A scientific approach to Operational Management

Rogier Lommers [0340812]

December 16, 2008

Contents

Abstract

I have chosen to use the picture at the cover of this thesis because I think it is a good metaphor for the contents. In a relay race, all the runners try to reach the next runner from their team as fast as possible. But the actual point of interest in relay races is the moment in time the baton needs to be given to the next runner. In other words, everyone can run, but it is very important to optimize the time and effort between two runs. This conclusion also refers to the case of improving business processes. By optimizing the efficiency between processes, the total run time can be decreased, which results in better and improved business processes.

Preface

You are reading my master thesis from the research on finding a scientific approach to Operational Management, which is written with the purpose to round off my education Computer Science at the Radboud University Nijmegen, faculty of Science. I performed this research at BI4U, which is a consultancy agency specialized in Business Intelligence solutions.

There are a couple of people I would like to thank. First of all my supervisor Theo van der Weide. At the moments when I thought I was stranded and did not exactly know how to continue, a short meeting with Theo did really help me a lot. His academic attitude is very comforting and whenever needed, he always managed to bring me back on track.

I would also like to thank the following people for their interest and time spent during my graduation. Off course my parents, for their infinite support, my fellow internship room mates for all the fun and serious conversations we had and in no particular order: Marcel van Gerven, Monique Lommers, Robert Kok and last but definitely not least Eva Koppelman. Whenever I came home after a day of work, bad-tempered because things did not go as I expected them to go she always managed to cheer me up. Eva, thanks for everything you are.

Enjoy reading this thesis, I enjoyed writing it.

Rogier Lommers
rogier@lommers.org

1 Introduction

Data warehouse systems collect records from one or more information systems. The source systems have their own way of writing the data to the medium. By using transformation methods, it is possible to extract the information from the source into a common collection of data. By using this collection, one can generate reports and analysis in order to make operational or strategic decisions. This is exactly the core business of BI4U. The companies founder asked me to write my Master Thesis at his company. I think this is a wonderful opportunity to get in touch with a commercial company and therefore, I want to thank BI4U for giving me this opportunity.

This master thesis discusses an approach of operational management. Chapter one begins with the background of BI4U, introduces the problem statement and subject question of the research and determines its scope. The second chapter explains the basics of data warehousing. The steps needed for a operational database to be transformed into a data warehouse and from a data warehouse into the final reports are discussed. Chapter three is a definition study in patient logistics. In order to reason about it, we need to know what the characteristic are and how to recognize them. Chapter four introduces the business data models which enables us to measure patient logistics and chapter five uses this knowledge and extracts the data from the source and data warehouse systems. The next step in this thesis is to adapt data mining techniques on the acquired data. Chapter six discusses the various data mining methods which are applied in chapter seven. Finally, the conclusions are presented in chapter eight.

1.1 The background of BI4U

BI4U was founded in 2002 and currently (2008) fifty employees are working at the company. The core business used to be detachment, but since two years, the companies focus has shifted to offering consultancy services and implementing their own data warehousing solutions. These solutions are based on in-house development data models. One of the solutions is called the Intelligence Factory, previously known as the Software Factory. The Intelligence Factory can be described as a working method or a set of knowledge, methods, models and modules that support (partly) automatic generation of a data warehouse. An example of a data warehouse, generated by using the Intelligence Factory framework is Ariadne. Ariadne is a generic business data warehouse model, which can be implemented at different hospitals. In other words: when a hospital decides to purchase services from BI4U, it purchases the software package Ariadne. Then, Ariadne will be adapted to fit in the hospital's source information systems in order to generate management reports. Ariadne is the link between operational systems

and management reports built up from historical data. The Intelligence Factory is based on the Kaizen philosophy, which stands for *Continuous Improvement*, see figure 1. By using this framework, one is able to implement software products like Ariadne in a generic way and the corresponding processes will be improved during time exceeds.

Figure 1: Kaizen Logo

1.2 Problem Statement

By writing this master thesis, I want to examine and improve the quality of the Ariadne model, focused on the patient logistics part. The research question can be described as follows:

> *A scientific approach to operational management. A case study with the purpose to be able to reason about an organization with the aim of improving the business processes.*

In order to reason about a subject, it is necessary to make a definition study about it. This will be the first subquestion from this thesis. What is patient logistics exactly? Therefore, we will start by defining the concept of patient logistics. What does the customer's view of the concept look like. Chapter three discusses this matter. The next question then arises: how can we measure patient logistics? In chapter four, we introduce business data models which help us to measure patient logistics, one is able to reason about the quality of the processes. This master thesis can be described as the total of methods which enables someone to reason about the quality of the patient logistics processes. It provides a framework which helps in measuring and optimizing the processes related to patient logistics by using data warehousing and data mining techniques. This brings us the third subquestion, discussed in chapter seven: how can we adapt data mining in order to improve business processes? A summarizing of the research question with the corresponding subquestion is as follows:

- Research question: A scientific approach to operational management. A case study with the purpose to be able to reason about an organization with the aim of improving the business processes.

    1 What is patient logistics exactly?

2 How can we measure patient logistics?

3 How can we adapt data mining techniques in order to improve business processes?

As a product of this thesis, we will deliver a report which enables one to reason about patient logistics. This includes the definition of patient logistics, the indicators which enables us to measure patient logistics and corresponding business data models. Furthermore, the case study of applying data mining on the gathered data is included at the end of this thesis.

1.3 Area of Knowledge

Research will be done in the discipline of Information Science. On one side, the subject has affinity with the technical aspects like data mining and data warehousing. On the other side, it touches the business side, since it is impossible to find a generic way to improve logistic processes without having the knowledge about these processes. The scope of the research will be limited to the level of patient logistics within a medical center. With a view to the given time and resources, it is more feasible to concentrate on a concrete subject, other then focusing on the whole process. Figure 2

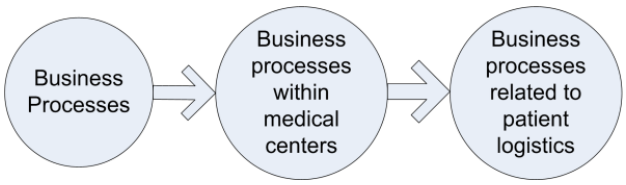

Figure 2: Scope of research

Assume, there are a lot of records, pieces of information, belonging to a certain disorder, for example the medical removal of someone's meniscus. With this information, we are going to research if it is possible to use data warehousing and data mining techniques to determine if the process of removing the meniscus is efficient or not and we will try to find a way of improving these processes.

1.4 Research Model

By using the approach of Verschuren, a research model for this master thesis was developed [22]. His approach introduces a standard way of how to start a research. During the creation of the action plan, it was necessary to draw the model several times, in order to get the research questions clear. The model can provide a broader insight regarding the relation of the research questions with the main goal of the research and provides a step-by-step approach of the research. See figure 3 for a schematic view of the research

model. We will now discuss the required steps in order to complete this thesis.

- The first step of the thesis is to get a clear view of what patient logistics exactly is. In order to verify if a process is efficient and optimized for use, one has to know what the subdivided parts exactly do and how they can be optimized and improved. In order to improve a process, it is necessary to link the process to measurable indicators. These steps will be the first part of the thesis. First, study what patient logistics is and secondly, create indicators which enables us to measure the processes within a hospital.

- Measuring processes is related to creating reports which enables the management to make strategic decisions. These reports are built up out of pieces information, originating from the data warehouse or from the source databases. By creating business data models, one is able to get a clear understanding of what data is needed in order to measure the indicators mentioned in step one. The business data models are also a good way to verify the information needed with the customer; in this case the domain expert. After creating the data models, the required data can be extracted from the source systems by using SQL statements, which stands for structured query language.

- At this moment, we have the required data out of the source information systems or the data mart staging area presented in tables. See chapter two for a definition of this area. The data and the data models are linked to each other and therefore, by using the data, one is able to reason about the quality of patient logistics.

- The next step is to retrieve additional, hidden information out of the above mentioned pieces of information. An example will be a table of patient information, linked with disorders. Imagine we have a huge amount of records containing all kind of patients suffering all kind of disorders. Together with these pieces of information are the date of appointments and the date of the first intervention. Example: Patient W suffers disorder X and has an appointment on date/time Y. The actual date/time of the intervention is Z. Now we can use data mining techniques to cluster/classify certain groups of patients and new, hidden information can be discovered. For example, it is possible that waiting times from patients suffering a certain disorder is bigger then other disorders. This is an example of hidden information, originating from a huge amount of other information.

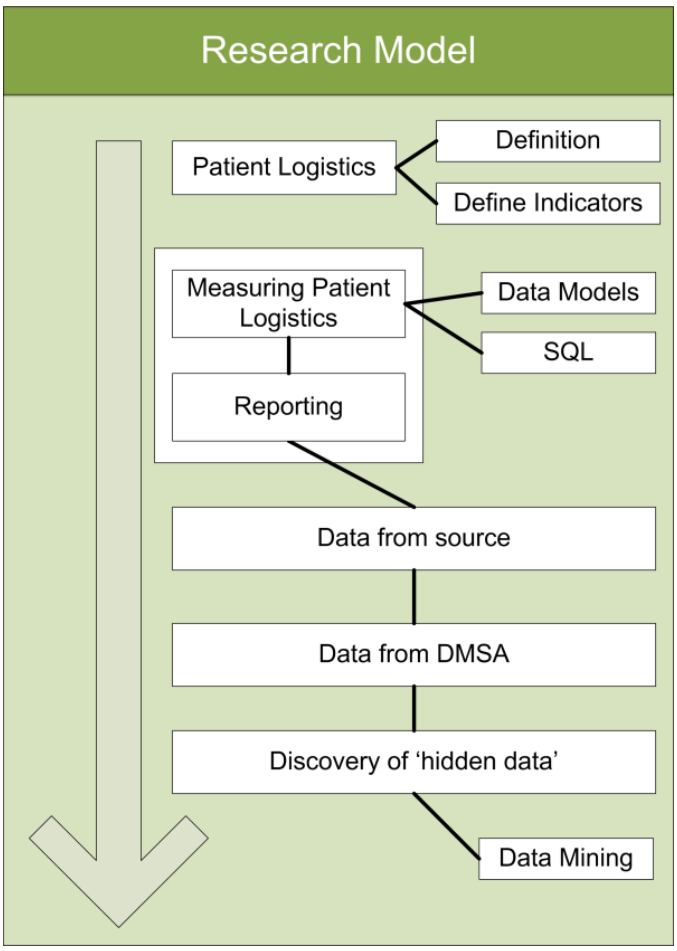

Figure 3: Research model

During this thesis, we use information originating from database systems. This information is stored by operational applications like financial systems, staff information systems, planning, production systems where patient information is stored at, and so on. The next chapter discusses how different databases/information systems are being established into a data warehouse system.

2 Data Warehousing Explained

BI4U is, among other things, active in several medical centers. The information systems BI4U delivers to the customers are based on data warehousing techniques. The research question for this master thesis fits the need for a more effective way of recognizing processes. Data warehousing is a relatively new technique which is applied since the early 90's. By sorting through large amounts of data and picking out relevant pieces of information, companies

can create analysis and reports which help in making the right strategic and operational decisions. During this thesis, we put the focus on a specific case study, namely the case of a hospital's patient logistics processes. It would be very helpful for a hospital if their processes would be as effective as possible. The Ariadne business data warehouse [1] includes information about a variety of objects like patients, medical interventions, operations, operation room history and financial information. Besides the business data warehouse, various templates for database structures and packages have been defined to be able to link the data warehouse with the various sources of a specific customer. Figure 4 provides a graphical interpretation of how a data warehouse is created.

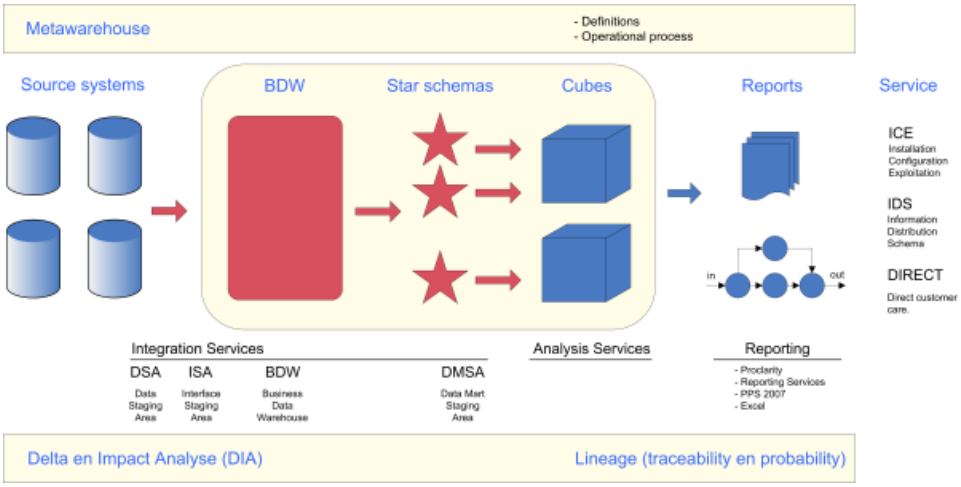

Figure 4: Schema of Intelligence Factory

In this section the data warehouse model will be explained. In [17], Sen and Sinha describe several possible architectures for a data warehouse: data marts, enterprise, hub-and-spoke data marts, enterprise with operational data store and distributed. The architecture at BI4U is of the type Enterprise Data Warehouse Architecture, because all the meta data is stored in the data warehouse itself. Meta data is data-over-data. It contains information about the fact that there is data, not the data itself. For example, when a patient's last name is changed in the system, the date of change is stored as meta data.

[1] BDW is short for Business Data Warehouse

In this type of architecture, there is a central data warehouse on which data analysis is performed which is then stored in relational database management system and/or data marts through which it can be used for reports. See figure 5 for a schematic view of this architecture. The data warehouse model of BI4U consists of several stages:

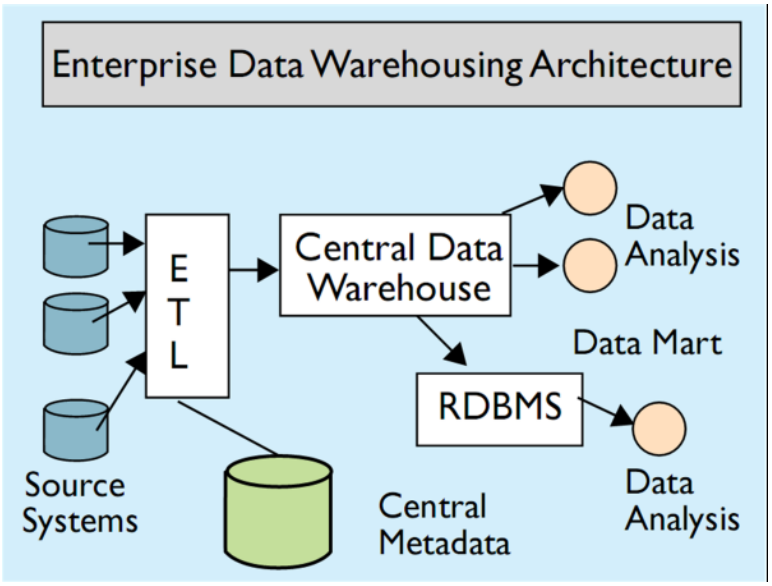

Figure 5: Enterprise data warehouse architecture

- Data staging Area: a system that stands between the legacy systems and the analytics system. Here, the source data is copied to the data warehouse.
- Interface staging area: the area where the data from the several source systems is being filtered and integrated.
- Business data warehouse: Here, all the relevant data that can be used for analysis will be stored.
- Data mart staging area: Optimized information for reporting services is stored in the data mart staging area.
- Cubes: Data stored in a summarized way to enable a quick way to view the overall data.
- Reports: The final reports, used for making operational and strategic decisions.

In order to provide a usable outcome to the end user, the data moves through all these stages. The different stages will be discussed more thoroughly in

the next two paragraphs. The first paragraph discusses the way data is being transformed through the first four phases and the second paragraph discusses how the data is transformed through the last four phases.

2.1 From Source to Business Data Warehouse

The first three steps (source, data staging area, interface staging area and business data warehouse) can be recognized as the process of extraction, transformation and loading (ETL), which is a common approach in the data warehousing field. When a new project is born, it starts with making an inventory of the available source information systems. These operational data are stored in specific database server systems, the so called source systems. The Ariadne model supports a wide range of source systems, including Microsoft SQL Server, oracle, comma separated value-files, Excel, Access and so on [4].

ETL processes involve the cleaning, transforming, combining, duplicating and structuring the source data in order that it can be used in this data warehouse environment for analysis [16]. In general, the term wrapper is used in data warehouse literature for the element which connects the sources and the data warehouse. The wrapper takes care of connecting a source system to the generic model of the data warehouse. Relating this to the situation at BI4U, this involves the data staging area database and the source-to-DSA package.

Between the wrapper and the actual data warehouse something often described as a middle layer is present. The middle layer takes care of the filtering and integrating of information from multiple sources. Relating this to the situation at BI4U this involves the interface staging area database and the data staging area to interface staging area packages packages. Regarding to the ETL, the implementation of the first phase concerns the extracting part. The goal of the data staging area is to extract data from the source information systems with a minimal burden on the source systems. The new database is first emptied and then filled with data from the source system by source-to-DSA packages. The data is extracted without modifying the structure or content. It could be that the entire source database is being copied to the data staging area database, but it is also a possibility that irrelevant tables are not extracted. This depends on the information which is needed in the data warehouse. A domain expert decides the need for data. There is a homomorphism from the data staging area into the sources, meaning that there is a function that relates DSA information to source information such that the structural properties from DSA are related consistently to those in the source. One could describe the database (both content and structure) as DSA $\subseteq$ *Source*.

This means that the transformation is valid, but does not imply that all information from the source information systems is present in the data stating area database. The next phase, from the data staging area to the interface staging area is about restructuring the data so that it can be compared with the current data warehouse content and that new data can be inserted in the data warehouse. The data is transformed to fit the structure of the defined business data model, whose structure is also used for the business data warehouse. As discussed previously, this generic business data model can be used for multiple customer implementations and therefore the structure, naming and type of attributes in the interfase staging area can, and probably will be different from the source system and the data staging area. It can also result in certain data present in the data staging area not being transferred to the interface staging area, because the data is not necessary. Another option is that calculated data on data originating from the data staging area is only copied to the interface staging area, not together with the original data.

The business data warehouse phase concerns the loading part. The goal is to store all the relevant data that can be used for analysis, based on the generic business data model of entities and relations. The database is the core of the data warehouse. It contains all the data that was loaded into the data warehouse at earlier times and it is filled with new data by the interface staging area to business data warehouse packages based on the data in the interface staging area database. The structure of both databases are the same, there are no modifications made to the data. In cases where this is relevant, the data is compared regarding slowly changing dimensions and possibly enriched with time information. The main purpose of this is to be able to keep track of history records.

2.2 From Business Data Warehouse to Reports

The other steps involve the modification of the data from the data warehouse so that it is optimized for analysis and can be used for creating reports. BI4U develops generic business data models that can be used for multiple data warehouse implementations in similar organizations. This results in a generic structure of the interface staging area, the business data warehouse and the data mart staging area. The packages used in the first three stages are customer specific [5]. From the business data warehouse to the data mart staging area phase, the data is reorganized from a relational model to a dimensional model. The goal is to restructure the data to improve analytical performance. In the relational model of the business data warehouse, analysis would require going through many relations with joined queries, which is very intensive to process to the database management system. For optimization reasons, in the data mart stating area, the data is related into star schemas, with fact and dimension tables in order to structure the data for

optimal analysis performance. It is structured in a way that is is easier and faster to relate certain interesting information (facts) to certain dimensions, decreasing the required time for calculations.

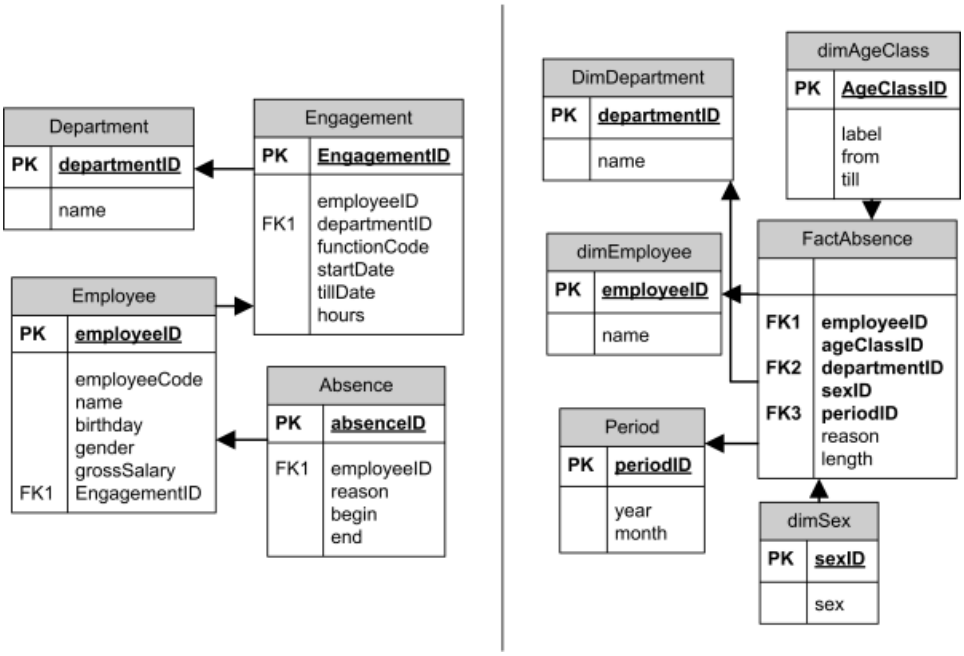

Figure 6: Example data warehouse dimensions

The left part of figure 6 shows an example of the transformation of the previous phase. All the data is stored in a operational way. When we look at the right part, the transformation step from business data warehouse to the data mart staging area has been made. All the data is now stored in such a way that it is related to a specific absence fact entry. Since it is interesting to be able to show reports on, for example, the basis of departments, sex and age classifications, these are added as dimensions. During the transformation from business data warehouse to the data mart staging area, for each absence entry the age and sex can be determined through the relevant employee and the department can be determined through the engagement of the employee. This step in the process is important, since it gives us the possibility to track historical information into a dimensional model, therefore this figure is added to this thesis.

The next step is the transformation from the data mart staging area phase to the cube phase. Now, the data which is stored in a relational database like the one of figure 6 is transformed and stored into a multi-dimensional database. During this transformation, much of the analysis that will eventually be used in the reports is done. The data stored in the cubes can be

used to display the desired information in the reports, or in fact, the reports will query the cubes for the required information. In the report functionality, all that has to be done is to define which dimensions should be related to each other. Any change in the data mart staging area will have an impact on one or more cubes and reports. A multi-dimensional cube exists of several elements: measures, dimensions and attributes. Measures are the objects of the analysis and are derived from fact tables in the data mart staging area. Dimensions provide the context for measures and are derived from dimension tables in the data mart staging area. Attributes define a dimension, and can form a hierarchy of the dimension. A hierarchy of the dimension time can, for instance, be made out of the attributes day, month, year. Harinarayan [10] mentions three ways a cube can be implemented.

- The first option is to physically materialize the whole data cube. This implicates in the best performance for the end-user. All possible calculations are performed and stored, even before the cube has been accessed. This is also called the pre-calculated cube. A said before, one big advantage is the speed. But, on the other hand, there are also disadvantages liked to this option. First, there is the age of the data. Because the cubes are pre-computed, only information till the moment of computation is absorbed into the cube. A second disadvantage is storage. Calculated data requires a huge amount of storage units.

- The second option is no materialization. This delivers the worst performance to the end-user with regard to response time. However it is cheap with regard to computational expenses and no storage of pre-calculated information is required.

- The third and last option is to partly materialize the data cube. Only those parts of the cubes that are most important are materialized. This provides a more feasible solution to pre-calculating, since this way it is possible to balance end-user performance with the costs of computational calculations and storage of materialization.

Like most things, it is clear that one has to make a trade-off between on one hand the costs of computation and storage and, on the other hand, the response time for the end user when requesting reports and querying on the data warehouse. The Ariadne model uses the third method: the most-accessed reports are updated every night and caching this data for use in reports results in an optimal balance between speed and costs.

3 Definition Study: Patient Logistics

As mentioned in the research model, the second part of this thesis will be to describe what patient logistics is. A lot of people talk about logistics, patient logistics and efficiency, but what is it exactly? How can patient logistics be measured? This chapters' goal is to make a definition study of patient logistics and to define indicators which make it possible to measure the quality of a hospital's patient logistics.

Well funded logistic processes offer better service for patients and a more efficient run of the medical center. This is equivalent to financial advantage, better production and a decrease in the pressure of work. The current health care can be characterized as one-step-logistics. This means that for a patient, every logistic step is organized. This way of working is also known as the push-model, because the caretaker *pushes* the patient through the system. The other way around is by using the pull-model, which is focused on the organization, divided into clusters for the patients. The route/steps a patient undergoes, is determined in advance. By using the pull model, all attention can be focused on the care itself, rather then on arranging the care.

An optimal patient logistics requires the management of variation in supply and demand. For example, in a week, twice as much surgery hours are booked as the week after. Reducing this variation and to be able to anticipate better is the focus for improving the patient logistics. Due to the unpredictability of the supply, overcapacity cannot be avoided. This not necessarily results in a lack of efficiency; important is the focus on an optimal usage of this unpredictability. We can define three types of patient logistics.

- The first one is called unit logistics. This type of logistics can be characterized by the management of supply and demand of one department. Each department (unit) has its own inputs and outputs.
- The second type of logistics is focusing on specific groups of patients. New types of organizations, designed for clustering of patients are typically identified by this type of logistics. Clinical paths are an example of improving an organization in health care in an integrally way. The systems are optimized for certain types of, for example, disorders. The term clinical paths are discussed later.
- The third type is called chain logistics. This type manages the complete streams of patients, improving an health care institution in an integrally way, focusing on capacity management.

3.1 Case: Brake-A-leg

We now discuss a case of a list of the processes which occur when someone, a patient, breaks his or her leg. It emphasizes the need for smoother integration of the business processes. Imagine, a men breaks his leg on the street. He is not in (mortal) danger and acts calm and is approachable. An ambulance transports him to the hospital where he will undergo a treatment.

Time	View of Patient (1)	View of Hospital (2)	Time of Doctor
10.30	Intake doctor	Doctor diagnoses patient	10 minutes
11.00	Making picture of leg	Using scan room and lab	25 minutes
11.15	Transport to operation Room		
11.45	Operation of leg	Operation room in use, staff, recovery room	
13.00	Stay in hospital	Bed and room in use	

(1): The patient's perception of the treatment
(2): What kind of recourses are necessary

It is clear we can distinguish two views of these sequence of processes. One, there is the view of the patient. This includes the perception of the patient, everything he undergoes and is aware of. The second view is the view of the hospital. The use of equipment, rooms, staff, and so on. In order to get optimal patient logistic processes, we have to make a trade-off between the patient's view and the hospital's view of the problem. Combining the best of both worlds would result in an optimal patient logistics. Production control and capacity management are the two key elements in achieving a good service, accessible services, short waiting periods and effective usage of available recourses.

Another issue is the tendency of more and more cost-efficiency health care processes, due to the impact of a system called free marketing. Production control and capacity management are relatively new points of interest. For a hospital, it is common to use information systems for specific departments, each running its own information system (unit logistics). A coordinating information system is not present [12]. This results in problems in coordination between departments. For example, in a lot of hospitals the operation room management information system keeps track of a lot of detailed information. By using this information it is possible to generate actual reports about the operation room occupation between peak and lower hours. Adjoining departments like the intensive care, outpatients' treatment or nursery do not have this variety in information. This often results in forced adaptation of these departments to the operation room department without

having the possibility to give an adequate answer. This eventually results in a lack of efficiency. For example, when we look at the operation room, it can be very handy to plan the one-day-patients in the afternoon, while the outpatients' treatment prefers the morning. Hoorns conclusion is relatively simple [12]: due to its complexity, it is impossible to model and implement an information system which makes it possible to create a department-wide information system for planning and control (chain logistics). By shifting the horizon to a more tactic and/or strategic view, it is easier to model these problems into an information system. With this new view, questions like these become more and more relevant:

- What are the consequences on the waiting times, use of waiting rooms and operation room when we extend the amount of places for specialist?

- What are the consequences when we make a change in the planning of the operation room and which changes lead to a better flow and usage of the available resources?

- What can be useful in order to gain a more flexible planning for staff so the effort of manpower is in better harmony with the fluctuation in supply of work.

The coherence between the departments in the logistical chain needs some kind of control. Another recent study shows the a solution, together with corresponding problems in current hospitals [20]. The recommendations are quite simple: distinguish the different groups of patients and build a work flow for every group. The next chapter discusses these groups of patients.

3.2 Three types of Patients

Like mentioned in the previous chapter, patients can be divided into three groups:

1. Critical patients. This group of patients need acute health care due to their life critical status.
2. Elective patients. These patients do not need urgent health care. Recent study shows that 80% of the patients in hospitals is part of the elective group.
3. Polyclinic health care. This group includes patients which total stay time is no longer then one day.

Have a look at the following diagram (figure 7) which emphasizes the flow of patients into the system of health care.

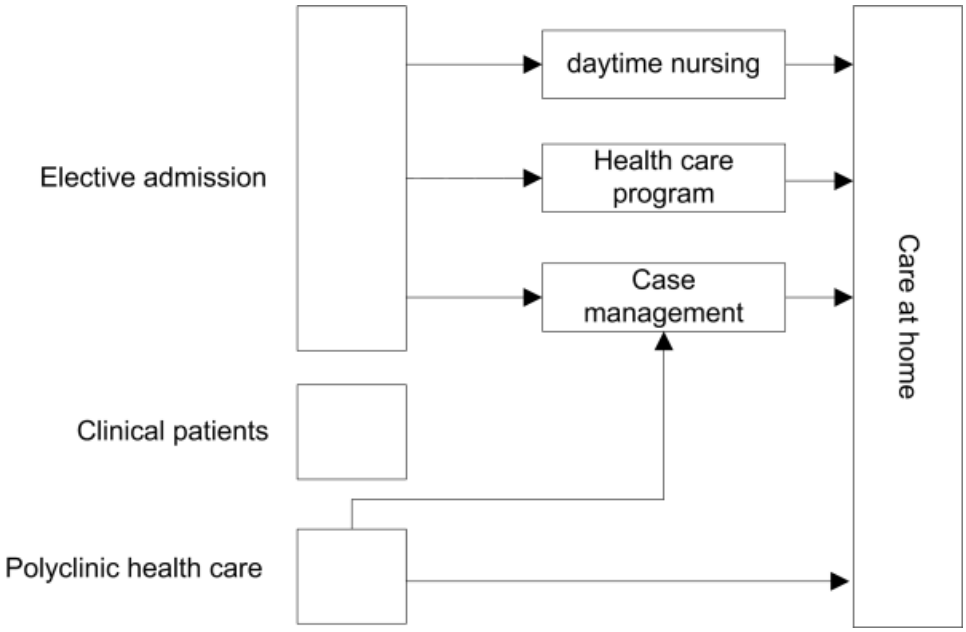

Figure 7: Three types of Patients

3.3 Clinical Pathways

A clinical pathway, also known as a critical pathway, is a multi-disciplinary tool which describes routine interventions for a group of patients with similar needs. Multi-disciplinary means that the patient needs medical care for more then one discipline/department. Clinical pathways are focused on the patient and includes the expected outcomes at each step. Their origin comes from planning techniques, developed in the 50's. An example is the program evaluation and review technique, short as PERT. By using these methods, one is better able to plan and control complex processes like huge building projects, space traveling projects and so on. The clinical path method makes an inventory of the critical points between starting point A and ending point B. The remarkable thing on clinical paths is the fact that the shortest path is not automatically the clinical path, like in the transport business. The shortest path describes the critical points which cause a delay on the way to the endpoint, for example when an pre-operative examination is delayed by one day, the total amount of days will also be delayed by one day. There is a point where the whole process stagnates until the pre-operative examination has taken place. This means that the pre-operative examination is a critical point in the overall path and has to be controlled and planned in an accurate way.

In health care, people have to deal with a lot of complex situations and processes. In order to deal with these situations by using information systems, it is necessary to gain some support by the people which are working in health care. Factors like uncertainty, unpredictability, the need for individualization and therapeutical freedom are arguments against systematical planning and control systems [18]. Therefore, medical centers need a new way of working, called "from fragmentation to collaboration" [15]. This means that the patients need to be a starting point where the medical care and the structure of a medical center begins. The clinical process is the central axis from which the processes of the organization are built. The desired situation can be described as a situation where a complete program can be offered to the patient, instead of a succession of stand-alone, non-planned and controlled interventions. Imagine, 25 activities for one patient a day. This includes medication, care of wounds, examinations, testing, meals, etc. In a medium sized hospital, the old-fashioned way of working demands somewhere between 10.000 and 15.000 clinical orders fired at the hospitals organization. It is not a surprise that a lot of time and effort is spent on communication, conflicts, daily control, the search for the size of the capacity that fits the need, etc.

3.3.1 Clinical/critical Pathways as a Process

The process of the development of clinical paths in an organization is very useful [18]. Study shows that the process of developing is marked as much more valuable as the clinical paths on itself. Developing this path requires teamwork, examining the population of patients, the team, the current situation, and so on. Discussing the daily work flows as a team has a lot of advantages. Even people who are working for years in the same team discover that, during the development of the clinical paths, implicit assumptions not always end up in the correct way. The reasons are quite varied: unclear targets, incomplete job descriptions, and so on. The process of developing clinical paths by using the 30-steps plan [21] is therefore a successful way to implement.

3.3.2 Clinical Pathways as a Method

Clinical pathways make use of a variety of health care control methods. A central indicator in creating clinical paths is the predictability of the problem, together with the solution. In general, there are three methods to be distinguished:

- Standard clinical paths. This type of clinical paths is used when the process of health care can be predicted easily and interventions and targets are in a standard way set up in a time-task matrix.

- Patient specific clinical paths (customized). Clinical paths are not being developed for a group of patients, but for one person in particular. A certain path is used when the time it takes cannot be predicted for a group of patients, but it can be predicted for one special case: one patient.

- Case management. Besides standard clinical paths and patient specific clinical paths there are the case management paths. Groups of patients which are not very predictable, are qualified for this group. Examples are patients, accommodating the intensive care with a non-routine check-up. Daily interdisciplinary consideration is necessary in order to give good care. The aim of case management is to tune the team into the specific population of this group of patients. This method of clinical paths is also known as “a living and breathing clinical path”.

The variety in the three methods is as follows. The first one, the standard clinical paths, is covered by 60% of the population of patients. The patient specific clinical paths (the customized way) is covered by 20%, even does the third method, the case management. As a conclusion of this chapter, we can say that it is evident that clinical pathways are a lot more then a piece of paper, containing the total amount of appointments for (a group of)

patients. It can be described as the complete action plan in order to create an approach for handling health care, focusing on a certain population of patients. Emphasis is placed on creating the clinical path by using the 30-steps plan, together with the methods used. In order to work with critical paths, one required condition is the way the path can be predicted.

3.4 The push and pull methods

There are two types of marketing strategies (methods) to be distinguished. Scientific articles do compare the differences between both methods with low and high tide. No matter where and what someone is selling, it is always necessary to apply some kind of marketing strategy in order to get the customers/patients into the shop/basket/care center. Relevant to the health care system are the push and the pull method. The next sub chapters discuss both methods, starting by the push method.

3.4.1 The Push Method

The push method can be characterized as the marketing method which enables the seller to focus more and more on approaching the customer. On a very low level, the salesman in the shop which asks the customer: "Can I help you" is an example of the push method. The strategy makes use of a companies sales force and trade promotion activities to create consumer demand for a product. Another good example of push selling is in the mobile phones business, where the major handset manufacturers such as Nokia promote their products via retailers such as Belcompany. Personal selling and trade promotions are often the most effective promotional tools for companies such as Nokia, for example offering subsidies on the handsets to encourage retailers to sell higher volumes.

Linking to the health care, the bottleneck in the current health care system is the way the patient processes are being introduced. Dependent on the patient's urgency and the available recourses, patients are pushed into the process, therefore it is called the push method. This implicates certain logistic disadvantages, namely:

- The patient is not the central subject in the logistical design. This means that before a process starts, he or she does not now how long it takes, when the process will be finished and so on. Questions like how long the treatment will take and when will the patient meet whom, stay unanswered. One can say that the patient experiences the whole process as a black box.

- The available capacity determines the turnaround time. The turnaround time is equivalent to the time between the first and the last step of the

treatment. When the turnaround time of a single process is unclear, the total process cannot be streamlined and synchronized.

- Unhappy staff. Unpredictability from the turnaround time of the processes implies unpredictability in the need of staff. So not only the patients, but also the staff experiences the process as a black box. This results in unclear schedules, dynamic and unpredictable working times and therefore unstable loans.

3.4.2 The Pull Method

The in the previous chapter mentioned problems can be solved by implementing the pull method, which can be characterized as tempting the customer to buy the product and/or service. Classic examples are special discounts like "Now for half of the price" and "Hand in this voucher and receive an extra discount of 5 percent". The pull method requires high spending on advertising and consumer promotion to build up consumer demand for a product. If the strategy is successful, consumers will ask their retailers for the product, the retailers will ask the wholesalers, and the wholesales will ask the producers.

To return to the problems of patient logistics, the transformation from push to pull leads to an improvement in quality and saving in costs. Deciding in a pull method is the discipline to execute each step, in a structural way, focusing on the improvement of the process. The pull model distinguishes the complete processes into five steps, introduced by Bakker [3]. These five steps can be described as a framework which enables us to implement the pull method in medical centers. Aim is to centralize the patient in the complete process of health care, just like the customer who wants to buy a product. Figure 8 is a schematic view of the framework.

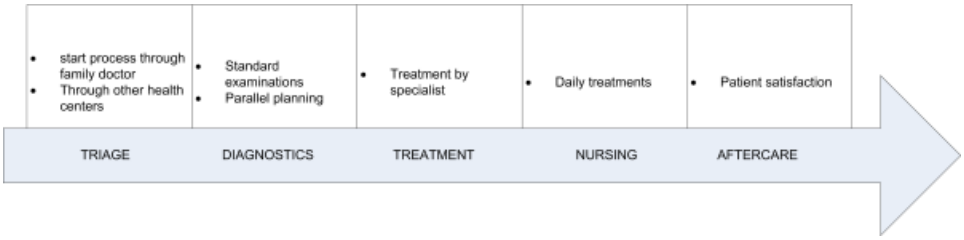

Figure 8: Framework for Implementation of the Pull Method

In order to implement the pull method, the article [3] prescribes that the organization, in this case a hospital or a medical center, has to meet certain preconditions, namely:

- First of all, the article recommends to divide the flow of patients in groups with each their own predictability (also have a look at the chapter about clinical pathways). By clustering patients on the type of care they need, segments of homogeneous groups of patients are created, whose logistic processes can be standardized. This means that activities are being done in a best practice way and in time further optimized.

- The second condition is about turnaround time and the corresponding planning and control. The supply of care can be predicted by using statistical information from the past. The amount of elective treatments can be predicted, just like the amount of critical treatments. This statistical information is very valuable in creating and optimizing processes.

- Measurements. It is important to measure key performance indicators (KPI's) like the quality of processes (turnaround time of each step in the process), but also the quality of the final result of the process. Examples are the amount of recovered patients, complications, etc.

3.5 An example: from Push to Pull

The article from Dekker called From Push to Pull [6] describes how to transform an organization from working with the push strategy to working with the pull strategy by introducing a case. The case is about the department Information Technologies Logistics from Philips Lightning Electronics. The company produces so called linking devices. These are small devices which take care of the power flow inside strip lights. Customers of the company include the internal business department called Luminair, but also some external manufacturers. Till 2001, the production was planned by using long term predictions, the forecasts. This is part of the push method. The information system contained twelve months of forecast, for each product. The whole business model was based on these forecasts. Production was planned on a period of time of eight weeks in the future. This resulted in slow reactions on marked fluctuations and almost no control of the stock used. The answer of the problems was to transform the push method into a pull method. So, instead of letting the production push the chain by using the forecasts, the transformation focused on the customer demands, which from that moment on determines the production.

After implementing this pull method, the amount of stock was reduced significantly and was no longer an uncontrollable spin-off product due to a bad forecast, but it was transformed into a planned buffer. The big advantage of a planned buffer is the ability to bring the amount of stock to a minimal level. When a product is sold internal or external, then, the amount of stock is filled up. When a product is not sold, the amount of stock stays equal. The ability to planning and control of the amount of stock is raised. In this particular case, stock is reduced by 20 percent. This finally resulted in lower costs of the product for the customer.

4 Measuring the Quality of Patient Logistics

The use of information systems has grown the last decades. Especially when different applications are integrated. For example, a common database with product information can be accessed by a variety of applications like cash points and financial applications. Databases have to reflect more facets of the whole application area than databases for a single application. This not only affects the size of the database, but it also makes the structure more complex. Another point of interest is the way business rules are integrated into the software. Database designers need to translate the real world into a database, with other words: he or she needs to model the properties of objects from the real world in the domain of discourse and not the functionality of the various application programs. In the past decades, several steps have been developed, in order to make the process of database modeling as smooth as possible. The following steps can be distinguished:

1. Requirements Analysis
2. Conceptual Database Design
3. Logical Database Design
4. Physical Database Design

Step 2, the conceptual database design, is the most demanding step, because the later phases are merely transformation steps. The conceptual schema is the first transformation from the real world into a formalized description about the system functionality. This chapter describes the possibilities to measure the quality of patient logistics.

In order to create a conceptual model, corresponding to the objects and relationships and with the aim of enhancing the patient logistics, we need to know how we can measure the quality of the processes. To be able to clearly reason about the indicators, we use the above mentioned schema which describes the five steps for implementing a pull model.

During a three-days workshop with staff of a Dutch hospital, a process of requirements engineering has been performed. A variety of staff members were present, including the board of directors, departmental management, application management (also responsible for planning and control) and the manager of clinical paths. Several indicators have been formed. These indicators are useful, because they show the information which is needed to determine the quality of the processes, linked to patient logistics. The next five paragraphs discusses the steps needed to implement the pull method, starting at the triage step.

4.1 Triage

The first part of the model is about finding out what patients one is dealing with. The total flow of patients in a hospital can be divided into two types [2].

- First, there is the mono-disciplinary type. This group of patients is the most easy one, hence their complaint can be investigated and treated by one specialist in one department.
- The second type, the hard one, is called the multi-disciplinary type. The last type requires the involvement of different specialties for their medical treatment. It is clear that this type requires more effort regarding the coordination of the available resources. One solution is to create special units, designed to process the above mentioned multi-disciplinary patients. When one wants to implements those units, he first needs to identify those patients and categorize them. This categorization can be done in several ways, for example an categorization by utilization, reimbursement, quality assurance and management applications.

Measuring the processes which occur during the triage phase can be done by the following indicators, which are applicable for the polyclinic and elective patients only, because the way critical patients need to be treated is way more different. Also have a look at paragraph 3.2 for a common way of dividing different groups of patients.

4.1.1 Waiting Time

This indicator, shown at figure 9, measures the amount of time a patient has to wait in the waiting room before he or she may enter the process. When a patient has an appointment at, for example, 14:00 hours, how long does it take when he or she is being helped. The business data model can be described as follows.

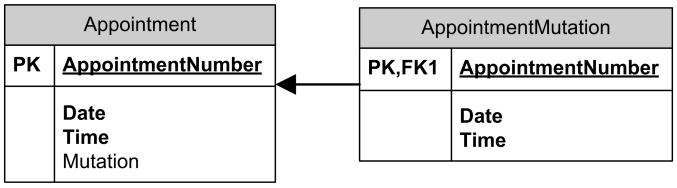

Figure 9: Waiting Time

An appointment has a certain date and time and eventually a mutation. This depends on the fact that a patient is treated later then the initial appointment is planned.

4.2 Diagnostics

When the patient is categorized as either elective, polyclinic or critical, the next step is to perform the diagnostics part of the process as efficient as possible. This can be achieved by making a standard way of diagnosing those patients with the same complaints. For example, an elective patient is suffering cataract, which is a problem with the eyes. It is relative simple to diagnose this defect and a standard way of diagnose can help optimizing this process. The diagnose phase for a cataract patient always contains a couple of tests, eye measurements, creating an ECG scan and a blood sample is needed. Those steps are the same for every cataract patient and therefore can be planned integrally. The indicators can be described as follow.

4.2.1 Amount of re-admissions for each Disorder

This indicator measures the amount of patients who enter the diagnostics phase for the second time or more for the same disorder. When a patient's diagnostics is not correct, he or she needs to go back into the diagnostics phase, which is inefficient. When a certain disorder has a lot of re-admissions, the hospital needs to make some interventions in order to optimize the process. Have a look at figure 10.

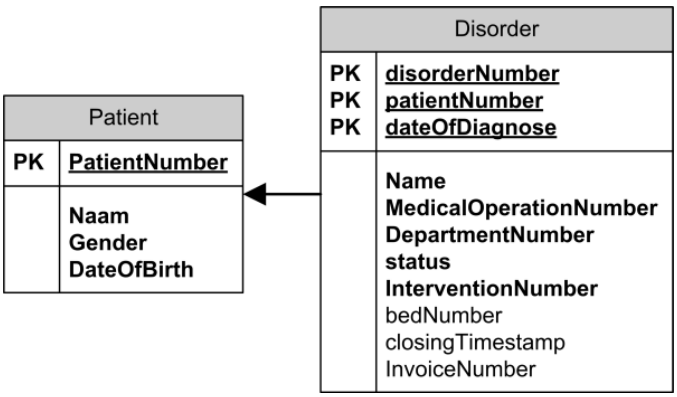

Figure 10: Amount of Re-Admissions for each Disorder

Each disorder has a primary key, which contains the disorder number, patient number and the date when the diagnose took place. On this way, it is possible to track which patients have multiple diagnoses for the same disorder.

4.2.2 Amount of Diagnostic Interventions specialist/disorder

By measuring this indicator, is it possible to track specialists who need more then the average amount of diagnostic interventions for one disorder (in Dutch: Diagnose Behandel Combinatie). For example, two specialists diagnose patients with cataract. The first one uses a significant higher amount of diagnostic interventions then the second one, which can implicate that the first one is working inefficiently or something else is wrong. This observation can be achieved by using this indicator, whose business data model is shown in figure 11.

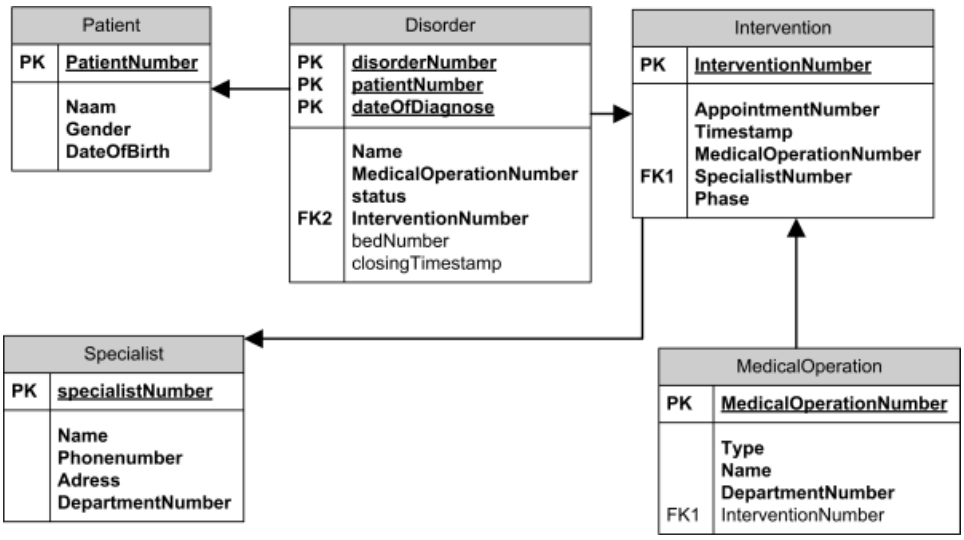

Figure 11: Amount of Diagnostic Interventions specialist/disorder

A specialist performs an intervention in order to diagnose a disorder. Each interventions belongs to a phase (triage, diagnostics, treatment, nursing, aftercare). If the amount of interventions is bigger then another specialist, then his process is not optimized. Of course other factors like the age of patients must be considered and therefore, must been grouped by in the select statement.

4.2.3 Diagnose Time

This indicator, shown in figure 12, measures the amount of time it takes from the first consult until a valid formation of the diagnose has been made.

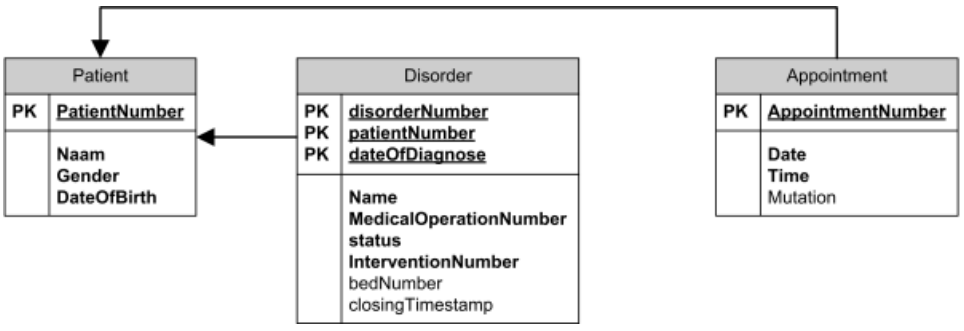

Figure 12: Diagnose Time

The moment of first contact is called “TimestampFirstContact”. From that moment, it is interesting to see how long it takes until the specialists have formed a final diagnoses.

4.3 Treatment

After the diagnostics phase, the patient undergoes one or more treatments in order to heal his/her complaints. It is possible to measure this treatment phase, by using the following indicators.

4.3.1 Intervening Time Diagnostics - Treatment

After successfully diagnosing the disorder at a certain time stamp, the patient goes to the next step of the process. The time it takes for the patient from leaving the diagnostics phase until entering the treatment phase is called the intervening time and the business data model is shown in figure 13. By using this indicator, one is able to measure the intervening times for disorders, group of patients, etc.

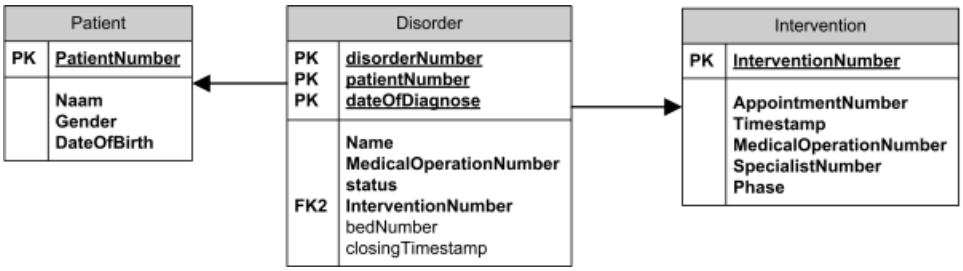

Figure 13: Intervening Time Diagnostics - Treatment

A patient suffers some kind of disorder. At a certain date, identified by a time stamp, the disorder has been diagnosed and the patient enters the treatment phase. The time it takes from leaving the diagnose phase and entering the treatment phase can be calculated by using an attribute called intervention. For each intervention, the phase it is part of is saved. When the date of the last intervention in the diagnostics and the first intervention in the treatment phase are known, it is possible to calculate the intervening time.

4.3.2 Amount of Interventions specialist/disorder

Just like in the diagnostics phase, it is possible to track specialists who need more then the average amount of therapeutical interventions for one disorder/dbc. Differences between the way specialists work can be found. For example, if one specialist uses two times as much therapeutical interventions as another specialist, something can be wrong. Have a look at figure 14 for the business data model.

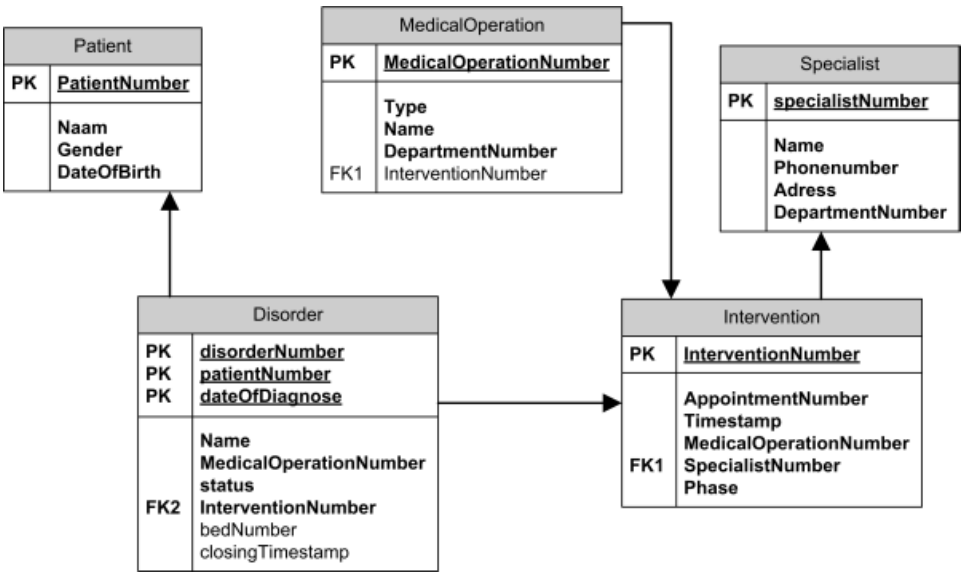

Figure 14: Amount of Interventions Specialist/Disorder

4.3.3 Undergoing a Medical Intervention twice or more

This indicator shows the percentage of patients who need to undergo a second (or more) medical intervention during the treatment phase for the same complaint due to an error at the first attempt. For example, when someone breaks his leg, it has to be set to the right angle. When later on, someone finds out that the leg is set to a wrong angle, the same medical intervention has to be done again. By using this indicator, it is possible to track the specialists who need more extra medical interventions then the average usage. It is also possible to determine for which disorders extra interventions are often needed, and therefore are inefficient and produces waste. Figure 15 shows the business data model for the indicator “Undergoing a Medical Intervention twice or more”.

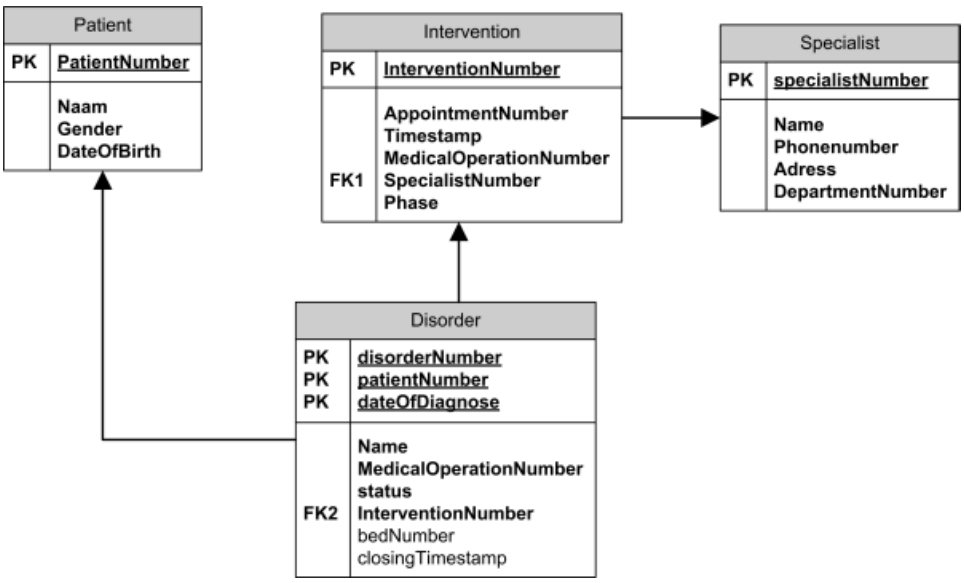

Figure 15: Undergoing a Medical Intervention Twice or More

A patient suffers a disorder. This disorder is treated through some interventions, made by a specialist. When a patient undergoes an intervention twice, this can mean a non-optimal efficiency and by using this indicator, one is able to reduce the amount of waste.

4.4 Nursing

When a patient is being treated or undergoes a treatment, he or she needs medical attention in the form of nursing. To determine the quality of this part of the total process, we need to measure the following indicators.

4.4.1 Intervening Time Treatment - Nursing

After a successful treatment, the time stamp of the last intervention belonging to the treatment phase is known. The next step is to enter the nursing phase. The time it takes for the patient from leaving the treatment phase until entering the nursing phase is called the intervening time. By using this indicator, shown at figure 16, one is able to measure the intervening times for disorders, group of patients, etc.

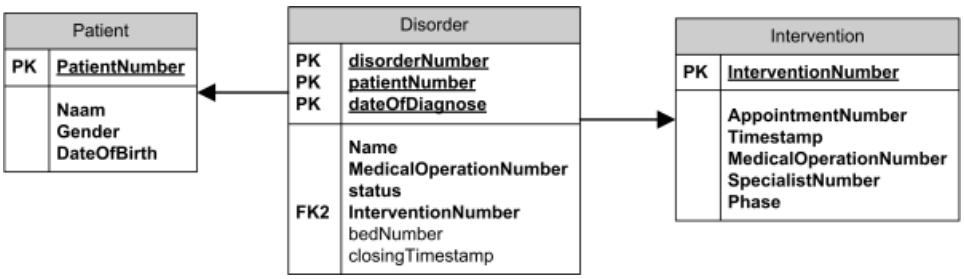

Figure 16: Intervening Time Treatment - Nursing

A patient suffers some kind of disorder. At a certain date (time stamp), the disorder has been treated and the patient enters the nursing phase. The time it takes from leaving the treatment phase and entering the nursing phase can be calculated by using an attribute called intervention. For each intervention, the phase it is part of is saved. When the date of the last intervention in the treatment and the first intervention in the nursing phase are known, it is possible to calculate the intervening time between both phases.

4.4.2 Bed occupation: Percentage Occupied Beds/Department

Patients make use of beds. Therefore, the way bed are being used in a hospital is crucial for optimized patient logistics processes. The indicator is called percentage occupied beds per department and enables the management to track the departments which beds are not optimized in usage. If the value is low, it is possible that the amount of beds available is to high (see figure 17).

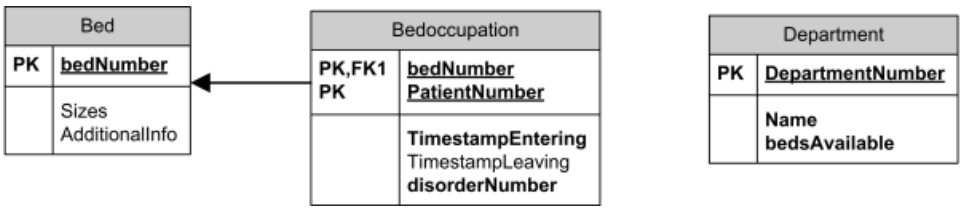

Figure 17: Bed occupation: Percentage Occupied Beds/Department

4.4.3 FTE's for Each Department

FTE is short for Full Time Equivalent and shows the involvement of a person in a treatment. An FTE of 1.0 means that the person is equivalent to a full-time worker, while an FTE of 0.5 signals that the worker is only half-time. In the nursing phase, it is important to measure the average amount FTE's each disorder requires. This can be accomplished by using the following indicator, shown as figure 18.

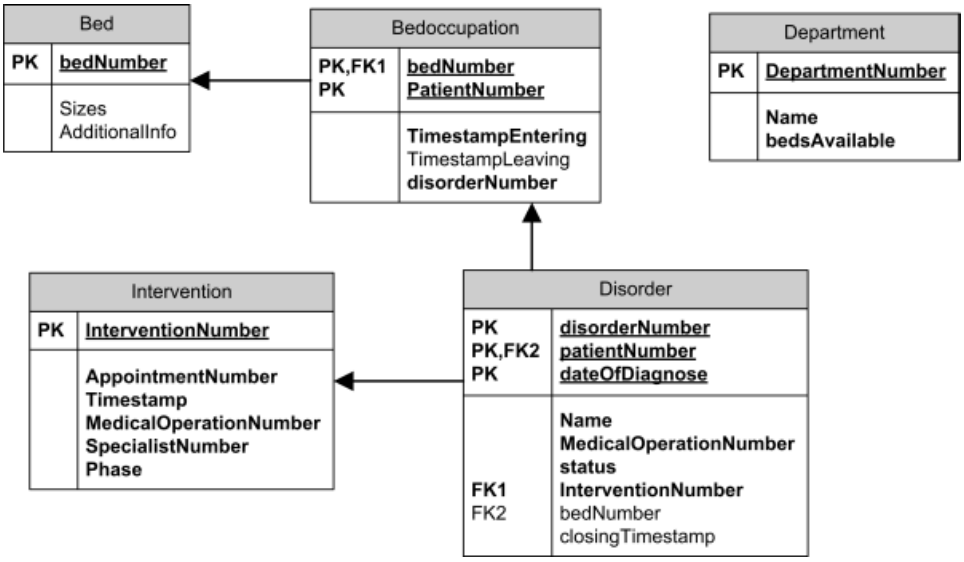

Figure 18: FTE's for Each Department

Every intervention a nurse executes is stored in the table 'intervention', together with the corresponding phase. This results in the ability to measure the average FTE for each group of disorders and/or department.

4.5 Aftercare

After completing the time a patient spends in the hospital, the last phase will become active. This phase takes care of the way the patient is treated at the moment he is no longer present in the hospital. The next three indicators belong to the aftercare phase.

4.5.1 Intervening Time Last Intervention - Closing Disorder

When a patient is discharged from the hospital, his or her disorder must been set to "closed" into the system. If the time stamp of the last intervention differences a lot with the time the disorder is closed, then the process includes waste. By using this indicator, it is possible to track this waste. See figure 19 for the business data model.

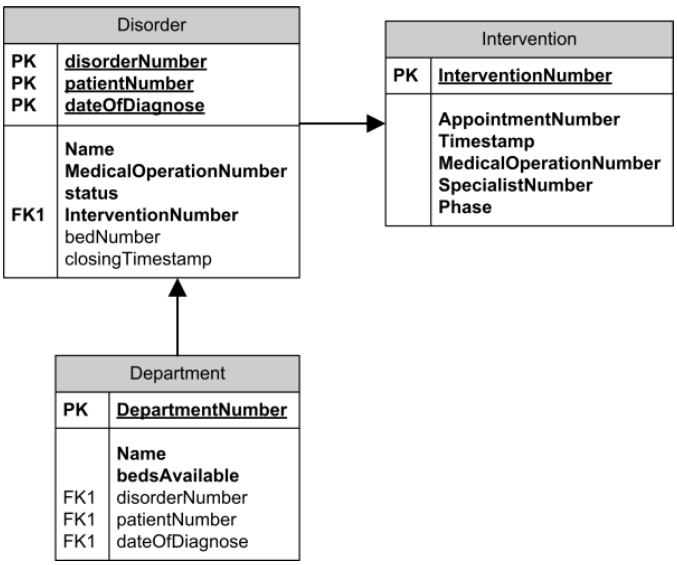

Figure 19: Intervening Time Last Intervention - Closing Disorder

The attribute "status" in the table disorder contains whether a disorder is still active or has already been closed. The date/time this happens is saved in the attribute "closingTimestamp". Together with the table intervention we can measure the time between closing a disorder and executing the last medical intervention.

4.5.2 Invoice Transaction Time

When a patient's disorder status is set to "closed" and the invoice is sent and paid, we can calculate the time it took between closing the disorder and receiving the money by using the business data model shown in figure 20.

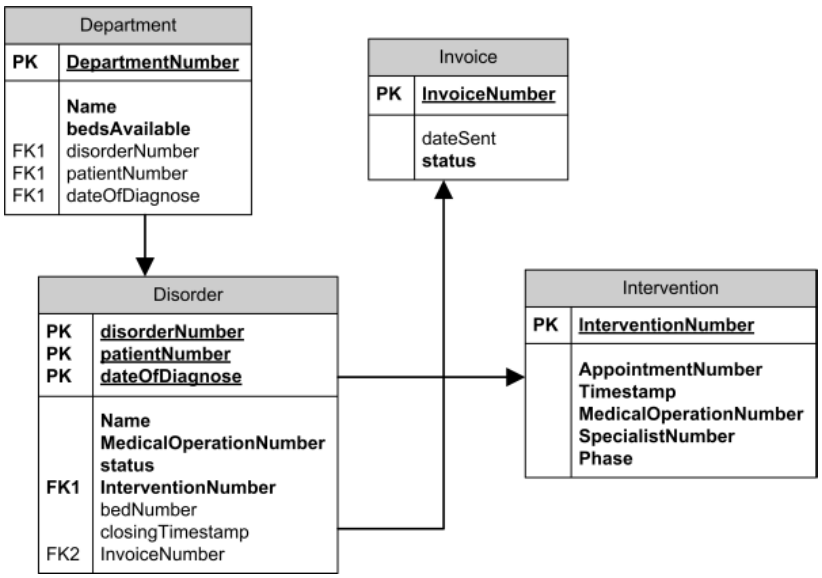

Figure 20: Invoice Transaction Time

Now we can see several financial values, including a list of departments which are late in sending the invoices and therefore, generating waste.

4.6 Overview of Indicators

This chapter shows an overview of the indicators which will be used in order to measure the quality of the processes belonging to patient logistics.

- Triage
    1. Waiting time
- Diagnostics
    1. Amount of re-admissions for each disorder
    2. Amount of diagnostic interventions specialist/disorder
    3. Diagnose time
- Treatment
    1. Intervening time diagnostics - treatment
    2. Amount of interventions specialist/disorder
    3. Undergoing a medical intervention twice or more
- Nursing
    1. Intervening time treatment - nursing
    2. Bed occupation: percentage occupied beds/department
    3. FTE's for each department
- Aftercare
    1. Intervening time last intervention - closing disorder
    2. Invoice transaction time

5 Linking Business Data Models - Data warehouse

5.1 Extracting the data from the Source

The next step in this thesis is to link the business data models which are introduced in the previous chapter with the current data warehouse, which is a part of the process of implementing Ariadne. The five-phase model from Bakker [3] (triage, diagnostics, treatment, nursing, aftercare) are measurable due to the 12 indicators which were set up in the previous chapter. To make the next step more clear, have a look at the following schema (figure 21), which represent this steps schematically.

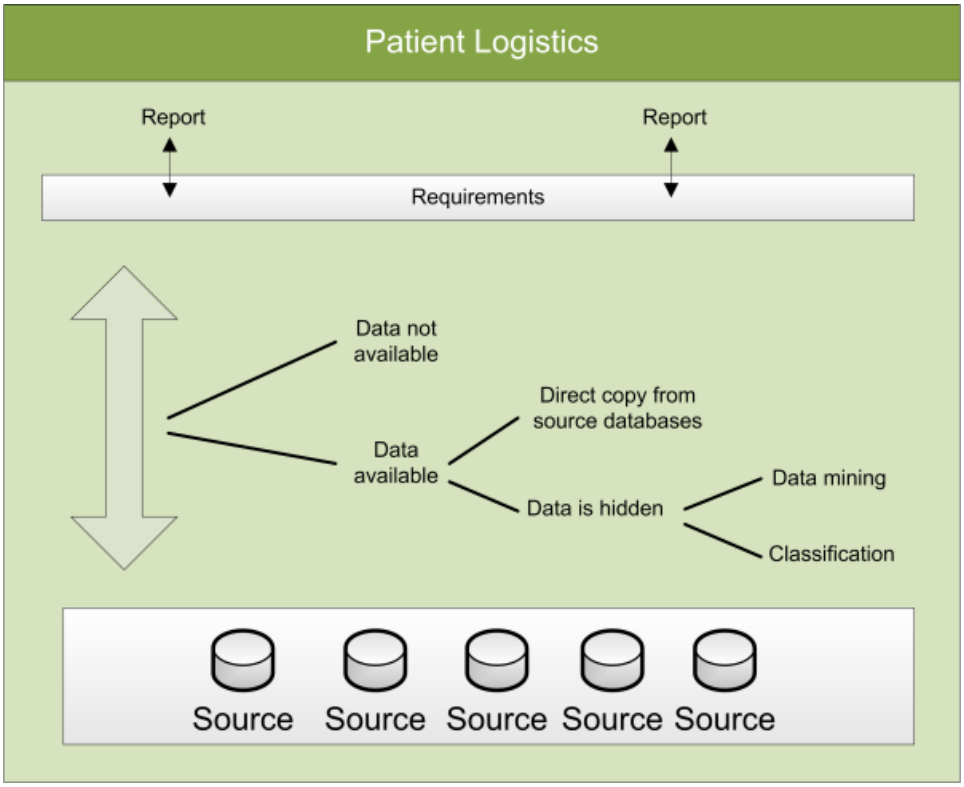

Figure 21: Linking Business Data Models - Data warehouse

The business data models explained in the previous chapters are part of the requirements. In other words they describe the amount of information and data which is needed in order to create the desired reports. Data can be derived from several sources. A possibility is that the data does exists in the source. This set of data can be divided into two categories: direct copy and hidden data. The "direct copy" data can be used directly from the source and by using a SQL query [2] the data is being revealed. The hidden data is

[2]Structured Query Language

available, but only after some data mining techniques like classification and clustering are used. This chapter describes both ways. The first paragraph shows how the required data is extracted from the source databases. In the next paragraph, we will try to discover new information, by applying data mining techniques on the data sets, originated from the source databases and extracted by SQL query's.

5.2 Direct Copy from Source Databases

As mentioned above, this chapter discusses how the required information can be extracted from the source databases and the data warehouse. Each sub chapter discusses an indicator, corresponding the ones of chapter 4.

5.2.1 Waiting Time

The indicator "waiting time" from the triage phase can be copied directly from the source with the following query:

```
USE CS_EZIS_SZ_49
go

SELECT      am.AFSPRAAKNR AS AppointmentID
,           am.DATUM AS Date
,           am.TIJD AS Time
,           aa.DATUM AS ActualDate
,           aa.TIJD AS ActualTime
FROM  AGENDA_MUTATIE AS am
INNER JOIN  AGENDA_AFSPRAAK AS aa
      ON    am.AFSPRAAKNR = aa.AFSPRAAKNR
AND am.DATUM = aa.DATUM
AND am.TIJD <> aa.TIJD
```

An example of the outcome can be found here:

AppointmentID	Date	Time	ActualDate	ActualTime
85061447	3-12-2003	10:40	3-12-2003	10:50
85065003	16-2-2004	12:40	16-2-2004	13:50
85077501	12-12-2003	13:51	12-12-2003	14:21
85064770	25-11-2003	16:00	25-11-2003	15:30
85077245	8-1-2004	11:49	8-1-2004	15:30
85141961	26-1-2004	14:00	26-1-2004	13:50
85010089	7-5-2004	15:20	7-5-2004	13:45
85078081	19-12-2003	20:00	19-12-2003	20:20
85062833	11-12-2003	14:30	11-12-2003	15:30

5.2.2 Amount of re-admissions for each Disorder

The amount of re-admissions for each disorder within a specific range of time can be found by using the following query. To keep the results reasonably limited, only the re-admissions in the year 2006 are shown.

```
select dbc1.patientnr AS PatientID
,      dbc1.hoofddiag AS DiagnosisID
,      dbc1.specialism AS SpecialismID
,      dbc1.begindat AS StartDateDBC1
,      dbc1.einddat AS EndDateDBC1
,      dbc2.begindat AS StartDateDBC2
,      dbc2.einddat AS EndDateDBC2
,      dbc1.dbcnummer AS DBCnumber1
,      dbc2.dbcnummer AS DBCnumber2
,      dbc1.episode AS ClinicalPathDBC1
,      dbc2.episode AS ClinicalPathDBC2
from (
      select patientnr,
             episode_dbcper.begindat,
             episode_dbcper.einddat,
             dbcnummer,
             episode_episode.episode,
             episode_dbcper.hoofddiag,
             episode_dbcper.specialism,
             episode_dbcper.behcode
      from   episode_dbcper
      inner join  episode_episode
            on episode_episode.episode = episode_dbcper.episode
      where year(episode_dbcper.einddat) = 2006
) dbc1
inner join (
      select patientnr,
             episode_dbcper.begindat,
             episode_dbcper.einddat,
             dbcnummer,
             episode_episode.episode,
             episode_dbcper.hoofddiag,
             episode_dbcper.specialism,
             episode_dbcper.behcode
      from  episode_dbcper
      inner join episode_episode
            on episode_episode.episode = episode_dbcper.episode
      where year(episode_dbcper.einddat) = 2006
) dbc2
on    dbc1.patientnr = dbc2.patientnr
-- Re-admission in the same specialism, diagnosis, treatment
and   dbc1.hoofddiag = dbc2.hoofddiag
and   dbc1.specialism = dbc2.specialism
-- Not the same DBC or Clinical Path
and   dbc1.episode <> dbc2.episode
and   dbc1.dbcnummer <> dbc2.dbcnummer
```

A part of the outcome looks like this:

PID	SpID	DID	StDateDBC1	EndDateDBC1	StDateDBC2	EndDateDBC2	DBC#1	DCB#2	DBC1route	DBC2route
195	8	22	11-7-2006	27-7-2006	7-9-2006	12-10-2006	493123	516440	347073	361500
195	8	21	7-9-2006	12-10-2006	11-7-2006	27-7-2006	516440	493123	361500	347073
225	7	101	5-4-2006	5-4-2006	6-7-2006	6-7-2006	449005	491117	318880	345944
225	7	101	6-7-2006	6-7-2006	5-4-2006	5-4-2006	491117	449005	345944	318880

5.2.3 Amount of diagnostic interventions specialist/disorder

To see which specialists need more then an average amount of interventions for a DBC, we need the following query:

```
USE CS_EZIS_SZ_49;
go

SELECT      b.casenr AS DBCNumber
,           b.patientnr AS PatientNumber
,           b.afdeling AS Department
,           b.uitvoerder AS Specialist
,           b.soort_beh AS TypeOfTreatment
,           b.datum AS DateOfIntervention
,           a.code AS InterventionID
,           b.bron AS Source
FROM  faktuur_verrsec a
JOIN  faktuur_verricht b
      ON     a.id = b.id
WHERE  b.afdeling <> 'dbc'
            AND    b.casenr <> ''
            AND    a.type_code = 'D'
            AND    year(b.datum) = 2007
            AND    month(b.datum) = 2
            AND    day(b.datum) = 14
ORDER BY 1,2,3
```

An example of the output:

DBC#	Patient#	Dpt	Spec.	TypeOfTreatment	DateOfIntervention	InterventionID	Source
423913	385038	REU	00501	PB	2007-02-14 00:00:00.000	10001	AGEN
423913	385038	REU	00501	PB	2007-02-14 00:00:00.000	90013	AGEN
426774	3079522	KL20	51002	KB	2007-02-14 00:00:00.000	70489	LBSZ
426774	3079522	KL20	51002	KB	2007-02-14 00:00:00.000	70610	LBSZ

5.2.4 Diagnose Time

The query which displays the time it takes between the first consult and the time a valid diagnose has been formed can be described as follows.

```
use SZAriadneDMSA
go

SELECT verr.PatientID,
       verr.OpnameID AS admissionID,
       verr.VerrichtingDatumID AS interventionDate,
       verrType.TariefGroepNaam AS TaxCodingName,
       (
        SELECT    min(VerrichtingDatumID) AS DateFirstContact
        FROM      dbo.FactVerrichting
        WHERE     PatientID = verr.patientID
        GROUP BY  PatientID
       ) as FirstContact

FROM  dbo.FactVerrichting verr
      INNER JOIN dbo.DimVerrichtingTypering verrType
      ON verrType.VerrichtingTyperingID = verr.VerrichtingTyperingID
WHERE verr.PatientID IS NOT NULL
ORDER BY PatientID, OpnameID
```

An example of the outcome:

PatientID	AdmissionID	InterventionID	TaxCodingName	DateFirstContact
16	164178	20070131	Laboratoriumonderzoeken	20070131
16	164178	20070201	Laboratoriumonderzoeken	20070131
16	164178	20070131	Kaarten	20070131
16	164178	20070131	Kaarten	20070131

```
16        164178      20070131       Laboratoriumonderzoeken     20070131
16        164178      20070131       Medisch spec. behandelingen 20070131
```

5.2.5 Intervening Time Diagnostics - Treatment

The time it takes from leaving the diagnostics phase until entering the treatment phase can be calculated by the following query.

```
use SZAriadneDMSA
go

with
FirstContact as (
 SELECT DimPatient.PatientNummer
 ,      MIN(DimDatum.Datum) AS MinDatum
 FROM   FactVerrichting INNER JOIN
 DimDatum ON FactVerrichting.VerrichtingDatumID = DimDatum.DatumID LEFT OUTER JOIN
 DimPatient ON FactVerrichting.PatientID = DimPatient.PatientID
 GROUP BY DimPatient.PatientNummer
),

LastContact as (
 SELECT DimPatient.PatientNummer
 ,      MAX(DimDatum.Datum) AS MaxDatum
 FROM   FactVerrichting INNER JOIN
 DimDatum ON FactVerrichting.VerrichtingDatumID = DimDatum.DatumID LEFT OUTER JOIN
 DimPatient ON FactVerrichting.PatientID = DimPatient.PatientID
 GROUP BY DimPatient.PatientNummer
)

SELECT      DimPatient.PatientNummer AS PatientNumber
,           factVerrichting.OpnameID AS admissionID
,           dimDatum.Datum as interventionDate
,           DimVerrichtingTypering.TariefGroepNaam AS TaxCodingName
,           fc.MinDatum AS FirstDate
,           lc.MaxDatum AS LastDate
,           DATEDIFF(day, fc.MinDatum, lc.MaxDatum) AS Duration
FROM  FactVerrichting LEFT OUTER JOIN
DimPatient ON FactVerrichting.PatientID = DimPatient.PatientID LEFT OUTER JOIN
DimVerrichtingTypering ON FactVerrichting.VerrichtingTyperingID =
DimVerrichtingTypering.VerrichtingTyperingID
LEFT JOIN DimDatum ON FactVerrichting.VerrichtingDatumID = DimDatum.DatumID

join FirstContact fc on fc.PatientNummer = DimPatient.PatientNummer
join LastContact lc on lc.PatientNummer = DimPatient.PatientNummer

ORDER BY DimPatient.PatientNummer, factVerrichting.OpnameID
```

An example of the outcome:

```
Pat# Addm# Interv.Date Taxcoding              FirstDate LastDate  Duration
9060 15381 11-7-2006   Laboratoriumonderzoeken 19-5-2006 23-3-2007 308
9060 15381 11-7-2006   Laboratoriumonderzoeken 19-5-2006 23-3-2007 308
9060 15381 11-7-2006   Laboratoriumonderzoeken 19-5-2006 23-3-2007 308
9060 15381 11-7-2006   InstellingstarievenAlg. 19-5-2006 23-3-2007 308
```

5.2.6 Amount of Interventions Specialist/Disorder

To see which specialists need more then an average amount of interventions for a DBC, we need the following query:

```
USE CS_EZIS_SZ_49;
go

SELECT      b.casenr AS DBCid,
            b.patientnr AS PatientID,
            b.afdeling AS Department,
            b.uitvoerder AS SpecialistID,
            b.soort_beh AS TypeOfTreatment,
            b.datum AS DateOfIntervention,
            a.code AS InterventionID,
            b.bron AS Source
FROM  faktuur_verrsec a
JOIN  faktuur_verricht b
      ON     a.id = b.id
WHERE  b.afdeling <> 'dbc'
            AND    b.casenr <> ''
            AND    a.type_code = 'D'
            AND    year(b.datum) = 2007
            AND    month(b.datum) = 2
            AND    day(b.datum) = 14
ORDER BY 1,2,3
```

An example of the outcome:

```
DBCid  Patient# Dep  Spec# Type DateOfIntervention      InterventionID Source
423913 385038   REU  00501 PB   2007-02-14 00:00:00.000 10001          AGEN
423913 385038   REU  00501 PB   2007-02-14 00:00:00.000 90013          AGEN
426774 3079522  KL20 51002 KB   2007-02-14 00:00:00.000 70489          LBSZ
426774 3079522  KL20 51002 KB   2007-02-14 00:00:00.000 70610          LBSZ
```

5.2.7 Undergoing a medical intervention twice or more

Amount of patients who need to undergo a second (or more) medical intervention during the treatment, executed by a specialist.

```
USE SZAriadneDMSA;
go

SELECT      verr.UitvoerendArtsID AS SpecialistID
,           verr.PatientID AS PatientID
,           verr.DBCID AS DBCid
,           count(verr.UitvoerendArtsID) as AmountOfInterventions
,           verr.opnameID AS AdmissionID
,           spec.SpecialismeNaam AS Specialism
FROM  FactVerrichting verr
LEFT JOIN dbo.DimSpecialisme spec
      ON    verr.DBCSpecialismeID = spec.SpecialismeID
WHERE DBCID!=''
GROUP BY verr.UitvoerendArtsID
       , verr.PatientID
       , verr.DBCID
       , verr.opnameID
       , spec.SpecialismeNaam
ORDER BY count(verr.UitvoerendArtsID) DESC
```

This query shows the amount of patients who undergo a medical intervention twice or more. An example of the output is as follows:

Spec#	PatientID	DBCID	AmountOfInterventions	AdmissionID	Specialism
4364	172782	337199	1422	122276	Heelkunde
4364	269681	62947	1418	102545	Heelkunde
4364	247357	188536	1117	134653	Interne geneeskunde
4364	395092	364214	1103	4611	Orthopaedie
4364	150737	28722	845	191869	Heelkunde

5.2.8 Intervening Time Treatment - Nursing

This indicator measures the time it takes from the last intervention in the treatment phase until the first intervention in the nursing phase and can be described as follows:

```
use SZAriadneDMSA
go

with
FirstContact as (
 SELECT DimPatient.PatientNummer
 ,      MIN(DimDatum.Datum) AS MinDatum
 FROM   FactVerrichting INNER JOIN
 DimDatum ON FactVerrichting.VerrichtingDatumID = DimDatum.DatumID LEFT OUTER JOIN
 DimPatient ON FactVerrichting.PatientID = DimPatient.PatientID
 GROUP BY DimPatient.PatientNummer
),

LastContact as (
 SELECT DimPatient.PatientNummer
 ,      MAX(DimDatum.Datum) AS MaxDatum
 FROM   FactVerrichting INNER JOIN
 DimDatum ON FactVerrichting.VerrichtingDatumID = DimDatum.DatumID LEFT OUTER JOIN
 DimPatient ON FactVerrichting.PatientID = DimPatient.PatientID
 GROUP BY DimPatient.PatientNummer
)

SELECT      DimPatient.PatientNummer AS PatientNumber
,           factVerrichting.OpnameID AS admissionID
,           dimDatum.Datum as interventionDate
,           DimVerrichtingTypering.TariefGroepNaam AS TaxCodingName
,           fc.MinDatum AS FirstDate
,           lc.MaxDatum AS LastDate
,           DATEDIFF(day, fc.MinDatum, lc.MaxDatum) AS Duration
FROM  FactVerrichting LEFT OUTER JOIN
DimPatient ON FactVerrichting.PatientID = DimPatient.PatientID LEFT OUTER JOIN
DimVerrichtingTypering ON FactVerrichting.VerrichtingTyperingID =
DimVerrichtingTypering.VerrichtingTyperingID
LEFT JOIN DimDatum ON FactVerrichting.VerrichtingDatumID = DimDatum.DatumID

join FirstContact fc on fc.PatientNummer = DimPatient.PatientNummer
join LastContact lc on lc.PatientNummer = DimPatient.PatientNummer

ORDER BY DimPatient.PatientNummer, factVerrichting.OpnameID
```

An example of the output:

```
Pat# Addm# Interv.Date Taxcoding              FirstDate LastDate  Duration
9060 15381 11-7-2006   Laboratoriumonderzoeken 19-5-2006 23-3-2007 308
9060 15381 11-7-2006   Laboratoriumonderzoeken 19-5-2006 23-3-2007 308
9060 15381 11-7-2006   Laboratoriumonderzoeken 19-5-2006 23-3-2007 308
9060 15381 11-7-2006   InstellingstarievenAlg. 19-5-2006 23-3-2007 308
```

5.2.9 Bed occupation: Percentage Occupied beds/department

To get a clear view of how optimized a department uses the assigned beds, one can use the following query:

```
use CS_EZIS_SZ_49
go

SELECT      afdeling.omschr AS Department
,           count(bedden.afdeling) AS AmountOfBeds
,           afdeling.bedaant AS CalculatedAmountOfBeds
,           afdeling.kostenpl AS TaxCodingName
FROM        opname_bedden bedden
LEFT JOIN   opname_afdeling afdeling
      ON    bedden.afdeling = afdeling.code
GROUP BY    bedden.afdeling
          , afdeling.omschr
          , afdeling.bedaant
          , afdeling.kostenpl
```

It shows the amount of beds actual in use for each department and the amount of beds which are allocated to the apartment. Have a look at an example of the output:

Department	AmountOfBeds	CalculatedAmountOfBeds	TaxCodingName
Neurologie	35	34	6611
Vaatchirurgie / Urologie	35	30	6612
Kindergeneeskunde	35	17	6620
Neonatologie	19	13	6625
Cardiologie	47	36	6621

5.2.10 FTE's for each department

To get a view of the degree how different disorders are being treated by the hospital's personnel, it is interesting to know how many FTE (Full Time Equivalent) for each department can be registered. it is possible to calculate this value by using the following query:

```
use SZAriadneDMSA
go

SELECT      verrichting.OpnameID AS AdmissionID
,           dbc.DiagnoseNaam AS Diagnosis
,           verrichting.PatientID AS PatientID
,           verrichting.VerrichtingDatumID AS DateOfIntervention
,           verrichting.DBCBeginDatumID AS DisorderBeginDate
,           verrichting.DBCEindDatumID AS DisorderEndDate
FROM  factVerrichting verrichting
LEFT JOIN   dimDBC DBC
  ON        dbc.dbcID = verrichting.dbcID
WHERE verrichting.OpnameID IS NOT NULL
  AND (verrichting.DBCBeginDatumID IS NOT NULL
   OR verrichting.DBCEindDatumID IS NOT NULL)
GROUP BY    verrichting.OpnameID
,                 dbc.DiagnoseNaam
,                 verrichting.PatientID
,                 verrichting.VerrichtingDatumID
,                 verrichting.DBCBeginDatumID
,                 verrichting.DBCEindDatumID
ORDER BY verrichting.OpnameID
```

An example of the output can be described as follows:

```
Adm# Diagnosis                  PatID  DateInt. DisorderBeginDate DisorderEndDate
2    gezond / geen pathologie   338635 20060929 20060929          20061010
10   benigne adnexafwijking     105778 20050808 20050807          20050926
10   benigne adnexafwijking     105778 20050807 20050807          20050926
10   overige (buik)kl algemeen  105778 20050807 20050807          20050807
12   angina pectoris, onstabiel 39774  20050108 20050107          20050429
```

5.2.11 Intervening Time Last Intervention - Closing Disorder

The time between the last intervention and the moment of closing the disorder can be categorized as waste. The following query gives a result of this information:

```
use SZAriadneDMSA;
go
with
FirstContact as (
 SELECT DimPatient.PatientNummer
 ,      MIN(DimDatum.Datum) AS MinDatum
 FROM   FactVerrichting INNER JOIN
 DimDatum ON FactVerrichting.VerrichtingDatumID = DimDatum.DatumID LEFT OUTER JOIN
 DimPatient ON FactVerrichting.PatientID = DimPatient.PatientID
 GROUP BY DimPatient.PatientNummer
),
LastContact as (
 SELECT DimPatient.PatientNummer
 ,      MAX(DimDatum.Datum) AS MaxDatum
 FROM   FactVerrichting INNER JOIN
 DimDatum ON FactVerrichting.VerrichtingDatumID = DimDatum.DatumID LEFT OUTER JOIN
 DimPatient ON FactVerrichting.PatientID = DimPatient.PatientID
 GROUP BY DimPatient.PatientNummer
),
patient as (
SELECT DimPatient.PatientNummer AS PatientNumber
,      factVerrichting.OpnameID AS admissionID
,      dimDatum.Datum as interventionDate
,      DimVerrichtingTypering.TariefGroepNaam AS TaxCodingName
,      fc.MinDatum AS FirstDate
,      lc.MaxDatum AS LastDate
,      DATEDIFF(day, fc.MinDatum, lc.MaxDatum) AS Duration
FROM FactVerrichting LEFT OUTER JOIN
     DimPatient ON FactVerrichting.PatientID = DimPatient.PatientID LEFT OUTER JOIN
     DimVerrichtingTypering
ON   FactVerrichting.VerrichtingTyperingID = DimVerrichtingTypering.VerrichtingTyperingID
     LEFT JOIN DimDatum ON FactVerrichting.VerrichtingDatumID = DimDatum.DatumID
join FirstContact fc on fc.PatientNummer = DimPatient.PatientNummer
join LastContact lc on lc.PatientNummer = DimPatient.PatientNummer
)
SELECT      DISTINCT faktuurnum AS InvoiceID
,           patientnr AS PatientID
,           debiteurnr AS debtID
,           bedrag AS Sum
,           patient.LastDate AS DateLastIntervention
,           faktuurdat AS DateOfInvoice
FROM             CS_EZIS_SZ_49.dbo.FAKTUUR_NOTAS
  LEFT JOIN      patient
  ON             patient.PatientNumber COLLATE SQL_Latin1_General_CP1_CI_AS =
CS_EZIS_SZ_49.dbo.FAKTUUR_NOTAS.patientnr
WHERE patient.LastDate <= faktuurdat
```

The following is an example of the outcome:

InvoiceID	patientID	DebtID	Sum	DateLastIntervention	DateInvoice
99080416	3057923	527	2722,8	2007-08-14 00:00:00.000	2002-03-31 00:00:00.000
98040429	848453	527	147,7	2007-06-04 00:00:00.000	1998-03-31 00:00:00.000
98564819	391003	527	21,42	2007-04-24 00:00:00.000	2000-04-30 00:00:00.000
98711509	973118	527	48,6	2004-12-29 00:00:00.000	2000-12-31 00:00:00.000
99522279	745166	745	166	2007-03-26 00:00:00.000	2004-04-21 00:00:00.000

5.2.12 Invoice Transaction Time

The time it takes from the last intervention until the invoice has been sent can be calculated by using the following query:

```
use CS_EZIS_SZ_49
go

SELECT      patientnr AS PatientNr
,           ontslagdat AS EndDateAdmission
,           ontslwijze AS ReasonEndingAdmission
,           faktuurdat AS InvoiceDate
,           afdeling AS Department
,           faktuurnum AS InvoiceID
FROM  FAKTUUR_VERRICHT
WHERE ontslagdat IS NOT NULL
```

The following is an example of the outcome:

```
PatientNR EndDateAdmission        Reason InvoiceDate Department InvoiceID
1744465   2004-01-06 00:00:00.000 0      2004-02-13  OPNN       99513644
1744465   2004-01-06 00:00:00.000 0      2004-02-13  OPNN       99513644
1744465   2004-01-06 00:00:00.000 0      2004-02-13  OPNN       99513644
407562    2004-01-12 00:00:00.000 0      2004-02-13  OPNN       99513646
```

6 Data Mining Explained

The growth of enormous database systems the last past decades has led to a the demand for a new, powerful tool for turning data into useful, task-oriented knowledge. In order to satisfy this need, researchers have been exploring ideas and developed methods in machine learning, pattern recognition, statistical data analysis, data visualization, and so on. A new research area called Data Mining and Knowledge Discovery was born. Data mining can be described as a variety of mathematical methods and software techniques used for finding patterns and regularities in sets of data.

The technique is first initialized in the 1960's and 1970's. But in a period of three decades, a lot of things changed. In the old days, computing power was very limited and expensive. Nowadays, computing power of regular desktop computers are similar to the power of the mainframes from back then, and even more. Another big change are the costs of data storage. Data storage is cheap and the information that can be mined from it very important. The accumulation of data also grew explosively in the past decades. Due to the introduction of electronic data gathering devices, an immense amount of data is available, ready to be mined. So, as a conclusion, the gathering of data is applied more and more. Therefore, the use and advantages of data mining and knowledge discovery grow in a similar way.

6.1 Knowledge Discovery in Databases (KDD)

A lot of research on knowledge discovery has been done and is still going on. The definition of knowledge discovery can be described as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [8]. But there are some issues. The fundamental input to a

discovery system is the raw data present in the source databases. The main problem of those sources is the fact that they come from the real world, including it's dynamic, incomplete, noisy and large character. Other concerns include whether the database contains adequate information for interesting discovery and how to deal with the overabundance of irrelevant information.

In today's database systems, content is always changing. Data can be time sensitive, and the discovery of new data is affected by the timeliness of data observations. Some data values such as social security numbers are timeless, but other values like someone's height or weight differ from time to time. Another important thing in discovering new data is the relevance of data. For example, an item of data is relevant to the current focus of discovery. When a patient database is being explored for interesting patterns of symptoms and diagnoses, non-medical data such as a patient's name and zip code are irrelevant. The presence or absence of values for relevant data attributes can effect the discovery of new data. Imagine, a patient was comatose at the time his diagnosis is made. This piece of information is so important that is does not allow the substitution of a default value: less important missing data can be defaulted.

When knowledge is discovered from a database, there are three assets we have to deal with: the form, the representation and the degree of certainty [8]. The form of a discovered piece of knowledge can be categorized by the type of data pattern it describes, namely interfield patterns and interrecord patterns. The first one are values of fields in the same record, for example a procedure called 'surgery' implies that the amount of days the patient is in the hospital equals or is greater then five. Interrecord patterns relate values aggregated over groups of records. An example is the fact that diabetic patients have twice as many complications as non diabetic patients. An example of interfield pattern searching is when a patient X undergoes several interventions in a range of data. When the time between the first and the last intervention is greater then five days, there is a chance that the patient has undergone an medical operation. For interrecord pattern searching, two or more tables are needed. If a patient needs more then the average amount of interventions for one disorder, he probably suffers diabetics. Have a look at figure 22, which is a schematic overview of these two types of data mining.

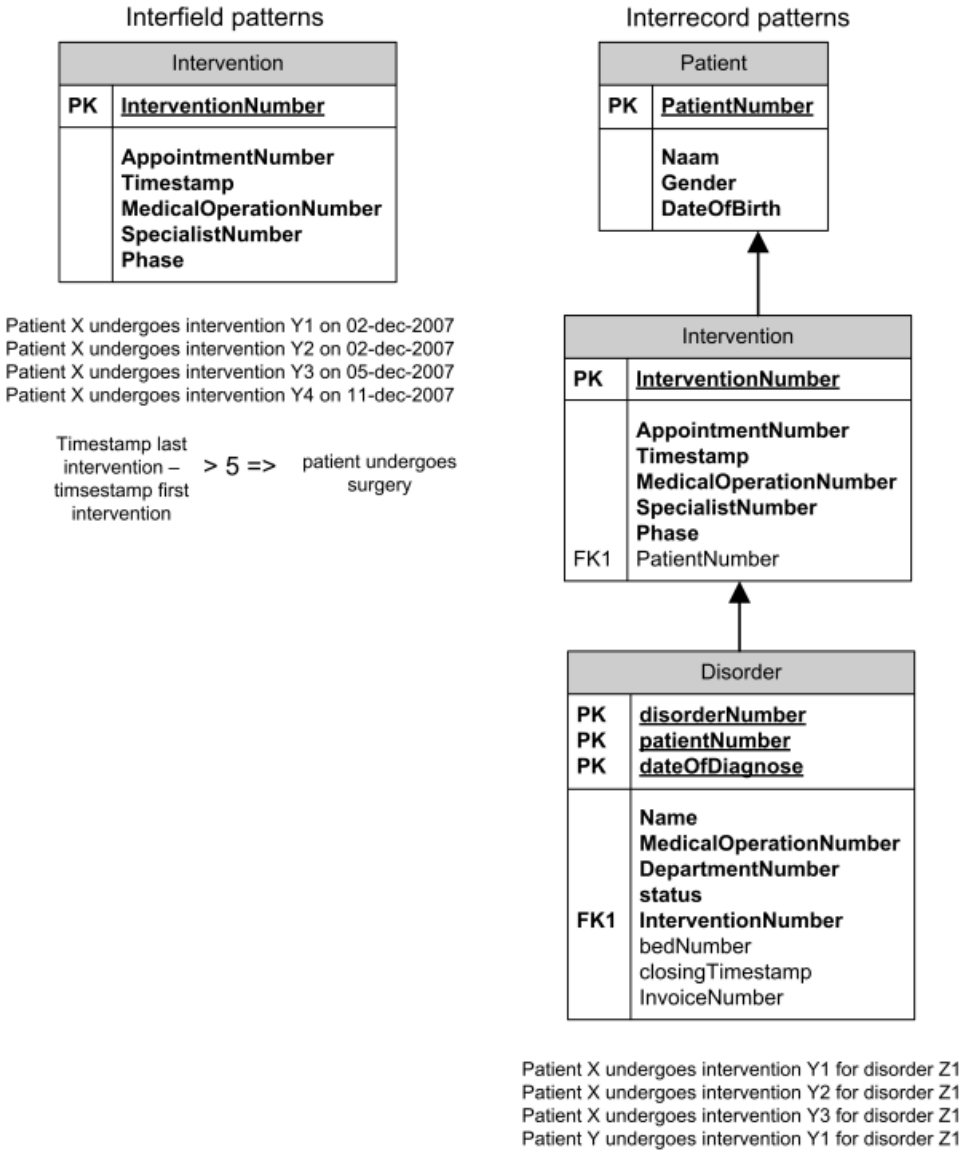

Figure 22: Interfield and Interrecord Data Mining

Discovery of interrecord patterns is a form of data summarization in time-dependent data, interrecord relationships can also identify interesting trends. By using discovery algorithms it is possible to extract knowledge from data. Knowledge discovery process is an interactive and iterative process, defined in nine steps [7]. For a schematic overview, see figure 23.

1 Developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer's viewpoint.

2 The next step is creating a target data set. This is done by selecting a data set or focusing on a subset of variables or data samples, on which discovery is to be performed.

3 Third is data cleaning and preprocessing. Basic operations like the removal of noise.

4 Reduction and projection. Finding features in order to represent the data, depending on the goal of the task.

5 Linking the goal of the knowledge discovery process to a specific data mining method. Examples are summarization, classification, regression and clustering.

6 The sixth step is to exploratory analyze the model and hypothesis selection. In other words, during this step, the data mining algorithms and selecting method(s) to be used for searching data are chosen. The process includes deciding which models and parameters might be appropriate and matching a particular data mining method with the overall criteria of the knowledge discovery processes.

7 Data mining: searching for patterns of interest in a particular representational form or a set of such representations.

8 Interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.

9 The last step is acting on the discovered knowledge by using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to people who are interested.

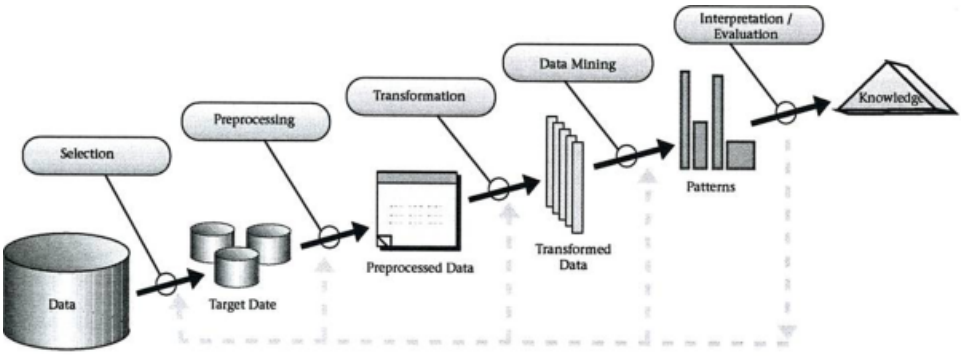

Figure 23: Nine steps of Knowledge Discovery

6.2 Data Verification and Discovery

Knowledge discovery is related to two goals [11]. The first goal is about verification of data and the second goal is about discovery. With verification, the system is limited to verifying the user's hypothesis. With discovery, the system tries to find new patterns. This discovery goal can furthermore be divided into two categories called prediction and description. Prediction tries to find patterns for predicting the future behavior of some entities, while description tries to find patterns for presentation to a user in a human-understandable form. In this master thesis, We will use the discovery-oriented data mining technique. For example, imagine a database containing several tables. One table records transactions in a supermarket. Another table called "items" records the products each transaction includes. So, one customer generates one transaction and this transaction generates an X-amount of items belonging to that transaction. From every amount of a customers who buy product A, an amount of b also buys product B. We now speak of an association rule a - b, with a probability of b/a. Example: it could be that 80% of the customers who buy butter and cheese, also buy bread. By using a fuzzy logic table, association rules are being generated. Have a look at the table from our supermarket example.

	1	2	3	4	5	6	7	8	9	ITEM
1	1	1	0	0	1	1	1	1	0	
2	1	0	1	0	0	0	0	1	1	
3	0	1	1	0	1	0	1	0	0	
4	1	0	1	0	1	1	0	1	1	
5	0	0	0	0	1	0	0	0	0	
6	0	1	0	0	1	0	1	0	0	
TRANSACTION										

When we look at the table, it becomes clear that the combination 2,5,7 is being sold by a minimum of 50% of the customers (transactions). The supermarket can leap on to this behavior of the customer by, for example, placing the products in one shelf or combine product into one special discount.

6.3 Data Mining Methods

The data mining component of the above mentioned nine steps of knowledge discovery involves repeated iterative application of particular data mining methods. This chapter presents an overview of the primary goals of data mining and a description of the methods used to address these goals, together with a description of the data mining algorithms that support these methods. Data mining is about fitting models to or determining patterns from a large amount of (observed) data, mostly originating from database systems [19]. The fitted models play the role of inferred knowledge, because only a human being can judge about whether the models reflect useful or

interesting knowledge. Two primary mathematical formalisms are used in model fitting, the statistical and the logical approach. The statistical approach allows non-deterministic effects in the model, whereas a logical model is purely deterministic. This thesis uses the statistical approach to data mining, which is the most widely used way of knowledge discovery and uses statistic functions like classification, clustering and regression. Have a look at figure 24, which shows a two-dimensional data set, consisting of 23 cases [7].

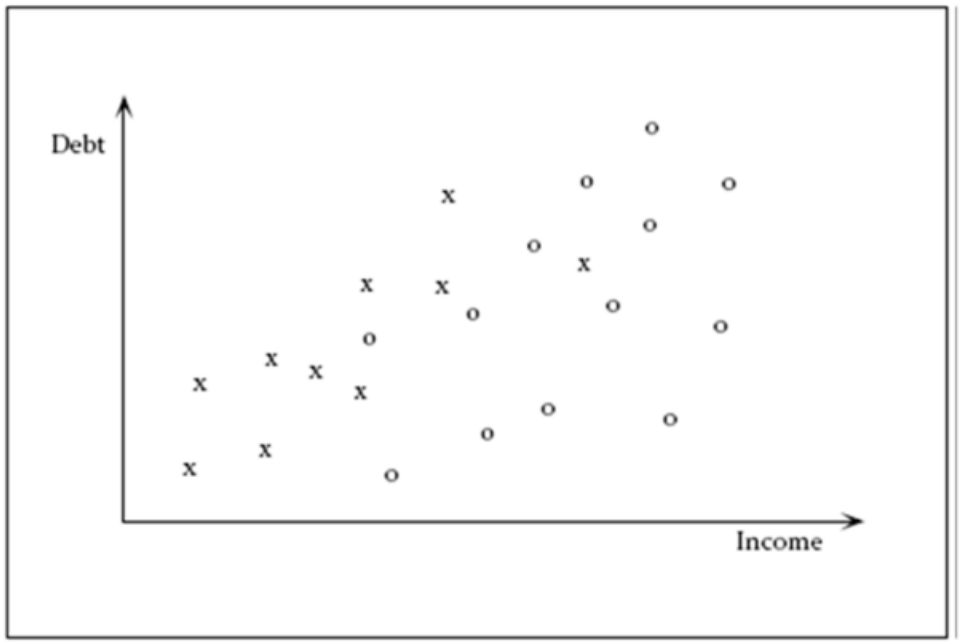

Figure 24: Example two-dimensional data set

Each point on the graph represents a person who has been given a loan by a particular bank at some time in the past. The horizontal axis shows the income of the person, the vertical axis shows the total personal debt of the person. The processed data has been categorized into two groups. The X's represents persons who have defaulted (unpaid loans for example) on their loan and the O's represent persons whose loans are in good status with the bank. By using this data set, one is able to generate useful information for the bank. By using this information, the bank can make better decisions whether someone is qualified to receive a loan.

As mentioned before, the two primary goals of data mining in practice are prediction and description. Prediction uses variables or fields from the database in order to find new information or future values of other variables of interest, while description focuses on finding human-interpretable

patterns describing the data. The goals of prediction and description can be achieved by using a variety of particular data mining methods.

6.3.1 Classification

By using classification, one is able to map a data item into one of several predefined classes [Weis]. Examples of classification methods used as part of knowledge discovery applications include the classification of trends in financial markets and the automated identification of objects of interest in large image databases. Figure 25 shows a simple partitioning of the previous mentioned loan data into two relevant classes.

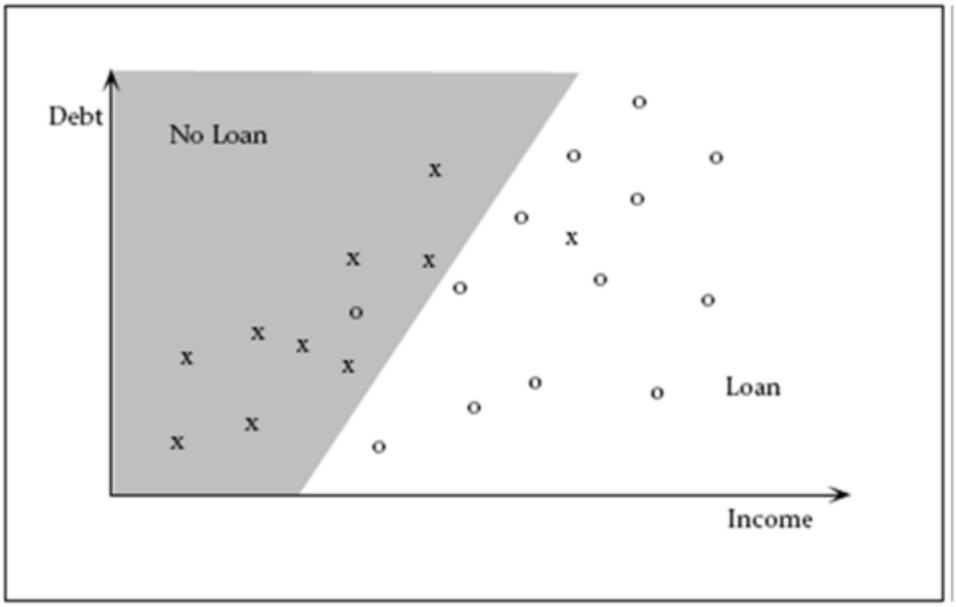

Figure 25: Example of Linear Classification Boundary

By using this classification, the bank is able to automatically decide which customers are qualified for getting a loan and which customers are not. Classification involves finding rules that partition the data into disjoint groups, for the classification is the training data set, whose class labels are already known. Classification analyzes the training data set and constructs a model based on the class label, and aims to assign a class label to the future unlabeled records. Since the class field is known, this type of classification is known as Supervised learning. A set of classification rules are generated by such a classification process, which can be: to classify future data and develop a better understanding of each class in the database. This is also a form of supervised learning.

There are several classification discovery models [9]. They are: the decision tree, neural networks, genetic algorithms and the statistical models like linear/geometric discriminates. Consider the following example. The domestic flights in a country were at one time only operated by a certain airline company called X. Recently, many other private airlines began their operations for domestic travel. Some of the customers of X started flying with these private airlines and, as a result X lost these customers. Let us assume that X wants to understand why some customers remain loyal while other leave. Ultimately, the airline wants to predict which customers it is most likely to lose to its competitors. Their aim to build a model based on the historical data of loyal customers versus customers who have left. This becomes a classification problem. It is a supervised learning task as the historical data becomes the training which is used to train the model. The decision tree is the most popular classification technique.

6.3.2 Regression

The regression method can be described as a learning function that maps a data item into a real-valued prediction variable. There are many regression applications. For example, predicting the amount of biomass present in a forest given remotely sensed microwave measurements. Another example is estimating the probability that a patient will survive given the results of a set of diagnostic tests or, a third example, giving a forecast of the customer's demand for a new product as a function of advertising expenses. Figure 26 is an example of a simple linear regression tool and shows the total debt as fitted as a linear function of income.

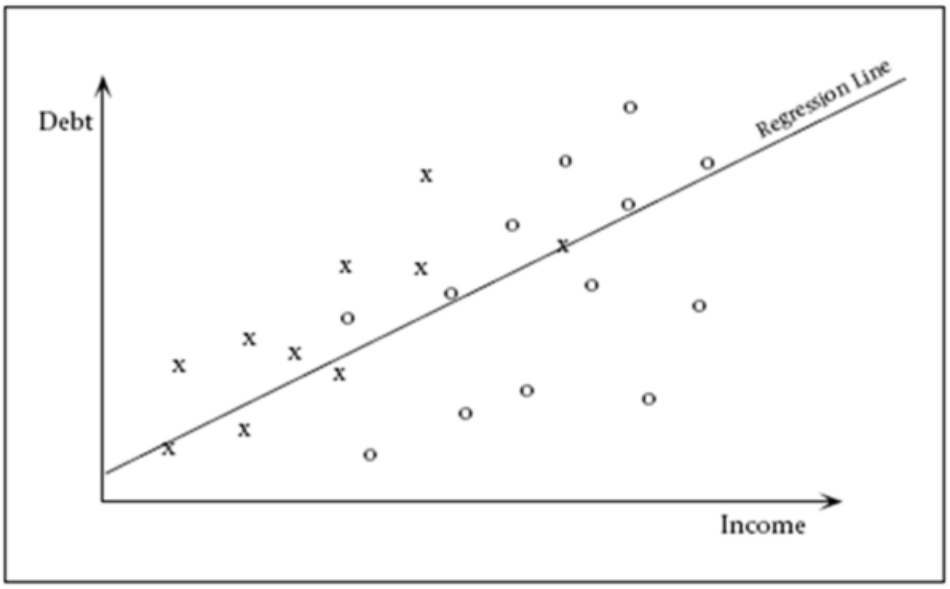

Figure 26: A simple linear regression tool

6.3.3 Clustering

Clustering is part of the category descriptive methods and enabled the user to identify a finite set of categories or clusters to describe the data. It is a method of grouping data into different groups, so that each group shares similar trends and patterns. These categories can be exclusive or consist of a richer presentation, such as hierarchical or overlapping categories. An example of clustering in knowledge discovery is the discovery of homogeneous subpopulations for consumers in marketing databases. When we apply the clustering technique on our example of loans and debts, results like figure 27 will be generated. Clustering according to similarity is a concept which appears in many disciplines. If a measure of similarity is available then there are a number of techniques for forming clusters. Another approach is to build set functions that measure some particular property of groups this latter approach achieves what is known as optimal partitioning.

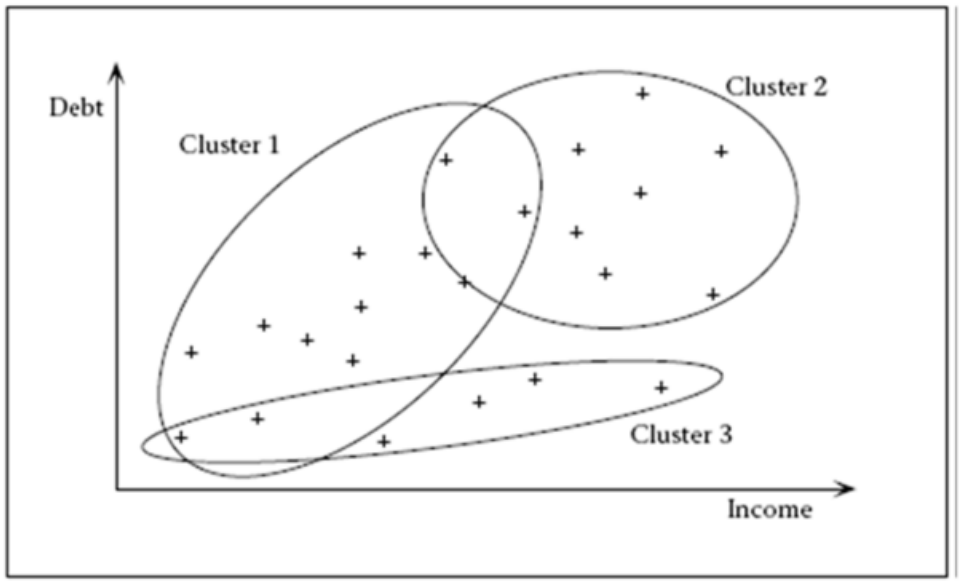

Figure 27: Clustering applied

Note that the original classification of the X's representing persons who have defaulted and the O's representing persons whose loans are in good status is deleted, because the classes are no longer "known" at the moment, hence we are using the clustering method to find these classes. Objectives of clustering are the uncovering of natural groupings and initiating a hypothesis about the data. For example, a retailer may want to know where similarities exist in his customer base, so that he can create and understand different groups. He can use the existing database of the different customers or, more specifically, different transactions collected over a period of time. The clustering methods will help him in identifying different categories of customers. During the discovery process, the differences between data sets can be discovered in order to separate them into different groups, and similarity between data sets can be used to group similar data together.

6.3.4 Dependence Modeling and Summarization

The dependency modeling technique tries to find a model that describes dependencies between variables. There are two levels possible. First, called the structural level of the model. This level describes which variables are dependent on each other, often in a graphic form. The second method, the quantitative level, specifies the strengths of the dependencies using some numeric scale. The summarization method involves a way of finding a compact description for a subset of data. An example of summarization is the discovery of functional relationships between variables. Summarization techniques are often applied to interactive exploratory data analysis and automated re-

port generation.

6.3.5 K-means Algorithm

Live mentioned in the previous chapter, there are a couple of clustering algorithms. Clustering can be described as the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets called clusters). In every subset, the data shares some common property. In this thesis, I use the K-means algorithm (MacQueen, 1967), which is one of the most simple unsupervised learning algorithm that solves the well known clustering problem. The algorithm follows a simple and easy way to classify a given data set through a certain number of clusters fixed a priori. Assume an amount of K clusters. The main idea is to define k centroids, one for each cluster. These clusters must be placed in a smart way, because the different locations causes different results. The best way to achieve this is to place them as far as possible away from each other. The next step is to take a watch point belonging to a given data set and associate it to the nearest point. When no point is pending, the first step is completed and the first cluster has been formed. At this point, the algorithm re-calculates k new centroids as the centers of the clusters resulting from the previous step. Now the algorithm is situated in a loop. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. Now, the quality of the clustering can be calculated by the following error function, see figure 28:

$$E = \sum_{j=1}^{k} \sum_{i_l \in C_j} \mid i_l - w_j \mid^2$$

Figure 28: Function which calculates the quality of the Clusters

The appropriate choice of k is problem and domain dependent and generally a user tries several values of k. Assuming that there are n different patterns, each of dimension d, the computational cost of a direct k-means algorithm per iteration (as part of the above mentioned loop) can be described intro three parts:

- The time required for the first “for” loop in figure 28 is O(nkd).
- The time required for calculating the centroids (second “for” loop in figure 28) is O(nd).
- The time required for calculating the error function is O(nd).

Figure 29 shows the algorithm in pseudo code.

function Direct-k-means()

Initialize k prototypes $(w_1, \ldots, w_k)$ such that $w_j = i_l,\ j \in \{1, \ldots, k\},\ l \in \{1, \ldots, n\}$

Each cluster C_j is associated with prototype w_j

Repeat

for each input vector i_l, where $l \in \{1, \ldots, n\}$, *do*

Assign i_l to the cluster C_{j*} with nearest prototype w_{j*} (i.e., $\mid i_l - w_{j*} \mid \ \leq \ \mid i_l - w_j \mid,\ j \in \{1, \ldots, k\}$)

for each cluster C_j, where $j \in \{1, \ldots, k\}$, *do*

Update the prototype w_j to be the centroid of all samples currently in C_j, so that $w_j = \sum_{i_l \in C_j} i_l / \mid C_j \mid$

Figure 29: The K-means algorithm in pseudo code

The following two images (figure 30 and 31) demonstrate the k-means clustering algorithm in action, for a two-dimensional case. The initial centers are generated randomly to demonstrate the stages in more detail.

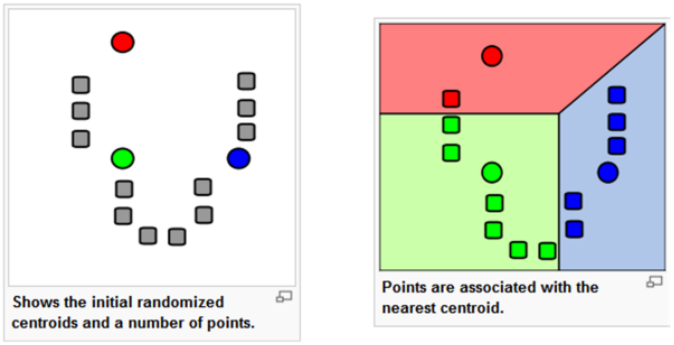

Figure 30: The first two steps of the K-mean algorithm

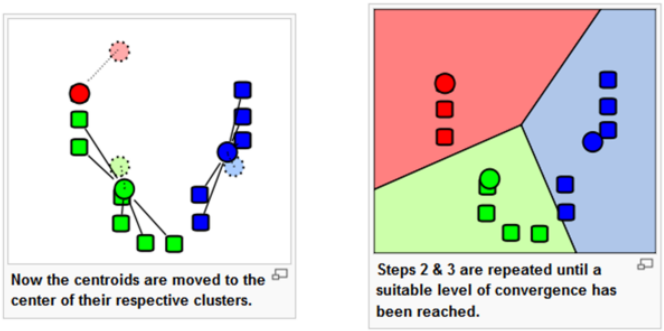

Figure 31: Steps three and four of the K-mean algorithm

7 Logistic-Based Patient Grouping

Present-day health care shows a growing demand for the coordination of patient care and logistics [14]. This is needed especially in those cases in which hospitals have structured health care into specialty-oriented units or like nowadays, when governments forces medical centers to use diagnostic treatment combinations (DBC's, short for "Diagnose Behandel Combinaties"). In this thesis, we will investigate the possibility of building an alternative, logistic-driven clustering and classification system for medical multi-disciplinary patients with the aid of machine learning techniques.

Patients who require the involvement of different specialties are hardly a new phenomenon in the Dutch health care. Commonly spoken, one can say that because of the increasing specialization of doctors within the hospital and an aging population this group of patients is increasing. Recent studies [1] in the Netherlands show that approximately 65% of the patients visiting a hospital are multi-disciplinary. This tendency can be compared with cross-marketing, where, for example, customers that purchase tires and auto accessories also get automotive services done. Therefore it would be nice if the process of providing medical care to the group of multi-disciplinary patients is improved. In this chapter, I will try to manage this. The first two chapters are about finding meaningful groups (clusters) of records, originating from the indicators we have discussed in chapter 6.

In order to apply the data mining algorithms, we need a software tool. There is a variety of tools available, for instance the tool Weka [24] [3]. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. The big advantage of Weka is that the above explained K-mean algorithm is part of the standard installation of Weka.

7.1 Amount of Medical Interventions

This test tries to find out if it is possible to find clusters of specialists, linked to a certain specialism, who use a certain amount of interventions in order to treat a disorder. When this test succeeds, the gathered information enables the strategic and operational management to adapt on the current situation with the aim of improving the processes.

Within Weka, we import the comma separated value (CSV) dataset called

[3] http://www.cs.waikato.ac.nz/ ml/weka/

"Undergoing a medical intervention twice or more". This set contains 30535 records, each containing six attributes. After loading the dataset, we choose to open the Weka Experimenter, which enables us to apply data mining algorithms on the dataset. We want to investigate if it is possible to automatically create clusters of specialists, identified by unique specialist ID numbers, who approximately need the same amount of interventions within a specialism. After choosing the K-Means algorithm, we have to identify the attributes which need to be investigated by the algorithm. We choose SpecialistID, AmountOfInterventions and Specialism; the remaining attributes can be ignored and therefore are added to the ignore list. The results are shown in the next sub chapter.

7.1.1 Output: Amount of Medical Interventions

```
=== Run information ===
Scheme:       weka.clusterers.SimpleKMeans -N 6 -S 10
Relation:     07_undergoing_a_medical_intervention_twice_or_more
Instances:    30535
Attributes:   6
              SpecialistID
              AmountOfInterventions
              Specialism
Ignored:
              PatientID
              DBCid
              AdmissionID
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======
Number of iterations: 16
Within cluster sum of squared errors: 23428.990723679384

Cluster centroids:

Cluster 0
Mean/Mode:  spec4364  71.7013 Gastro-enterologie
Std Devs:   N/A       35.3739 N/A
Cluster 1
Mean/Mode:  spec4709  31.3789 Interne geneeskunde
Std Devs:   N/A        8.0889 N/A
```

```
Cluster 2
Mean/Mode:  spec4364 217.4802 Interne geneeskunde
Std Devs:   N/A      123.142  N/A
Cluster 3
Mean/Mode:  spec4364  38.3528 Interne geneeskunde
Std Devs:   N/A       15.1255 N/A
Cluster 4
Mean/Mode:  spec4364  49.0605 Kindergeneeskunde
Std Devs:   N/A       15.1274 N/A
Cluster 5
Mean/Mode:  spec219  46.8653 Interne geneeskunde
Std Devs:   N/A       25.9066 N/A

Clustered Instances

0       4383 ( 14%)
1       2774 (  9%)
2        783 (  3%)
3      16887 ( 55%)
4       4179 ( 14%)
5       1529 (  5%)

Class attribute: SpecialistID
Classes to Clusters:

     0     1     2     3     4     5  <-- assigned to cluster
  2182 10381  4279   765  1447  5538 | spec4364
    65   485   190    23    72   316 | spec219
     1     4    12     1     0     1 | spec3173
     2     3     7     2     1     2 | spec2334
     0     7     4     1     0     1 | spec1072
     0     3     5     2     1     4 | spec2050
     1     6     6     1     0     2 | spec870
     1     1    64     0     0     0 | spec3375
     1     2     2     1     0     2 | spec255
     0    54     1     1     0     0 | spec1590
    58   407   120     5    81    26 | spec4709
     3    76     2     2     2     3 | spec5136
    32   240   142     7    14    54 | spec5190
     2     3    11     1     0     7 | spec5008
     3    11     1     1     0     0 | spec54
     0   157     2     8     0     0 | spec1491
     0    91     0     4     0     0 | spec3616
     5    24     5     1     2     1 | spec1160
```

```
    0     0     0     0     4     0 | spec5000
    0     0    70     0     0     0 | spec3813
    3    25     1     1     1     1 | spec3756
    3    65     6     2     2     6 | spec2991
    1    18     4     0     1     0 | spec1612
    1    11     0     1     0     0 | spec4168
    2    63     1     0     4     2 | spec5074
    0    67     0     1     0     0 | spec456
    6    52     5     0     1     8 | spec398
    0    69     0     1     0     0 | spec1527
    0     1    41     0     0     0 | spec961
    1     0     0     0     0   137 | spec4588
    0     7     0     1     0     0 | spec4363
    .     .     .     .     .     .       .
    .     .     .     .     .     .       .
    .     .     .     .     .     .       .

Cluster 0 <-- spec4778
Cluster 1 <-- spec4364
Cluster 2 <-- spec5190
Cluster 3 <-- spec1491
Cluster 4 <-- spec4709
Cluster 5 <-- spec219
```

As the output above shows, the algorithm tries to generate clusters of the input data. It finds six cluster centroids, starting with the specialism "Gastroenterologie". Figure 32 displays the results in a graph. It is easy to see which clusters are found, namely the specialists, operating under a certain specialism, doing an x amount of interventions. Please see chapter 7.3 for an explanation of why this outcome is obvious.

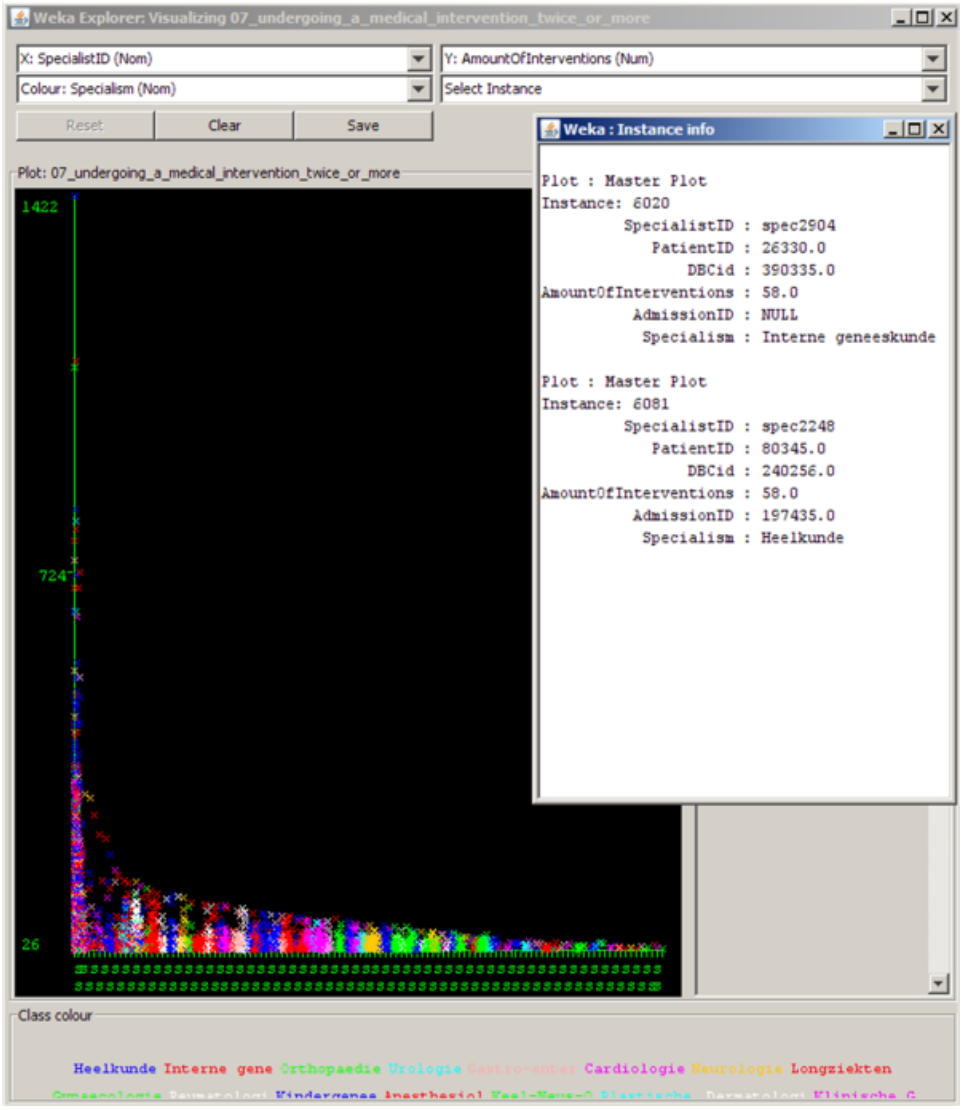

Figure 32: Output of clusters test 1

7.2 Duration of Treatment

The second test discusses the test set "Intervening Time Treatment - Nursing" and it tries to generate clusters of tax coding (expertise) areas in order to give an expected duration of a disorder within an tax coding area. Have a look at the output of the experimenter in Weka.

7.2.1 Output: Duration of Treatment

```
=== Model and evaluation on training set ===
```

```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 29166.62442140495

Cluster centroids:

Cluster 0
Mean/Mode:  Laboratoriumonderzoeken 1064.2342
Std Devs:   N/A        143.9715
Cluster 1
Mean/Mode:  Laboratoriumonderzoeken  492.4702
Std Devs:   N/A        216.7187

Clustered Instances

0      44196 ( 67%)
1      21339 ( 33%)

kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 1107.6244214062704

Cluster centroids:

Cluster 0
Mean/Mode:  1064.2342
Std Devs:    143.9715
Cluster 1
Mean/Mode:   492.4702
Std Devs:    216.7187

Clustered Instances

0      44196 ( 67%)
1      21339 ( 33%)

Class attribute: TaxCodingName
Classes to Clusters:
```

```
    0     1  <-- assigned to cluster
 9083  2999 | Kaarten
 1623   965 | Beeldvormende diagnostiek
 2187   867 | Tariefgroep 14
24200 13276 | Laboratoriumonderzoeken
 3530  1192 | Medisch specialistische behandelingen
  332   137 | Klinisch tarief
 2502  1618 | Instellingstarieven algemeen
   84    12 | Nucleaire geneeskunde
  186   102 | Pathologie
   87    42 | Ligdagen
  171    30 | Tariefgroep 15
  191    45 | Tariefgroep 0
   14    44 | Kaakchirurgie
    0     2 | Tariefgroep 16
    5     8 | Inlichtingen en rapporten
    1     0 | NULL

Cluster 0 <-- Laboratoriumonderzoeken
Cluster 1 <-- Kaarten

Incorrectly clustered instances : 38336.0  58.497  %
```

Just like the previous test, the results of this test are shown in figure 33. The graph shows a clustering in area's of specialties on the x-axis which form homogeneous groups by looking at the amount of interventions. The colors indicate the disorders, identified by a unique identification number. By using the clusters, it is possible to get information about how long the treatment of a certain disorder will take. The explanation of chapter 7.3 is also applicable on this indicator.

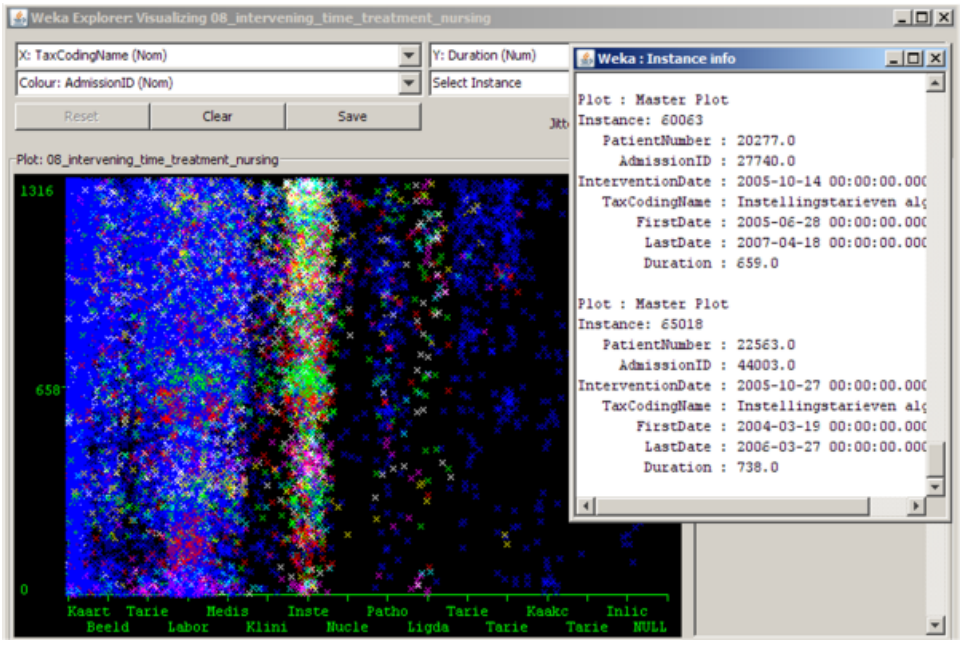

Figure 33: Output of clusters test 2

7.3 Inadequacy of Data

After evaluating the output, it becomes clear that the test set did not meet the required qualities of a good test set. The amount of records was sufficient, but during the formulation of the indicators in chapter 4, we emphasized on attributes containing dates. For example, the indicator "Waiting time" focuses on the time it takes for a patient between entering the hospital and his or her first treatment. The query which generates the data delivers attributes like AppointmentID, AppointmentDate, AppointmentTime, ActualDate and ActualTime. This information can be of great use in strategic and operational planning, but unfortunately not for data mining algorithms. The reason of this is that it makes no sense to find patterns from two date fields. It is possible to use date attributes, but they need to have some kind of meaning. For example, if we have a date field which shows the date of a treatment, then there is no use of data mining. However, if we can determine how long a treatment takes from the date of admission, we can use data mining to calculate the average time it takes until the treatment starts. By using clustering/classification it would be possible to discover patterns in this information. For example: patients who undergo a treatment with ID x and suffer symptoms with ID y1, y2 and y3 have an average waiting time of Z.

As a conclusion we can say that it would be of more value if we could

adapt our indicators to ones which are focused on the time between two dates. This would result in the ability to apply clustering and/or classification on this data. Due to a limited amount of time in writing this thesis, I have decided to apply the two small tests of data mining techniques from the previous sub chapters.

8 Conclusion

By writing this thesis, we tried to create a way how hospitals can improve the quality of the processes, focused on patient logistics. But in order to improve processes, one first needs to contextualize the subject. By developing a framework which enables us to measure the phenomenon patient logistics, it is possible to improve it. As a consequence of the obligatory implementation of the new way of health care, hospitals and other medical centers are forced to optimize their processes. The so called DBC's (Diagnose Behandeling Combinatie) are the ways of how medical care is financed. Medical centers are forced to implement care in digestible chunks, which results in the need of optimizing the processes. On this way, the free market systems is being stimulated. In the past, each hospital had it's own price for, for example, removing someone's meniscus. Now the Dutch government forces the medical centers to treat it's patient for a fixed price, by creating a standard set of marginal values. By introducing DBS's, for every disorder the patient pays a fixed price for the total amount of care he or she uses. On this way, the government stays responsible for the quality of health care and assures that medical care is accessible for everyone. It is clear that the need for optimized processes is getting higher due to this system. By using this thesis' approach, a hospital can improve it's processes, with the focus on patient logistics.

The research question was about finding an approach to operational management with the purpose to reason about an organization with the aim of improving the business processes. The first subquestion about the definition study of patient logistics is answered in section three. How can we describe this phenomenon and, more important, on what way can we improve the quality of patient logistics? A variety of articles discuss this subject and all the articles I have read share one conclusion: in order to improve logistical processes within a hospital, it takes a lot of time and effort to achieve this. When a hospital is undergoing some improvements in the processes, the support from the staff is very important. The old-fashioned process is divided into subprocesses. Theoretically spoken, the aim is to make these processes as efficient as possible, creating the highest possible return from human activities. But the staff needs to see the advantages of the changes in the processes, otherwise they see the phenomenon "to work" not as a so-

cial matter, but just as a way to make money. This can result in situations were employees are forced to make certain decisions against their professional ethics, with the "simple" reason that these decisions are cheaper and more efficient. Therefore I think that it is very important to create a high rate of acceptance from the employees. This can be achieved by involving them in the process of changing the business processes. The article Better Health Care from less money [3] even concludes a saving in costs of 2.5 billion a year. Key to this saving is to make the employees motivated. When employees have more control over their tasks and are being stimulated in a positive way, the advantages of improving the processes become clear.

By improving the processes, a better and more efficient way of working can be realized, resulting in a decrease of errors and a higher and better health care for the patients. This thesis can help in this process of improvement. The second subquestion is answered by delivering a framework which helps in improving processes. It makes it possible for a medical center to reason about their logistic processes and improve them by making use of the indicators of chapter 4. A hospital is able to categorize the events which occur when a patient enters the hospital in five categories (triage, diagnostics, treatment, nursing and aftercare). After categorizing, one can measure the processes by using the indicators belonging to each of the five categories. The link between the data warehouse and the measurements improves the quality, because with the gathered information, the hospital can generate reports from recent information and compares it with information from the past. The differences can be interpreted and used for improvement.

Another aspect is the data mining technique, explained in chapter six and adapted to our own business models in chapter seven, which discusses two indicators from chapter 4 and applies the data mining clustering technique in order to generate new information from already known information from the source databases and/or the data warehouse. So, as an answer to subquestion three, we can say that it indeed is possible to optimize the phenomenon patient logistics by using data mining techniques. Unfortunately, the chosen attributes on the business data models were incorrect, resulting in an useless data set. By using a correct data set, it is possible to generate new data out of data originating from the data warehouse system and/or the source database systems.

8.1 Ethical discussion

During the writing of this master thesis, I spoke with my sister who is a family doctor. After explaining things like the subject and area of expertise she came up with something I never realized before. She told me about the fact that processes can be optimized more then one desires. At first,

I thought this was nonsense, but then she started explaining. She told me about a woman with the age of twenty-one. She went to the family doctor, because she felt something in one of her breasts. The family doctor did a quick examination and came to his/her conclusion: there was a small lump in the breast, which is a symptom of cancer. Now, imagine the health care processes are very well optimized. In the most "ideal" situation, this young woman enters the hospital at the same day she heard of the small lump. In a few hours, doctors were able to investigate her and the radiography department did made the required photos. The conclusion is sad but true, her breast needs to be amputated. Due to the optimized processes, she is scheduled at the operation room the very next day and her breast is being amputated. So, the time it takes from the first consult at the family doctor until amputating her breast is no longer then 24 hours. I don't think this is the desired situation. Optimizing processes is good, but there is always some kind of trade-off to be made. The young woman needs some time to get used to the fact that she suffers breast cancer. Removing a woman's breast has a huge impact on her life. Situations like these need to be considered at the time a hospital is improving their processes; health care is about taking care of people, not about improving processes. I thought this ethical discussion would be nice to mention in this thesis. I think it is definitely something to keep in mind during this period when operational business processes are being improved.

8.2 Implementing Weka in Data Warehouse

Data warehouses contain a lot of information from a specific subject. In this thesis' case, a lot of data, originating from source information systems at hospitals or data warehouses are being used. A data warehouse is a standard way of working with data, starting at making an inventory from the source data, creating the data staging area, etc. Finally, reports are created from the derived data. The tool used during this thesis offers the ability to run automatically, directly linked to the data originating from the data warehouse. In this way, data mining algorithms can be accessed on the fly, fully integrated with the data warehouse. Results are shown just like the conventional reports. By offering the end-users of the data warehouse a relative easy way of applying data mining on their data, a big advantage can be achieved. A small search shows that fully integrated data mining techniques are not yet common in data warehousing and therefore, implementation can be interesting. If one manages to implement those techniques within his data warehouse, it is in his advantage against the competition.

List of Figures

References

[1] R. Agrawal, R. Srikant, *Fast Algorithms for Mining Association Rules*, IBM Almaden Research Center, University of Wisconsin.

[2] Aruster, Weijters, Vries de, Bosch van den, Daelemans, *Logistic-Based Patient Grouping for Multi-disciplinary Treatment*, Technical University Eindhoven, Tilburg University, 2003.

[3] P. Bakker, *Het kan echt: betere zorg voor minder geld*, Sneller Beter, De Logistiek in de Zorg, TPG Nederland, 7 juli 2004.

[4] Business Intelligence For You (BI4U), *BI4U Website*, http://www.bi4u.nl, viewed: 09/15/2007.

[5] Business Intelligence For You (BI4U), *Naslagwerk Integration Services*, Portal from BI4U, September 2007.

[6] H. Dekker, *From Push to Pull*, IT Logistics, June 2002, year 6, number 6, p30.

[7] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *From Data Mining to Knowledge Discovery in Databases*, American Association for Artificial Intelligence, Fall 1992.

[8] J. Frawley, G. Piatetsky-Shapiro, C.J. Matheus, *Knowledge Discovery in Databases: An overview, The American Association for Artificial Intelligence*, Fall 1992.

[9] Jiawei Han, Kamber, *Data Mining Concepts and Techniques*, 1-55860-901-6, Second Edition, Elsevier Inc.

[10] V. Harinarayan, A. Rajaram, *Implementing Data Cubes Efficiently*, Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Canada.

[11] M. Holshemer, A. Siebes, *Data Mining: The Search for Knowledge in Databases*, CS-R9406, Centrum voor Wiskunde en Informatica, 1991.

[12] J.W. Hoorn, J. van der Eijk, *Simulatie als Learning tool bij hardnekkige Logistieke problemen*, Vreelandgroep Organisatieadvisuers, 2005, http://www.vreelandgroep.nl/pdf/Simulatie.pdf, viewed: 08/17/2007.

[13] Karsten M. Decker, Sergio Focardi, *Technology Overview: A Report on Data Mining*, CSCS TR-95-02, Swiss Scientific Computing Center, 1995.

[14] L. Maruster, T. Weijters, G. de Vries, *Logistic-Based Patient Grouping for Multi-disciplinary Treatment*, Technical Universiry Eindhoven.

[15] H. Mintzberg, *Toward Healthier Hospitals*, Health Care Manage Rev., 1997.

[16] J. Mundy, W. Thornthwaite, R. Kimball, *The Microsoft Data Warehouse Toolkit*, Wiley, USA, 2006, ISBN: 0-471-26715.

[17] A. Sen, A.P. Sinha, *A Comparison of Data Warehousing methodologies*, Communications of the ACM, March 2005, Vol. 38, No 3, p79-84.

[18] W. Sermeus, K. Vanhaecht, *Wat zijn Klinische Paden*?, Acta Hospitalia 2002/3.

[19] Sumathi, Sivanandam, *Introduction to Data Mining and its Applications*, 3-540-34350-4, Springer Berlin Heidelberg New York.

[20] J. Valkenborgh, *Samenspel in Patientenlogistiek*: Naar integrale proces- en capaciteitssturing in het ziekenhuis van morgen, Atrium Sante, VU Medisch Centrum, 6 juni 2006.

[21] K, Vanhaecht, W. Sermeus, G. Peeters, *Ontwikkeling en gebruik van klinische paden (clinical pathways) in de gezondheidszorg.* Tijdschrift voor Geneeskunde, 2002.

[22] P. Verschuren, *Het ontwerpen van een onderzoek*, Lemma, Utrecht, The Netherlands, 1998, ISBN 90-5189-707-3.

[23] Weis, Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods From Statistics, Neural Networks, Machine Learning and Expert Systems*, San Francisco, Mogan Kaufmann.

[24] Weka 3: *Data Mining Software in Java*, Waikato University, http://www.cs.waikato.ac.nz/ml/weka, viewed: 01/22/2008.