Tampereen yliopisto

Anniina Aitalaakso

SELECTION CRITERIA AND THE NEEDS FOR PREDICTIVE ALGORITHMS

Faculty of Engineering and Natural Sciences
Master's thesis
August 2019

ABSTRACT

Anniina Aitalaakso: Selection criteria and the needs for predictive algorithms
Master's Thesis
Tampere University
Information and Knowledge Management
August 2019

Predictive analytics has not been widely adopted to use in healthcare. As information systems supporting the use of predictive models get more common and the amount of data collected gets bigger, the potential to use predictive models also grows. As the implementation of Apotti-system moves forward and the amount of collected data grows, the users of Apotti-system also have the possibility to utilize predictive analytics in their work. This research was about finding out the needs for predictive models and the selection criteria formed based on these needs as well as the needs and limitations healthcare data creates for the use of predictive algorithms.

Theory to support the understanding of the subject has been formed based on literature. Different reference databases have been used to find source material. Empirical part, that is defining the needs for predictive algorithms, was done via internal workshops. The participants of the workshops were divided into three groups, depending on their background. These groups were: social services, primary healthcare and special healthcare. Every group thought of situations in which they would have needed information, that involved predicting future and the needs were formed based on these situations. All the needs were collected together and criteria for the selection of algorithms were formed. The research helps with the selection of the algorithms through defining the current situation of the organization and recognizing situations, where it would be beneficial to use predictive models. The research also reveals elements, which are affecting in the background of algorithms helping to understand the factors that affect the planning and choosing of algorithms.

The special needs healthcare data creates for algorithms were recognized through literature. It was discovered, that algorithms that are used in the treatment of patients, must be approved for medical devices, before they can be used in the treatment process. What also came up, was that since healthcare data is complex by structure, simple algorithms should be used to benefit the most. What affects the sharing of data between different devices, is the fact, that there are no general standards in use in healthcare for data sharing.

The other research questions were answered through the empirical work. In total, 13 needs were collected from all groups. Based on the targets of the algorithms, four categories were recognized. These categories are: segmentation, operational, effectiveness and health. The selected algorithms should therefore make the grouping of patient possible to allow more individual treatment options for patients', help planning the daily operations, including better allocating of resources, measure the effectiveness of the services and treatments offered to improve satisfaction and resource management as well as helping with the preventative care of patients' and clients'.

Keywords: Predictive analytics, algorithms, big data, healthcare

TIIVISTELMÄ

Anniina Aitalaakso: Ennustavien algoritmien valinta ja tarpeet
Diplomityö
Tampereen yliopisto
Tietojohtamisen tutkinto-ohjelma
Elokuu 2019

Ennustava analytiikka ei ole vielä kovin yleisesti käytössä terveydenhuollossa. Ennustavia malleja tukevien tietojärjestelmien yleistyessä ja kerätyn datan määrän kasvaessa myös potentiaali ennustaville malleille ja niiden hyödyntämiselle kasvaa. Apotti-järjestelmän käyttöönoton edistyessä ja kerätyn datan määrän kasvaessa, myös järjestelmän käyttäjillä on mahdollisuus hyödyntää ennustavia algoritmeja työssään. Tutkimuksessa selvitettiinkin, mitä tarpeita ennustaville malleille on, millä kriteereillä niitä pitäisi valita käyttöön sekä millaisia tarpeita ja rajoitteita terveydenhuollon data luo ennustaville malleille.

Teoriapohja aiheen ymmärtämiselle on luotu kirjallisuuden avulla, ja erilaisia viitetietokantoja on käytetty lähdemateriaalin löytämiseen. Empiirinen osuus, eli ennustavien mallien tarpeiden määritys, toteutettiin sisäisten työpajojen avulla. Työpajojen osallistujat jaettiin kolmeen ryhmään: sosiaalihuolto, perusterveydenhuolto sekä erikoissairaanhoito. Jokainen ryhmä esitteli tilanteita, joissa heille olisi hyötyä siitä, että voitaisiin ennustaa tulevaa ja niiden perusteella määritettiin jokaisen käyttäjäryhmän tarpeet. Kaikkien ryhmien esiin nostamat tarpeet koottiin yhteen ja niistä muodostettiin kriteerit ennustavien mallien valinnalle. Tutkimus auttaa algoritmien valinnan kanssa määrittämällä organisaation tämän hetkisen tilanteen ja tunnistamalla tilanteita, joissa ennustavista malleista olisi hyötyä. Tutkimus myös avaa tekijöitä, jotka vaikuttavat algoritmien taustalla, auttaen siten ymmärtämään algoritmien suunnitteluun ja valintaan vaikuttavia tekijöitä.

Terveydenhuollon datan erityisvaatimukset algoritmeille tunnistettiin kirjallisuuden avulla. Havaittiin, että potilaiden hoitoon käytettävät algoritmit täytyy hyväksyä lääkinnällisiksi laitteiksi, ennen kuin niitä voidaan hyödyntää hoidossa. Esille nousi myöskin se, että terveydenhuollon datan ollessa monimutkaista rakenteeltaan, sen kanssa toimii paremmin yksinkertaiset algoritmit. Huomattavaa on myöskin se, että terveydenhuollossa ei ole käytössä yleisiä standardeja datan jakamiselle, mikä vaikeuttaa datan jakamista eri laitteiden välillä.

Muihin tutkimuskysymyksiin saatiin vastaukset empirian kautta. Tarpeita kerättiin kaikilta ryhmiltä yhteensä 13. Tarpeista tunnistettiin neljä kategoriaa, joihin kerätyt tarpeet voidaan luokitella. Nämä kategoriat ovat: ryhmittely, operatiivinen, vaikuttavuus sekä terveys. Käyttöönotettavien algoritmien tulisi siis mahdollistaa potilaiden ja asiakkaiden ryhmittely paremman hoitopolun mahdollistamiseksi, auttaa päivittäisten toimintojen suunnittelussa ja muun muassa tehostaa resurssien käyttöä, mitata vaikuttavuutta ja sitä kautta parantaa asiakastyytyväisyyttä ja resursointia sekä mahdollistaa potilaan hoidon ja terveydentilan ennaltaehkäiseviä toimia.

Avainsanat: Ennustava analytiikka, algoritmit, big data, terveydenhuolto

PREFACE

I would like to thank everyone from Apotti who participated in this process, especially Ulla Voutilainen for offering this opportunity, Anu Lankinen for all the help with the workshops and Jari Renko for all the ideas brought throughout the process. I am also thanking everyone who participated the workshops and helped me to discover the world of algorithms. I also want to thank professor Samuli Pekkola for comments throughout the whole writing process.

And, of course, thank you Jossu for making me laugh with your comments, onion is indeed a real thing. The biggest thanks belong to the whole community of Tampere University of Technology and everyone I have met during this journey. Thank you for helping me to find my own path.

Helsinki, 13.8.2019

Anniina Aitalaakso

CONTENTS

1. INTRODUCTION 1
1.1 Research Background 1
1.2 Research Problem and Research Questions 3
1.3 Research Limitations and Scope 4
1.4 Research Structure 5
2. BIG DATA 7
2.1 Characteristics 7
2.2 Process 9
2.2.1 Data Management 9
2.2.2 Analytics 10
2.3 Big Data in Healthcare 11
3. PREDICTIVE ANALYTICS 14
3.1 Levels of Data Analytics 14
3.2 Algorithms 15
3.3 Analysis 17
3.3.1 Concepts 17
3.3.2 Methods 18
4. RESEARCH METHODS 22
4.1 Methodology 22
4.2 Workshops 24
4.2.1 Arrangements 25
4.2.2 Social Services and Primary Healthcare 26
4.2.3 Special Healthcare 27
4.3 Analysis 28
5. EMPIRICAL FINDINGS 29
5.1 Analysis of Needs 29
5.2 Selection Criteria 31
6. DISCUSSION 35
7. CONCLUSIONS 36
7.1 Summary 36
7.2 Research Evaluation and Limitations 39
7.3 Suggestions for Future Research 40
REFERENCES 41

LIST OF SYMBOLS AND ABBREVIATIONS

CDS	Clinical Decision Support
EHR	Electronic health record
EMR	Electronic medical record
FDA	Food and Drug Administration
HDFS	Hadoop Distributed File System
k-NN	K-nearest neighbor algorithm
MEDEV	The Medicine Evaluation Committee
NLP	Natural Language Processing

1. INTRODUCTION

Finland has long been leading the way in adopting social- and healthcare information systems. This has led to a situation, where many of the now used systems are outdated and are not supporting doctors' work as hoped any more. Because of this, many organizations are planning to replace their old systems in forthcoming years. The systems in use are mostly used to write down patient related information, thus missing features to manage information and resources. (Deloitte 2014) One organization to take the challenge is Apotti, which is developing new social- and healthcare information and enterprise resource planning system. Apotti identified a need for research regarding predictive analytics, hence giving the subject for this thesis. The subject of the thesis is *Selection Criteria and the Needs for Predictive Algorithms*. The goal is to recognize the needs Apotti has for predictive algorithms and then create a prioritized list of the needs to help selecting the algorithms.

This chapter introduces first the background for this thesis and Apotti, which is the organization this thesis is made for. Apotti develops a social and healthcare information system for a group of municipalities in Uusimaa region. Also, the most important stakeholders, which are included in the development process, are introduced. The research questions are formed and the reasons why they were chosen will be discussed. At the end, the structure of the thesis is presented as well as the general overview of each chapter's content.

1.1 Research Background

In organizations, analytics lands under business intelligence processes, which aims to understand and analyze business. Analytics is usually used for bringing value to organizations. This is achieved through improving decision making and the way of working by analyzing available data and using it with statistical and quantitative analysis. The usage of analytics in an organization is relatively easy to start, the process can begin if there is enough data available and powerful hardware exists. (Davenport et al. 2007, pp. 26, 39; Marr 2017, pp. 8) Value bringing information is real, coherent, actual, complete, controlled and it is in real context (Davenport et al. 2007, pp. 207-208). These features have to be found from the information used in decision making to make it reliable.

Data science is a field of science, that uses different methods and processes for supporting and guiding the extraction of information and knowledge from data (Waller & Fawcett 2013; Sanchez-Pinto et al. 2018). The field has evolved from data learning, and it emphasizes more the preparation and cleaning of data than data modelling, as the traditional approach suggests. Data scientist is a professional, who uses methods and processes related to data science in order to gain valuable information from data. (Donoho 2017) Data analysis is a process, which aims to collect information from available data (Cao 2017). This information can be used for example to create predictive models to support decision making.

Predictive analytics means utilizing algorithms for predicting the future, hence the aim is to get an answer to the question 'what will happen' (Waller & Fawcett 2013; Cohen et al. 2014; Hagerty 2016). Predicting the future is done with the help of data collected over a long period of time (Persson & Kavathatzopoulos 2018). Therefore, the models predict, what is most likely going to happen based on previously occurred incidents and the relations between them (Cohen et al. 2014). Algorithms are used to automate the data analysis process and to find variables that might have an impact on future events.

Algorithms produce a correct answer for a problem building it step by step (Kingston 1990). It has been noticed, that the simplest algorithms used for big data are more effective than complex algorithms used for a small amount of non-complex data. Hence, it is important, that the algorithms made for big data can handle large and complex datasets, rather than being complex by structure. (Chen, M. et al. 2014) The largest and fastest expanding datasets can be found from healthcare. The data comes mainly from electronic health records (EHR), which includes for example the information about patient's reception visits and their treatments as well as the medication they have used. The growing rate of the data collected from healthcare is estimated to be as big as 2.4 exabytes per year. (Kambatla et al. 2014)

Social- and healthcare information systems are intended for handling and storing patient related documents and the information they contain. The producer of the system is responsible for proving, that the system meets all the demands set to information systems in question. (Valvira) Social and healthcare information systems have no established standards in Finland. Hospitals and health centres have been able to use information systems of their choosing. This has led to a situation, where patient data is spread over different systems nationally and locally. (Lehto & Neittaanmäki 2017) Interoperability between different systems is weak, which affects the data available for treatment decisions, making doctors' work more challenging.

The collected patient-related data is now only used in its original purpose, which is to solve patients' problems, that are usually unexpected and acute in nature (Lehto & Neittaanmäki 2017). In the future, the same data could be used preventatively, focusing on keeping patients healthy and fully functional for longer. However, this requires interoperability between social- and healthcare information systems and that the data is coherent in all used systems. Several social- and healthcare development processes have been started to improve the situation, Apotti being one of them.

This research is made for Oy Apotti Ab. It is a company that was created for the need of building a shared social- and healthcare information and enterprise resource planning system for municipalities in Uusimaa region (Helsinki, Vantaa, Kirkkonummi, Kauniainen, Kerava, Tuusula) and HUS Helsinki University Hospital. The project started in 2013 and the goal is to be ready with all the deployments in 2020. The project is now in the implementation and deployment phase, implementation is about to be ready, and deployment has just started. Deployment happens in stages; the first deployment was in November 2018 in HUS Peijas hospital. (Apotti internal 2019) Apotti's value promises are presented in Figure 1.

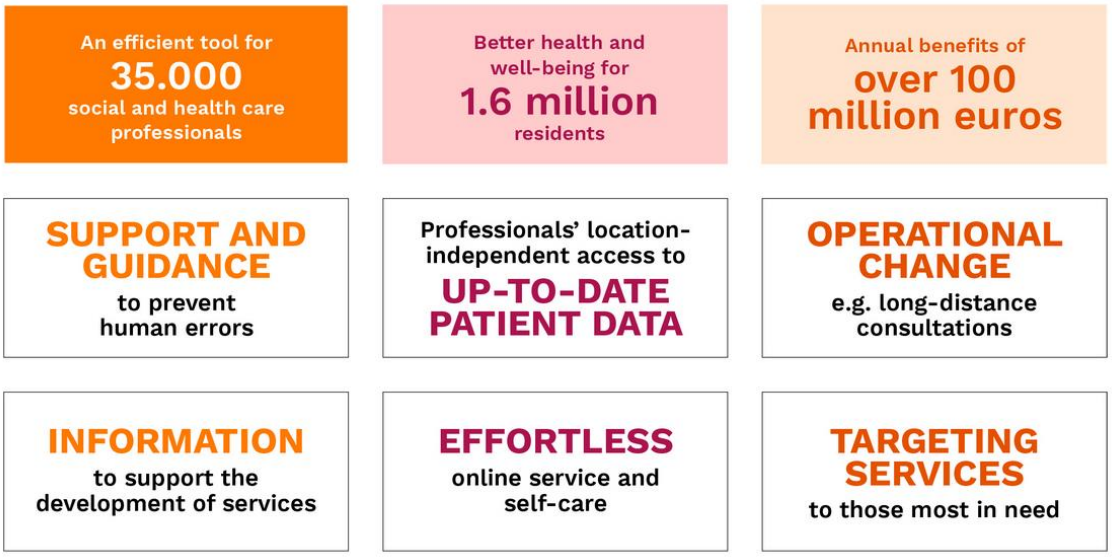

Figure 1 Value promises (Apotti)

As a project, Apotti's goal is to develop social- and healthcare services and adopt a common information system for these sectors. Currently there are several different information systems in use, which do not work together or support the daily work. Apotti will combine social- and healthcare information, making the information available in real-time for all users. Its advantages, compared to current situation, are for example saving employees' time as all information can be fetched from the same place, as can be seen from Figure 1. It is also estimated to bring over 100 million euros' savings annually. (Apotti internal 2019) Being able to access all patient-related information from the same system makes the treatment of a patient easier and helps to form comprehensive picture of the patient's health. Thus, doctors are not dependent on patient's own description of past appointments.

The information system vendor is American Epic Systems Corporation, which won the competitive bidding and signed a contract with Apotti in April 2016 (Apotti internal 2019). Epic Systems Corporation is a privately-owned company founded in 1979. It has over 9 000 employees working in the headquarters in Verona, Wisconsin. (Epic) Globally over 350 organizations are using Epic's information system, some being bigger and more complex than Apotti-system is. Accenture is the subcontractor in the project. Accenture was founded in 1989 and it is providing services in strategy, consulting, digital, technology and operations. It has over 450 000 employees in over 120 countries. (Accenture) Accenture supports Epic Systems and Apotti in the implementation and maintenance of the software. On top of that, Accenture will plan and implement additional features to the system.

The purpose of Apotti is to integrate social- and healthcare information systems used in Uusimaa region, to allow better data flow and the utilization of collected data. The goal is to develop locally integrated information system to allow continuous data exchange between different hospitals and health centers using Apotti-system. The implementation of Apotti-system is currently in deployment phase, which means that the future and the opportunities a new information system brings, once fully implemented, can already be started to think about. Thus, new functionalities the new information system enables can already be considered and planned. The contract made with Epic includes a few predictive models Epic has developed, that can be chosen from a list of completed models that are already in use in other organizations that use Epic's information system. This forms the base for this research as the idea is to examine the possibilities the new information system and more extensive data collection brings, focus being in predictive analytics.

Large-scale utilization of predictive models is not relevant yet, because of the lack of proper data available. However, predictive algorithms are a part that is being considered an important feature in the new information system. It was decided that it is time to start investigating the needs Apotti-system users have for predictive models. The aim of this thesis is to find out these needs and based on them, create a prioritized list of criteria to help to choose the models. The intention is to collect the needs without too many limitations. It is not purposeful to let the users select from the ready models. The determination of the needs is supposed to work as a guideline when determining which of the models are the most useful for the users, not forgetting patients and clients either.

1.2 Research Problem and Research Questions

The use of predictive analytics has been noticed to have several possible benefits in healthcare. These include things, that make the job of healthcare personnel easier, such as easy access to patient data and fast availability of test results. Also, things that affect patients' health are included, for example improved accuracy of clinical decisions and predictions of different treatment's effectiveness for individuals as well as proactive treatments aiming to maintain the state of health. (Wang, Y. & Hajli 2017) To get an opportunity to pursue these benefits, one has to find out, what kind of needs the users have for predictive models and if it is possible to fulfil these needs with the existing models.

The objective of this thesis is thus to provide information about the current situation in Apotti and define the needs Apotti-system users have for predictive algorithms. To reach these goals, research questions must be formed to guide and define the research and to make it easier to

control. It also makes reaching the set goals easier when they are properly outlined. To further improve the manageability of the subject, the problem is divided into smaller units, which form the base for the thesis. Dividing the problem into smaller pieces also helps to understand the big picture and see how the issues are related together. These smaller units form the research questions for this thesis. The aim of the questions is to give information about what is needed to be able to pick the most useful algorithms and understand the mechanisms that affect behind them. The research questions of the thesis are:

1. *What are the needs for predictive algorithms?*
2. *What are the criteria for selecting predictive algorithms?*
3. *What needs, and limitations healthcare data creates for predictive algorithms?*

The goal is to find out, what kind of needs Apotti has for predictive algorithms and how the needs could be prioritized, to be able to decide between different algorithms. Because Apotti works in the healthcare industry, they also collect a massive amount of data which contains a lot of confidential information. This information cannot spread outside the organization and only authorized people can have access to it. Hence, the research is also about finding out, if healthcare data creates any special needs or limitations for using predictive algorithms.

The answers to these questions are achieved first by creating enough theory background to support the understanding of the research questions and secondly, holding workshops to collect data about the situation and needs in Apotti. After the needs are clear and in comprehensive form, they can be prioritised, thus providing criteria for selecting predictive algorithms to use. The prioritizing is made based on categories recognized from the collected needs. These categories are then analysed further, and a recommendation is made for what kind of algorithms should be chosen for use based on the needs.

The aim of the thesis is to help Apotti to realize the potentials of using predictive analytics and to find out the needs they have regarding algorithms. Also, the potential benefits of using predictive models in patient care are supposed to come up while building the theory background. Knowing the possible outcome helps to guide the implementation of algorithms and to convince doctors and other healthcare personnel of the benefits of using predictive analytics in patient care. This makes it easier to plan the future usage of algorithms and see, how preventive care could be used to benefit patients the most.

1.3 Research Limitations and Scope

Demands for this thesis come from both the company and its objectives for the work, as well as the limitations and requirements set for a master's thesis. The company gave a loose framing for the thesis, defining the subject but giving space to decide the best approaches to reach the wanted results. The goal Apotti wanted to reach was to define, how algorithms can be selected and what is the best way to solve, what kind of algorithms are needed. We decided, that the best way was to arrange a workshop, where users from different backgrounds would be able to discuss together about their work, making it possible to get ideas covering the whole organization.

It was decided to collect the needs without too much defining or limitation at first. This is because most of the workshop participants do not have basic knowledge of algorithms and predictive analytics, and they are not sure what can be achieved with these. It was better to let the participants throw ideas without too much thinking by giving them only loose frames for the assignment. There was a risk, that the workshop participants would get stuck too much on thinking 'is this possible to reach with predictive analytics?'. This was wanted to be avoided to benefit the most from the workshops. It was decided to do the elimination of the needs, that are not possible to reach with predictive analytics, in the analysis phase.

Because of the academic nature of the thesis, there are also limitations to the subject and the scope in which it can be handled. Attention had to be paid to the validity and reliability of the results. Also, the research had to be planned so that it would be able to be conducted in a reasonable timeframe. This affected the time available for analysis, and the workshops had to be

timed so, that as many of the wanted participants could participate, but at the same time giving enough time to prepare the workshops.

The schedule of the thesis was planned so, that the data collection phase took place in April. Apotti-system's deployment in the beginning of May in Vantaa brought additional challenges to the schedule of the workshops as the participants were busier the closer the deployment was. This was causing challenges in finding dates for workshops where everyone could participate. Also, the planned implementation of the workshop had to be moved to an earlier date because of the deployment. This ended up being a poor time, because of another deployment-related event. It was decided to arrange a second workshop for the ones, who were not able to participate the first one because of this.

Originally, a survey was planned to be carried out to be able to prioritize the needs that were collected from Apotti-system users. This was decided to be left out from the research after the first workshop, as the answers collected were mostly group specific, meaning that for example the needs collected from social services were specific to their work, and they were not adaptable to fill the needs special healthcare has. Hence, the respondents would most likely have picked their own answers to be the most important needs, if they were to answer a survey.

1.4 Research Structure

This chapter introduces the structure of the thesis chapter by chapter. Also, the contents of the chapters are presented briefly but comprehensively. This chapter forms the base for understanding how the subject is built in this research and what topics are handled. This is a multi-methods study, and workshops were used to collect data. Theoretical background is built with the help of literature, to form a base for understanding the subject. The theory is then utilized in the planning of workshops, to lead the participants in the topic of predictive analytics. Figure 2 presents the content of this thesis.

Figure 2 The content of the thesis

The thesis starts with an introduction to the topic, as can be seen from Figure 2. The thesis starts with introducing a general background for the thesis as well as important information for understanding the subject and why it was chosen in chapter 1. It also consists of the introduction of Apotti, the research questions and problems and the limitations and the scope of the research as well as the overall structure of the thesis.

Chapters 2 and 3 form the theoretical background for the research and help to understand the subject and the research questions. Chapter 2 handles big data and its characteristic features, as well as the progress of the analysis process with big data. Also, Hadoop's MapReduce is introduced, it being one of the most used tools with big data. At the end of chapter 2, healthcare data and its characteristic features are introduced. Chapter 3 introduces the levels of data analytics, especially predictive analytics is introduced in a deeper level. Also, algorithms and their use in predictive analytics is discussed. At the end, some analysis methods used with predictive analytics and algorithms are presented.

Chapter 4, research methods, introduces the methodology used through Saunders et al. (2007) research onion. Also, the preparations, content and progress of the workshops are presented. At the end of the chapter 4, the data analysis process is introduced. The results of the analysis process are introduced in chapter 5. It collects the results of the workshops together, analyzing the needs and answering research questions related to the needs for predictive algorithms and the way they should be chosen.

Chapter 6 consists of general discussion around the subject and the research decisions. Chapter 7, conclusions, collects together all the results and gives a brief summary about the thesis. Also, the key findings and results are presented and evaluated.

2. BIG DATA

Big data definitions have been changing rapidly but size, as the name suggests, is something that always comes to mind when talking about big data (Gandomi & Haider 2015). Big data is usually described as large amount of data that cannot be handled with using traditional data-processing tools and systems (Zikopoulos et al. 2012; Huang et al. 2017). More broadly, it is a dataset whose size and complexity requires the development of new methods and processes, because traditionally used ones cannot handle them. These new tools make it possible to store, analyse, and use all the available, complex data, which is acquired from multiple sources. (Belle et al. 2015; Carter & Sholler 2016)

This chapter starts with an introduction to big data, presenting its characteristic features and a brief history of the development of big data. After the base is constructed, it is time to move on to the process of big data analytics and go through the steps it consists of. Also, the most important and frequently appeared methods in literature related to big data are introduced. At the end of this chapter, healthcare data is being discussed, which is one of the fastest growing datasets.

2.1 Characteristics

In 2013 SINTEF, an independent research organization, estimated that 90 % of world's data was created within the previous two years (Dragland) and another research mentioned, that 90 percent of the produced data is in digital format (Banerjee et al. 2013). This gives a clear picture of how fast data is being generated and stored these days. The development of collected data can be seen from Figure 3.

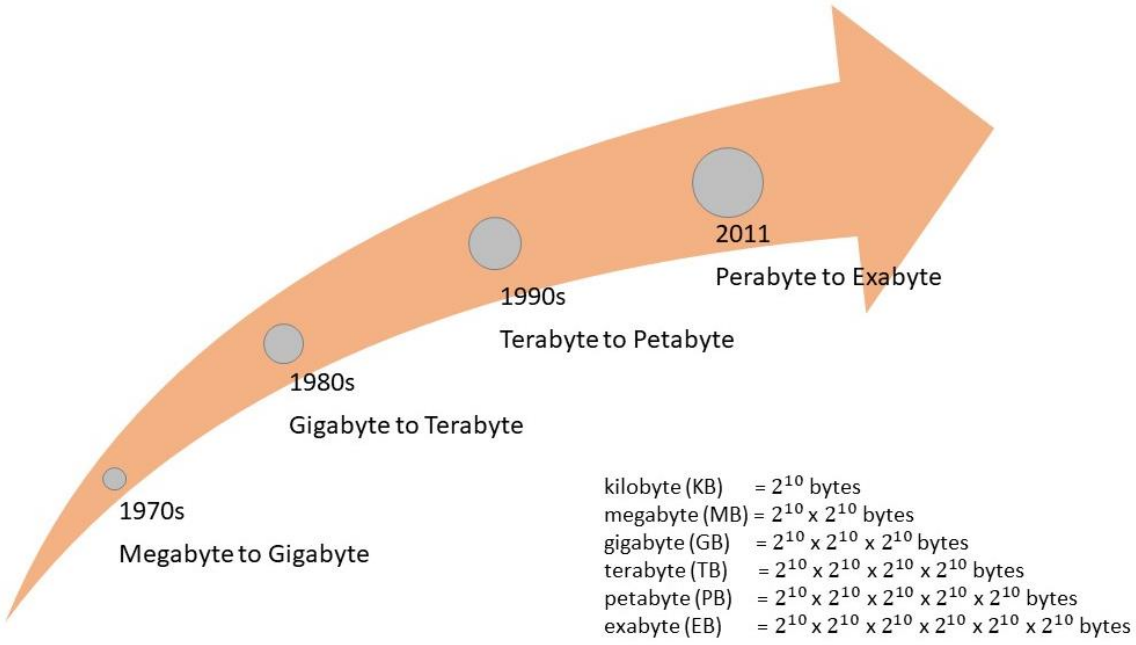

Figure 3 History of big data (adapted from Hu et al. 2014)

As presented in Figure 3, the history of big data can be divided roughly into four parts based on the growth of the data. Each of these stages lead to the development of new technologies, because the old ones were not enough after the growth. In the 1970s, collected data was mainly related to business operations and was managed through databases and relational queries. Development of digital technology caused the data volume to grow from gigabytes to terabytes in the 1980s and the data was stored in parallel databases. (Hu et al. 2014) The development of internet and all the websites created, and queries made inspired by this, led to the growth of data in the 1990s. To be able to handle and analyse this kind of data, Google developed Google File System (GFS) and MapReduce models. (Chen, M. et al. 2014) Currently the size of data is measured in exabytes, but new, ground-breaking technology has not been introduced.

Apart from the obvious, big data has more characteristic features than just the fast growth of data collected. Usually three things come to mind when talking about big data. These are volume, variety, and velocity, the three V's. (McAfee & Brynjolfsson 2012; Elgendy & Elragal 2014; Janssen et al. 2017) Volume describes the fact, that the data increases fast, as pointed out. The problem with the growing speed is, how to properly utilize the data collected and how it is possible to find the information needed. On the other hand, it is also problematic to figure out what information can be found from the data. (Zikopoulos et al. 2012; Gandomi & Haider 2015) Data is being collected from different sources and formats, which leads us to the fact, that the data collected is complex and semi structured or even unstructured. This is meant when talking about the variety of data. It is hard to clean and analyse complex and unstructured data with traditional tools and processes and new ones need to be developed to be able to properly handle big data. (Zikopoulos et al. 2012, pp. 5-9; Elgendy & Elragal 2014) Velocity part consist of the speed with which data is being collected and how fast the data can be processed from changing sources (Janssen et al. 2017; Huang et al. 2017). The target is to collect and analyse data almost in real-time to benefit the most from it.

In addition to the three V's, there has been discussion about adding another three characteristics to the list: veracity, variability, and value. Veracity describes the unreliability of the data, consisting of imprecise and uncertain data. This kind of data comes for example from social media, where all kinds of data and information can be shared, some being true and some not. Variability derives from the variation in the data flow rates, which is not dependent of factors such as time of day or month but is rather random. (Gandomi & Haider 2015; Marr 2017) Marr (2017) argued, that also value should be added to the list of characteristics. He wanted to add value to the list, because being able to work with vast amount of data is useless, if it does not result in value to the organization. (Marr 2017, pp. 86-88) These all represent well the features of big data and would be a good addition to the list.

Big data is hardly ever structured, but it can be a combination of structured and unstructured, semi-structured data. Structured data is something that can be arranged into a structured form whereas unstructured data does not fit into these forms. A typical place to store structured data is a database, where each column represents the same type of structured data. An example of unstructured data is an e-mail conversation, which cannot be put in a structured format. Semi-structured data can be for example an image fetched from internet with tags. (Elgendy & Elragal 2014; Marr 2017, pp. 86-88) The image itself is unstructured data, but the tags used with it make it possible to classify the image.

Typically, big data is acquired from several sources and then combined. This data in different repositories is also usually collected for a certain need or purpose, and data quality varies between different sources. There are no specific procedures or standards on how to combine data from multiple sources and how to be assured, that the data is suitable for ones needs. This can induce problems with integration of the data. (Bates et al. 2018; Wahyudi et al. 2018) It is also found problematic to find skilled users who can work with big data toolsets and are able to find meaningful information from the data (Larose & Larose 2015; Rahman & Slepian 2016). As data is being generated more and faster, and storing it is becoming cheaper all the time, it is important to have skilled workers, who can recognize the important parts from all the noise data contains (Banerjee et al. 2013).

2.2 Process

Data analytics process usually starts with identifying a problem that needs to be solved (Donoho 2017). Solving a problem with big data differs with this, since data is constantly being collected from multiple sources. This already collected data is then used to gain insights and solve problems. The process of handling big data can be divided into two parts: data management and analytics, as presented in Figure 4.

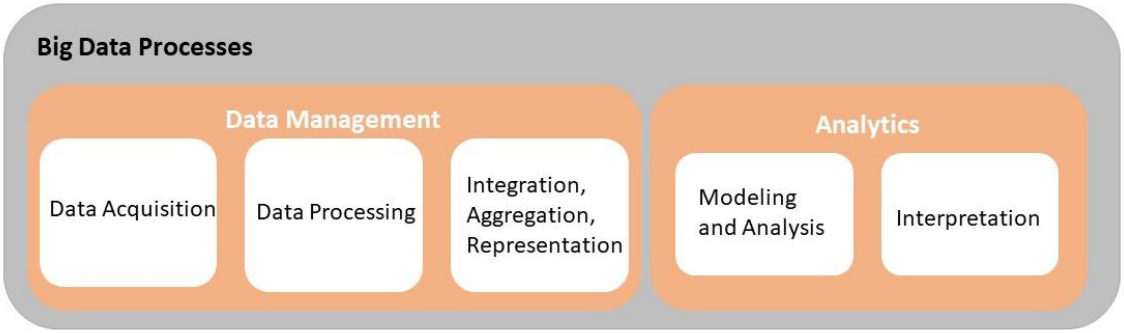

Figure 4 Big Data process (adapted from Gandomi & Haider 2015)

The management part presented in Figure 4 consists of techniques and processes which are used to collect and store data as well as the ones used to prepare the data for analysis. The analytics part is a collection of techniques used for analysing big data and retrieving information from it. (Gandomi & Haider 2015) The tasks to be performed depend on the set goals and the available data. Once the developed models are tested and they satisfy the demands, they can be deployed. This chapter is divided into two parts following the big data processes. The first part is about data management and the second one about analytics and tools that can be utilized in the process.

2.2.1 Data Management

Big data analytics process starts with data management tasks, the first one being gathering the needed data. Data can be collected and searched either from external or internal sources. External can be acquired from outside the organization. It can be publicly available, for example government data or privately owned such as data gathered from social media. Internal data comes from the organization itself and can be for example information collected from customers. (Marr 2017; Wahyudi et al. 2018) Internal data is important, because that is what separates you from your competitors. If everyone has the same data, the analytics are also going to be similar. (Davenport et al. 2010) It is important to think what kind of data is possibly needed now and in the future, it is insufficient to collect data just because it is available.

After gathering the data, it must be processed further before any analytics is possible. This step involves tasks such as extracting information, cleaning, and annotation (Gandomi & Haider 2015; Wahyudi et al. 2018). The target of these tasks is to improve data quality and thus the value it can create (Hu et al. 2014). The cleaning of data is important, because data usually comes from different sources and in different formats. It is needed to be unified before the data can be analysed. (Davenport et al. 2007; Wahyudi et al. 2018) Categorizing and annotating data can be used to make the data more useful. These refer to assigning different labels to the attributes in data and then categorizing them based on these labels. This makes it easier to manage the data and to find certain information. (Hu et al. 2014) All this needs to happen automatically to be able to utilize big data efficiently and in real-time.

The last step of data management involves integration, aggregation and representation of the collected data (Gandomi & Haider 2015). Since the data comes from multiple sources, data sources need to be collected together and processed so that the data is in a unified form. Also, all the findings can be summarized, and representations made. At this point, the data is processed to a form in which it can be stored to a wanted storage space. (Hu et al. 2014) In this phase, all the processing tasks are already made, and this phase concentrates mostly on preparing the data for storage.

Traditional data storage places are, for example, databases and data warehouses. Adding information to these happens through extract, transform, load (ETL) process. It is a workflow, which consists of tasks that perform the extraction of data from wanted sources, loading it to a wanted infrastructure and transforming it to fit the wanted needs. (Vassiliadis 2009) The extraction step connects to multiple source systems to collect all the needed data for analysis. This data is then reformatted to a standard form, with a set of rules, in the transformation step. At the load step, transformed data is imported to a correct space in data warehouse, where the data will be processed. (Hu et al. 2014)

Because big data is mostly unstructured, it is not convenient to use the same transforming process that is used with traditional, structured data. Thus, a new workflow has been introduced to be specifically used with big data. This is called magnetic, agile, deep (MAD). Big data is collected from various sources with different formats, and the data is not cleaned or unified before storing as traditional data is, magnetic reflects this. The data also needs to be easily produced and quickly adapted, and the storage place must adapt to these different needs, hence agile. The last step, deep, refers to the fact, that big data uses complex statistical methods and the storage place must adapt to past queries. (Cohen, J. et al. 2009)

2.2.2 Analytics

After data management, it is time to start the analysis. *"Extracting data is not the same as extracting information"* (Rahman & Slepian 2016). This needs to be kept in mind when analysing big data. All the available tools are not useful, if one does not know how to use them or how to interpret the results. The analytics part of big data processing utilises different methods and tools to process data to extract value from it. The chosen methods to further analyse, model, and visualize the gathered data depends purely on the target, what is wanted to achieve using analytics. The target is to find useful information from the collected data, such as patterns or correlations, which can then be used to improve the quality of decision making and obtain value for business. (Hu et al. 2014; Janssen et al. 2017; Wahyudi et al. 2018) Selecting the best model for use requires performance estimation of different models. After selecting a model for use, it can be assessed by estimating its prediction error. (Hastie et al. 2016, pp. 222) With proper preparation, the information obtained from the data can bring a valuable advantage to organizations. Different analysing methods used with predictive analytics are further discussed in chapter 3.3.2. This chapter focuses on introducing MapReduce framework, which is one of the most used tools for processing and analysing big data.

Data intensive computing was developed to be able to process big data. One of the most used data processing paradigms using distributed data processing is MapReduce framework. (Wang et al. 2013) Google published the MapReduce framework in 2004. It provides a parallel processing model to process big data. In short, the process is divided in two steps, Map and Reduce. The queries are split and handled in different nodes simultaneously, this is Map step. After processing the queries, they are gathered back together in Reduce step. (Rahman & Slepian 2016) MapReduce model is presented in Figure 5.

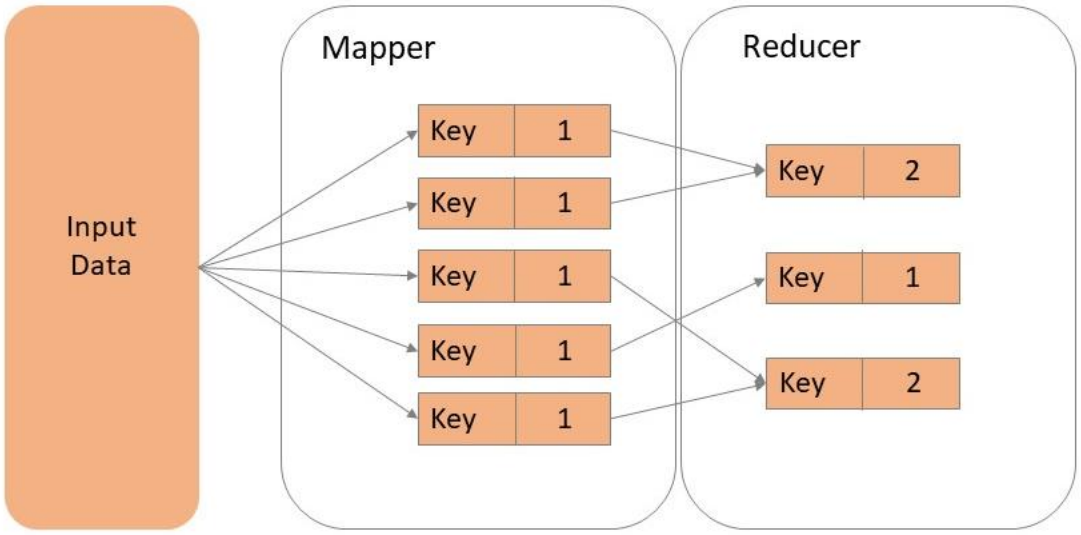

Figure 5 MapReduce process (adapted from Padhy 2013)

Apache made the first open-source implementation using MapReduce framework in 2005 called Hadoop MapReduce (Rahman & Slepian 2016). Apache Hadoop is a collection of open-source software modules used in storing and managing massive data. Hadoop uses commodity servers and creates clusters to simultaneously process multiple datasets. (Hu et al. 2014) Hadoop's MapReduce is one of the most used ones. Besides MapReduce, an important module is Hadoop Distributed File System (HDFS), which allows a fast fetching of data from the repository. (Wang et al. 2013; Chen, M. et al. 2014) It is responsible for the access and management of data (Nesi et al. 2015).

In Figure 5, input data can be thought to represent the HDFS, where the data is stored. After fetching the data, it is filtered and sorted into key and value pairs. This happens simultaneously in multiple clusters. After this, the data is collected back together, and reducer performs a summary operation, counting together all key values. In its simplest form, the keys can be each individual word in the input data. Reducer would then count, how many times different words appear in the text.

HDFS consists of NameNodes and DataNodes. Each cluster has one NameNode, which manages metadata and the access to the files. There is also several DataNodes, where the actual data is located. Similar to HDFS, Hadoop MapReduce consists of a single master JobTracker and multiple TaskTrackers. JobTracker shares tasks between TaskTrackers, manages the execution of the tasks and makes sure the failed tasks are processed again until they are successfully completed. (Wang et al. 2013; Padhy 2013; Chen & Zhang 2014) HDFS and MapReduce run on the same set of nodes. This allows the tasks to be performed on the nodes, where the needed data is already present. (Hu et al. 2014)

2.3 Big Data in Healthcare

The biggest and fastest growing data sets come from the healthcare industry. They consist mainly of the data collected from electronic medical records (EMRs) and electronic health records (EHRs). These are used to improve patient care and to lower the costs of healthcare. (Cohen et al. 2014; Kambatla et al. 2014) EHRs are patient data storages in digital format, where storing and transferring data is secure and accessible only by authorized personnel. The main purpose of EHRs are to support continuous, effective and high-quality healthcare. (Häyrinen et al. 2008; Ward et al. 2014) EMR is integrated into EHR. It includes information about patient's appointments as well as examinations and treatments made by medical professionals. (Callen et al. 2013)

Healthcare data is a good example of the three V's associated with big data. The data is spread over multiple healthcare systems, and besides EMRs, it can be collected from various sources, such as smartphone applications and wearable devices. This means that new data is being generated and stored constantly in different formats. Most of the data repositories don't have the capability to interact directly with each other, making the exchange of data hard and non-transparent. (Belle et al. 2015; Hernandez & Zhang 2017) This makes it harder to diagnose a patient, since all crucial information might not be available. The analytics workflow using healthcare data is presented in Figure 6.

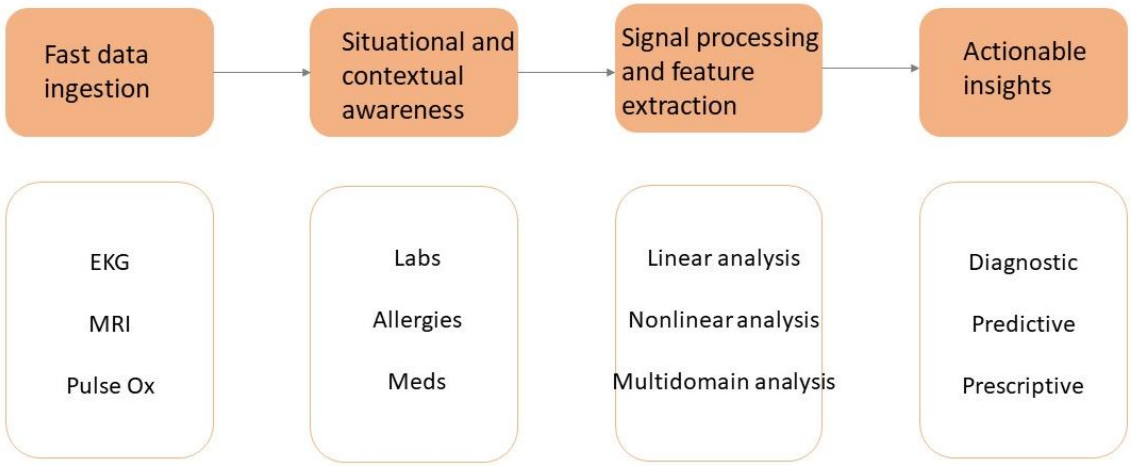

Figure 6 Generalized healthcare analytic workflow (adapted from Belle et al. 2015)

As can be seen from Figure 6, the amount of data collected in healthcare organization is massive and it is being generated constantly from EKGs, MRIs and other devices connected to patients and clients. When making analytical decisions, the data collected from different devices is contextualized and combined with previously collected data from doctor's visits and other available resources. Then all this data is analysed with a selection of signal processing methods and machine learning tools are used to produce insights from this data. The information gathered from the analytics process can be diagnostic, predictive or prescriptive in nature (Ward et al. 2014; Belle et al. 2015), producing different insights about patient's health.

Healthcare data can vary from being only a few megabytes to several hundreds of megabytes. Thus, it requires large storage capabilities, especially if stored for a long time. (Belle et al. 2015) Security is something that needs to be considered when storing data, especially since healthcare data differs materially from many other forms of data. Unintended release of healthcare data can have an impact on people's lives both psychologically and materially, more than even the release of their financial information. This is challenging, because the data owners do not control the data collected from them. (Bates et al. 2018)

Distributive data networks are one way of controlling the security of the data. Used with healthcare data, they allow the confidential data to be kept with the original data holders, the same people who are most familiar with the data. These networks reduce the risk of connecting the data with individuals the data is collected from. Establishing these types of connects is complex and adoption of data standards are needed to ensure fluent data flow. (Bates et al. 2018) The redeeming feature of distributive networks, is that the memory for each computer is private, and a permission is needed in order to gain data from other computers (Cormen 2009, pp. 772).

Algorithms used for analysing and producing insights from constantly growing healthcare data need to be fast and accurate (Belle et al. 2015). It has been noticed, that the simplest algorithms used for big data are more effective than complex algorithms used for small amount of non-complex data. Hence, it is important, that the algorithms made for big data can handle large and complex datasets, rather than being complex by structure. (Chen, M. et al. 2014)

As promising as the analytical future for healthcare data is, the lack of skilled employees has been slowing down the adoption of data analytics in healthcare (Bates et al. 2018) and common standards and processes have not yet been adopted to support integration. The lack of standards and protocols in healthcare industry makes the analytics process ineffective because all available data cannot be properly utilized (Belle et al. 2015). This might lead to a situation, where all medications of a patient have not been considered in the analysis process, making the result not desirable. Information systems in healthcare mostly require data entry screens and forms to be filled in a certain format manually (Sanchez-Pinto et al. 2018). This exposes the data to human errors affecting also the analytics result. Nonetheless, using big data in healthcare has many possible benefits. It can help saving lives, improve care delivery and help limiting the healthcare costs if used correctly and utilized in right places. (Belle et al. 2015)

3. PREDICTIVE ANALYTICS

The term statistical learning refers to a set of tools, that are created to better understand data and make predictions of it (James et al. 2013, pp. 1). Predictive analytics uses diversely different statistical techniques to form algorithms that analyse data and predict future and unknown events. These techniques consist of data mining, predictive modelling and machine learning methods that are used to analyse data in real time. (Cohen et al. 2014; Hernandez & Zhang 2017) Building a predictive model consists of getting the right data, building the model and validating the completed model before it can be used. (Cohen et al. 2014)

In a broader context, predictive analytics is only one part of data analytics. The others are descriptive, diagnostic and prescriptive analytics and they all have a different target on what comes to analytics. All these levels are introduced broader later in this chapter. After introducing the levels, algorithms and their usage are discussed. At the end of the chapter, different methods utilized in predictive analytics are presented.

3.1 Levels of Data Analytics

Predictive analytics is one of the levels in data analytics. Other levels are descriptive, diagnostic, and prescriptive analytics. These are presented in Figure 7. Descriptive analytics consists of data visualization and reporting, diagnostic analytics looks for reasons behind the occurred events and analyses them, predictive analytics is all about data mining and predicting the future, and prescriptive analytics uses simulation and optimization to improve efficiency in the organization (Hu et al. 2014; Hagerty 2016). The levels move from figuring out what has happened to answering how to make something happen.

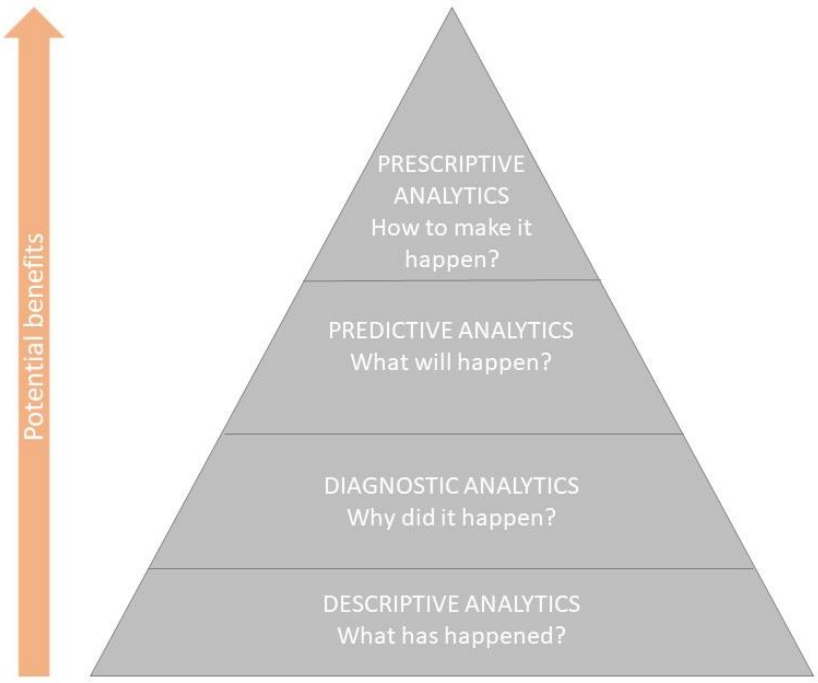

Figure 7 Data analytics maturity (adapted from Hagerty 2016)

As can be seen from Figure 7, the potential benefits of using a certain method grows while moving up in the pyramid. At the same time, the ways to execute the targets of analytics get harder. Descriptive analytics is at the bottom of the pyramid, answering the question 'what has happened'. The next step moving up is diagnostic analytics, which explains why something happened. Predictive analytics analyses data to tell what will happen, and at the top level, prescriptive analytics, finds out how to make something happen. The pyramid describes the focus of the levels. It gets more and more focused towards the top, moving from what has happened to how to make something happen and getting more complex in the way up.

Exploitation of analytics in an organization starts with descriptive analytics. It uses historical data to describe what happened and it is a way to get familiar with the available data. Answering this question will help to understand what has been done, what are the events that lead to the situation and how to learn from the past. (Hu et al. 2014) It consists of categorizing, classifying, combining and analysing data so that it is in understandable format, such as simple visualizations, and ready to use when making decisions. Descriptive analytics helps to find trends and patterns from data, and this can help organizations for example to classify clients into segments and this way provide more efficient marketing. (Davenport et al. 2007; Raghupathi & Raghupathi 2013) After finding out what has happened, the logical next step is to find out what caused it, starting the transition to move up in the pyramid.

The next step moving up in the pyramid, is diagnostic analytics. In addition to finding out what has happened, also the reasons behind the events are investigated. (Hagerty 2016) Diagnostic analytics focuses more on the events that lead to the situation rather than the outcome. The aim is to find out all the different actions that caused the event as well as the probabilities for those actions to be carried out. This gives information about what should be fixed or made better and it is a step towards predictive analytics. (Banerjee et al. 2013) After finding out the actions that caused an event, the prevention of these events can be investigated and little by little adopt predictive analytics tools for use.

Predictive analytics requires usually more developed data and analysis methods than descriptive or diagnostic analytics (Davenport et al. 2007, pp. 222). It concentrates on predicting future events based on the data collected, answering the question 'what is likely to happen'. Predictive analytics pursues to uncover patterns, using for example linear regression, from data and extrapolate them to the future to make predictions. It also aims to capture relationships in data, for example with data mining, and makes predictions based on the found relationships. (Raghupathi & Raghupathi 2013; Hu et al. 2014; Gandomi & Haider 2015; Wang, Y. et al. 2018) As predictive analytics concentrates on forecasting the future, it is also possible to investigate, how to end up in a desirable situation. This way of thinking moves us closer to the last analytics step.

The last and most complex level is prescriptive analytics, which tries to figure out how to make something happen. In this step, data is analysed automatically and in real-time to compare and predict how different things affect each other. Based on these results, the best action to perform is discovered. (Evans & Lindner 2012) Prescriptive analytics has an impact on decision making and efficiency in the organization. It uses decision analysis, which consists of tools such as simulation and optimization. Simulation can be used to identify issues and optimization to find optimal solutions for a certain problem. (Banerjee et al. 2013; Hu et al. 2014)

3.2 Algorithms

An algorithm is a computational procedure that produces a solution to a problem, proceeding one step at a time towards the correct answer, transforming given inputs into wanted outputs. In many situations, there are more than one correct solution. In these cases, an algorithm can produce any of them. An algorithm is considered correct, if it produces a correct answer every time it solves a given problem. (Kingston 1990, pp. 2; Cormen 2009, pp. 5-6) Deep understanding of algorithms is not required to be able to use them in predictive analytics as long as the chosen model is suitable for the task it is performing.

In predictive analytics, algorithms are used to support decision making and automated decisions. Algorithms are iterative and they automate the optimal solution finding process. The decisions algorithms make are based on historical data. Hence, the answer an algorithm produces, is the likeliest thing to happen based on previous events. (Kotu & Deshpande 2014, pp. 4-5; Persson & Kavathatzopoulos 2018) Predictive analytics does not consider different circumstances or human errors that lead to the event either. The solutions are purely based on facts, not taking into consideration the events that lead to the observation, possibly resulting in discrimination. (Sloan & Warner 2018)

Many algorithms are recursive, repeating the same step, until a limiting condition is met, and the solution can be combined by following these steps back. Selecting an appropriate algorithm to use depends on the type and structure of the available data, the objective of data mining, number of records and attributes used, available computing power as well as the possible outliers present in the data. (Kotu & Deshpande 2014, pp. 10) Outliers are observation points, that do not fit the model. They are usually considered as measurement errors. (Outliers 2014) They might also be a sign, that the algorithm is outdated or that it is not considering every possible occasion. Outliers are usually ignored and the analysis concentrates on the data points that fit in the model instead. (Persson & Kavathatzopoulos 2018) The available computing power must be considered, because it affects the efficiency of algorithms.

The analysis of algorithms refers to the resources that are required to use an algorithm. Resources can be for example the needed memory, communication bandwidth and the time the algorithm needs to perform an operation. Also, data structures and the content of available data are important to consider when selecting an algorithm, because none of them work with all purposes so their strengths and limitations play a big role in the selection process. Even with all necessary resources available, algorithms take different time to perform the analysis. (Cormen 2009, pp. 9-12, 23) Running time and performance of algorithms can be evaluated with probabilistic analysis. It is used to estimate the complexity of an algorithm, giving an estimation of the most suitable algorithm for the task. (Mitzenmacher & Upfal 2005, pp. 20)

The process of designing algorithms starts with a clear definition of a problem. Based on this, the needs the problem creates for algorithms can be figured out. It is not relevant to define small details, such as the form of the input data, at this point. Instead, it should be focused on the essential parts of the problem first. (Kingston 1990, pp. 62) The target is usually to minimize computational steps and the memory needed when performing the algorithm to make it efficient (Mitzenmacher & Upfal 2005, pp. 5). In order to use predictive models, the data has to be well structured and either in categorical or numerical form (Sloan & Warner 2018). This makes data categorizing straightforward and finding dependencies easier from input variables.

High error rates when using predictive algorithms are still common. There has been a great progress in the fields of machine learning and predictive analytics, but there is still a lot of work with discovering the aspects of how to make a good prediction. (Sloan & Warner 2018) In many occasions, the outputs of an algorithm are just accepted without further analysis of why things are happening and what consequences this might have. People also tend to collect data from a few sources only, which distorts algorithms. (Persson & Kavathatzopoulos 2018) It is not necessary to understand how exactly do algorithms work, but it would be a good thing to know, why algorithms give the answers they do. This clarifies the analysis process and the effects of the decision made based on analysis are clearer.

An algorithm can be classified as a computer program or a hardware design (Cormen 2009, pp. 6). That being said, if an algorithm is used in the diagnosis or treatment of a patient, it is classified as a medical device. Medical devices have to be approved, before they can be used to treat patients. Medical device is a product, that is used in the diagnosis, mitigation, therapy or prevention of a disease (Baura & Baura 2012, pp. 1-2). These are exactly the targets of predictive models. If a medical device is used in Europe, it needs to have a CE mark. CE marks are producer's confirmation, that the product measures up to all the regulations and standards required from the device. (Suomen Standardisoimisliitto SFS ry)

Medical devices are divided into three groups depending on the risk to users: Class I, Class II and Class III. In Europe, Class II is divided in two parts: Class IIa and Class IIb. Devices in Class

I have a low risk for patients using them, and Class III has the highest risks. The Medicine Evaluation Committee (MEDEV) regulates medical devices in Europe and Food and Drug Administration (FDA) in America. The group where the device belongs to defines the actions that are needed to get the device into commercial use. The higher the device is classified, the more risks it has for patients and thus more actions are needed to get the device approved. It is not excepted to prove that the device works perfectly, instead it is needed to show that the device is tested accurately and in correct clinical settings, so it has as low risks to patients as possible. Algorithms used in diagnosing patients are called Clinical Decision Support Software (CDS). (Harvey) Algorithms, that are not integrated into a physical medical device are usually classified as Class II medical devices. (Huss)

The first step to get the approval for medical device, is to define the intended use for the product. This contains for example the name and model as well as the intended purpose of the device. It is important to describe the usage clearly and as precisely as possible, not exaggerating or embellishing it. This helps not only with the regulation process but also with the development team getting the idea of what they are working with. (Harvey) If algorithms are used in patient care, doctors have to have an understanding of the analysis process in order to explain to patients the outcome of the algorithm and why a certain treatment is recommended. This adds doctors' workload and the need to develop tools, which explain the prediction process in clear terms which are easy to understand. Also, doctors' attitude towards using automated systems in patient care can be negative and a lot of effort is needed the convince them about the positive effects of using predictive algorithms with patients. (Huss) The way to do to this, is to provide reliable results of algorithms used with real patient data.

3.3 Analysis

Predictive analytics uses machine learning, data mining and statistical models to predict future events. Statistical models contain tools which aim to understand data. Also, machine learning takes an advantage of these tools when analysing and constructing models. The idea behind machine learning is to teach a task to a computer with a training set containing examples. After studying these examples, the computer can perform the task with new data. (Louridas & Ebert 2016) Data mining uses machine learning algorithms in order to gain information from large datasets (Sanchez-Pinto et al. 2018)**.** This chapter introduces some of these tools used to analyse data.

3.3.1 Concepts

Methods used with analytics can be either supervised or unsupervised (Yaqoob et al. 2016). Supervised learning means the building of a model in order to reveal relationships between inputs and outputs, when the output variables are known. The difference with unsupervised learning is, that the outcomes are unfamiliar. Instead of finding relationships between variables, it can be used to find naturally occurring patterns. (Han et al. 2011, pp. 330; Sanchez-Pinto et al. 2018) In other words, for each input value (s) x_i, i = 1,…,n there is an associated response measure y_i in supervised learning. The aim is to predict accurately the outcome for future observations or to better understand the relationships between each variable. With unsupervised learning, the input values x_i exist, but no matching response measures y_i. (James et al. 2013, pp. 26-27) The difference between supervised and unsupervised learning is presented in Figure 8.

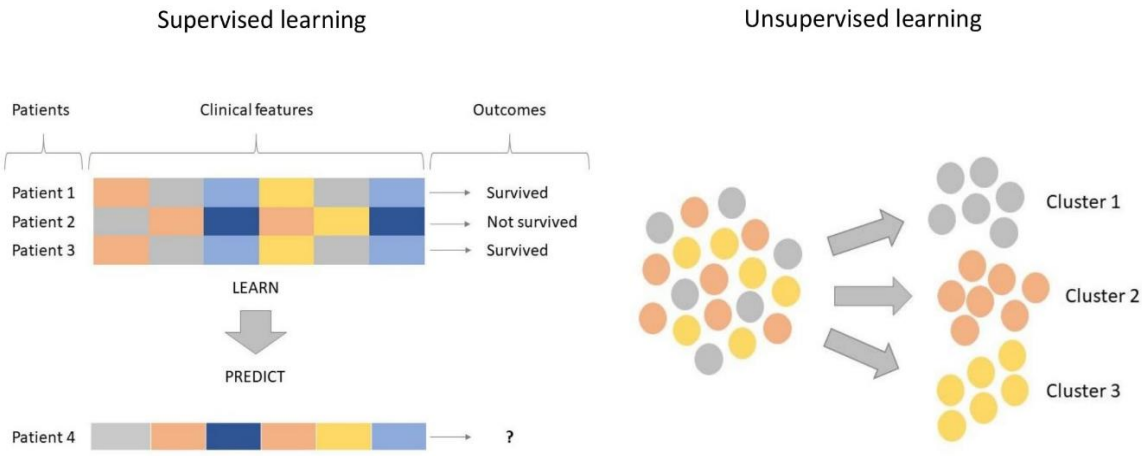

Figure 8 Supervised and unsupervised learning (Sanchez-Pinto et al. 2018)

Supervised learning can be used for example to predict clinical outcomes as the example shows in Figure 8. Patients past symptoms and related outcomes are studied, and the information gained from this can be used to predict future outcomes based on patient's symptoms. When the outcome of an event is not known, the learning is unsupervised. It can be used for example to group similar features together and then study these groups to find similarities in behaviour. (Sanchez-Pinto et al. 2018)

These inputs and outputs can be either qualitative or quantitative. Qualitative values are numeric, whereas quantitative attributes take a value from a certain class or category. An example of qualitative value can be person's age or height, and quantitative values can be things such as eye colour or blood type. (Han et al. 2011, pp. 42-43; James et al. 2013, pp. 28) The used methods depend on which values are used as inputs. There are two approaches for selecting the used methods: regression and classification. Regression can be used with quantitative values and classification with qualitative and they are both supervised learning. (Hastie et al. 2016, pp. 10)

3.3.2 Methods

One of the most used statistical method with predictive analytics is linear regression. Linear regression is a good tool to study quantitative values. Simplest form of linear regression has only one predictor variable Y. Based on that value, a quantitative response Y is predicted and a linear relationship between X and Y is approximated in equation

$$\boldsymbol{Y \approx \beta_0 + \beta_1 X,} \qquad (1)$$

where β_0 and β_1 are unknown constants, representing the intercept and slope terms in the model. However, there is usually more than one variable, making simple linear regression un-useful. A better way is to adapt the model to allow multiple predictors by giving each predictor a separate slope coefficient. The multiple linear regression model,

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,} \qquad (2)$$

where X_j is the predictor and β_j represents the correlation between the response and the variable. (James et al. 2013, pp. 59-61, 71-72; Bajpai 2017, pp. 462-467)

Classification is used to predict qualitative outputs. The most used methods in this approach are logistic regression, linear discriminant analysis and K-nearest neighbours. (James et al. 2013, pp. 127-128) In predictive analytics, classification is used to predict output variables based on given input variables. This is achieved by learning relationships between the predicted output and

input variables from a known data set. (Han et al. 2011, pp. 327-330; Kotu & Deshpande 2014, pp. 13)

K-nearest neighbour (k-NN) algorithm memorizes all attributes and their labels from training data. When unlabelled attributes occur, the closest labelled attributes are used to label these new ones. All the attributes of a data set can be set to n-dimensional space, where n refers to the number of attributes. The k in k-NN algorithm means the number of labelled attributes that are considered when labelling the new attribute. (Liu et al. 2013; Kotu & Deshpande 2014, pp. 99-102) Usage of k-NN algorithm can be seen from Figure 9 with examples when k = 1 and k = 3.

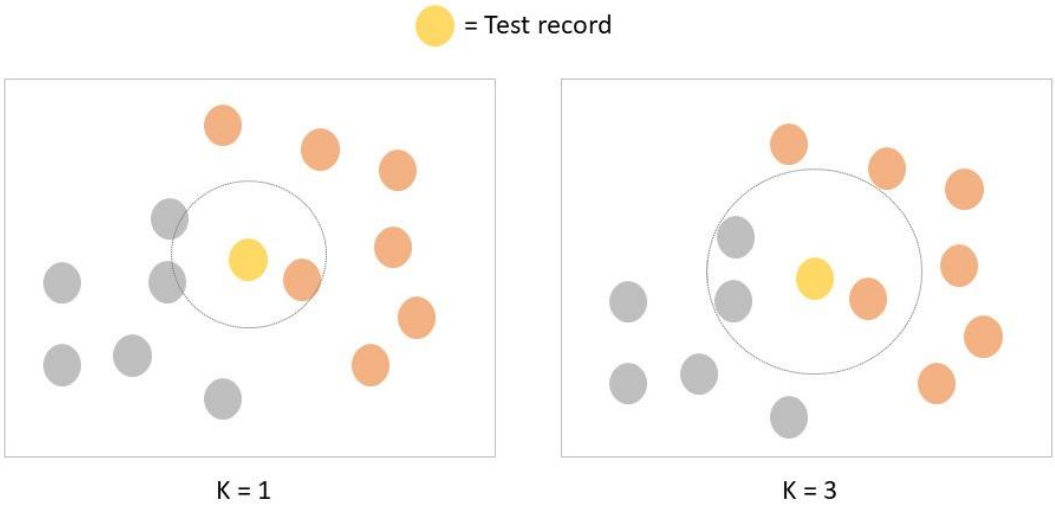

Figure 9 Data set with unlabeled record (adapted from Kotu & Deshpande 2014, pp. 102)

The used k in k-NN algorithm is usually an odd number. This way it is easier to decide between labels, as it is more unlikely to get the same amount for different labels. (Kotu & Deshpande 2014) As can be seen from Figure 9, using k = 1 is quite uncertain, because it considers only the closest attribute when labelling a test record. Labelling is quite straightforward after the closest attributes have been determined. Test record is always labelled as the one which has more labelled attributes present. For example, in Figure 9 when k = 1, the test record gets labelled as orange but when considering 3 closest attributes, it changes to a grey one because of the number of labelled attributes inside the circle.

Another commonly used classifier is Naïve Bayesian. It utilizes the probabilistic relationships between the input attributes and the outcome. (Kotu & Deshpande 2014, pp. 112-114) It simplifies presumptions by assuming that input variables are conditionally independent, meaning that if X_a is true, is does not affect the probability of X_b (Larose & Larose 2015, pp. 424). Naïve Bayesian states that

$$\boldsymbol{P(Y|X)} = \frac{\boldsymbol{P(Y)} * \boldsymbol{P(X|Y)}}{\boldsymbol{P(X)}}, \qquad (3)$$

where X = input attribute set $\{X_1, X_2, \ldots, X_n\}$, Y = outcome, P(X) is the probability of the evidence, P(Y) the probability of outcome, called prior probability, P(X|Y) the class conditional probability presenting likelihood of the event and P(Y|X) conditional probability providing the probability of an outcome when X is known. (Wang, Q. et al. 2007; Kotu & Deshpande 2014; Natingga 2018, pp. 29-30) The Naïve Bayesian algorithm differs from other classifiers because it does not require much work up-front, such as training the model as in k-NN algorithm.

A method used with unstructured data is Natural Language Processing (NLP). It is a method developed for analysing narrative data, which makes it a useful tool for analysing clinical notes,

such as notes from patients' doctor's appointments. (Sanchez-Pinto et al. 2018; Shah et al. 2018) NLP makes it possible to extract information and knowledge from unstructured natural language and converts it to a machine-readable form, making the additional analysis, with for example Hadoop, possible (Nesi et al. 2015). The process in its simplest form, is to assign a label to every word in the document and then use classification algorithms, such as Naïve Bayesian, to analyse these labels and the content of the document (Collobert et al. 2011).

If there is enough data available, the best option for both regression and classification, is to divide the dataset into three parts: one for training, one for validating and one for testing the chosen model. Typically, data is divided so that training data makes up 50 % of the data, and the rest of the data is divided between validating and testing, 25 % for each. To benefit the most from this division, the test set should be kept away from the model until the analysis is ready. This ensures the most precise outcome and correct evaluation of the model. (Hastie et al. 2016, pp. 222) The target is, that the model performs well with both training and test data (James et al. 2013, pp. 128). Test data is used to study if the model is trained properly, and whether the outcome is precise enough.

Neural networks are nonlinear statistical models, that can be used with both supervised and unsupervised learning. The most used one is a single layer perceptron, also known as the single hidden layer back-propagation network. Neural networks were first used to study human brains, hence the name. (Hastie et al. 2016, pp. 392-394) A general network diagram of a single layer perceptron is presented in Figure 10.

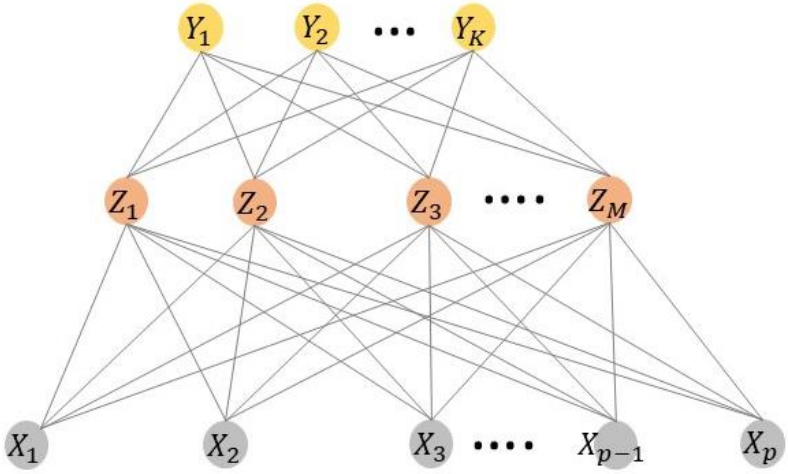

Figure 10 A single layer perceptron network (Hastie et al. 2016, pp. 393)

Figure 10 presents a general network diagram where X_p is the input variable, Z_M different combinations created from inputs and Y_K the target. Z_M is also called a hidden unit, because the values are not presented directly. There can be more than one hidden unit in every diagram. Usually K = 1 when used with regression, meaning there is only one output Y_1. However, the network can handle multiple outputs, that is why the network is presented in its general form. (Schuller & Batliner 2013, pp. 251-252; Hastie et al. 2016, pp. 392-393)

Unsupervised learning uses clustering to divide data into groups. The target is to group observations into homogenous clusters, where observations in each cluster have similarities between each other and dissimilarities between observations in different clusters (Hastie et al. 2016; Bajpai 2017, pp. 660-662), as showed in Figure 8. The clustering process has two steps: assigning observations to their closest centre and redefining the centres. These steps are repeated, until the cluster centres are not changing anymore, and all observations have an assigned cluster. (Tzortzis & Likas 2014; Natingga 2018, pp. 102-110)

The most popular clustering algorithm is k-means clustering, which aims to divide the dataset into clusters, where each observation is closest to the mean of the cluster (Louridas & Ebert 2016; Natingga 2018, pp. 102-110). The k-means equation is

$$\boldsymbol{\varepsilon_{sum}} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{M}} \boldsymbol{V_k} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{M}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\delta_{ik}} \, ||\boldsymbol{X_i} - \boldsymbol{m_k}||^{\boldsymbol{2}}, \qquad (4)$$

where M is the number of clusters, $(C_k)_{k=1}^{M}$ equals the clusters, $V_k = \sum_{i=1}^{N} \delta_{ik} ||X_i - m_k||^2$, and $m_k = \frac{\sum_{i=1}^{N} \delta_{ik} X_i}{\sum_{i=1}^{N} \delta_{ik}}$ are the variance[2] and the centre of the k:th cluster, and δ_{ik} is a cluster indicator variable with $\delta_{ik} = 1$ if $x_i \in C_k$ and 0 otherwise. K-means clustering is a fast and simple algorithm, but its biggest disadvantage is the fact, that the number of formed clusters must be defined beforehand. Therefore, the algorithm should be performed several times, changing the number of clusters each time, to achieve the best result. (Tzortzis & Likas 2014)

4. RESEARCH METHODS

Methodology is a theory of methods and the situations they are appropriate for. It describes what can be achieved with different methods and what kind of philosophy of science is included. Methods consist of chosen techniques as well as the ways to apply those methods. Ontology and epistemology create the methodological assumptions for a research. Ontology is a branch of science which examines things that exist and what kind of conceptions there are for these existing things, whereas epistemology is a study of information and knowledge. (Saunders et al. 2007, pp. 3, 110-113)

This chapter introduces the methodology of the work and research methods selected to support the achieving of a set objective for the research. After introducing the base for conducting a research, the selected methods for collecting data are presented. Also, the analysis process will be gone through.

4.1 Methodology

The objective of this thesis is to find out, what needs there are for predictive algorithms and how models can be selected based on these needs. The algorithms will have an impact on operational activities and patient care. This chapter further elaborates why certain research decisions have been made to give the best support to the realization of the objective of the thesis. The framework for the thesis is created based on the research onion Saunders et al. (2007) introduced. The onion consists of layers, that specify the research process while moving towards the middle layer.

After defining the subject for a research, next important step is to consider the appropriate research methods and ways to collect and analyze data. Used tools and methodologies are chosen in a way they support scientifically valid and reliable realizations of the research the best. The research onion is used to present the demands for the research. It consists of layers which are used to determine the strategy for the research. The outermost layer of the onion consists of the philosophy of the research, which has an effect to the rest of the layers and selections to be made. Innermost layer describes techniques and procedures used for data acquisition. As a whole, the onion has six layers: research philosophy, approach, strategy, choice, time horizon and used techniques and procedures for collecting data. (Saunders et al. 2007, pp. 20-21, 108-109) Each layer can have several selections, and a decision made in one level can have an impact on the available options in inner levels as well. The research onion gives guidelines to perform a research and helps with determining the demands for a research. The research onion and the selections made for the thesis are presented in Figure 11.

Figure 11 Selections for research (Saunders et al. 2007, pp. 108)

Selected research philosophy has an effect to the whole research. Saunders et al. (2007) divides research philosophies to positivism, realism, interpretivism and pragmatism. These can be further divided into ontology, epistemology and axiology. Positivism describes real, already occurred events and ignores uncertain and hypothesized things. Realism emphasizes things, that can be measured and observed, but they have to be in the right context. Interpretivism underlines the role of humans in the research context and their behavior must be understood. Pragmatism emphasizes research questions' role and human experience when determining the research philosophy. The most important thing is to be able to answer research questions. (Saunders et al. 2007, pp. 109-110, 119) Pragmatism was chosen to be the directive philosophy for this research as can be seen from Figure 11, together with all the other choices made for this research.

The second layer defines the approach for a research. These can be divided into two groups: inductive and deductive. Deductive approach is typical for natural sciences, where laws and definitions create a base for explaining events and observations. It starts by studying the theory behind an event and moves to making observations from surroundings, linking the observations to the studied theory. Inductive approach has an opposite way of approaching the situation. It starts with observing the situation and based on the findings, aims to create new, generalized theory. (Saunders et al. 2007, pp. 124-125) This research is deductive in nature, because the purpose is not to create new theory, but to observe the situation Apotti has and make decisions based on that.

The next step is to choose an appropriate strategy for the research. This step describes the ways of how data is collected and defines the form in which the data should be collected as well as the way the research is intended to put into practice. The layer consists of three levels: the meaning of the research, the format of the data and the strategy itself. The most important thing to consider when selecting a strategy, is which one supports the answering of research questions the best, as the philosophy was chosen to be pragmatism. (Saunders et al. 2007, pp. 141-145) Survey was chosen to be the main strategy in this thesis. The data collected describes the present situation in the organization, thus being descriptive in nature.

The *choices* layer has three options: mono method, multi-method and mixed methods. Mono method, as the name suggests, means that only one method is used to collect data for the research. If only one method is used, the collected data is usually qualitative. Multi-method approach uses several methods, either quantitative or qualitative, for collecting data and mixed methods can have both qualitative and quantitative methods in use. Mixed methods are considered to be a form of multi-methods approach. Multi-method or mixed methods are generally used with pragmatism. (Saunders et al. 2007, pp. 151-155) This research collects categorical information, thus making the thesis a multi-method research.

The fifth layer defines the time horizon for the research. There are two options: cross-sectional and longitudinal. Longitudinal research follows the research subject for a certain period of time, ranging from weeks to years, and observes for example changes in behavior during that time. Longitudinal approach allows to find patterns and factors influencing the situation. Whereas cross-sectional study concentrates on understanding phenomena and events at a certain moment, making it possible to analyze the situation thoroughly, but not the underlying reasons. (Saunders et al. 2007, pp. 155) This thesis is a cross-sectional study aiming to define the current needs and the situation in general in Apotti.

The innermost level consists of selecting the used techniques and procedures for collecting data. The methods available are dependent on the choices made in earlier layers and it is important to keep in mind the validity and reliability of the research when choosing the methods for use. Pragmatism does not create any limitations for the needed data or the methods to collect it. Data can be qualitative or quantitative, as long as the methods support the answering of the research questions the best way possible. (Saunders et al. 2007, pp. 156-160) This research uses workshops to collect organizational data about the current situation. Techniques and the data collecting process are further introduced in the next chapter.

The theory was built based on literature to support the empirical research. The aim was to build sufficient theory background to understand the subject and the research questions of the thesis as well as to help with outlining the subject and research questions. The theory background makes it possible to recognize the most important points and trends related to the subject and presents them in a logical way, making it possible to combine relevant topics to your own research. Therefore, it is important to assess references in a critical way when collecting them. (Jankowicz 2005; Gall et al. 2007, pp. 61-66) The aim was to choose references which are peer reviewed, scientific publications with citations from other researches. Sometimes compromises were needed when evaluating the criteria, because all matters, that were wanted to be handled, did not have enough proper researches available.

Creating the theory background proceeds as Saunders et al. (2007, pp. 60-61) introduced. It starts with defining the research subject and creating the research questions, which were presented in the earlier chapter. Based on the research questions, important subjects and terms were identified and search phrases were combined to be used in making searches from different databases. Google Scholar, TUNI's Andor and Scopus were used in this research. Based on the searches and evaluations, a group of references were found to form the base for the theory part. New needs were identified while building up the theory, and new search phrases were created and carried out to meet these needs.

4.2 Workshops

The main method for collecting data in this research is a workshop. The aim of it is to recognize defects, which can be solved with the help of algorithms. The participants were selected in a way, that they represent all the groups which the algorithms have an effect to. Workshop was selected for the main method because it was a good way to collect needs from several users at the same time and bring different user groups together to discuss about their ideas. It was also seen as an easy way to get people from different user groups to interact together and possibly to give them an opportunity to think about the subject from a broader perspective.

Originally the idea was to hold only one workshop for all experts, but due to another event, some of the participants had to cancel their sign-up for the first workshop. It was decided to organize another for them. Due to this, the number of participants in each workshop stayed small, 3 or 4 participants per each workshop. This was noticed to be suitable for the nature of the workshops and gave an option to interact and discuss more with the participants. The workshops were held during April 2019 and the length varied between 60-90 minutes. The arrangements made for the workshops as well as the progress of both workshops are introduced in the following chapters.

4.2.1 Arrangements

The workshops were held to people, who have been participating in Apotti-system's development and have experience working in healthcare. The participants were divided into three groups, representing either special healthcare, primary healthcare or social services, depending on their area of expertise. These are also representing the user groups Apotti-system will have once fully implemented. Workshops gave the participants a possibility to interact with each other and discuss about the needs together as well as share their ideas about the possibilities of predictive analytics.

For the beginning of a workshop, a presentation was prepared to give a brief introduction about the subject. The presentation started with introducing analytics and its usage in general. After that, the presentation moved to handle predictive analytics and its possibilities in healthcare. This was achieved through presenting two case examples of predictive algorithms and factors influencing them. The purpose of the introductory part was to give examples and ideas to participants about where predictive analytics could be used in Apotti and to give the right mindset for participants before starting to work on the given assignment.

The data collecting part of the workshop was achieved through an assignment. The assignment was done in groups, which were formed based on participants' experience in special healthcare, primary healthcare or social services. The purpose of the assignment was to find out, what needs different groups and Apotti-system users have for predictive analytics. Four important subjects were identified regarding the needs. These subjects are: the need, meaning for work, process or workflow and the outcome. These subjects were arranged into a fourfold table, which is presented in Figure 12. The purpose of the table was to survey, what needs there are in different groups and how they could be solved with the help of predictive algorithms.

TARVE ***Need***	**MERKITYS TYÖHÖN** ***Meaning for work***
PROSESSI / TYÖNKULKU ***Process / workflow***	**LOPPUTULOS** ***Outcome***

Figure 12 Assignment for workshop

A fourfold table presented in Figure 12 was handed out to every participant. They were asked to write a need for predictive analytics they have recognized in their own work to the *need* box. The next box, *meaning for work*, was to be filled with subjects, the need has an effect to. *Process or workflow* box contains information about how the need affects process or workflow. The last box, *outcome*, has information about how the need affects the overall situation. The most important part for the thesis is the *need* column, as it will answer the research question *What are the needs for predictive algorithms.* The other parts were considered important to include to be able to understand the context and the elements of which the change would affect.

The researcher participated to the workshops and was actively taking part in the conversation by giving ideas of what to discuss about when thinking about the needs. The other organizers were in charge of answering questions about Apotti and the current situation, while the researcher concentrated on observing and making notes. During the assignment, the researcher also focused on observing the situation and making notes about the conversation the participants were having, thus being an observer as a participant. This means that the participants were aware of the researcher's role and the researcher did not participate to the assignment the same way as the invited participants. (Saunders et al. 2007, pp. 219-222)

4.2.2 Social Services and Primary Healthcare

The first workshop was held in April 5th. The participants were Apotti experts, with background in working in different roles in healthcare. They have also been participating in the development of Apotti-system. We invited persons we thought have a good understanding of what the needs in their area are, and who had been addressing an interest towards predictive analytics, to the workshop. Participants were divided into three groups, depending on what their background was: special healthcare, primary healthcare and social services. A few participants had to cancel in the last minute, because of another event. This caused the lack of special healthcare representatives in the first workshop and answers were collected only from the social services and primary healthcare's point of view. List of participants and their area of expertise is presented below in Table 1.

Table 1 Participants in the first workshop and their background

W1	Participant 1	Social services, drug and alcohol abusers
	Participant 2	Primary healthcare, preventative medicine
	Participant 3	Primary healthcare, preventative medicine

The workshop started with a brief introduction to analytics, covering its utilization possibilities in general and an overview of predictive algorithms and possible cases they could be used in healthcare and with Apotti-system, before giving instructions about the assignment. We reserved 1,5 hours in total for the workshop. The first 45 minutes were dedicated to introducing the purpose of the workshop and an introduction about predictive algorithms, as well as practical applications of predictive algorithms and the last 45 minutes to work with the assignment and introducing the results to others.

Every group got a fourfold table presented in chapter 4.2.1, and a few questions to help them to go through the needs they have in their everyday work for predictive analytics and write them down to the fourfold table. The questions were as follows:

- *Do you get enough information from reporting that allows you to, for example, predict future?*
- *Have you felt the need for information that predicts future to help you to plan your work?*
- *What kind of information do you need to measure the effectiveness of your work?*
- *Have you noticed any situations you could have prevented, if you had proper information available?*

The participants had already thought of the subject before they came to the workshop. They told, that they had recognized needs during a longer period of time, so it was easy to think of them when someone asked about the matter. What proved to be difficult, was to write them down in a way that everyone could understand them. At first, the participants did not need the questions

above, as they were writing down needs they had recognized beforehand. There was a little time at the end, when the questions were assessed and discussed about, and needs were recognized with the help of these questions.

There was a lot of conversation and questions already in the introductory part, where predictive analytics was presented. The conversation was mainly about the situation in Apotti at the moment and questions about if it was possible to achieve the things, that the participants find important, with the new Apotti-system. The answers we got from the assignment part were good, clearly presented and they had obviously been considered beforehand. The needs social services and primary healthcare had, were not so much health related as operational activities and the needs arising from them.

The participants were happy with the content of the workshop. There was a lot of conversation about the subject during the whole workshop, which means that there is interest towards predictive analytics and its possibilities in the organization. The participants also evaluated the workshop being useful for them as well, as they got information about the subject and situation in Apotti while being able to get their voices heard. Based on the received feedback, it was decided to arrange a new workshop for those who were unable to attend to the first one. The second workshop will focus especially on special healthcare experts and their needs for predictive analytics, as they were not represented in the first workshop.

4.2.3 Special Healthcare

The second workshop was held on 17th of April. The participants were experts in special healthcare, and they have been a part of developing Apotti-system. The participants' background is presented in Table 2. All the invited persons were able to join this time. The structure of the workshop was the same than in the first one, but this one lasted only for one hour, because of the short notice to arrange a second workshop. Because of this, the presentation part was compressed into half an hour and the other half was reserved for the assignment. The conversation had to be led back to the assignment a few times, to be able to handle all the necessary subjects during the reserved time. Other than this, the workshop worked well despite of the shorter time.

Table 2 Participants in the second workshop and their background

W2	Participant 1	Special healthcare, oncology
	Participant 2	Special healthcare, cardiology
	Participant 3	Special healthcare, emergency medicine
	Participant 4	Special healthcare, supervisor in health centers

The progress of the second workshop was a little different than the first one's. Because of the shorter time, the assignment part was decided to carry out so, that the participants were discussing about the items in the fourfold table presented in chapter 4.2.1. The researcher was writing down the discussion and needs the participants had, and filled the fourfold table afterwards based on the conversation. Because of this, observing the situation did not get as much attention as in the first workshop. Instead of observing, the researcher was steering the conversation to the right direction to be able to keep to the schedule.

There was no need for questions to help the participants to outline their thoughts and to prepare the ground for discussion in this workshop. As in the first one, the participants had already thought about the subject beforehand in this workshop as well. The needs raised were clearly formed during a longer period and they were seen as important improvements regarding the participants' work and the health of patients. The whole time reserved for the assignment, was used

to discuss about these already recognized needs. Rather than giving ideas about what to discuss, the discussion needed to be restricted at times to be able to deal with every raised issue.

There was a lot of conversation already during the presentation, and the time reserved for it was exceeded a little because of this. The conversation concentrated at first on medicine and their price and efficiency on treatments. Also, artificial intelligent and machine learning was discussed and seen as an interesting opportunity in healthcare. The desire to move from basic reporting to using predictive analytics was also raised. When discussing about the needs, the conversation shifted towards outcome measures and how these could be selected in a way, that health could be measured.

Mentioning predictive analytics moved the conversation to doctors' attitude towards using artificial intelligent in patient care. A concern, that it might not be clear, how healthcare and patients could benefit from using predictive analytics was raised. Because of this, the attitudes can be negative towards new ways of making diagnoses. Apotti-system is using structured documentation when making notes about patients' visits to doctors. Doctors find using structured documentation complicated and time-consuming. Because the benefits are not clear to them, they find it useless. Thus, motivating doctors and demonstrating the benefits of structured documentation is an important step towards adapting to new ways of working.

When discussing about the need of predicting complications in treatments, pharmacogenomics was brought up and its possibilities in selecting the right medication for patients. Pharmacogenomics means utilizing the information in genome to predict the impact of medication to a patient (Earley 2015). At the moment, doctor's overall assessment is used as a measure of possible complications, but a better way would be to use biological measures for predicting, whether a certain medication would be useful to a patient or not.

4.3 Analysis

The data gathered from workshops was collected in Finnish, either in written format or writing down the conversations between participants. Collected data was based on fourfold table that was shared to participants and it was used to outline answers. In addition to this, notes about the conversation during workshops were utilized in the analysis process. The data was analyzed with the help of MS Excel. Because of the nature of the workshops and collected data, the workshops were not recorded or transcribed.

All the materials produced in the workshops were collected and cleaned for analysis purposes. The data was collected in Finnish from the workshops. Thus, the answers have been translated to be able to use in this research. Other than this, the answers were not cleaned much before analyzing. The data collected from different workshops was collected together and the answers were grouped based on the target of the needs.

Originally the idea was to send a survey to the participants of the workshops and ask them to arrange the collected needs in order of importance. After the first workshop it was noticed that the participants were mainly focused on needs that were specific to their work and not adaptable to everyone. Because of this, it was decided not to do a survey, because the survey results would have reflected the distribution of the participants and their area of expertise. It was decided to group the needs based on shared features and create categories that would help when selecting the algorithms for use.

The needs from both workshops were collected together to be able to outline them and to start the analysis. The analysis was made following qualitative data analysis method, where the data is first divided into more manageable parts and then identifying relationships between these parts (O'Gorman & MacIntosh 2015, pp. 140 - 143) This resulted in groups, consisting of similar types of needs. Altogether four groups were recognized from the needs: segmentation, operational, effectiveness and health. These reflect the overall situation in the organization and represent all the needs recognized. These groups were used as a starting point when considering the algorithms to be selected.

5. EMPIRICAL FINDINGS

The usage of data-analytics in healthcare will not remove the need for doctors, but it will alter their work. In the future, doctors will spend less and less time diagnosing patients and concentrating more on instructing and taking care of them. Predictive analytics will be used for making diagnoses. Analytics can also be used with personalized medicine. This is based on data-intensive pharmacogenomics, which concentrates on understanding molecules in order to develop more advanced treatments. (Earley 2015) Apotti is only starting to explore the possibilities of predictive analytics, but their possibilities are already being recognized and different algorithms are even asked for.

The needs for predictive analytics were defined with the help of workshops. They were thought to be the best way to get experts from different fields together to discuss about the possibilities of predictive analytics in their work and collect ideas about future possibilities. The needs were collected through an assignment and a fourfold table presented in chapter 4.2.1 was distributed to the participants to help with collecting the needs. Only the first column, need, of the fourfold table was used when analyzing the results. The other columns are used to understand the subject and topics it has an effect to.

This chapter answers the first and second research questions, which were about the needs and selection criteria for predictive algorithms. This chapter introduces the results of the workshops, which were held in the need of collecting information about the current situation in Apotti regarding predictive analytics. The results are analyzed and introduced more closely to form a picture of what is needed form predictive analytics in healthcare. At the end, the results are gathered together and reflected on how to pick predictive algorithms.

5.1 Analysis of Needs

The needs collected from primary healthcare concentrated on segmentation and the possibilities it brings, such as doctor's appointments in different lengths for different patient groups or directing patients to self-care services or to use electronic services instead of appointments. Social services needs were rather scattered, and one big theme around the subject was not found. They also raised the need for better allocating of resources, which would be accomplished for example by predicting the recovery time from substance abuse. Special healthcare needs were more focused to patients' health and treatment options, but also a need for resource management was raised, especially related to doctors' workload and the ways to unburden it and to share the resources more even in general.

All the collected needs from workshops are presented below in Table 3. The needs are grouped based on whether they were collected from primary healthcare, social services or special healthcare experts.

Table 3 Needs collected from workshops

Primary healthcare	Assessing primary healthcare and its effectiveness (how to improve the effectiveness of the service)
	Recognizing patients / clients, who will benefit from self-care services and the points, when face-to-face contacts are needed
	Recognizing customer segments
	Recognize patients / clients, who would benefit from using electronic services

Social services	Gain information on social- and healthcare services' effectiveness compared to uncoordinated services (for example the elderly, disabled)
	The impact of the ones using many services to the total costs / their experience in getting the service they need
	Occupancy rate for alcohol and drug rehabilitation in institutions (detoxification, rehabilitation)
	Prognosis for recovery time (from substance abuse)
Special healthcare	Predicting the outcome of cardioversion
	Predict the cause for patient going to ER or hospital ward because of complications in treatment
	Measuring demand
	Measuring the load rate (for example during epidemic, phone calls to ER)
	Measuring the work load of doctors in ER's. Following the number of patients in different times and how many patients a doctor can treat in a certain time

Primary healthcare needs, as can be seen from Table 3, centered around grouping patients and clients into segments, to make the offered care more personalized and to offer different treatment and service options for different groups. Also, a need for appointments in different lengths for patients and clients in different groups and with different symptoms was raised in the conversation. It would be important to recognize patients and clients who would benefit more for example from electronic services than doctors' appointments, to save the appointment slots to the ones needing them. The grouping of patients and clients would also help with allocating available resources better, where and when they are needed and, to plan the daily work to support the needs arising from the patients and clients better.

Also, a need for better allocating of resources came up during the conversation with primary healthcare, as well as better understanding of patients and clients' needs and hopes, to be able to organize services in a way the users are happy with them. This would require finding out, what factors make a good care and if there are differences in the way different groups see the quality of the care. One question is, if common features can be found from the way people assess the evaluation of care or if it is dependent only on individual's viewpoint. Better resource management would have an effect to the cost structure of healthcare as well as the satisfaction of patients and clients towards the services they receive.

The needs in social services were more scattered, and there was not one common theme around them. Social services also raised a need to allocate resources better. A way to solve this for them, is to better asses the occupancy rate for alcohol and drug rehabilitation institutions. An important factor in this is to find out how different public holidays, especially the ones having different dates yearly (such as Easter), affect the occupancy rates and if extra resources are needed in certain times. Also, being able to give an estimation of clients and patients recovery time from substance abuse, would be a helpful tool to help dividing resources better and more effectively.

Social services had a need to also find out, how patients and clients estimate the overall process related to their treatment, and if they find the services and treatments they got, helpful. This would help them to develop and plan the work of social services more precisely and to offer services, that would benefit patients and clients the most. Information about the impact of coordinated services and patients and clients using many services to the total costs was also raised. This means, that there is a need to investigate the workflow and see, if it is helpful to the ones receiving it and to treat patients and clients' different problems and symptoms simultaneously as a whole, not in separate processes. Also, the ones receiving many services and the overall cost of their services is wanted to be evaluated and find out, if the clients and patients find this process useful.

The answers got from special healthcare indicate, that they want the work to be more seamless, because the wellbeing of a patient is always their priority. The better the supportive tasks are planned, the more time doctors have for taking care of the patients. The need for predicting the outcome of cardioversion is raised from this. Now the treatment is to perform a surgery and see if it helped the patient or not. The need is to predict before the surgery whether it would be helpful for the patient or not and make the decision, whether to do the surgery based on that. There is always a risk for patients in surgeries and this way unnecessary surgeries could be avoided and make the waiting lists for surgery shorter.

The needs in resource management for special healthcare concentrated on finding out the work load of doctors and recognizing the situations, when the work load is bigger and divided unevenly between health centers. One solution to this would be to show patients and clients the busiest moments of the day, so the ones without acute symptoms would be able to choose a quieter slot for their visit. So, there is a need to discover the busiest days and hours and how different holidays affect the number of patients. This would affect the way healthcare professionals evaluate their workload and stressfulness as well as the length of waiting lists and available appointment slots for patients to see doctors.

Some of the needs, like the process of the ones needing many services and segmentation of patients and clients, are not directly related to predictive analytics as they are more about finding out the current situation. Being able to truly understand the factors, that lead to the current situation, helps with planning future work and making changes to make the process better and more effective. Even though the needs themselves are not necessarily predictive analytics, the reasons behind them can be identified to be related to future prediction. For example, the intention behind segmentation is more effective resource planning and thus predicting the rate of occupancy in different situations and with different services.

All in all, better allocating of resources came up from all the groups, albeit the ways to achieve better resource management were different. Otherwise, the needs raised were quite specific to the areas the participants were most familiar with. Thus, the needs themselves cannot be generalized, but there are different categories that are recognizable from the needs. All in all, the needs could be categorized easily with a little generalization and it was noticeable, that there were four bigger themes around them. These themes form the base for the categorization, which are discussed in the following chapter.

5.2 Selection Criteria

It is mentioned in chapter 3.2, that factors affecting the choosing of algorithms include data type and structure, the objective, number of records and attributes used, available computing power and outliers in the data. Because of the nature of this research, the only thing affecting the selection is the objective, that is to answer the needs raised in the organization as well as possible. The other factors are considered when deploying the chosen algorithms. Although, it can be assumed, that some of these factors have been considered in the development process, because the algorithms are being developed in the same organization from whom the information system is acquired from.

When assessing the collected needs, it was noticed, that they can roughly be divided into four categories based on what is wanted to achieve with them. These categories are segmentation, operational, effectiveness and health. Segmentation and health categories were easy to recognize, and it was obvious they would form two of the categories. The other two categories were formed based on the needs left and these were thought to best represent the needs as a whole. These recognized categories, as well as a summary of the included needs, are introduced below in Table 4.

From the needs primary healthcare raised, a first category was clearly visible, the need to group patients and clients into groups to be able to personalize their treatment recommendations. Thus, the first category was *segmentation*. Another clear category formed based on special healthcare's need to predict the outcome of treatments. This category was named *health*. There

was also a need to measure the effectiveness of services and it forms one category, *effectiveness*. The rest of the needs consisted of varying needs that affected daily operations, consisting mainly from cost and resources related needs. These needs were decided to leave together to form *operational* category.

Table 4 Categories identified from collected needs

Segmentation Dividing patients into groups based on their needs	**Operational** Needs affecting daily operations
Effectiveness Measuring the effectiveness of services	**Health** Predicting the outcome of different treatments

The segmentation category in Table 4 consists of needs, which are aiming to divide patients and clients into groups, in which they would have similar features, to offer them more individual treatment options. This also has an effect to resources as they can be planned more precisely according to patient's and client's needs. The needs in the segmentation category are:

- Recognizing patients / clients, who will benefit from self-care services and the points, when face-to-face contacts are needed
- Recognizing customer segments
- Recognize patients / clients, who would benefit using electronic services

The needs in this category were straightforward and related to grouping patients and clients. The purpose of this is to recognize different customer groups based on their common features, to offer patients and clients services that would be suitable for them and their symptoms. Because of this, the same treatment options are not offered to everyone, making healthcare more personalized and effective. This also means, that there is no need to reserve an appointment slot for everyone, as their situation might be treated without medication.

The operational category consists of needs, that have an effect to daily operations. The target of these needs is to find out the work load of single staff members or the whole health center as well as to find out the effect of a certain treatment process to overall costs and patient's or client's experience about the treatment. The needs are:

- The impact of the ones using many services to the total costs / their experience in getting the service they need
- Occupancy rate for alcohol and drug rehabilitation in institutions (detoxification, rehabilitation)
- Measuring the work load of doctors in ER's. Following the number of patients in different times and how many patients a doctor can treat in a certain time
- Measuring demand
- Measuring the load rate (for example during epidemic, phone calls to ER)

As mentioned before, this group consists of varying needs and there is not one common feature to describe all of them. Instead, all these needs have an effect to the daily operations of health centers and hospitals. These needs affect mainly to healthcare costs and the allocating of resources, especially through occupancy rates, but they also improve patient's satisfaction and the well-being of employees.

The needs grouped under effectiveness are aiming to measure the effectiveness of offered services, including different treatment options. Effectiveness in this context means measuring how patients and clients experience the service they have received and if they find it useful, and what kinds of measures should be observed and data collected to measure the effectiveness. The needs categorized under effectiveness are:

- Assessing primary healthcare and its effectiveness (how to improve the effectiveness of the service)
- Gain information on social- and healthcare services' effectiveness compared to uncoordinated services (for example the elderly, disabled)

All the needs in this group are targeting to measure the effectiveness of offered services. Recognizing the features affecting to effectiveness would affect to patients' medication and the treatments offered to them. Knowing how they evaluate the effectiveness, would help to offer more suitable treatment options for them and to develop the offered services so that they are evaluated to be more useful. This affects directly to patients and clients' health and satisfaction of offered treatments and indirectly also to employees' well-being at work through balancing their workloads.

The health category consists of needs that aim to predict the outcome of a treatment to prevent unnecessary operations and medications. This also allows the treatment to move more towards personalized healthcare, predicting the outcome based on patient's former symptoms. The needs categorized under health are:

- Predicting the outcome of cardioversion
- Predict the cause for patient going to ER or hospital wards because of complications in treatment
- Prognosis for recovery time (from substance abuse)

These needs are affecting directly to patients' health and well-being and they reduce the number of experimental operations made to patients, where the outcome is not certain, because there are no guarantees of some treatments working for every patient. There is also a need to find out the elements affecting recovery times, especially when recovering from substance abuse and if these elements have an effect to the time spent in rehab.

All these needs can be presented through the category they are grouped to as well as the field of expertise the need is collected from. In Figure 13 the needs are presented through these features. There are two things, that can be read from the figure. Firstly, the different categories each group has related to their needs and secondly, the number of needs in different categories. The needs are divided between the expert groups in which they were raised. For example, we can see that special healthcare raised five needs in total. Three of these needs were operational-related and two of them health-related.

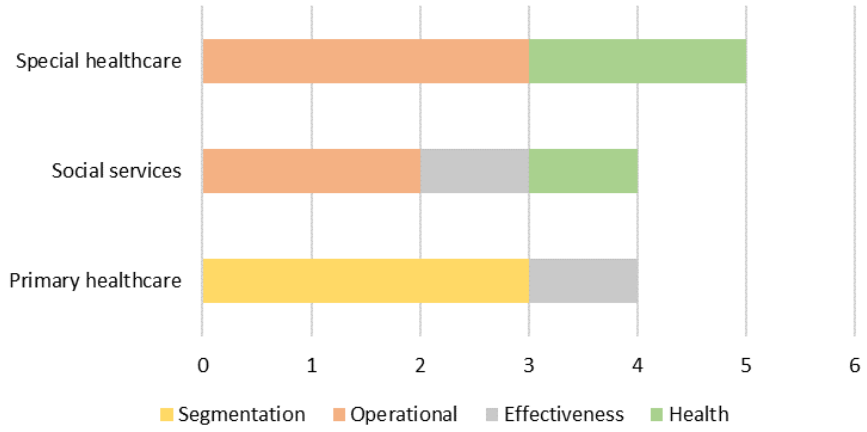

Figure 13 Needs grouped by field of expertise

None of the groups had the need for improvements in all four categories, as can be seen from Figure 13. Social services raised needs from three different categories, whereas the rest of the groups had needs from two of the identified categories. Primary healthcare was the only group to have needs regarding segmentation. Effectiveness was important to social services and primary healthcare, operational needs was raised from special healthcare and social services and health-related needs were raised by special healthcare and social services.

When looking at the ready-made algorithms, from which the final selection is made from, it can be noticed, that there are no effectiveness-related algorithms available. Thus, only three of the four identified categories are relevant when making the final selection. These categories, segmentation, operational and health, mean that the algorithms selected for use should support daily operations and the allocating of resources as well as grouping patients and clients to better respond to their needs regarding their health and treatment options.

Because effectiveness is left outside from the criteria, it can be seen from Figure 13, that segmentation is the only category to answer the needs for primary healthcare. The remaining categories, operational and health, are divided between social services and special healthcare. Operational needs make up roughly half of the needs when combining the selected three categories, and the remaining needs are divided evenly between the other two categories. To fill the needs for every group, algorithms from each of the three categories should be chosen to use, if possible.

There is an additional need for algorithms which measure effectiveness, as there are currently none available to be selected. Needs to measure effectiveness were raised from two different user groups, social services and primary healthcare, so they cannot be left unnoticed. At first, it should be studied what kind of factors are influencing the experience of effectiveness, and this way select proper measurements to be able to collect the needed data. Based on the influencing factors and available data, the related variables can be selected, and an algorithm built.

The algorithms are chosen for use so, that first the needs are compared to the available algorithms to see, if there are any developed for the wanted purpose. After this, the categories and the targets of them are checked, and algorithms to help to achieve these goals are selected to reach the wanted number of algorithms. So, all the selected algorithms do not have to fill a need, as long as they support segmentation-, operational- or health-related goals. Primarily algorithms are chosen to respond to a need and secondary to fit categorical goals. After checking the needs and categories and comparing them to the available algorithms, the points of improvements should be recognizable, meaning situations where a need occurs but no algorithm is available to respond to it.

Because Apotti is not using any predictive algorithms yet, they should at first concentrate on algorithms, that are general, focusing on the big picture instead of one, specific area. Using more general algorithms makes more sense, because it allows to survey the situation and to see, if the algorithms are suitable to be used in Finland with the selected variables. It needs to be kept in mind, that Epic is an American company, meaning that the data used to develop algorithms is most probably also from America. Thus, the variables might be somewhat different in Finland.

When algorithms are being deployed, it is important to pay attention to the chosen variables and that they fit the data. Also, following the work of algorithms and the accuracy of their predictions needs to be followed to make sure the algorithms are suitable to Finnish environment and that the algorithms have been configured properly. Before deployment, algorithms should be trained with Finnish data. Because Apotti is a relatively new information system, there is not much data either, thus the suitableness of the algorithm cannot be directly based on the available data and the results of the training data.

6. DISCUSSION

Predictive models are not generally used in healthcare in Finland. This is because different systems do not communicate well together, and data collected from them is in a form that cannot properly be utilized in data analysis. As Apotti-system is relatively new, large amount of data has not yet been collected. As the implementations are proceeding, more and more data will be collected and with time, Apotti-system will be learned to be utilized better, leading to increasing the amount of data. The implementation-phase of the algorithms is demanding, as there is a lot of work with validating the algorithms and verifying the fittingness of the variables. After implementation, it should be followed, if the algorithms meet the target.

All levels of data analytics were presented in this thesis. The research focused mainly to predictive analytics, as the target was to recognize the possibilities of predictive models in healthcare. The purpose was not to handle healthcare data and the situation through all levels of data analytics, but to create a comprehensive picture to readers about them to help to understand the overall process and utilization of results. The focus of data analytics now is in prescriptive analytics in Apotti, but there is a wish to start using predictive analytics in some areas in the foreseeable future.

The criteria to choose algorithms obtained through this research describe the present situation, and areas that should be developed further in the organization. The needs collected were about predicting the future, improving workflows and making the work of professionals more seamless. It is possible to make a decision about which algorithms to use based on the recognized needs. Implementing the selected algorithms also open new research possibilities. It can be for example found out, how well the selected algorithms support the needs recognized in this research in the organization.

The number of people the needs were collected from were quite small. It would be interesting to see, how the results would change, if more workshops were held to include more people within the same organization. Now the results represent all the user groups Apotti-system has, but they come from a limited perspective. However, the groups created based on the needs, would probably not change, since they present well the overall situation in the organization.

Apotti deals with healthcare data, which brings special requirements for data protection, because the information handled is very sensitive in nature and cannot be read by anyone who does not have the right to do so. These factors must be kept in mind when using and analyzing the data. In top of that, big data brings its own challenges for the processing and utilization of data. These characteristics have to be considered when planning the implementation of algorithms and made sure the process offers proper protection procedures for personal data and its management.

If there was a corresponding algorithm to every need raised and they were selected to be used, it would have an effect to patients as well as healthcare professionals. At first, using the algorithms and learning where they can be utilized requires some learning, but after they have been in use for a while, they will make the workflow more seamless, help diagnosing patients and clients as well as help to plan the schedule so, that needed professionals would be available when they are needed.

7. CONCLUSIONS

The aim of this research was to recognize needs Apotti-system users have for predictive analytics and what kind of criteria for the selection can be formed based on these needs. Research questions of this thesis were answered through a theory part based on literature, helping to understand the subject, and empirical research, which was used to specify the needs of the organization.

This chapter presents a summary of the whole work as well as the key findings and results which were achieved through this research. The findings are also connected to theory and they are discussed in a broader context. After this, the research is critically evaluated and identified limitations are presented. At the end of this chapter, future research possibilities are identified and discussed.

7.1 Summary

The aim of the research was to find out, what kinds of needs Apotti has for predictive analytics and what kinds of algorithms should be selected for use based on these raised needs. At first, theory background about big data and predictive analytics based on literature was formed and a base to understand analytics and predictive algorithms usage in healthcare. The needs were defined with the help of workshops, that were held to healthcare professionals who also had prior experience about Apotti-system and its development. Based on the findings from workshops, the needs were recognized, and a suggestion was made about which algorithms to choose.

Literature was used to study big data and predictive algorithms. Big data is fast growing data collected from multiple sources in different formats, that cannot be analyzed with traditional methods and tools. This requires the adopting of new working methods and learning new skills to be able to utilize big data in business related decisions. The process of big data utilization can be divided into two parts: data management and analytics. Data management consists of data gathering, processing, unification and storage related activities. Whereas analytics part includes different tools and methods that can be used to model and analyze big data. One of the most used tools for this is Hadoop's MapReduce.

A good example about big data is healthcare data, which is one of the fastest growing and biggest datasets. Data is being collected from several sources, including information systems like EMR and EHR, which collect patient related information in electronical form. In addition to this, data is being generated constantly from devices attached to patients. What needs to be considered with healthcare data and algorithms, is that if algorithms are used related to patient's treatment, they are classified as medical devices. This means that they have to be approved and proved to be safe and meet other demands before they can be used.

The understanding of predictive analytics starts with the levels of data analytics. These are descriptive, diagnostic, predictive and prescriptive analytics. Before adopting predictive analytics, the processes of finding out what happened and why, must be under control and in use. When this base has been constructed, the process of predicting the future and finding out related events to build algorithms can be started. There is still one level after predictive analytics, prescriptive analytics. It is the hardest one to achieve. It starts with defining a point, which is wanted to achieve, and the tasks required to reach that point.

Algorithms are logical answers to a problem, proceeding step by step towards the correct answer. In predictive analytics, algorithms are used to support decision making and automating decision making. Assessing of algorithms means the needed resources, such as the objective, available computing power and outliers, to use an algorithm. These resources must be considered also when building an algorithm. There are several different ways and methods to use when building an algorithm, which can be divided to supervised and unsupervised learning methods.

The most used methods with predictive algorithms are linear regression, k-nearest neighbor and k-means clustering.

The empirical part of the research was implemented through workshops, from which data was collected to support the selection and implementation of algorithms. The intention was to define the needs for Apotti-system with the help of conversation and an assignment, to form criteria to guide the selection. This was achieved rather well, the participants were interested about the subject and its possibilities in their work and they had been thinking of the needs they have beforehand. The rest of this chapter concentrates on introducing the research questions and their answers.

What needs, and limitations healthcare data creates for predictive algorithms?

This research question was answered through literature. The aim was to recognize factors, that have an effect to algorithms used and developed with healthcare data. Few factors worthy of attention were recognized, including one bigger factor that must be considered already in the development process.

One thing to be considered when developing algorithms is that if they are used in the treatment process of patients', they are classified as medical devices. This creates restrictions and demands for the usage of the algorithm and for proving its safety. These algorithms need a CE mark, which is approved by MEDEV, when used in Europe. FDA regulates medical devices and their usage in America. Medical devices are divided in three groups, which define the tasks that are needed to get an approval for use. The higher the device is classified, the more proves are needed to get the approval. Algorithms that are not integrated to a device, usually belong to class II.

Several smaller things were recognized to be considered with algorithms, but which are not affecting the development or selection. What is noteworthy is for example that with complex data, simple algorithms are noticed to be more efficient and work better than complex algorithms. Healthcare data is usually collected from several sources, making the structure varying and complex, thus the used algorithms are recommended to be simple by structure. There are no common standards for sharing and storing data in healthcare. This makes data sharing between different devices challenging affecting also to the usage of algorithms, because the available data might not be cohesive, or all information is not available.

What is remarkable in this research, is that the usage of healthcare data between different countries in prediction might not be directly comparable. The factors between a disease can vary between countries, in which case the variables used with algorithms need to be adjusted to fit the new data. Notable is also that it is recommended to use distributed data networks with healthcare data. Data is then located in different nodes, and control of the data is with the original owner and if required, it can be shared with others.

What are the needs for predictive algorithms?

The needs Apotti-system users have, were collected in workshops, where experts from the same field had an opportunity to discuss together and consider what kind of need they have in their works for predictive algorithms. The participants were divided into three groups, representing social services, primary healthcare and special healthcare. The needs between different user groups are quite different and they are not directly generalized to be utilized with all the groups. In Table 5 these needs are collected grouped by topics.

Table 5 The needs

Recognizing patients / clients, who will benefit from self-care services and the points, when face-to-face contacts are needed
Recognizing customer segments
Recognize patients / clients, who would benefit using electronic services

The impact of the ones using many services to the total costs / their experience in getting the service they need
Occupancy rate for alcohol and drug rehabilitation in institutions (detoxification, rehabilitation)
Measuring the workload of doctors in ER's. Following the number of patients in different times and how many patients a doctor can treat in a certain time
Measuring demand
Measuring the load rate (for example during epidemic, phone calls to ER)
Assessing primary healthcare and its effectiveness (how to improve the effectiveness of the service)
Gain information on social- and healthcare services' effectiveness compared to uncoordinated services (for example the elderly, disabled)
Predicting the outcome of cardioversion
Predict the cause for patient going to ER or hospital wards because of complications in treatment
Prognosis for recovery time (from substance abuse)

All the needs collected from workshops have been collected to Table 5. As mentioned before, the needs were different between different expert groups, but for example better allocating of resources was raised from more than one group. There were different ways to achieve this result, for example primary healthcare has the need to group patients and clients into segments, social services would benefit from finding out the occupancy rate for alcohol and drug rehabilitation centers and special healthcare would measure the workload of doctors and health centers. The common feature for all of these is the target to plan work and working hours better, to have the needed resources always available.

In addition to this, a need to make the treatment of patients' and clients' more personalized, for example through segmentation or figuring out the factors affecting the recovery time from substance abuse, was raised. Allocating and the planning of treatments would be more effective for different patients because of this. This also has an effect to resources and the better allocating and sharing of them. The segmentation of patients' and clients' would allow offering different treatment options for different groups and doctors' appointments in different lengths for the ones needing them.

Also needs regarding the effectiveness of the treatment were raised as well as finding out how patients' and clients' evaluate the treatment and services they received and whether the received treatments or services respond to their needs or not. Also, factors affecting directly to patient's health were raised, including things like predicting the outcome of cardioversion and complications in treatment, especially the ones where patient is in danger of going to ER or hospital ward.

What are the criteria for selecting predictive algorithms?

The criteria for choosing algorithms was determined by identifying different categories from the collected needs. The defining of categories was straightforward, consisting of dividing the needs in similar groups by identifying the targets. Four categories were identified based on the needs. These categories are:

- *Segmentation*
- *Operational*
- *Effectiveness*
- *Health*

Based on the needs and categories identified from them, there are needs for algorithms that will group patients' and clients' based on shared features, allowing to plan treatments and services targeted to these groups. Also, algorithms that help with every day work, such as predicting occupancy rates, prognosis for recovery times and measuring demands in different times, are needed. In addition, algorithms that help maintaining patient's health and predict their probability to some events are considered an important target of application with predictive analytics.

There is also a need to measure, how patients' and clients' evaluate the treatment and services they receive, as well as the effectiveness of them. It was noticed, that there are no algorithms developed to serve this purpose, and this category was left out from the criteria to select algorithms. Still, these needs must be acknowledged, and new algorithms developed to help to measure effectiveness.

Primarily, algorithms are selected based on the collected needs and how well they respond to those needs. Secondarily, to make the number of algorithms complete, algorithms that answer to the targets of the identified categories are chosen. At the end, there is a group of algorithms, that support the needs raised and that will make the work more effortless. The result of the work and answers to the research questions were achieved through literature and empirical work. The outcome was a proposal of how to select algorithms and what kind of algorithms are needed. The proposal is based on needs collected in workshops from Apotti-system users and the analysis made from the needs in question.

7.2 Research Evaluation and Limitations

The evaluation of empirical research can be divided into four parts: construct validity, external and internal validity and reliability. These have been used to evaluate also the reliability and quality of this research. (Yin 2014, pp. 45-46) It is important to include the evaluation of all the abovementioned parts to form a comprehensive view of the validity of the research. There are several ways to achieve reliability through these parts, the ways used in this research are presented in this chapter.

Construct validity evaluates the validity of the research through the used source materials (Yin 2014, pp. 46-47). The validity of this research was achieved through including several references when constructing the theory and assessing all the used sources. Sources that were found unreliable, were not used. Also, the relevance of the information was confirmed from multiple places. The theory part for example, consists of multiple references to form a cohesive study that reflects the subject from multiple viewpoints.

Internal validity concerns all the events related to a certain condition or event and is usually used with explanatory studies (Yin 2014, pp. 46-48) In general, it is about explaining the findings of a study and finding reasons for them to happen and ruling out others. This is not that relevant concerning this research, as the findings are not related and there are not cause and effect relationships between them.

External validity deals with the generalization of the research (Yin 2014, pp. 46-48). The research and its implementation itself are able to be generalized. The research setting allows a broader data collection from bigger groups, but with the used sample size and selected participants, generalization is not possible to other areas. The research and achieved results concern strictly to areas using Apotti-system. The generalization of this research suffers from the fact, that the subject is quite specific, and the used sample size is rather small. It is suitable for reaching the goals of this research, but the results cannot be utilized with the whole country without future research.

Reliability means the documentation of the research execution in a level, that it can be repeated with the same results, without forgetting the quality of the research (Yin 2014, pp. 46-49). The progression of the workshops is documented with details, so that it would be able to be repeated and similar results could be collected. Similar way, the research settings are described as detailed as possible, to ensure the quality of the research. Some limitations to reliability exist.

The articles found about using predictive analytics in healthcare were mainly about the possibilities of adopting certain methods in healthcare, not that much about experiences of successfully implementing and using them. Therefore, the possibilities of predictive analytics in healthcare should be approached carefully keeping in mind, that the results might take time to appear and that a lot of work is required, before the algorithms are successfully implemented and the full potential of the algorithms are released.

7.3 Suggestions for Future Research

Analytics is developing fast and the usage of predictive analytics is also becoming more common. At the moment, it is still in early development, but the industry is taking big steps forward. Along with new information systems and the data collected, the opportunities to utilize predictive analytics are also increasing. However, this all takes time, as does the validation of predictive models. Thus, the benefits of predictive algorithms take time to be visible and justifying the usage for the users might be challenging because of this.

This research could be carried out with a broader sample size and a different approach to the used methods. A survey would be a good tool to collect a large amount of data with a reasonable effort (Saunders et al. 2007, pp. 93-94). The scale of this research is suitable for the targets of the study, but to get more generalized answers, a bigger sample size is needed. It would also be interesting to see, if the sample size affects for example the categories recognized. Also, the needs in different user groups could be evaluated more closely and to do a follow-up interview or questionnaire to observe, if new needs have occurred.

After implementing the algorithms, a possible follow-up research can be done, focusing on finding out, how well the algorithms respond to the needs raised in this research. The new research could include finding out if the algorithms used are useful for Apotti-users or do the users think some of the algorithms should be decommissioned. It could also be studied, if new needs have been raised and how the algorithms have been received for use in general. Also, an important follow-up research subject would be, what kinds of attitudes the users have for predictive algorithms, do they find them useful or a waste of time and are the results of algorithms trusted and included in the decision-making process.

Another interesting follow-up research is related to the healthcare data variation between countries. Implementing these algorithms would allow a research to be conducted about how the chosen variables perform with data collected from Finnish patients and clients and if there has been a need to re-configure some of the variables. The affecting factors behind diseases might differ in some level between countries, but additionally, it can be found out if, for example the algorithms helping with resource management, have been implemented with the same variables in every country.

REFERENCES

Accenture About Us web page. Available (accessed 20.2.2019): https://www.linkedin.com/company/accenture.

Apotti internal (2019). Internal presentation.

Apotti Apotti as a Companyweb page. Available (accessed 19.2.2019): https://www.apotti.fi/en/oy-apotti-ab-2/.

Bajpai, N. (2017). Business Research Methods, 2nd Edition, 2nd ed. Pearson India.

Banerjee, A., Bandyopadhyay, T. & Acharya, P. (2013). Data Analytics: Hyped Up Aspirations or True Potential? Vikalpa: The Journal for Decision Makers, Vol. 38(4), pp. 1-12.

Bates, D.W., Heitmueller, A., Kakad, M. & Saria, S. (2018). Why policymakers should care about "big data" in healthcare, HEALTH POLICY AND TECHNOLOGY, Vol. 7(2), pp. 211-216.

Baura, G.D. & Baura, G. (2012). Medical Device Technologies, Academic Press.

Belle, A., Thiagarajan, R., Soroushmehr, S.M.R., Navidi, F., Beard, D.A. & Najarian, K. (2015). Big data analytics in healthcare, BioMed Research International, Vol. 2015 pp. 370194-16.

Callen, J., Paoloni, R., Li, J., Stewart, M., Gibson, K., Georgiou, A., Braithwaite, J. & Westbrook, J. (2013). Perceptions of the Effect of Information and Communication Technology on the Quality of Care Delivered in Emergency Departments: A Cross-Site Qualitative Study, Annals of Emergency Medicine, Vol. 61(2), pp. 131-144.

Cao, L. (2017). Data Science: A Comprehensive Overview, ACM Comput. Surv., Vol. 50(3), pp. 43:1–43:42. http://doi.acm.org/10.1145/3076253.

Carter, D. & Sholler, D. (2016). Data science on the ground: Hype, criticism, and everyday work, Journal of the Association for Information Science and Technology, Vol. 67(10), pp. 2309-2319.

Chen, M., Mao, S. & Liu, Y. (2014). Big Data: A Survey, Mobile Networks and Applications, Vol. 19(2), pp. 171-209.

Chen & Zhang, C. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, Vol. 275 pp. 314-347.

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M. & Welton, C. (2009). MAD skills: New analysis practices for big data, Proceedings of the VLDB Endowment, pp. 1481-1492.

Cohen, Amarasingham, R., Shah, A., Xie, B. & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care, Health affairs (Project Hope), Vol. 33(7), pp. 1139-1147. https://www.ncbi.nlm.nih.gov/pubmed/25006139.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011). Natural language processing (almost) from scratch, Journal of Machine Learning Research, Vol. 12 pp. 2493-2537.

Cormen, T.H. (2009). Introduction to algorithms, 3rd ed. MIT Press, Cambridge, MA.

Davenport, T.H., Harris, J.G. & Emberson, C. (2007). Competing on analytics: the new science of winning, 322-324 p.

Davenport, T.H., Harris, J.G. & Morison, R. (2010). Analytics at work: smarter decisions, better results, Harvard Business Press, Boston (Mass.).

Deloitte (2014). Kohti älykkäitä ja yhteentoimivia ratkaisuja. Sosiaali- ja terveydenhuollon asiakas- ja potilastietojärjestelmät Suomessa, https://www2.deloitte.com/content/dam/Deloitte/fi/Documents/life-sciences-health-care/Deloitte%20Sote%20asiakas-%20ja%20potilastietoj%C3%A4rjestelm%C3%A4t%20Suomessa.pdf.

Donoho, D. (2017). 50 Years of Data Science, Journal of Computational and Graphical Statistics, Vol. 26(4), pp. 745-766. http://www.tandfonline.com/doi/abs/10.1080/10618600.2017.1384734.

Dragland, Å Big Data – for better or worseweb page. Available (accessed 25.2.2019): http://www.sintef.no/en/latest-news/big-data-for-better-or-worse/.

Earley, S. (2015). The Promise of Healthcare Analytics, IT Professional, Vol. 17(2), pp. 7-9.

Elgendy, N. & Elragal, A. (2014). Big data analytics: A literature review paper, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 214-227.

Epic About Us web page. Available (accessed 20.2.2019): https://www.linkedin.com/company/epic_163658.

Evans, J.R. & Lindner, C.H. (2012). Business analytics: the next frontier for decision sciences, Decision Line, Vol. 43(2), pp. 4-6.

Gall, M.D., Gall, J.P. & Borg, W.R. (2007). Educational research: an introduction, 8.th ed. Pearson/Allyn & Bacon, Boston.

Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, Vol. 35(2), pp. 137-144. https://www.sciencedirect.com/science/article/pii/S0268401214001066.

Hagerty, J. (2016). 2017 Planning Guide for Data and Analytics, Available: https://www.gartner.com/doc/3471553/-planning-guide-data-analytics.

Han, J., Kamber, M. & Pei, J. (2011). Data mining: concepts and techniques, 3rd; 3 ed. Elsevier, Burlington, MA.

Harvey, H. How to get clinical AI tech approved by regulatorsweb page. Available (accessed 12.3.2019): https://towardsdatascience.com/how-to-get-clinical-ai-tech-approved-by-regulators-fa16dfa1983b.

Hastie, T., Tibshirani, R. & Friedman, J. (2016). The elements of statistical learning: Data mining, inference, and prediction. 2nd ed., corrected at 11th printing, Springer, New York.

Hernandez, I. & Zhang, Y. (2017). Using predictive analytics and big data to optimize pharmaceutical outcomes, American journal of health-system pharmacy: AJHP: official journal of the American Society of Health-System Pharmacists, Vol. 74(18), pp. 1494-1500. https://www.ncbi.nlm.nih.gov/pubmed/28887351.

Hu, H., Wen, Y., Chua, T. & Li, X. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, IEEE Access, Vol. 2 pp. 652-687.

Huang, S., McIntosh, S., Sobolevsky, S. & Hung, P.C.K. (2017). Big Data Analytics and Business Intelligence in Industry, Information Systems Frontiers, Vol. 19(6), pp. 1229-1232.

Huss, M. Challenges of implementing AI in healthcare, web page. Available (accessed 12.3.2019): https://peltarion.com/article/challenges-of-implementing-ai-in-healthcare.

Häyrinen, K., Saranto, K. & Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: A review of the research literature, International journal of medical informatics, Vol. 77(5), pp. 291-304.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R, Springer, New York.

Jankowicz, A.D. (2005). Business research projects, 4th ed. Thomson, London.

Janssen, M., van der Voort, H. & Wahyudi, A. (2017). Factors influencing big data decision-making quality, Journal of Business Research, Vol. 70 pp. 338-345.

Kambatla, K., Kollias, G., Kumar, V. & Grama, A. (2014). Trends in big data analytics, Journal of Parallel and Distributed Computing, Vol. 74(7), pp. 2561-2573. https://www.sciencedirect.com/science/article/pii/S0743731514000057.

Kingston, J.H. (1990). Algorithms and data structures: design, correctness, analysis, Addison-Wesley, Sydney.

Kotu, V. & Deshpande, B. (2014). Predictive analytics and data mining: concepts and practice with rapidminer, Elsevier Inc, Waltham, Massachusetts,

Larose, D.T. & Larose, C.D. (2015). Data mining and predictive analytics, Second; 2 ed. John Wiley & Sons, Hoboken, New Jersey.

Lehto, M. & Neittaanmäki, P. (2017). Suomen terveysdataympäristö , Informaatioteknologian tiedekunnan julkaisuja, Vol. 35https://www.jyu.fi/it/fi/tutkimus/julkaisut/it-julkaisut/suomen-terveysdataymparisto-verk.pdf.

Liu, Z., Pan, Q. & Dezert, J. (2013). A new belief-based K-nearest neighbor classification method, Pattern Recognition, Vol. 46(3), pp. 834-844.

Louridas, P. & Ebert, C. (2016). Machine Learning, IEEE Software, Vol. 33(5), pp. 110-115.

Marr, B. (2017). Data strategy: how to profit from a world of big data, analytics and the internet of things, Kogan Page Limited, London, United Kingdom.

McAfee, A. & Brynjolfsson, E. (2012). Big data: the management revolution, Harvard business review, Vol. 90(10), pp. 128.

Mitzenmacher, M. & Upfal, E. (2005). Probability and computing: randomized algorithms and probabilistic analysis, Cambridge University Press, Cambridge.

Natingga, D. (2018). Data Science Algorithms in a Week - Second Edition, 2nd ed. Packt Publishing.

Nesi, P., Pantaleo, G. & Sanesi, G. (2015). A hadoop based platform for natural language processing of web pages and documents, Journal of Visual Languages and Computing, Vol. 31 pp. 130-138.

O'Gorman, K. & MacIntosh, R. (2015). Research methods for business & management: a guide to writing your dissertation, Second ed. Goodfellow Publishers Ltd, Oxford, England.

Padhy, R.P. (2013). Big Data Processing with Hadoop-MapReduce in Cloud Systems, International Journal of Cloud Computing and Services Science (IJ-CLOSER), Vol. 2(1), pp. 16.

Persson, A. & Kavathatzopoulos, I. (2018). How to Make Decisions with Algorithms: Ethical Decision-making Using Algorithms Within Predictive Analytics, SIGCAS Comput. Soc., Vol. 47(4), pp. 122–133. http://doi.acm.org/10.1145/3243141.3243154.

Outliers (2014). in: A Dictionary of Epidemiology, 6th ed., Oxford University Press.

Raghupathi, W. & Raghupathi, V. (2013). An overview of health analytics, J Health Med Informat, Vol. 4(132), pp. 2.

Rahman, F. & Slepian, M.J. (2016). Application of big-data in healthcare analytics — Prospects and challenges, 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, pp. 13-16.

Sanchez-Pinto, L.N., Luo, Y. & Churpek, M.M. (2018). Big Data and Data Science in Critical Care, Chest, Vol. 154(5), pp. 1239-1248.

Saunders, M., Lewis, P. & Thornhill, A. (2007). Research methods for business students, 5th ed. Prentice Hall, Harlow.

Schuller, B. & Batliner, A. (2013). Computational paralinguistics: emotion, affect and personality in speech and language processing, First ed. John Wiley & Sons, Hoboken, New Jersey,

Shah, N.D., Steyerberg, E.W. & Kent, D.M. (2018). Big Data and Predictive Analytics: Recalibrating Expectations, JAMA, Vol. 320(1), pp. 27-28. https://jamanetwork.com/journals/jama/fullarticle/2683125.

Sloan, R.H. & Warner, R. (2018). When Is an Algorithm Transparent? Predictive Analytics, Privacy, and Public Policy, IEEE Security Privacy, Vol. 16(3), pp. 18-25.

Suomen Standardisoimisliitto SFS ry Standardit, direktiivit ja CE-merkintä web page. Available (accessed 22.3.2019): https://www.sfs.fi/julkaisut_ja_palvelut/standardi_tutuksi/standardit_direktiivit_ja_ce-merkinta#CEmerkint.

Tzortzis, G. & Likas, A. (2014). The MinMax k-Means clustering algorithm, Pattern Recognition, Vol. 47(7), pp. 2505-2516.

Valvira Tietojärjestelmätweb page. Available (accessed 27.3.2019): https://www.valvira.fi/terveydenhuolto/terveysteknologia/tuotteen_markkinoille_saattaminen/tietojarjestelmat.

Vassiliadis, P. (2009). A survey of extract-transform-load technology, International Journal of Data Warehousing and Mining, Vol. 5(3), pp. 1-27.

Wahyudi, A., Kuk, G. & Janssen, M. (2018). A Process Pattern Model for Tackling and Improving Big Data Quality, Information Systems Frontiers, Vol. 20(3), pp. 457-469.

Waller, M.A. & Fawcett, S.E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management, Journal of Business Logistics, Vol. 34(2), pp. 77-84.

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy, Applied and Environmental Microbiology, Vol. 73(16), pp. 5261-5267.

Wang, Y. & Hajli, N. (2017). Exploring the path to big data analytics success in healthcare, Journal of Business Research, Vol. 70 pp. 287-299.

Wang, Y., Kung, L. & Byrd, T.A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations, Technological Forecasting & Social Change, Vol. 126 pp. 3-13.

Wang, Tao, J., Ranjan, R., Marten, H., Streit, A., Chen, J. & Chen, D. (2013). G-Hadoop: MapReduce across distributed data centers for data-intensive computing, Future Generation Computer Systems, Vol. 29(3), pp. 739-750.

Ward, M.J., Marsolo, K.A. & Froehle, C.M. (2014). Applications of business analytics in healthcare, Business Horizons, Vol. 57(5), pp. 582.

Yaqoob, I., Hashem, I.A.T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N.B., Vasilakos, A.V., Luleå tekniska universitet, Institutionen för system- och rymdteknik & Datavetenskap (2016). Big data: From beginning to future, International Journal of Information Management, Vol. 36(6), pp. 1231-1247.

Yin, R.K. (2014). Case study research: design and methods, 5th ed. SAGE, Los Angeles.

Zikopoulos, I., Paul, Eaton, C., Zikopoulos, P., Eaton, C., Zikopoulos, P. & Books24x7, I. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Publishing, Emeryville.