the numerical values will simplify the LSI process because will diminuish the noise around the other important words.

`private void CreateGraph()`

It is a method that allows us to create a two dimensional graph so we can present the clusters and see the tendency of the results in the visualization. This method is in ZedGraph[35] and I am using it to visualize the projection and the correlations scatter plots.

```
private void CreateGraph(ZedGraphControl zg1, double[] ukX, double[] ukY, double[]
vkX, double[] vkY)
        {
            GraphPane myPane = zg1.GraphPane;// get a reference to the GraphPane
            myPane.Title.Text = "Visualazation Data"; // Set the Titles
            myPane.XAxis.Title.Text = "Dimension 2";
            myPane.YAxis.Title.Text = "Dimension 3";
            int rowPos=-1;
            foreach (KeyValuePair<string, int> kvp in myIntList)
            {
                rowPos += kvp.Value;
                double[] newvkx = new double[kvp.Value];
                double[] newvky = new double[kvp.Value];
                if (rowPos < 1)
                {
                    Array.Copy(vkX, 0, newvkx, 0, kvp.Value);
                    Array.Copy(vkY, 0, newvky, 0, kvp.Value);
                LineItem curveVK = myPane.AddCurve("1st patients' data ", newvkx,
newvky, Color.White, SymbolType.Circle);
                    curveVK.IsOverrideOrdinal = true;

                    curveVK.Symbol.Fill = new Fill(Color.Red, Color.White);
                    curveVK.Symbol.Fill.Type = FillType.GradientByZ;
                    curveVK.Symbol.Fill.RangeMin = 0.0;
                    curveVK.Symbol.Fill.RangeMax = (double)(300);
                }
```

`private void Run_LSI ()`

The most important class of the LSI technique is **`Run_LS().`** After building the matrix we run a mathematic technique called the Single Value Decomposition (SVD). In other words the SVD technique makes the best possible reconstruction of the matrix with the least possible information. To achieve this, it throws out noise and emphasizes strong patterns and trends. As described in section 5.4.1 SVD decomposes the matrix of terms and text from medical records. Further, it creates three matrices U, S and V. The matrix S is always a diagonal matrix with non-negative descending values, formally known as the singular values. Each non zero value represents a feature. The matrix VS describes the relation between documents (VS's rows) and features (VS's columns).

```
private void RUN_LSA (object sender, EventArgs e)
        {
            connectToDB();
            createWordList();
            //Singular Value Decomposition
            Matrix docMatrix = new Matrix(A);
            SVD svd = new SVD(docMatrix);
            //A = U S VT
            Matrix U = svd.U;
            Matrix S = svd.S;
            Matrix V = svd.V;
            Matrix VT = svd.VH;
            int k = int.Parse(textBox3.Text);
            //Dimensionality Reduction: Computing Uk, Sk, Vk and VkT
            Matrix Uk = new Matrix(U.ToArray(), U.Rows, k);
            Matrix Sk = new Matrix(S.ToArray(), k, k);
            Matrix Vk = new Matrix(V.ToArray(), V.Rows, k);
            Matrix VkT = Vk.Transpose();
            CreateGraph(zedGraphControl1, Vk.ColVector(1), Vk.ColVector(2),
Vk.RowVector(1), Vk.RowVector(2));
            SetSize();
        }
```

Even though the decomposition is expressed in terms of V transpose we'll usually talk about V so that the features are the columns in both U and V. The matrix product US describes the relation between terms (US's rows) and the features (US's columns) [32]. The trick while using SVD relies on figuring out how many dimensions or "concepts" to use when approximating the matrix [32]. This method gives the possibility of computing U_k, S_k, V_k, and V_k^T by selecting the *K* between a ranges of numbers from 3, which is the limit. According to [32] the best potential range is K = 300.

> *"Choosing to lower K did not provide sufficient representation of the relationship among terms and documents. Choosing too large a value introduced too much noise. Several experiments showed that a value of k = 300 produced optimal results for specific tasks"* [32].

6.3 Experiments and Investigation of Results

The section below presents some tests and investigation made with the system prototype in order to test the efficiency and to observe the scalability of the system. Furthermore, investigates the cluster visualization created and the system efficiency and scalability.

6.3.1 Experiments

The computer used for this test was a Dell Studio with 4 GB Ram and an Intel Icore7 processor 1.6 GHz.

During this process different tests were made with different number of patients. Every test carried out was execute for the same patient in order to investigate the differences. Further, the created visualizations showed how the number of patients used affects the cluster.

The first test was to analyze data from 500 patients from the selected dataset. During this test the system responded fast. It took 4.5 seconds. The figure below presents a screenshot of the visualization.

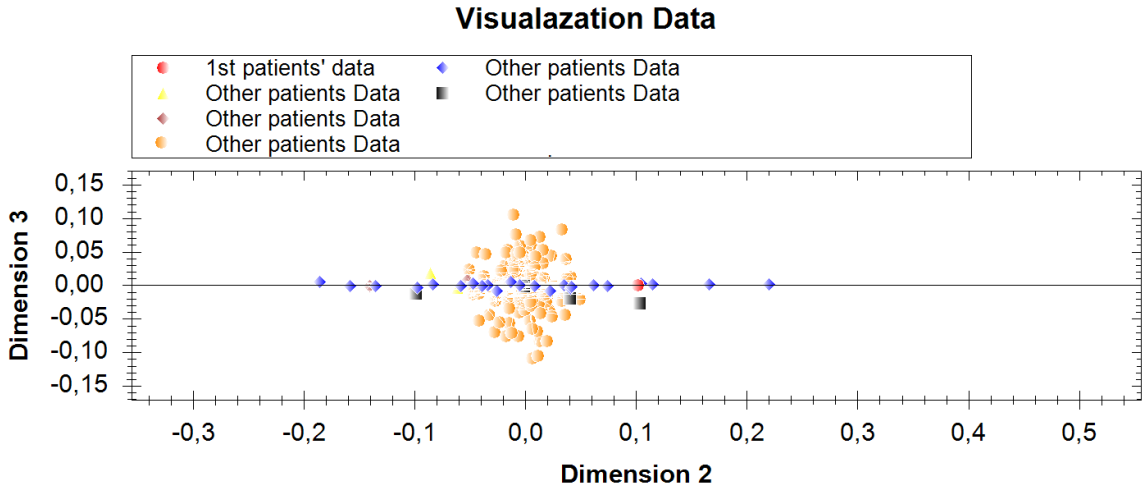


Figure 8. First experiment 500 patient data

Figure 8 shows the selected patient in red and other patients with other colors.

During the second test were analysed 1000 patient data. The response of the system was acceptable. It took 9.5 seconds. However, it was a bit slower than the first test, and this relates to the number of patients. The Figure 9 presents the results of the second test.

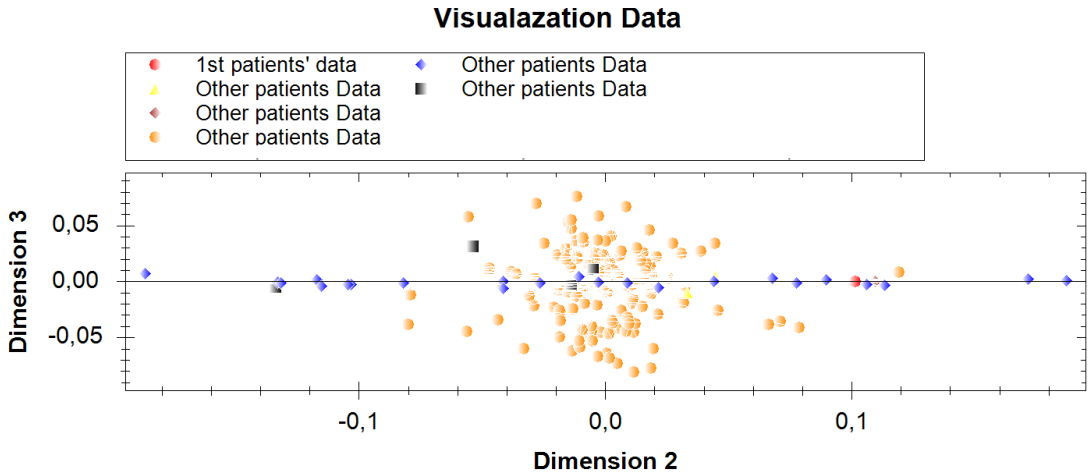


Figure 9. Second experiment 1000 patient data

Figure 9 makes clearer the position of the selected patient data and shows some extra colors than the first test because the number of patient data is greater.

The third test presented in Figure 10 shows the visualization and the results for 3000 patients. The execution time of this test took considerable time regarding to the two previous tests. It took 2 minutes and 15 seconds. This relates with the number of patients and their medical records. Further, it is related with the used personal computer.

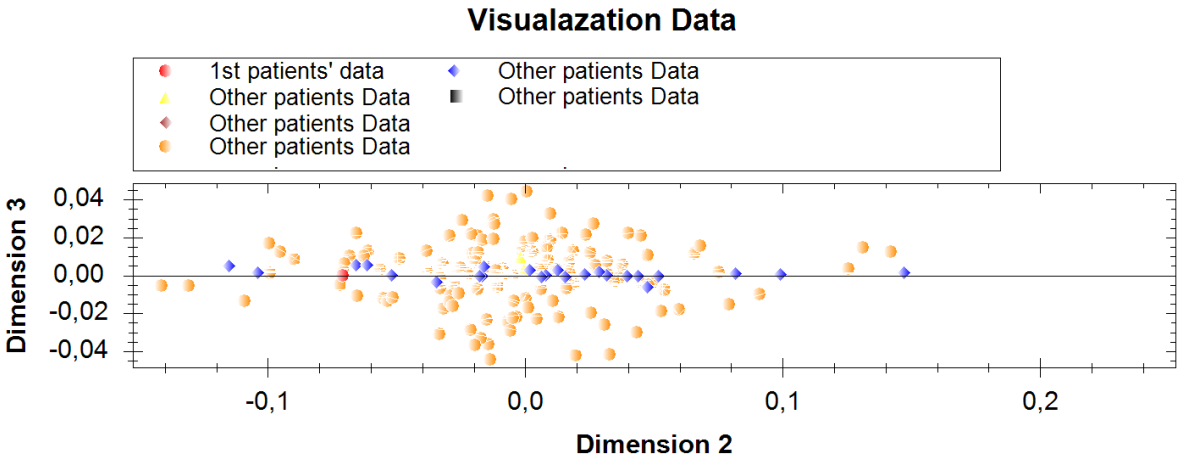


Figure 10. Third experiment 300 patient data

6.3.2 Reflections

During the test procedures and exploring of results, it can be said that visualizations were coherent during the different tests made. However, the patients' free medical text contains a large amount of data that has an effect on running time.

Furthermore, the large amount of abbreviations and misspelled words is tightly related with our results, because they have a negative impact on LSI. Medical record text also contains a lot

of numerical values such as the dosage of a specific drug in ml, dates, etc. These numbers were filtered during the pre-process cleaning of medical record text with the use of StopWords function. However, filtering this information is always debatable because we might lose information which may be useful to physicians. Considering the above issues and looking at current literature [5], it is feasible to discuss the efficiency and the scalability of the system in the future.

7. Conclusions and Future work

The aim of this work was to develop a procedure for analysing and processing large amount of medical data in order to help physicians to make better decisions during the diagnosis process and to forecast the high risk patients for a specific disease. In the following sections I present my conclusions and the outcomes of this work and suggesting future improvements that could be made.

7.1 Conclusions

The aim of this work was to develop a procedure for analysing and processing large amount of medical data in order to help physicians to make better decisions during the diagnosis process and to forecast high risk patients for developing a specific disease.

In order to accomplish the goal of this work an academic approach was followed. Primarily, I explored the problem area in order to identify the needs and to define the challenges faced by the medical domain while analysing medical data such as the patient medical records. Both primary and secondary sources were used to explore the problem domain. An interview with a physician was carried out to understand the process and the physician's needs during this process. In addition, a literature study was conducted. In the early phases, the literature study had a broad focus: identifying state-of-the-art big data techniques and technologies. In the later stages, the literature study focused more on identifying related work in the medical area carried out by academics as well as practitioners.

Secondly, an investigation of the selected dataset followed in order to get an overview of the medical data type. Moreover, a pre-process task was conducted to simplify and reduce the complex information presented into the dataset according to the literature study and physicians suggestions.

Thirdly, a detailed analysis of the problems and the challenges took place, succeeding into the identification of the most appropriate techniques that can be used to analyse and to process patient medical data in order to discover hidden patterns that can be used during the

diagnosis process. Finally, a solution was proposed in order to address the challenges and the problems faced and to answer the research question.

The solution proposed in this thesis is a procedure that combines different techniques used to pre-process, analyse, and visualize the patients' medical records in order to foresee the risk of a patient for a specific disease. More specifically, the procedure consists of the following techniques: MapReduce concept to pre-process the selected dataset, collaborative filtering to investigate the inter-family relations, latent sematic analysis of similar patients' medical record to predict the risk of a specific disease and clustering visualization method in order to visualize the LSI results.

It is in my belief that this work and the procedure followed made possible to answer the research question. The developed prototype constitutes a proof of concept: it demonstrates that actually it is possible to predict if a patient has a high risk of developing a specific disease based on patient inter and intra-family relations.

The development of this prototype was challenge which allowed me to understand and observe how close it is this system to be applicable in the real life?

During the overall process of this project all the assumptions, choices and decisions made were based on previous academic research in the medical area, the suggestions of the doctor made during the interview, as well as my point of view and understanding of the above. In the future, this work procedure can be used as a starting point for other researchers who investigate similar problems.

7.2 Future Work

The solution presented during this project requires further work to be carried out and in the future there is always space for improvement.

As explained before in section 7.1 some of the assumptions made throughout this process work are based on my understanding and my point of view. Other choices could have been made during the pre-processing, the analysis and visualisation of the dataset. However, according to [3] this depends on the perception and interpretation of the data scientist.

Since, inter and intra-family relations are very important, a better investigation could be made and give to each of these relation a weight that can be used during the LSI analysis. Making possible to analyse firstly the intra-family similarity patients and secondly the inter-family similarity and try to derive value by this comparison. Furthermore, other types of clustering techniques could be applied in order to visualise the LSI results in order to compare the new findings with the actual ones.

An important improvement that could be implemented in the future is the special case, where for a selected patient there are no medical records, since he has not been diagnosed with a specific disease. LSI makes still possible to solve this problem with the query implementation, allowing the physician to query the system prototype based on the symptoms of the selected patient. However, the above improvement needs more investigation.

The system prototype requires more improvements in the near future. A more in-depth research needs to be conducted with more physicians in order to define requirements specifications for the system making it more applicable in real life settings.

Due to scope, time and resources limitations it was not possible to implement these improvements in the current proposed solution. However, the implementation of these improvements could be a step further in the right direction to get much closer to a final solution.

References

[1] The Economist, "Data, Data Everywhere - A special report managing information," The Economist Newspaper LTD, 2010.

[2] Manyika, James ; Chui, Michael; Brown Brad; Bughin Jacques; Dobbs Richard; Roxburgh Charles; Byers Angela Hung, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, 2011.

[3] Frank Ohlhorst, Big Data Analytics: Turning Big Data into Big Money, New Jersey: John Wiley & Son Inc, 2013, pp. 9-25, 38-45, 46-50.

[4] D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási, "Time to CARE: a collaborative engine for practical disease prediction," *Data Min Knowl Disc*, vol. 20, no. 3, p. 388–415, May 2010.

[5] S. L. S. L. Kostas Pantazos, "De-identifying an EHR Database Anonymity, Correctness and Readability of the Medical Record l," MIE, 2011.

[6] Krist Wongsuphasawat, David H. Gotz, IBM T.J. Watson Research, "Outflow: Visualizing Patient Flow by Symptoms and Outcome," *IBM Research Paper*, 23 October 2011.

[7] M. J. S. J. C. v. B. E. F. S. M. C. S. a. J. A. K. Z. Afzal, "Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records," *BMC Medical Informatics and Decision Making*, vol. 13, no. 1, 2013

[8] WHO, "Classifications: Classification of Diseases (ICD)," 2013. [Online]. Available: http://www.who.int/classifications/icd/en/. [Accessed 20 October 2013].

[9] Zhuoran Wang, Anoop D. Shah, A. Rosemary Tate, Spiros Denaxas, John Shawe-Taylor, Harry Hemingway, "Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning," *PloS one*, vol. 7, no. 1, 2012.

[10] Nada Lavrac, "Selected techniques for data mining in medicine," *Intelligence in Medicine, Elsevier Science*, vol. 16, no. 1, p. 3–23, 7 July 1999.

[11] J. Preece, Y. Rogers, H. Sharp, "Interaction Design: beyond human-computer interaction," 3rd edition, John Wiley & Sons, Inc, 2011, pp. 389-432.

[12] Harrington Peter, Machine Learning in Action, Manning Publications Co, 2012, pp. 1-12, 206-210, 280-284.

[13] L. M. R. P. a. Z. Z. R. Kohavi, "Lessons and Challenges from Mining Retail E-Commerce Data," *Machine Learning*, vol. 57, no. 1-2, pp. 83-113, October 2004.

[14] Samuel Odei Danso, "An Exploration of Classification Prediction Techniques in Data Mining: The insurance domain," *M. Sc. Thesis, Bournemouth University, United Kingdom,* September 2007.

[15] T. Segaran, Programming Collective Intelligence, O'Reilly Media, August 2007, pp. 7-13, 31-55.

[16] Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman, "Mining of Massive Datasets," 2013. [Online]. Available: http://infolab.stanford.edu/~ullman/mmds.html. [Accessed 5 December 2013].

[17] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E.Gruber, "Bigtable: A Distributed Storage System for Structured Data," *ACM Transactions on Computer Systems (TOCS),* vol. 26, no. 2, 2005.

[18] D. Kaziukonis, "Context-aware recommender system based on graph theory," 2012.

[19] Riccardo Bellazzi, Blaz Zupan, "Predictive data mining in clinical medicine: Current issues," *international journal of medical informatics,* vol. 77, no. 2, p. 81–97, 17 November 2006.

[20] Igor Kononenko, "Machine Learning for Medical Diagnosis: History, State of the Art," *Artificial Intelligence in medicine,* vol. 23, no. 1, p. 89–109, August 2001.

[21] Darcy A. Davis, "Predicting individual Disease risk based on medical history," 20 August 2008.

[22] MyClinic A/S, "Medical record & electronic health record," [Online]. Available: http://www.myclinic.dk/index.php/en/myclinic/myclinic-for-almen-praksis/journal-og-ordinationer-. [Accessed 27 November 2013].

[23] IBM, "IBM Watson at work," [Online]. Available: http://www-03.ibm.com/innovation/us/watson/watson_in_healthcare.shtml. [Accessed 16 December 2013].

[24] Thomas Giles, Randall Wilcox, IBM Industry Solutions - Healthcare, "IBM Watson and Medical Records Text Analytics - HIMSS Presentation," IBM Corporation, 2011.

[25] M. B. Allan Møller, "Semantic Diagnosis System," Aalborg University of Copenhagen, Copenhagen, 2013.

[26] Michael Kleber, "The MapReduce Paradigm, Google, Inc.," Google Inc, 2008.

[27] Robert A. Hanneman, Mark Riddle, "Measures of similarity and structural equivalence," 2009. [Online]. Available: http://faculty.ucr.edu/~hanneman/nettext/C13_%20Structural_Equivalence.html. [Accessed 14 November 2013].

[28] Rational Software White Paper, "Rational Unified Process: Best Practices for Software Development Teams," IBM, White Paper 2001.

[29] ICPC-DAK-E, "De praktiserende lægers og regionernes fælles enhed for kvalitetsudvikling," [Online]. Available: http://www.dak-e.dk/flx/en/general-practice/icpc/. [Accessed 19 December 2013].

[30] E. H. S. L. a. C. L. G. D. M. Pennock, "Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach," *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence,* p. 473–480, 2000.

[31] Susan Portugal, "Data Mining Lecture - Data Mining - Assignment 3," 18 September 2008. [Online]. Available: http://csucidatamining.weebly.com/assign-3.html. [Accessed 30 October 2013].

[32] D. S. M. INDERJIT S. DHILLON, "Concept Decompositions for Large Sparse Text Data Using Clustering," *Machine learning,,* vol. 42, no. 1-2, p. 143–175, 2001.

[33] R. B. Bradford, "An empirical study of required dimensionality for large-scale latent semantic indexing application," *Proceedings of the 17th ACM conference on Information and knowledge management,* p. 153–162, 2008.

[34] A. R. L. M. R. Michael E. Wall, "Singular value decomposition and principal component analysis, A practical approach to microarray data analysis," *M. Wall,* p. 91–109, 2003.

[35] Sourge Forge, ZedGraph, "ZedGraph is a class library, user control, and web control for .net, written in C#, for drawing 2D Line, Bar, and Pie Charts. It features full, detailed customization capabilities, but most options have defaults for ease of use," [Online]. Available: http://sourceforge.net/projects/zedgraph/. [Accessed 16 December 2013].

[36] BluBit, "The Bluebit .NET Matrix Library (NML™) provides classes for object-oriented linear algebra in the .NET platform. NML™ can easily be used within the .NET framework and is an inexpensive solution to your object oriented linear algebra needs" [Online]. Available: http://www.bluebit.gr/NET/. [Accessed 17 December 2013].

Appendix 1

Results Messages

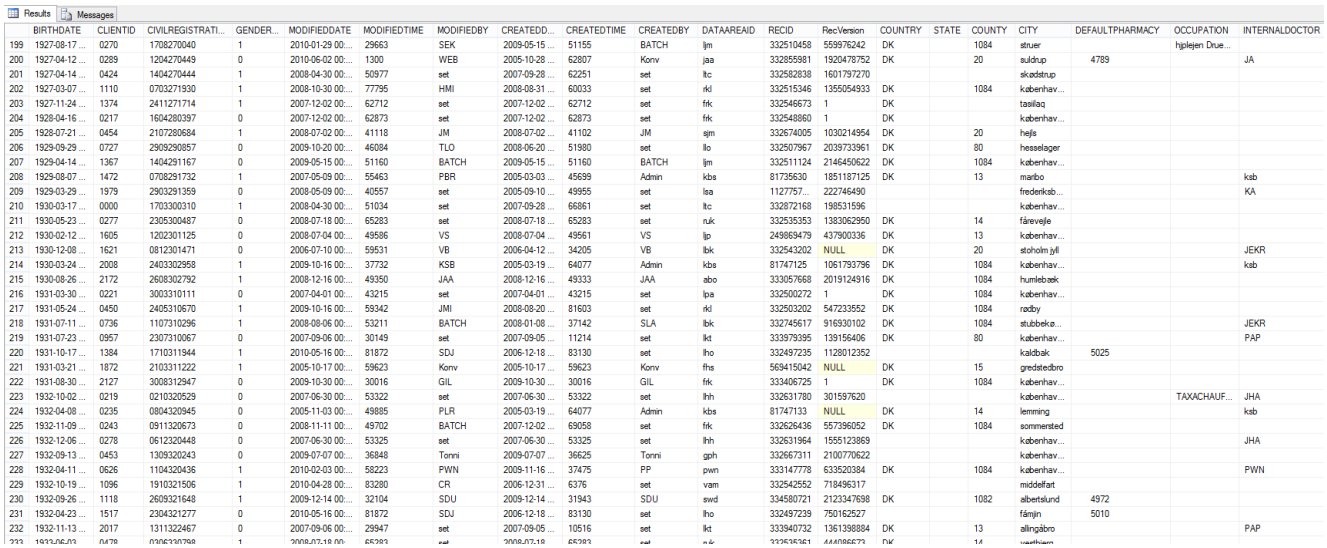
	CIVILREGISTRATIONNUMBER	MEDICALRECORDSUMMARY	ICD10ID	ICPCID	MEDICALRECORDID	CHRONICALLYDIAGNOSES	RECORDCATEGORYID	TREATMENTFLOWVERSIONID	DATAAREAID	RECVERSION	RECID
3...	2704462912	Hypertensionskontrol	i109	K86	00030195	0	Blodtryk	1.0	lvg	1	33646...
3...	0310672955	Basis			00030195	0	Alm. journal	1.0	ruk	2012783925	33286...
3...	2501241722	Vers. 2004	i109	K86	00030195	0	Blodtryk	1.0	vam	1587197853	33544...
3...	2603452944	Basis			00030196	0	Alm. journal	1.0	frk	2074523856	33272...
3...	1402106145	Basis			00030196	0	Alm. journal	1.0	lvg	1	33646...
3...	2010680711	Basis			00030196	0	Alm. journal	1.0	ruk	1	33286...
3...	1804391827	2003 udgave	E105B	T90	00030196	0	Diabetes journal	1.0	vam	1	33544...
3...	1906400757	Basis			00030197	0	Alm. journal	1.0	frk	302614031	33272...
3...	0905841838	Basis			00030197	0	Alm. journal	1.0	lvg	1	33646...
3...	1903682173	Basis			00030197	0	Alm. journal	1.0	ruk	1	33286...
3...	1210662706				00030197	0	D+R	1	vam	744305085	33544...
3...	2506722113	Basis			00030198	0	Alm. journal	1.0	frk	116715726	33272...
3...	0402821412	Svangrekontrol	Z349	W78	00030198	0	Svangrejournal	1.0	lvg	1044786402	33646...
3...	0606700326	Basis			00030198	0	Alm. journal	1.0	ruk	1	33286...
3...	1608621298	2005	z124	A98	00030198	0	Smear	1	vam	534854490	33544...
3...	2012862974	Basis			00030199	0	Alm. journal	1.0	frk	633274106	33272...
3...	1104783358	Svangrekontrol	Z349	W78	00030199	0	Svangrejournal	1.0	lvg	1181983198	33646...
3...	1106700160	Basis			00030199	0	Alm. journal	1.0	ruk	1	33286...
3...	1607551987			A98	00030199	0	Forebyg	1	vam	1926763813	33544...
3...	2409762962	Basis			00030200	0	Alm. journal	1.0	frk	842113018	33272...
3...	0112500950	Basis			00030200	0	Alm. journal	1.0	lvg	1	33646...
3...	0107701032	Basis			00030200	0	Alm. journal	1.0	ruk	1	33286...
3...	1810661164	2005	z124	A98	00030200	0	Smear	1	vam	1283026859	33544...
3...	0208791411	Basis			00030201	0	Alm. journal	1.0	lvg	1	33646...
3...	2101640316	Basis			00030202	0	Alm. journal	1.0	frk	1419530899	33272...
3...	1506321993	Diabetes type 2		T90	00030202	0	Diabetes journal	1.0	lvg	1	33646...
3...	2412712832	Basis			00030202	0	Alm. journal	1.0	ruk	1	33286...
3...	3008622033	Konverterede notater			00030202	0	Konv. lange notater	1.0	skk	1580839802	83781...
3...	2405742214	Basis			00030203	0	Alm. journal	1.0	frk	1493860753	33272...
3...	1603721589	Basis			00030203	0	Alm. journal	1.0	ruk	1	33286...
3...	1003871417	Basis			00030203	0	Alm. journal	1.0	vam	1	33544...
3...	0109703546	Basis			00030204	0	Alm. journal	1.0	lvg	1	33646...
3...	0802722354	Basis			00030204	0	Alm. journal	1.0	ruk	1	33286...
3...	0510411598	Basis			00030204	0	Alm. journal	1.0	vam	1	33544...

Figure 11 Instance of medical data stored in Medical Record Table

Results Messages

	MEDICALRECORDID	MEDICALRECORDTEXT	ICD10ID	ICPCID	MEDICALRECORDSUMMARY	MEDICALRECORDLINEID	RECORDSUBCATEGORY	DATAAREAID	RECVERSION	RECID
9...	00029050	*** Udskrivningsepikrise *** ...	g431		Hemicrania med aura (klassisk ...	329128	Ud.Epi	skk	1	83822...
9...	00019155	jk: Øm i ve sinus frontalis i forb. med catarralia. Næ...	j069	R74	Øvre luftvejsinfektion/forkølelse	329129	Kontinuation	skk	1	83822...
9...	00020849	Strep-A:neg.Virusinf.Må ikke flyve i øjeblikket. ...	j03	R76	Akut tonsillit strep-A	329130	Kontinuation	skk	1	83822...
9...	00021956	Samsø Danmark as Den 6. november 2001 Kæ...	f03	P70	Senil demens/Alzheimer o.p	329131	Am.Epi	skk	1	83822...
9...	00017389	jk: Postprandial køren og rumlen i maven. Lette syr...	r10	D01	Abdominale smerter/turevise	329132	Kontinuation	skk	1	83822...
9...	00022867	08-11-01 : trk:Fortsat på 14 dag,svingende febril,h...	j069	R74	Øvre luftvejsinfektion/forkølelse	329133	Kontinuation	skk	1	83822...
9...	00022036	Fluvac 01. sr	z269	A44	Vaccination	329134	Kontinuation	skk	1	83822...
9...	00022036	Bt 150/80. sr	i10	K86	Ukompliceret hypertension	329135	Kontinuation	skk	1	83822...
9...	00048450	19 Justine 96 Engerix 3. 20 MAR 96 Hæmatom i...			Kontinuation	329136	Kontinuation	skk	1	83822...
9...	00017728	Fluvac 01. sr	z269	A44	Vaccination	329137	Kontinuation	skk	1	83822...
9...	00024045	08-11-01 : trk:Tidl. prepulcid med effekt på maveg...	r10	D01	Abdominale smerter/turevise	329138	Kontinuation	skk	1	83822...
9...	00024925	08-11-01 : trk:5 uger i fin trivsel,fuldt ammebarn.	z001	A49	An forebyggende procedure [B...	329139	Kontinuation	skk	1	83822...
9...	00021619	08-11-01 : trk:hævet glandel bag ve.øre,i aftagend...	j069	R74	Øvre luftvejsinfektion/forkølelse	329140	Kontinuation	skk	1	83822...
9...	00019649	Vorte frys + lapis. sr	b07	S03	Virale vorter	329141	Kontinuation	skk	1	83822...
9...	00017728	jk: Bliver ved med at hoste. St.p: I.a. Er flere gang...	r05	R05	Hoste	329142	Kontinuation	skk	1	83822...
9...	00018880	bh: haft nogle dage med svien v. vandladning. U-s...	n30	U71	Cystit/anden urinvejsinfektion	329143	Kontinuation	skk	1	83822...
9...	00022996	08-11-01 : trk:opringning til pt, bedes bestille ny tid ...			Kons./Besøg	329144	Kontinuation	skk	1	83822...
9...	00017616	08-11-01 : trk:Statusattest til Mårupvej komm.			Kons./Besøg	329145	Kontinuation	skk	1	83822...
9...	00022788	Tandlæge ringer at der igen skal trækkes tænder ...			TK/TR	329146	Kontinuation	skk	1	83822...
9...	00017229	15 FEB 96 Fra gyn. Inge Saqib-Rasmussen :Vag. ...			Kontinuation	329147	Kontinuation	skk	1	83822...
9...	00048461	1 FEB 96 Fra kir. Finn Midtgaard Sørensen :M704...			Kontinuation	329148	Kontinuation	skk	1	83822...
9...	00018373	jk: Vag cyt tages om. Der fjernes subcutant fibrom i...	d239	S79	Anden godartet svulst i hud	329149	Kontinuation	skk	1	83822...
9...	00017549	08-11-01 : trk:Ikke indikation for rp. på proteinpulv...	z269	A44	Vaccination 2 ml B-combin	329150	Kontinuation	skk	1	83822...
9...	00019083	Der laves priktest.Kun reaktion på + ellers intet.De...	t784	A12	Allergi/allergisk reaktion Lisbet ...	329151	Kontinuation	skk	1	83822...
9...	00023927	jk:HIV neg.	z029	A62	Administrativ procedure [ATTE...	329152	Kontinuation	skk	1	83822...
9...	00022951	08-11-01 : trk:Kørekortsattest uden probl.vou 6/6,...	z029	A62	Administrativ procedure [ATTE...	329153	Kontinuation	skk	1	83822...
9...	00018579	2 små vorterester frys + lapis. ...	b07	S03	Vorter	329154	Kontinuation	skk	1	83822...
9...	00018195	jk: Amtskomm. vag. cyt. Vag.-slimhinder let atrofisk...	z124	X37	[CERVIX CYTOLOGI]	329155	Kontinuation	skk	1	83822...
9...	00024487	bh: Har fået 2 ugers ferie fra arbejde pga tilt. stress...	z299	A45	[SUNDHEDSOPLYSNING]	329156	Kontinuation	skk	1	83822...
9...	00019083	jk: Der er klart anstrengelsesudløst astma, med rev...	t784	A12	Allergi/allergisk reaktion Lisbet	329157	Kontinuation	skk	1	83822...
9...	00023057	08-11-01 : trk:ingen anfald siden sidst,velbehandle...	j45	R96	Astma	329158	Kontinuation	skk	1	83822...
9...	00020344	Lomudal ØJENDRÅBER 20 MG/ML ML10 1x4 ...			Kontinuation	329159	Kontinuation	skk	1	83822...
1...	00022961	08-11-01 : trk:Tager sin medicin,ingen anfald siden...	j45	R96	Astma	329160	Kontinuation	skk	1	83822...
1...	00020073	jk: Anfaldsvis svh. BT 105/70. Hb 10.0. Aterom i n...	r42	N17	Svimmelhed/usik føl ex H82	329161	Kontinuation	skk	1	83822...

Figure 12. Instance of EMR patient information, Structured and Unstructured



	BIRTHDATE	CLIENTID	CIVILREGISTRATI...	GENDER...	MODIFIEDDATE	MODIFIEDTIME	MODIFIEDBY	CREATEDD...	CREATEDTIME	CREATEDBY	DATAAREAID	RECID	RecVersion	COUNTRY	STATE	COUNTY	CITY	DEFAULTPHARMACY	OCCUPATION	INTERNALDOCTOR
199	1927-08-17 ...	0270	1708270040	1	2010-01-29 00:...	29663	SEK	2009-05-15 ...	51155	BATCH	ljm	332510458	559976242	DK		1084	struer		hjplejen Drue...	
200	1927-04-12 ...	0289	1204270449	0	2010-06-02 00:...	1300	WEB	2005-10-28 ...	62807	Konv	jaa	332855981	1920478752	DK		20	suldrup	4789		JA
201	1927-04-14 ...	0424	1404270444	1	2008-04-30 00:...	50977	set	2007-09-28 ...	62251	set	ltc	332582838	1601797270				skødstrup			
202	1927-03-07 ...	1110	0703271930	1	2008-10-30 00:...	77795	HMI	2008-08-31 ...	60033	set	rkl	332515346	1355054933	DK		1084	københav...			
203	1927-11-24 ...	1374	2411271714	1	2007-12-02 00:...	62712	set	2007-12-02 ...	62712	set	frk	332546673	1	DK			tasiilaq			
204	1928-04-16 ...	0217	1604280397	0	2007-12-02 00:...	62873	set	2007-12-02 ...	62873	set	frk	332548860	1	DK			københav...			
205	1928-07-21 ...	0454	2107280684	1	2008-07-02 00:...	41118	JM	2008-07-02 ...	41102	JM	sjm	332674005	1030214954	DK		20	hejls			
206	1929-09-29 ...	0727	2909290857	0	2009-10-20 00:...	46084	TLO	2008-06-20 ...	51980	set	llo	332507967	2039733961	DK		80	hesselager			
207	1929-04-14 ...	1367	1404291167	0	2009-05-15 00:...	51160	BATCH	2009-05-15 ...	51160	BATCH	ljm	332511124	2146450622	DK		1084	københav...			
208	1929-08-07 ...	1472	0708291732	1	2007-05-09 00:...	55463	PBR	2005-03-03 ...	45699	Admin	kbs	81735630	1851187125	DK		13	marbo			ksb
209	1929-03-29 ...	1979	2903291359	0	2008-05-09 00:...	40557	set	2005-09-10 ...	49955	set	lsa	1127757...	222746490				frederiksb...			KA
210	1930-03-17 ...	0000	1703300310	1	2008-04-30 00:...	51034	set	2007-09-28 ...	66861	set	ltc	332872168	198531596				københav...			
211	1930-05-23 ...	0277	2305300487	0	2008-07-18 00:...	65283	set	2008-07-18 ...	65283	set	ruk	332535353	1383062950	DK		14	fårevejle			
212	1930-02-12 ...	1605	1202301125	0	2008-07-04 00:...	49586	VS	2008-07-04 ...	49561	VS	ljp	249869479	437900336	DK		13	københav...			
213	1930-12-08 ...	1621	0812301471	0	2006-07-10 00:...	59531	VB	2006-04-12 ...	34205	VB	lbk	332543202	NULL	DK		20	stoholm jyll			JEKR
214	1930-03-24 ...	2008	2403302958	1	2009-10-16 00:...	37732	KSB	2005-03-19 ...	64077	Admin	kbs	81747125	1061793796	DK		1084	københav...			ksb
215	1930-08-26 ...	2172	2608302792	1	2008-12-16 00:...	49350	JAA	2008-12-16 ...	49333	JAA	abo	333057668	2019124916	DK		1084	humlebæk			
216	1931-03-30 ...	0221	3003310111	0	2007-04-01 00:...	43215	set	2007-04-01 ...	43215	set	lpa	332500272	1	DK		1084	københav...			
217	1931-05-24 ...	0450	2405310670	1	2009-10-16 00:...	59342	JMI	2008-08-20 ...	81603	set	rkl	332503202	547233552	DK		1084	rødby			
218	1931-07-11 ...	0736	1107310296	1	2008-08-06 00:...	53211	BATCH	2008-01-08 ...	37142	SLA	lbk	332745617	916930102	DK		1084	stubbeka...			JEKR
219	1931-07-23 ...	0957	2307310067	0	2007-09-06 00:...	30149	set	2007-09-05 ...	11214	set	lkt	333979395	139156406	DK		80	københav...			PAP
220	1931-10-17 ...	1384	1710311944	1	2010-05-16 00:...	81872	SDJ	2006-12-18 ...	83130	set	lho	332497235	1128012352				kalvbak	5025		
221	1931-03-21 ...	1872	2103311222	1	2005-10-17 00:...	59623	Konv	2005-10-17 ...	59623	Konv	fhs	569415042	NULL	DK		15	gredstedbro			
222	1931-08-30 ...	2127	3008312947	0	2009-10-30 00:...	30016	GIL	2009-10-30 ...	30016	GIL	frk	333406725	1	DK		1084	københav...			
223	1932-10-02 ...	0219	0210320529	0	2007-06-30 00:...	53322	set	2007-06-30 ...	53322	set	lhh	332631780	301597620				københav...		TAXACHAUF...	JHA
224	1932-04-08 ...	0235	0804320945	0	2005-11-03 00:...	49885	PLR	2005-03-19 ...	64077	Admin	kbs	81747133	NULL	DK		14	lemming			ksb
225	1932-11-09 ...	0243	0911320673	0	2008-11-11 00:...	49702	BATCH	2007-12-02 ...	69058	set	frk	332626436	557396052	DK		1084	sommersted			
226	1932-12-06 ...	0278	0612320448	0	2007-06-30 00:...	53325	set	2007-06-30 ...	53325	set	lhh	332631964	1555123869				københav...			JHA
227	1932-09-13 ...	0453	1309320243	0	2009-07-07 00:...	36848	Tonni	2009-07-07 ...	36625	Tonni	gph	332667311	2100770622				københav...			
228	1932-04-11 ...	0626	1104320436	1	2010-02-03 00:...	58223	PWN	2009-11-16 ...	37475	PP	pwn	333147778	633520384	DK		1084	københav...			PWN
229	1932-10-19 ...	1096	1910321506	1	2010-04-28 00:...	83280	CR	2006-12-31 ...	6376	set	vam	332542552	718496317				middelfart			
230	1932-09-26 ...	1118	2609321648	1	2009-12-14 00:...	32104	SDU	2009-12-14 ...	31943	SDU	swd	334580721	2123347698	DK		1082	albertslund	4972		
231	1932-04-23 ...	1517	2304321277	0	2010-05-16 00:...	81872	SDJ	2006-12-18 ...	83130	set	lho	332497239	750162527				fåmjn	5010		
232	1932-11-13 ...	2017	1311322467	0	2007-09-06 00:...	29947	set	2007-09-05 ...	10516	set	lkt	333940732	1361398884	DK		13	allingåbro			PAP
233	1932-06-03 ...	0478	0306330798	1	2008-07-18 00:...	65283	set	2008-07-18 ...	65283	set	ruk	332535361	444086672	DK		14	væthjem			

Figure 13 Overview of the information in Client Table

	UPRELATION	DOWNRELATION	FAMILYRELATIONTYPE	REFRECID	GENDERMALEFEMALE	DATAAREAID	RECID	RecVersion
76979	1808733539	2002771060	0	332528311	0	rkl	332528310	1031176163
76980	0203612466	0504902278	10	332676920	1	ruk	332676919	1031290452
76981	0212076301	0302933999	12	334017185	0	vam	334017184	1031410120
76982	0906902526	1506992825	12	332955929	1	lhh	332955928	1031555759
76983	0406472426	0608391013	3	333836178	1	lho	333836177	1031605188
76984	2106460809	0512400582	4	332893451	0	lhh	332893450	1031700733
76985	0308630324	1702932510	10	332504033	1	lve	332504032	1031730542
76986	0110671390	2612981672	10	333352294	1	lkt	333352293	1031739847
76987	0110671390	2612981672	10	333352294	1	lvg	333352293	1031739847
76988	0102852448	0102580214	0	333949944	1	lvg	333949943	1031814629
76989	2406961222	1107591634	2	333113663	1	llo	333113662	1031933856
76990	0907971523	2709753828	2	335315942	0	lvg	335315941	1032026686
76991	1702822785	0505861695	12	333756548	0	ljm	333756547	1032049607
76992	1407006088	2003713838	2	333901864	1	lkt	333901863	1032158175
76993	2104007253	2904971010	13	332539837	0	shh	332539836	1032284762
76994	0101682542	0907682313	3	332952538	1	lbt	332952537	1032345578
76995	1003056302	0308960860	13	334002193	1	lkc	334002192	1032403309
76996	2802046540	2804742586	2	1129153742	1	lsa	1129153741	1032467543
76997	2109066129	2309670704	20	140972853	0	kbs	140972852	1032561657
76998	2910251376	1410552111	3	333669112	1	lho	333669111	1032585365
76999	0803871067	1403571792	2	332770429	0	ruk	332770428	1032615818
77000	2908761905	0911711976	4	332949869	0	lbt	332949868	1032651762
77001	2412046303	1603663323	0	332886606	0	rkl	332886605	1032658742
77002	0407660682	1608903295	11	332813369	1	ruk	332813368	1032660316
77003	0810037555	0706663047	1	332517942	0	rac	332517941	1032834951
77004	1004772376	2508942865	11	332987156	1	lhh	332987155	1032917616
77005	0812503746	1502503643	3	333578886	1	abo	333578885	1033004214
77006	2812901605	1301620728	2	332763785	0	ruk	332763784	1033012134
77007	0603682707	0106751220	4	332734311	0	ruk	332734310	1033154975
77008	1407703444	2001046975	11	333975061	1	lkt	333975060	1033387861
77009	2807016075	2202752528	2	332780250	0	llo	332780249	1033555286
77010	0504641148	2610981757	11	332518921	1	vam	332518920	1033581101
77011	0211842398	2607532274	20	332842004	1	rkl	332842003	1033695344

Figure 14. Information presented in Family Relation Table

Appendix 2

Figure 15. An example of the tables available in the selected Database

Appendix 3

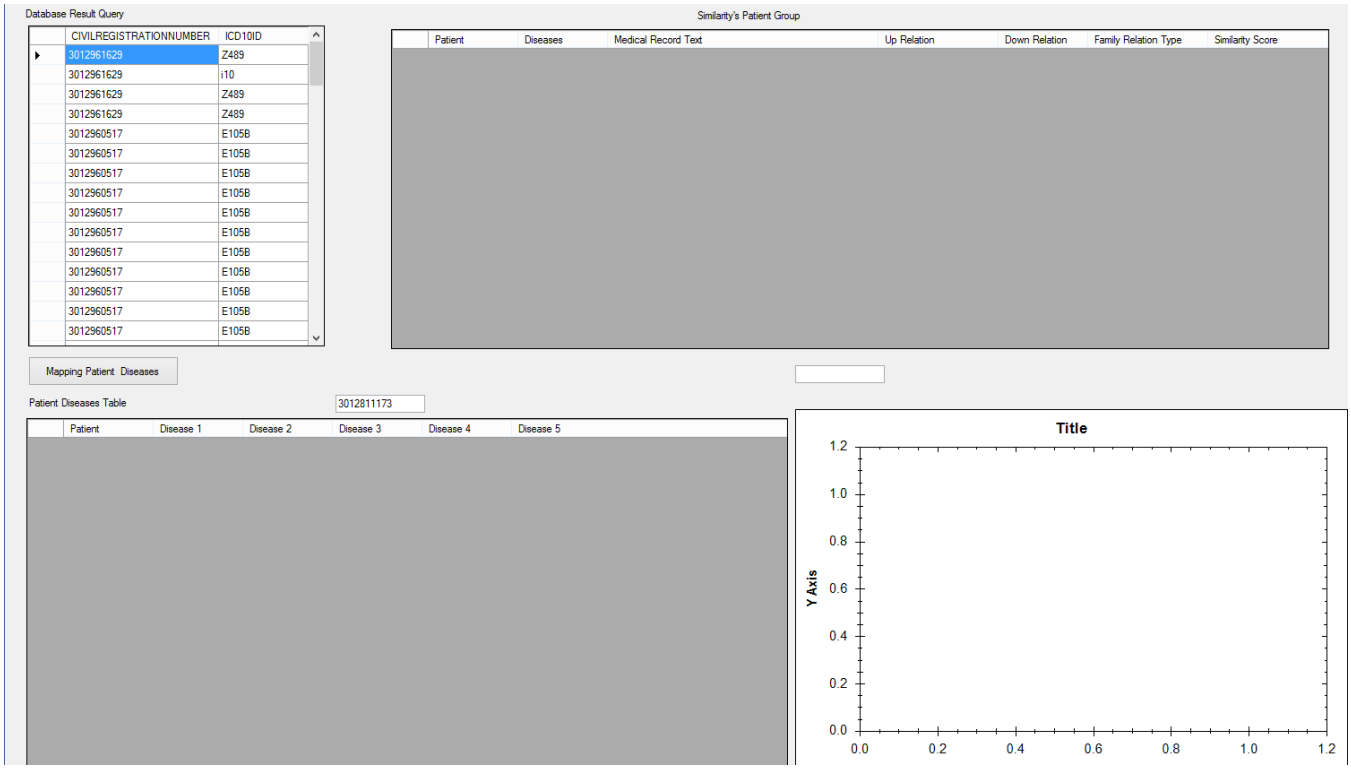


Figure 16. First step, patient and diagnosed diseases

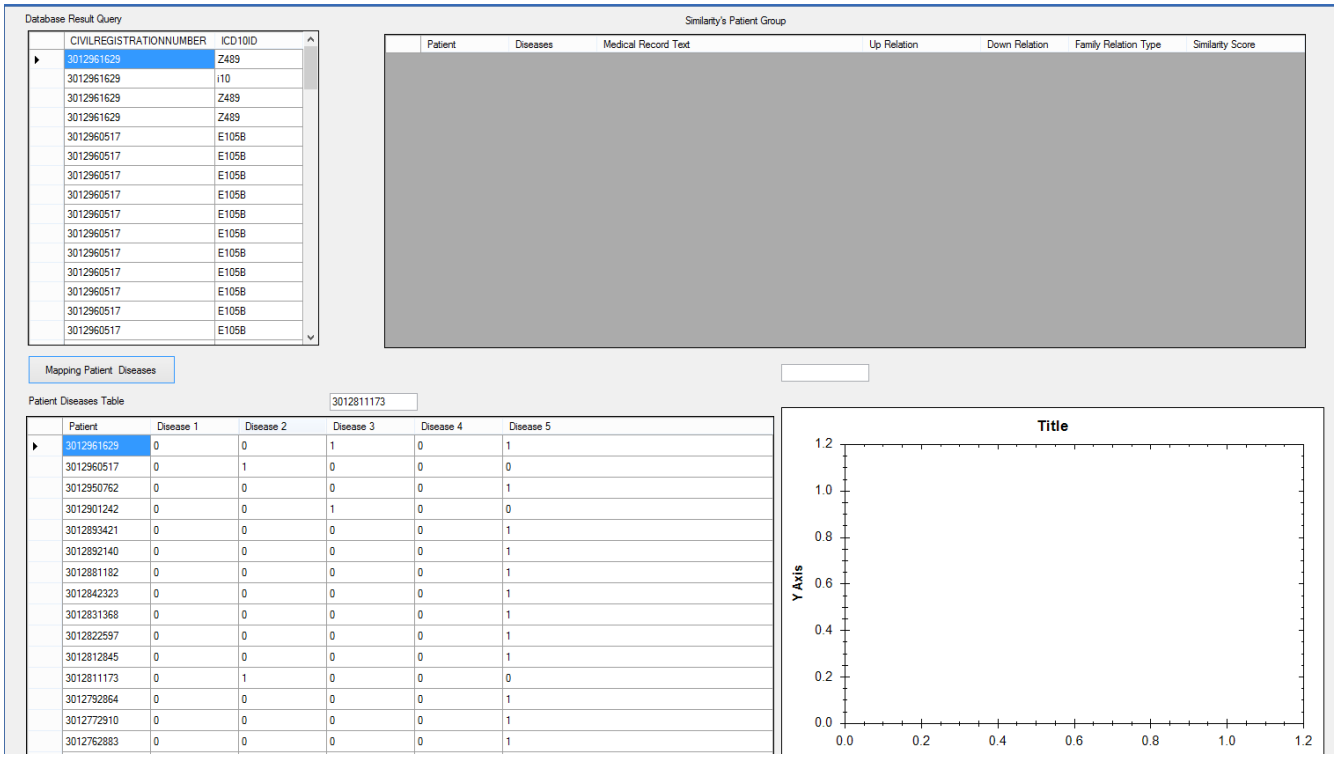


Figure 17. Second step mapping the above information

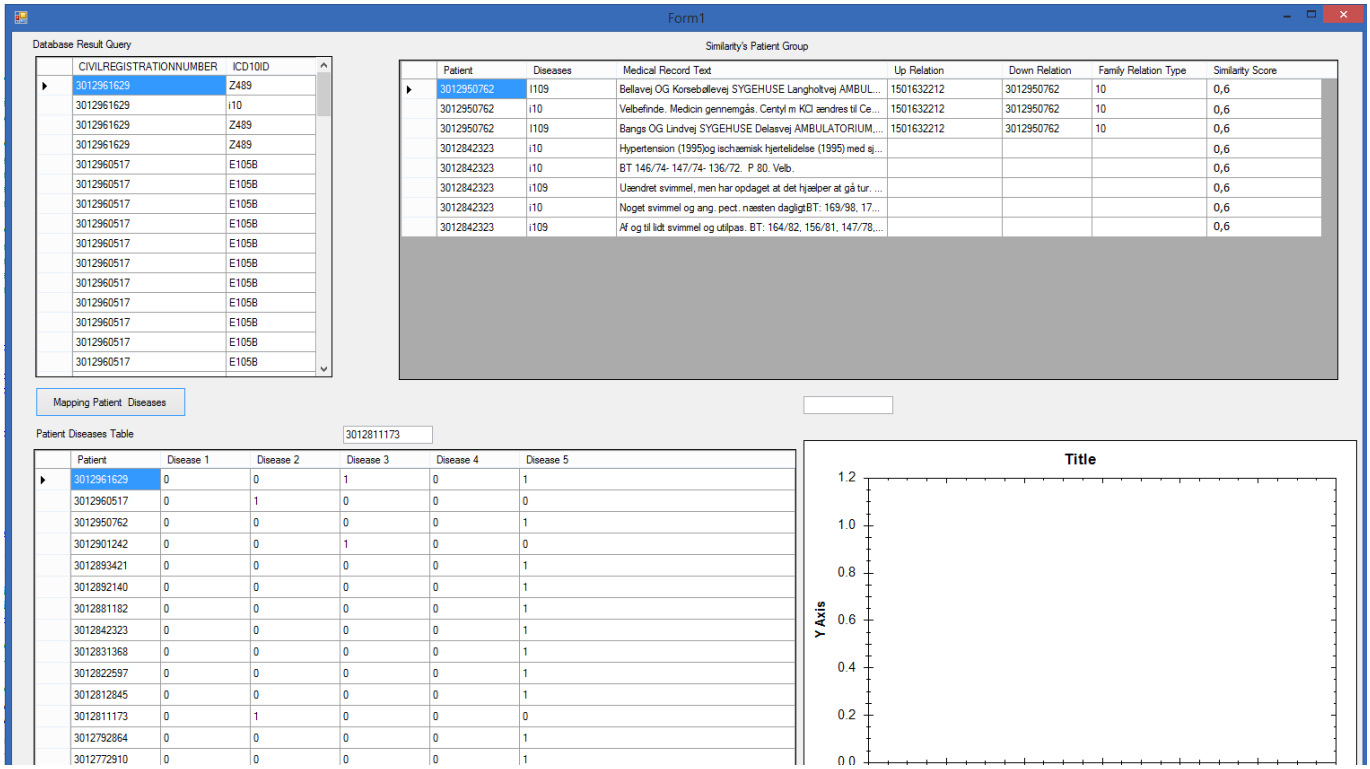


Figure 18. Third step, finding similarities