UNIVERSITY OF LJUBLJANA
SCHOOL OF ECONOMICS AND BUSINESS

MASTER'S THESIS

BIG DATA ANALYTICS IN THE HEALTHCARE SECTOR

Ljubljana, 1.12.2021 JANA PAJKOVSKA

AUTHORSHIP STATEMENT

The undersigned Jana Pajkovska, a student at the University of Ljubljana, School of Economics and Business, (hereafter: SEB LU), author of this written final work of studies with the title Big Data Analytics in the healthcare sector, prepared under the supervision of Jurij Jaklič.

D E C L A R E

1. this written final work of studies to be based on the results of my own research;

2. the printed form of this written final work of studies to be identical to its electronic form;

3. the text of this written final work of studies to be language-edited and technically in adherence with the SEB LU's Technical Guidelines for Written Works, which means that I cited and/or quoted works and opinions of other authors in this written final work of studies in accordance with the SEB LU's Technical Guidelines for Written Works;

4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;

5. to be aware of the consequences a proven plagiarism charge based on this written final work could have for my status at the SEB LU in accordance with the relevant SEB LU Rules;

6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;

7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained the permission of the Ethics Committee;

8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the SEB LU Study Information System;

9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically, and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;

10. my consent to the publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published.

Ljubljana, _____ Author's signature: _____

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AI- Artificial Intelligence

CDS- Clinical Decision Support

CGM- Continuous Glucose Monitoring

CT- Computerized Tomography

DM- Data Mining

DNA- Deoxyribonucleic Acid

ECG - Electrocardiogram

EHR- Electronic Health Records

EMR- Electronic Medical Record

FHIR- Fast Healthcare Interoperability Resources

GP- General Practitioner

ICU- Intensive Care Units

IMIA- International Medical Informatics Association

IoT- Internet of Things

IT- Information Technology

ML- Machine Learning

MRI- Magnetic Resonance Imaging

NHS- National Health Service

INTRODUCTION

By increasing the efficiency of many processes thus leading to increased customer satisfaction, technology has quickly become a crucial part of our everyday lives and is playing an important role in many business operations. Many industries benefit from the digitalization of their processes, and with healthcare being information-based science, it is not an exception. In practice, a majority of the clinical profession is gathering, synthesizing, and acting on data. The management and the usage of information in healthcare and biomedicine is a part of a field known as Medical Informatics (Hersh, 2002). Both private medical healthcare providers and public health institutions are implementing information systems and creating and sharing digital information. The purpose of this digital initiative is to improve the process of identifying, monitoring, treating, and if possible, preventing negative health outcomes (Gamache, Kharrazi, & Weiner, 2018). Various technologies, such as Electronic Health Records (EHR), communication technology, Data Mining (DM), Machine Learning (ML), and Artificial Intelligence (AI) are already utilized in the sector. There are many benefits derived from using this technology in medicine, such as collecting data from wearable devices and helping people to improve their health. However, there are also barriers, such as the quality of large amounts of data, especially when collected from multiple sources as it is uneven and usually not structured (Xu, Glicksberg, Su, Walker, Bian, & Wang, 2020).

The large amount of structured, semi-structured, or unstructured data that has been generated by different sources is referred to as Big Data. Many global enterprises such as Google, Facebook, or IBM collect this Big Data from their users, and extract valuable information from it (Kumar & Singh, 2019). Big Data technologies have shown great promise to society, such as discovering subtle patterns of human behavior and developing predictive models for future behavior. With this data, enterprises can predict the needs and preferences of their customers and offer personalized services, which leads to greater customer satisfaction. Nevertheless, some challenges have to be considered with these technologies. A major challenge is the ability to manage large amounts of data. Since the data is collected from different sources, tail behavior may appear during the analysis, however, with no real correlations. Furthermore, this type of data contains many errors and/or missing values (Fan, Han, & Liu, 2014). In many cases, the IT infrastructure needs to be upgraded and new management practices need to be implemented for an organization to completely embrace Big Data. Another challenge is that the employees that have the required skills to deal with Big Data are scarce. In general, many organizations lack the understanding of how Big Data can help their organization and business processes, which leads to resistance in acceptance within the organization (Alharthi, Krotov, & Bowman, 2017). Overall, there are many promising benefits and opportunities from using Big Data in the healthcare sector, but, there are also many barriers that need to be taken into a consideration.

The purpose of this Master's thesis is to contribute to the understanding of the impact of technology, more specifically, of Big Data in the healthcare sector. The research question of the Master's thesis is: What are the opportunities and the barriers of Big Data analytics in the healthcare sector? The first goal of this thesis includes collecting and analyzing data from various primary and secondary sources. Through the use of interviews, the primary information is based on first-hand experts and professionals in the field of healthcare. The goal extends to using secondary sources to compare previous findings and to complete a sentiment analysis. For the sentiment analysis, the data was collected by Text Mining from Twitter to gather people's opinions regarding Big Data usage in the healthcare sector. The second goal is to find out how people perceive Big Data usage in the healthcare sector. Specifically, to see whether they trust technologies such as AI or Machine Learning to make diagnoses and decisions instead of doctors. Finally, the third goal is to summarize the main opportunities and barriers of Big Data analytics in the healthcare sector, including actions on how the barriers might be overcome. Following this Introduction, which is the first chapter of the thesis, comes the Literature Review, which is the theoretical part of the thesis. This is the accumulation of secondary research and data to explain the topic and the main concepts. In the Adoption of Big Data subsection, the IT adoption factors and the opportunities and barriers of IT are explained, and the main concepts that are used throughout the thesis are defined. After providing short information on the evolution of Medical Informatics and Big Data in healthcare, this thesis explains how Big Data is collected, stored, analyzed, and used in healthcare and what are the benefits, barriers, and requirements from different usage of Big Data. Moreover, some practical examples of Big Data usage in healthcare, such as Data Mining, Machine Learning, and Artificial Intelligence examples are included. The next chapter, Methodology explains the primary and secondary research methods that are used for writing the thesis; in the Results chapter are explained the results that were obtained from conducting interviews with professionals or industry experts, working with Big Data in the healthcare sector, and their experiences on this matter. Additionally, to analyze the opinion of the general population on the topic of Big Data analytics in the healthcare sector, Rapid Miner was used for mining tweets on the topic and sentiment analysis on the results. In the Analysis and Discussion part, the contribution of this written work is described and the potential actions on how the barriers of Big Data analysis in the healthcare sector might be overcome are outlined. Furthermore, the thesis describes how the findings of this thesis relate to previous results found in the literature. In the Conclusion, the work that has been done throughout the thesis is summarized through a general overview. The Appendices part contains the questions that were used for the interviews.

1 LITERATURE REVIEW

1.1 Adoption of Big Data analytics in healthcare

The contribution of new information technologies does not begin with their invention, but rather with their acceptance and wide usage. This acceptance of new technologies is called diffusion; it is a series of individual decisions by the people to adopt these technologies. Those decisions are usually a product of comparing costs of adaptation versus the benefits that the new technologies claim to offer. The series of individual calculations of the potential benefits versus the potential costs are often characterized by uncertainty and limited information since the suppliers can often influence their decisions. The final diffusion rate is calculated by summing up the individual decisions regarding the innovation adaptation. Unlike the invention of IT, which is seemingly fast, their diffusion is a slow and continuous process. The adoption process is generally really slow at first, accelerating as it spreads to other potential adopters and then slowing down again as the population becomes more saturated. Until many users decide to use the new technologies, they can contribute very little, if at all, to people's well-being (Hall & Khan, 2003). The individual attitude towards accepting a new invention is based on personal silent beliefs regarding the consequences of the adoption and usage and the evaluation of those consequences. Certain innovation characteristics may affect the individual opinion of the technology and affect the rate at which innovation diffusion happens. Some of those characteristics are perceived usefulness, which is the degree to which innovation usage is perceived as better compared to the practice it replaces; image- the degree to which the adoption of the innovation enhances one's image or status in the social system; compatibility, that is the degree to which the innovation is compatible with what people do; complexity- the degree to which using the innovation is easy or complicated for usage; trialability, which is the degree to which one can experiment with the innovation on a limited basis, before making an adoption or rejection decision; visibility- the degree to which the innovation is visible in the organization and the result demonstrability, which is the degree to which the results of adoption the innovation are observable and communicable to others (Karahanna, Straub, & Chervany, 1999).

The attitude toward diffusion of rejection of innovation can be pre-adoption and post-adoption attitudes and can be formed based on information concerning past behavior, active information, and cognitive information. The pre-adoption attitude is usually based on indirect experience, while the post-adoption attitude is based on past and direct experience with the innovation. With the direct experience, the user gets more information regarding the innovation and can evaluate it more clearly and confidently. With the direct experience, the user gets familiar with the innovation, creates a stronger attitude-behavior, and can use that experience as a better basis for their diffusion/rejection decisions (Karahanna, Straub, & Chervany, 1999).

While the obvious information technology (IT) adoption factors are the benefits received by their users and the costs, other non-economic adoption factors can be as important in the determination of the new technology diffusion. Two of the most important adoption factors, besides the benefits and costs of new technology implementation, are the skill level of workers and the state of the capital goods sector. Both workers and capital goods are crucial for the successful implementation and operation of new technologies. If the technologies require complex skills which are costly and require time for the workers to acquire them, the adoption might be slow. That is why the overall available skills in the enterprise are important factors of diffusion. The technological capacity of an industry for adaptation is important as well since the implementation and operation of new technologies require appropriate technical capacities. If the new technologies are too advanced compared to the engineering capacity, their diffusion might take longer and cost more. Another important factor is customer commitment and relationships. To recover from expensive investments in new products and technologies, the companies need to be assured that there will be a future income that will help cover those expenses. The stability of the firms' relationship with its customers plays a great role in the diffusion process for new technologies. The bigger the guarantee for the presence of future demand, the more likely it is that the company will adopt the innovations (Hall & Khan, 2003).

Network effect due to technology is another important adoption factor because it has a high interrelation among technologies. The network effect in adoption is the value a user derives from a good or a service depending on the number of users. The network effect is usually positive, meaning that the more people join the network; the more value from a product can be derived. The new technologies and innovations have a low rate of diffusion because they have relatively poor performance and initial barriers. Another factor of innovation diffusion is that the new technologies can be significantly improved over time and their costs lowered, which can lead to their eventual acceptance or diffusion. While the innovations are imperfect in their early stage, the new technology improvements are an important factor for their adaptation. The reason behind this is that the technology efficacy rate is far bigger in their enhancement stage compared to the initial stage. Another factor is that when there is a new technology that is about to replace an already existing one, it is likely to provoke the providers to make changes and significantly improve the old technology. This will result in slowing down the diffusion of innovation. Besides that, the market structure and size need to be taken into consideration, as well as government regulations regarding the new technologies. All these factors play a great role in the new technology adaptation (Hall & Khan, 2003).

IT knowledge is a mandatory requirement in today's modern economic environment, bringing many benefits and opportunities to the companies implementing them. The IT adaptation has revolutionized the way business is conducted. The intense implementation of computers and data processing systems in manufacturing, service industry, and communications spread to governmental institutions, educational organizations, and finally

in the households. These adoptions are a significant driving force behind many socioeconomic changes in today's environment while generating new business opportunities and various benefits. Companies benefit from technology implementation by reinforcing their competitive position and improving productivity, which results in a bigger profit. Information technologies assist the companies through supplying the needed infrastructure and providing the right information at the right time. They allow integration between inter-organizational as well as intra-organizational functions and provide critical information. Information technologies are capabilities offered to businesses by computers, software applications, and telecommunications to deliver data, information, and knowledge to the company and its processes. These technologies are included in the operation, collection, transport, storage, access, presentation, usage, and transformation of information in any form. Their implementation is often followed by increased efficiency, lowered costs, and higher profit (Ghobakhlo, Sabouri, Hong, & Zulkifli, 2011).

Unfortunately, the adoption and implementation of information technologies come with challenges and barriers. The implementation of new technologies is especially challenging in emerging economies. The poor policies and insufficient infrastructure investment in the IT sector make it difficult for their adoption. While governments acknowledge this, little or no action has been taken in most of the cases for improving the situation and IT adoption. Since many developing countries do not have the resources to develop technologies, they depend on foreign aid to ensure IT development; however, developed countries have also not done much to help them. Many developing countries import new technologies without needing modifications to fit better their environment and culture. Poor IT infrastructure and the lack thereof is the major cause of IT stagnation in developing countries. Basic national infrastructure, access to all of the people with connection to the world is of utmost importance in any country and should be implemented everywhere. Moreover, developing countries lack skilled IT workers who can design, program, install, configure and maintain IT. While it is easier to relocate equipment, it is more difficult to relocate capacity which is human-embodied. Developing countries do not have enough graduates in the IT field and while there is growing awareness of the need and importance of this field, many universities and institutions have limited access to the internet and modern computing (Ejiaku, 2014). The developed countries, although not as severe, also face some barriers regarding IT implementation. The lack of IT skilled personnel is present in developed countries as well (Yu, Wu, Yu, & Xiao, 2006).

The process of IT adoption is critical for delivering the benefits of IT. The information technologies can be implemented in many industries and used for improving efficiency, lowering costs, increasing revenue, and accessing information. The processes and application of methods to improve the management and usage of patient data, medical knowledge, population data, and other information relevant to patient care are called Medical Informatics. While it is a relatively young science, which emerged after the invention of digital computers in the 1940s, it is also a multidisciplinary science,

interacting with various fields. This discipline interrelates to clinical sciences, public healthcare, computing and information sciences, statistics, and other similar fields, which means that the background of the Medical Informatics workers is quite diverse. Various branches have appeared emerging from Medical Informatics, such as public health informatics, consumer health informatics, clinical informatics, or bioinformatics (Wyatt & Liu, 2002). There are different definitions regarding this discipline. Some of them define Medical Informatics as a field dedicated to the systematic processing of data, information, and knowledge in medicine and healthcare (Haux, 2010). Others explain this discipline as a bridge to the gap between clinical care and medical research; because of the need for integrating medical research projects with data repositories built up during documentation of clinical care routines. While this field has many definitions and explanations, it plays an unquestionably important part in the healthcare sector as it is today (Prokosch & Ganslandt, 2009). This discipline is dedicated to the systematic processing of data and there are different technologies used in healthcare to collect and store large amounts of data. These massive datasets are called Big Data. As noted previously, they have a varied and complex structure and are therefore are difficult to store, analyze and visualize for further processes and results. The process of researching these massive amounts of data and finding the hidden patterns or possible correlations is called Big Data Analytics. These analyses help companies with gaining better insights and a competitive advantage, which is why these data analyses and implementations need to be executed as accurately as possible. While in the business world, an inaccurate analysis may cost money and resources, in the healthcare sector, this can mean loss of life (Sagiroglu & Sinanc, 2013).

The science behind collecting, preparing, analyzing, using, and understanding data is called Data Science. Data science uses Big Data in many different studies, methods, and technologies. For this, among others, data science uses different technologies such as Data Mining, Machine Learning, and Artificial Intelligence. Even though they are very similar to each other, sometimes used in literature with the same meaning, and they all derive from Big Data analysis, these techniques have differences (Konduforov, 2021). Data Mining is comprised of tools and techniques used for extracting new and useful information from large datasets and transforming it into a more understandable structure. It is also a tool used in data science for detecting correlations and patterns in the data sets, which were previously unknown. Data mining can be divided into two classes of tools: model building and pattern discovery. Model building is a descriptive summary of the data set, which in statistics includes regression models and cluster decompositions. These models describe the overall shape of the data, for example, pattern detection is seeking anomalies in the data sets. This means that Data Mining also comprises filtering and data reduction activities (Hand & Adams, 2015). Based on the patterns that can be found in a data set, the tasks in Data Mining can be divided into summarization, classification, clustering, association, and trend analysis (Fu, 1997). Besides finding patterns and structures, Data Mining is used for preparing the data for analysis and to make it more usable (Konduforov, 2021). Unlike Data Mining, Machine Learning is a set of tools, methods, and algorithms

used to train machines to analyze, understand and find hidden patterns in data sets, which can be used for making predictions. This technique studies the question of how to build computers that improve themselves automatically through experience. These tools aim at training machines on historical data, so they can learn and act upon new inputs based on the learned patterns without an explicit program that tells them to. This means that these machines learned how to perform a specific action based on the conducted training and not on manually written instructions (Konduforov, 2021). With Machine Learning, a certain computer program is told to learn from previous experience or historical data regarding a specific task, while measuring its performance at that task and with it, improving itself from experience (Zhang, 2020). Artificial Intelligence, on the other hand, is a broad term that cannot be defined accurately and unquestionably. AI refers to a science that deals with the creation of intelligent machines. These machines are called intelligent because they can think on their own and make decisions based on their calculations, like humans (Data Mining Vs Machine Learning Vs Artificial Intelligence Vs Deep Learning, 2021). Simply put, AI is an algorithm, code, or technique that enables machines to copy, develop and perform human cognitive behavior. AI tools are real-life data products, which are capable of performing tasks and solving problems the same way that humans do. AI systems can learn, plan, reason, make decisions, and solve problems by themselves. Most AI tools involve Machine Learning and Data Mining as well since these so-called "intelligent" machines require knowledge, which requires Data Mining research (Konduforov, 2021).

1.2 Medical Informatics and Big Data

Medical Informatics is still a relatively young science when compared to other medical disciplines. However, when talking about a systematic approach to information and data in healthcare, there is data going back for about 60 years. Within a few decades, the usage of IT in the healthcare sector tremendously increased, which lead to medicine and healthcare changes and improvements as well. Nowadays, it is difficult to imagine the healthcare system without these technologies, which are used for creating diagnostic tools and procedures. When referring to the beginning of Medical Informatics we go back to 1959, when in their paper, Ledley and Lusted were reviewing methods for handling information and making decisions under risk and uncertainty. In the paper, the authors were discussing comprehensively for the first time the need for computer-based decision-making in medicine. Fast forward 25 years, in the period 1984-1985, several papers regarding IT usage in medicine were published, and they strongly influenced the development of Medical Informatics and set the basis for understanding the need for methodological and technological knowledge in medicine. By the end of the 20th and beginning of the 21st century, the International Medical Informatics Association (IMIA) published the first international recommendations on education in Medical Informatics which were translated into many languages. IMIA supports and encourages the research, education, and practice of Medical Informatics, by bringing together scientists, researchers, and specialists in information sciences from all around the world, in order to improve the development,

implementation, and practice of healthcare informatics. This has led Medical Informatics to grow and take place in many countries. Today, Medical Informatics is one of the bases for medicine and healthcare (Haux, 2010).

Even more than just using IT for diagnosis and tracking diseases, the patients' health records also took a digital form as Electronic Health Records, making them and the patients' medical data easily accessible for the healthcare providers (Hoerbst & Ammenwerth, 2010). The implementation of the EHRs in healthcare resulted in generating large amounts of data. These records contain quantitative data such as laboratory results, and qualitative data: text-based documents written by healthcare specialists and transactional data such as records of medical delivery. The fast and simple accessibility to this useful medical data helps the healthcare providers to put forth an early and more accurate diagnosis of possible diseases (Murdoch & Detsky, 2014). However, as the nature of medical data evolves, so do the analytics techniques. The days of collecting data only from EHR are long behind us. The sources of medical data can be internal or external. The internal sources of health data are EHR, clinical decision support systems, clinical reports, receipts, and patients' medical history. The external sources can be located anywhere in the world and some examples of such sources are laboratories, pharmacies, governmental institutions, and insurance companies. Health data can be extracted from many different sources such as, but not limited to: web and social media- blogs, diet plans, or Smartphone apps; machine - reading remote sensors; big transactions data- healthcare claims; biometric claims- fingerprints, genetics, x-rays, and scans; human-generated data - emails, physicians' notes or paper documents (Raghupathi & Raghupathi, 2014). The number of sensors and monitors used in healthcare increases daily, generating more and more data which is later analyzed. This will eventually lead to improving the healthcare services and offer more accurate analysis and predictions to the patients. Many companies nowadays decide to store Big Data in the cloud, a virtual space where the data can be stored and easily accessed from different geographic locations. This alternative is cheaper compared to storing huge amounts of data on physical servers. Moreover, storing Big Data in the cloud allows Big Data systems to perform real-life analytics (Wang, Kung, & Byrd, 2016). As mentioned above, there are many sources of medical data: administrative claim records, medical records, EHR, data reported from the patients, biometric data, medical imaging, clinical trials, the internet, and biomarker data. Medical Big Data, if compared to collecting Big Data in other disciplines, is usually harder to access because of governmental and legal restrictions due to the risk of data misusage by third parties. Moreover, the medical data can be expensive because of the needed personnel and medical instruments used for collecting it. A big portion of the medical data also might have measurement errors, missing data, or coding errors from textual reports. Another feature of medical data is the important patients' characteristics, which have to be taken into consideration, as well as treatment information and the time of the treatment decision and change (Lee & Yoon, 2017).

The main sources of Big Data in the healthcare sector are the medical healthcare units. The data generated in these units is composed of datasets that are very big, fast, and complex, which makes it impossible for healthcare providers to use it with their existing tools. These datasets contain both structured and unstructured data. The data scientists who are analyzing Big Data are always trying to find hidden patterns in the health data, to provide improved care and treatments for the patients. These patterns lead to creating better treatment programs and lead to making more accurate decisions regarding the patients' health, which translates into saving many lives and decreasing the costs of the treatments. The most important and sensitive part of these datasets is the pictures, which are used by many physicians for diagnosing and checking the status of the patients' disease (Dey, Das, Naik, & Behera, 2019). The main sources for such data are computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and x-rays. With the help of different algorithms, the medical image data is processed and analyzed to extract as much as information possible for setting diagnosis. These algorithms have features to improve the contrast of the medical pictures, for noise reduction, sharpening of the edges of the pictures, and segmentation. These features make the manual diagnosis process automatic or semiautomatic. The growth of the patients' number makes the process of manual diagnosis too complex and unfeasible. These features are considered to be efficient, able to produce optimal analysis results, and help the healthcare providers to set accurate diagnoses. The analysis of the pictures, when integrated with other kinds of patients' data such as EHR and genomic (DNA) data improve the accuracy and decrease the time needed for diagnosis. The analysis of this data requires high processing, fast and optimal algorithms. Medical image processing plays a big part when determining the efficiency of Big Data analysis in healthcare (Bora, 2019).

Big Data technologies receive more attention because they can successfully handle high volumes of data compared to the traditional models. Unstructured, semi-structured, and structured data is supported by the Big Data framework while providing features such as predictive model design and Big Data mining tools. These features enable a better decision-making process through data processing and selecting the relevant information. There are two types of data processing: batch and stream processing. Batch processing processes high volumes of data by storing batches to analyze them and generate results. The stream processing performs simple and fast computations and reduces the computational time needed to analyze the data. In healthcare, because of the large amount of data produced continuously, stream processing is used, which is suitable for applications that require real-time feedback. Real-time applications tend to produce data containing missing values, which makes the analysis complicated. The Big Data approaches are essential for healthcare analysis. They allow real-time extraction of information from big amounts of data from the patients. The stream processing allows classification, frequent pattern mining, and extraction of relevant information from large data. When the prediction model detects possible complications, the alerts notify the healthcare providers. This prevents emergencies from occurring and aids medical care providers in the decision-

making process regarding disease diagnosis and recommendations. It is crucial for healthcare organizations and providers to detect as much as possible and as early as possible of any warning signs of a disease that a patient might show since it is easier and cheaper for the patient and the healthcare institution to treat the diseases at an early stage compared to if it was detected later (Aboudi & Benhlima, 2018).

This huge amount of different data generated fast poses a challenge on how to be handled. To make this large data available to the researchers and data scientists, the data needs to be saved in an easily accessible file format from where it can be read and effectively used for analysis. Another requirement for successful Big Data analytics in healthcare is the implementation of the needed technological tools and protocols in the clinical settings. The complexity of this setting requires experts from different fields such as biology, IT, mathematics, and statistics to work together. Because of its heterogeneity and big size, the collected data is available in cloud storage with pre-installed tools for Data Mining, Machine Learning, and AI functions developed by experts to convert the stored data into important information. The Machine Learning and AI algorithms are usually written in programming languages such as Python or R. To handle this Big Data in medicine a good knowledge of biology and computer sciences is required. Bioinformaticians are a good fit for this description (Dash, Shakyawar, Sharma, & Kaushik, 2019). An important challenge that needs to be addressed is the confidentiality of the data. As all of the medical data is highly sensitive, and in many countries, the data is owned by the patient and not the institution, the healthcare providers are required to respect patient confidentiality. To be used for setting a diagnosis, the data must be linked to the patients' identities. This means that the data cannot be fully anonymous and requires complex pseudonymization procedures to protect the patients' identities while analyzing and using it. Besides the above-mentioned, many other barriers need to be taken into consideration when using Big Data Analytics in the healthcare sector (Viceconti, Hunter, & Hose, 2015).

1.3 Big Data Analytics in healthcare

The healthcare providers and data experts need to continue improving and searching for new ways of Big Data usage in healthcare for better and faster-diagnosing diseases. However, while the usage of Big Data in the healthcare sector has been proven over and over again as an asset for early detection of diseases among patients, its implementation does come with its barriers and requirements. For successful implementation of Big Data in healthcare three teams need to be included in the process: the developing team, with IT skills that will develop the software, a medical team of doctors or experts in the medical field that will set the needs and the requirements for the Big Data solution that is being built and a Medical Informatics team, that works as a bridge between the IT and the medical team and understands both the IT possibilities and capacities and the medical terms and correlations. Besides only implementing, the new technologies need to be adapted to the already existing processes and company culture. Many standards need to be

taken into consideration and fulfilled before the new technologies can start operating in the companies as well (Viceconti, Hunter, & Hose, 2015).

1.3.1 Opportunities of Big Data

When talking about healthcare informatics and analytics, we are referring to the application of advanced technologies and data analytics in healthcare. Big Data analytics becomes more and more popular for practitioners as well as researchers in the healthcare sector. In the past 20 years, the usage of advanced IT, Artificial Intelligence, and data analytics, has pushed healthcare towards a more effective system (Pramanik, Lau, Azad, Hossain, Hossain, & Karmaker, 2020). The aging populations in many countries are pressuring the healthcare systems all around the world. With a systematic usage of health data, analytical, statistical, predictive, or cognitive models, the healthcare system will be able to predict epidemics, help cure diseases, improve the life quality and reduce the number of deaths that could have been prevented (Kankanhalli, Hahn, Tan, & Gao, 2016). Some of the most important opportunities for Big Data usage in the healthcare sector are:

Personalized medicine

Many of the diseases, if discovered early are preventable or at least treatable. Even more, if other characteristics of the patient are taken into consideration when diagnosing the disease and deciding on a treatment. However, the possible combination of risk factors is too complex, and it is difficult for a physician to fully analyze it during the patient interaction, and the growing number of patients makes this task impossible. Usually, medical providers take into account the patients' medical history and do selective laboratory tests for determining the health of the patient and the risk for future health problems. However, with the help of computing and analytics frameworks for processing and analyzing Big Data, healthcare providers can get a deep knowledge about the patients' health through similarities and connections. This allows them to provide personalized disease risk profiles for each patient, delivered not only from the patients' EHR but also from similarities, connections, and patterns seen at other patients. This provides an opportunity for proactive medicine and actively managing diseases. The focus of predictive medicine is on the genomic revolution. The idea is that, once all of the disease-related DNA mutations are known and cataloged, the healthcare providers will be able to detect the patterns and connections early and prevent or start treatment for the patients. Moreover, even if the patient carries the genetic risk of developing a certain disease, the likelihood is not always the same. The applications of the genome-based frameworks can understand the risks while taking into account other individual characteristics of the patient (Chawla & Davis, 2013). The similarities of a certain lifestyle, exposure to a similar risk or environment might have similar outcomes for the patients. These factors, combined with patients' medical history from EHRs and DNA allow creating a personalized disease risk profile for each patient. The possibility for a patient to walk out of their physician's office with a

personalized health assessment, tailored diagnostics, and recommendation for health and quality of life improvement, will reduce the risk of future diseases and/or severe disease symptoms greatly (Kim, 2018).

Clinical decision support systems

As previously mentioned, the medical data used to be collected by the healthcare professionals only, but that is not the case anymore. People started collecting their medical data while using smart watches or apps, which can be an important source of health data. The translation of the health data is a routine job for medical professionals, who have been trained to distinguish between relevant and irrelevant information. However, the huge amount of data generated constantly, and its potential value, surpasses the limits of human comprehension. Due to the overwhelming amount of data available, data analytics have developed systems that can obtain data-provided insights in real-time. These systems are called clinical decision support (CDS) and are defined as information systems designed to help with the decision-making process, by integrating different sources such as EHRs, laboratory test results, etc. The CDS systems can differentiate regarding their form and function, but they all generate clinically relevant outcomes based on input data. The output of a CDS system can be a generated prediction as an input for a clinical decision, for example, early warning messages or they can act without any human interference, an implantable cardioverter-defibrillator. The development and implementation of CDS systems is a very complex process and it is done by an interprofessional CDS team. The healthcare professionals participate only by setting the requirements of what they want the CDS system to do. The CDS systems are not designed to replace medical professionals, but rather to aid them in making decisions (Bezemer, et al., 2019). There are some examples of CDS systems where visual analytic methods combine evidence-based and data-driven approaches to improve clinical performance. The implementation of CDS systems provides an opportunity to support clinical diagnostics, therapy, and research. The systems may highlight clinical patterns, which were previously not considered. They allow effectively analyzing the routinely collected data and getting new insights about patients' health (Dagliati et al, 2018).

Modeling disease progression

Chronic diseases such as heart disease, diabetes, cancer, or obesity are the main reason for mortality in many countries in the world. These diseases are distinctive because they slowly progress over a long period and are viewed as a continuum. If the risk factors of the above-mentioned chronic progressive diseases can be controlled at an early stage, mortality or severe symptoms can be delayed or even prevented. An ECG (electrocardiogram) monitoring is a test used to check a patient's heart rhythm and electrical activity. When the ECG irregularities are detected in patients with heart problems, unfortunately, it is already too late and the patient may be already experiencing heart failure. If the risks are analyzed and tracked in time, heart failure can be successfully prevented by introducing lifestyle

changes. Modeling disease progression, including other factors such as the patients' lifestyle and the data from EHR will assess the patients' health and lead to early screening and tests. The early detection of chronic diseases will significantly reduce the time and costs needed for treatment and recovery, both for the patient and the medical institution. (Zhoua, et al., 2020)

Target identification

The pharmaceutical companies spend a huge amount of resources to develop new drugs, which is a very complex process, and yet, up to 90% of the tested new chemicals do not make it to the market. The whole process of developing new drugs starts with the identification of disease-relevant phenotypes (individual observable traits). This includes research, target identification, lead generation, preclinical testing, clinical trials in humans, and regulatory approval. Target identification or identifying drug targets for a disease is an essential part of the drug discovery and development process. The traditional pharmaceutical screening process for the identification of new drugs is expensive and takes time. Moreover, this traditional screening method is known to have thousands of failures per one successful drug candidate, which is why animals are used for testing the development of new drugs. However, they have a big disadvantage, since the diseases do not usually develop the same as in humans. This creates the need for new approaches regarding the screening process for drug development. The analysis of big health data may identify new, unknown connections and allow drug testing on human models with high certainty. Another benefit is that the analysis of biological and medical data will calculate the chances of success in drug discovery and development before even the stages of expensive preclinical and clinical trials are started. Human trials can be used to learn the differences in drug response and side effects. It is crucial to use human subjects to distinguish if the drug in question only affects a certain part of the population and why. Testing the new drugs on human models may double the success rate of the clinical development of drugs (Shilo, Rossman, & Segal, 2020).

There are also other benefits from Big Data usage in the healthcare sector such as Big Data-driven *population revealing patterns*, which is population medical analysis of patterns that might have been missed if the analysis were done on a smaller sample of the population; *disease phenotyping*, studying the variations of diseases and it progresses in different people, for better understanding the disease, the risks and possible to avoid or treat it; health process improvement, Big Data analysis plays a huge role in *optimizing the health processes* by reducing diagnostics and treatments errors, eliminate unnecessary tests, provide guidance for better resource distribution and reduce treatment costs; and *fraud detection and prevention*, lowering the risk of frauds in healthcare by detecting suspicious records of health data, or detecting healthcare insurance fraud (Shilo, Rossman, & Segal, 2020).

1.3.1.1 Data Mining in healthcare

The Data Mining concept first appeared in the 1990s for data analysis and knowledge discovery. Data Mining is defined as an analysis of big observational datasets, for finding unexpected relationships or patterns and summarizing the data, and making it more understandable and useful for its owner. The application of Data Mining in healthcare provides knowledge that can be used for clinical and administrative decision-making support (Yoo, et al., 2011). Once the data is collected from the patients or the medical institutions, it is selected according to previously set criteria. After the selection, the data goes to preprocessing stage, where all the unnecessary information is removed. After the preprocessing, the useful data is transformed and prepared for Data Mining. In this stage, the meaningful patterns of data are extracted from the rest of the data and are interpreted and evaluated, so they can later be used in a decision-making process (Ahmad, Qamar, & Rizvi, 2015). There are two types of Data Mining models: predictive and descriptive. The predictive models analyze datasets and find relationships and patterns and use them for predicting future unknown patterns and values, while the descriptive models apply learning functions for finding patterns that can describe the data and make it understandable for humans. In healthcare, more common are the predictive models, which are being used for predicting different diseases and helping the healthcare workers with the decision-making process (Jothia, Rashidb, & Husainc, 2015).

Treatment effectiveness is one of the benefits that Data Mining techniques can be used for. For example, they can be implemented for evaluating the effectiveness of medical treatments. Using datasets for comparing the cause and the symptoms of the treatments, Data Mining can analyze which treatment path leads to effectiveness. With a comparison of different outcomes from a group of people that suffer from the same disease, these techniques determine which treatments are most efficient and cost-effective. An example of this is united HealthCare, which has mined the treatment record data to find ways to cut costs and offer better healthcare quality. United HealthCare is a multinational healthcare and insurance company that has also developed clinical profiles of the patients and offers information about the healthcare workers' practice patterns while comparing them with other physicians' industry standards. *Healthcare management* is another usage of Data Mining. To help healthcare management, Data Mining techniques allow the development of applications that can improve the identifying and tracking of chronic diseases and high-risk patients. With their help, the appropriate measures can be taken and with it, the number of hospital admissions can be significantly reduced. American Data Network, a company for developing healthcare data applications, developed better diagnosis and treatment protocols by looking at readmission and resource usage while comparing it with current scientific literature. Moreover, the Group Health Cooperative, a healthcare non-profit organization, classifies the patients by demographic characteristics and medical conditions to determine which group uses resources the most and then develops programs for educating those patients and preventing their conditions. Blue Cross association of

health insurance companies uses Data Mining techniques for improvement of the outcome and cost reduction through better disease management. For example, it uses the data from the emergency department and hospitalization, the pharmacy records, and healthcare workers' interviews to identify new, unknown asthmatics and implement appropriate measures. Data Mining can be used for *customer relationship management.* As in any other industry, customer relationship management is an important part of healthcare as well. Data mining applications can be developed in healthcare for defining the preferences, patterns, current and future needs of the patients to improve the level of their satisfaction with the institution. Moreover, the answers to the surveys that the patient took can help with setting realistic expectations of waiting times, and finding new ways to improve the service. The pharmaceutical companies benefit from Data Mining as well. With tracking which healthcare providers prescribe which drugs, they can decide whom to target, and which physicians are fitted for which trials. *Fraud and abuse prevention* in healthcare can be much improved by Data Mining implementation. By identifying abnormal and unusual patterns of claims by healthcare providers, Data Mining applications can detect attempt fraud and abuse. For example, thanks to fraud and abuse detection, ReliaStar Financial Corp., a holding company that specializes in life insurance and related financial services businesses, has reported 20% higher annual savings, and the Australian Health Insurance Commission estimated tens of millions of dollars in annual savings (Koh & Tan, 2011).

1.3.1.2 Machine Learning in healthcare

While Data Mining is designed for extracting information from big quantities of data and has explanatory nature, Machine Learning teaches computers how to learn and understand given parameters (Hüllermeier, 2005). There are different types of Machine Learning, but most applications can be divided into supervised, unsupervised, semi-supervised, or reinforcement learning. With supervised learning, the data are labeled according to a specific outcome of interest (the patients are infected or not infected) (Wiens & Shenoy, 2018). This method of Machine Learning builds a map of the relationship between the input and the output using training data. Some examples of supervised Machine Learning are the classification of different types of lung diseases and recognition of different organs from medical images. Unsupervised Machine Learning is a method when the Machine Learning tools are using unlabeled data and similarity metrics for anomaly detection. Examples of unsupervised Machine Learning usage are a prediction of heart and hepatitis diseases. The semi-supervised learning method is when a small amount of labeled and big amounts of unlabeled data are available for training. This method is very useful, especially because acquiring a sufficient amount of labeled data in healthcare is quite difficult. Reinforcement learning is a method that learns a function based on a given set of observations, actions, and rewards as a response to an action taken. This method has great potential for transforming the healthcare sector as we know it, and it is used for checking symptoms for disease diagnosis (Qayyum, Qadir, Bilal, & Al-Fuqaha, 2020).

An example of Machine Learning usage in diagnosing Alzheimer's disease, which has become a growing problem and a common illness, especially with the older population. Alzheimer's disease is a terminal disease and a leading cause of death in the United States with no known cure to date; however, some treatments can slow down its progression. Age and family history are the main risk factors for this disease. *Early diagnosis* and treatment for this disease can significantly improve the health of the patient. With the usage of Machine Learning technologies, a model for early diagnosing Alzheimer's disease can be used, which will have a huge benefit for the healthcare institution and its patients as well. The model computes the probability of the presence of the disease in the patient, based on historical data, and gives results with the current medical state of the patient. (Maity & Das, 2017).

Besides that, Machine Learning techniques are usually used for diagnosing diseases such as heart disease, diabetes, thyroid disorder, and many more. The usage of Machine Learning can *reduce the costs for diagnosing diseases* and do so with relatively high accuracy (Shailaja, Seetharamulu, & Jabbar, 2018). There is a large amount of structured and unstructured medical data available. The structured data contains information that is easy to categorize in the database, such as patient weight, temperature, symptoms, and vitals. The unstructured data contains different notes, reports, conversations, images, videos, and audio recordings, which are very hard to quantify and categorize. Many methods and systems are available for analyzing structured data; however, currently, many innovators are focusing on unstructured data. With the help of Machine Learning, when applied correctly, healthcare providers can make near-perfect diagnoses and treatment decisions. This will lead to physicians giving the best medication to their patients for their conditions, determining high-risk patients early, and improving patients' general health while reducing cost. In US healthcare, for instance, 50% of the total costs come from 5% of the total patients, while the number of patients with chronic diseases grows gradually. With the help of Machine Learning, healthcare providers can help *diagnose diseases earlier and prevent emergency room visits*. This will not only reduce the costs for the healthcare institutions but the patients as well. With early diagnosis, the patients can start with the proper treatment and avoid expensive and time-consuming emergency care centers (Bhardwaj, Nambiar, & Dutta, 2017).

1.3.1.3 AI in healthcare

AI refers to IT, which can imitate human minds in learning and analyzing, and therefore work in problem-solving. From a software perspective, AI is a conceptual framework for executing algorithms and it is a mimic of the human brain, a network of neurons, which are connected and able to communicate. One neuron can change its state because of different inputs from the environment and cause a reaction to its neighbor neurons. As a response, this network of neurons can cause a response to an environmental factor, just as the human brain does. Because of its rapid development and implementation, AI is being used in

many technical fields such as IoT (Internet of things), autonomous driving, and robotics. While AI is different from Machine Learning because it is an algorithm created to solve problems on its own, Data Mining and Machine Learning are used many times for building AI systems (Rong, Mendez, Assi, Zhao, & Sawan, 2020). The potential of AI technologies in the healthcare sector is becoming more and more evident, which is why many researchers and scientists in the biomedical field have been trying to apply AI tools in medicine to improve the analysis and treatment outcomes, and with it, the efficiency of the medical industry. There are several areas where AI tools can be implemented: healthcare administration, clinical decision support, patient monitoring, and healthcare interventions (Reddy, Fox, & Purohit, 2019).

The main applications of AI in medicine can be divided into several categories. *AI for living assistance*: in places for assisted living for elderly and disabled people, these technologies are using smart robotic corresponding systems and improving the quality of life. For example, AI tools can be trained with image processing steps and be able to recognize different facial expressions and act upon them. This allows disabled people to move their wheelchairs without joysticks or sensors attached to their bodies. Moreover, AI can help blind people to live together with sighted people, through a smart interface assistant. AI tools can also be used for helping pregnant women through their entire pregnancy with dietary solutions, and other needed advice. *AI in biomedical information processing:* AI tools can help with natural language processing for biomedical information. Since large amounts of biomedical data have been collected over a long period, extracting useful information is a time-consuming and unsatisfying task performed by humans, which is why AI is a great solution for this. These intelligent tools can retrieve information from biomedical documents with an accuracy of about 90%, as accurate as a professional evaluator can do. *AI in biomedical research:* AI tools can act as an eDoctor for diagnosis, management, and prognosis. Furthermore, they can be used by biomedical researchers and scientists for summarizing the literature on a given topic, and screening figures of interest in a big volume of literature. *Disease diagnostics and prediction:* similar to Data Mining and Machine Learning, AI tools can be used for early diagnosing of diseases and their prediction and also, prediction of disease symptoms such as bladder failure or epileptic seizures (Rong, Mendez, Assi, Zhao, & Sawan, 2020).

1.3.2 Barriers of Big Data

While there are certain benefits and opportunities of usage of Big Data analytics in the healthcare sector, there are also some challenges or barriers that need to be taken into consideration. Since the usage of Big Data in healthcare is relatively new, the practical benefits of it are quite scarce. Moreover, there are some methodological issues regarding the quality, inconsistency, and instability of the collected data. There are limitations of observational studies, validation, some legal and ethical issues, and a necessity for improvement of data quality in the electronic healthcare records. Many technical issues

need to be solved and standardized before usage, such as issues of clinical integration and utility, which have been largely overlooked. These barriers and many more need to be solved before the implementation of Big Data analytics in healthcare (Lee & Yoon, 2017).

Data ownership and security

When talking about medical data, we have to point out that there are problems regarding data ownership. Big Data technologies allow companies and institutions to collect data continuously and from many different locations. While the processing and analysis of this data may provide great benefits to the healthcare sector, many different regulations around the world from where the data originated need to be satisfied so the data can be used. Sometimes it is not clear who is the owner of the gathered data, the medical institution, or the patient, and using the data without the patient's consent might lead to legal problems. Another barrier is the privacy and security of the data. In medicine, as well as in any other field, cyber-attacks are one of the biggest problems with Big Data analysis. As the data is usually distributed to different places, the vulnerability of cyber security is higher (Raguseo, 2018). Moreover, the data quality varies based on the source which collected it. The data from the EHRs are not the same quality as the data which is continuously collected from different devices. The collected data might have missing values and errors that might create patterns where they are none (Janke, Overbeek, Kocher, & Levy, 2016).

Information System structure support

Medical institutions first need to invest in appropriate infrastructure and IT before adopting Big Data analytics solutions. Even if the institution decides to do such an investment for Big Data analytics adaptation, there is no guarantee of return of that investment. Besides providing the IT and the needed infrastructure for Big Data analytics adaptation, the personnel must be trained accordingly, because of their minimal IT expertise, which is usually the case. Those training for the medical personnel will surely cost a significant amount of financial resources that not every medical institution can afford. Another issue with the personnel training is the reluctance to change and adapt to these new technological implementations. In some fields, new technologies have been implemented and the medical personnel are more eager to accept new inventions, however, this is not the case with every field (Raguseo, 2018).

Human supervision and patient's acceptance

The complex nature of the diseases and the number of other factors contributing to the health of the patient makes it difficult for technology to diagnose specific disease. Machine Learning and Artificial Intelligence, while effective in analyzing the data, still must rely on human judgment for final diagnosis. Furthermore, the complexity of the care process makes it yet more difficult to fully implement technology in healthcare. The healthcare team must work together to be able to give the patient the best quality care possible, coordinating the medical and psychological needs of the patients, the medical facilities,

communities, and the patients' families, and these factors are difficult to quantify (Wang, 2019). Another barrier that must be looked upon is the patients' acceptance of the prediction tools and how do they feel about a machine making decisions regarding their health. The data with which the predictive models work must be taken and stored anonymously, which makes it difficult for that data to be later used for analyzing patterns and predicting diseases for that same patient (Janke, Overbeek, Kocher, & Levy, 2016).

Doctor-patient relationship

Due to the implementation of technology in healthcare, the physicians lost control over the maintenance of the privacy of patients' data, which has had an impact on the trust that the patients used to have in their healthcare providers. The control over decision-making regarding diagnostics and treatments has been diminished as well, making it even harder for patients to place their lives in the hands of the medical institutions. Researchers have shown that patients are usually not aware of how their data is stored and shared for healthcare policy and research. Once they are informed and aware of the process they express alarm (Brill, Moss, & Prater, 2019). In addition, the doctor-patient relationship plays a big role while delivering healthcare services. The patients must be constantly informed on their health and treatments, educated, and consulted in the decision-making process. These complexities pose big challenges when it comes to the effectiveness of Big Data analytics in the healthcare sector (Wang, 2019).

Ethical challenges

The ethical challenges are also one of the most important barriers and have to be considered when using technology in healthcare. When building predictive models and frameworks, different factors are used together with the medical records and results. Some of those factors are lifestyle, environment, and wealth. However, these analyses are more complex than they were considered in the beginning. If the right factors are not implemented in the analysis, millions of data entries do not make sense and the data does not show the real picture. The ethical requirements include technical precision of data analysis and accurate statistical performance. Transparency of the data is necessary for understanding and interpreting the meaning of collected data. Lack of clarity of the data could lead to a lack of confidence. However, the transparency of the data threatens the right to confidentiality and health data privacy. Moreover, combinations of Big Data could lead to linking disease records with the location of environmental contamination. This can lead to establishing specific groups threatening the right of "not being profiled". The patients need to be asked for consent regarding the collecting, flow, and sharing of information as well as consent for obtaining the data (Leon-Sanz, 2019).

Other barriers such as *management and governance* of the Big Data, which raises the question of who is responsible for its collection, storing, sharing, and using; *predictive modeling for risk and resource use*, Big Data might not necessarily be enough for creating

predictive models for diseases and treatments and a collecting large datasets with no value cannot be analyzed for generating future insights; *disease and treatment heterogeneity*, there are too many factors that need to be foreseen when diagnosing and some think that human opinion must be present for making decisions; the *quality of care and performance measurement* and *drug and medical device safety surveillance*. (Shilo, Rossman, & Segal, 2020).

1.3.2.1 Data Mining in healthcare

Besides the previously mentioned benefits from Data Mining usage in healthcare, there are also some challenges. *Data quality*, since to retrieve useful and reliable information from a Data Mining process, the data quality needs to be good. However, that is not always the case. The collected data from different sources makes the extracted information quite complex and with many missing values. While the EHR system makes it easier and simpler to collect data for different patients, the data contains many missing values, incorrect information, or miscoding. When a patient is admitted to a hospital, they need to fill a form; however, due to the clinical state of the patient, they might not have the needed concentration and leave out valuable information or enter incorrect data. This makes dealing with the data quite complex and time-consuming. *Data sharing and privacy* is an ongoing issue since the collected medical data contains personal health information, which makes it difficult if not impossible to access that data. This creates a huge gap between the data that has been collected and the data that has been analyzed. In order to be able to access the needed health data for analysis, certain security protocols need to be implemented. For accessing the health data, it needs to be anonymized using de-identification techniques; however, it needs to be anonymized the right amount. If the data is de-identified to a great extend, it might lose its quality and become useless. *A variety of methods and complex math* is a barrier since the Data Mining techniques are quite complex. Many healthcare workers decide to continue working with traditional statistical methods instead. Some people consider *Relying only on predictive models* to be a barrier as well. The Data Mining techniques offer many benefits and help physicians with providing better healthcare service, but there should not be unrealistic expectations from these applications. These models usually have the accuracy, which is not 100%, since they are based on the outcomes of one or multi-Data Mining studies, and relying only on these predictive models when making clinical decisions can be dangerous (Tekieh & Raahemi, 2015).

1.3.2.2 Machine Learning in healthcare

The ethical challenges are also important when talking about Machine Learning usage in healthcare. Based on a survey that has been conducted in the United Kingdom, up to 63% of the adult population does not feel comfortable with sharing their medical data, to improve Big Data usage in healthcare and replace the doctors and nurses in tasks that they

usually perform. According to another study, which was performed in Germany, up to 83% of the medical students believe that Big Data technologies will improve the healthcare system and provide better services, however, students are more skeptical about how these technologies will be able to establish a final diagnosis. While many of the decision-makers in the United States that participated in the survey think that Big Data technologies will improve medicine, half of them think that there will be many fatal errors, that the tools won't work properly, and will not meet the expectations. These surveys show the ethical challenges surrounding Machine Learning in healthcare, especially data privacy, transparency, accountability, and liability. These challenges can be grouped into three categories based on how and where they rise: *Data protection and privacy, minimizing the effect of bias, being effectively regulated* and *achieving transparency* (Vayena, Blasimme, & Cohen, 2016). Similar to Data Mining, missing values are the biggest problem when it comes to clinical data which is used for building predictive models. Moreover, another issue is the unobserved or censored values, which are events that happen when the patient is not observed and that data has not been registered. Even with the needed data available, there is still a risk factor that the Machine Learning model did not estimate the situation that well, and it made a mistake. There are a finite number of clinical outcomes that were used when the predictive models were built and the fact that the model was created based on that finite number of clinical outcomes presents uncertainty. Furthermore, the data set in which the model is used may differ from the data set in which the model was built, which might lead to different results. Since these technologies rely on the data that they have been trained on, there is a possibility for *biased Machine Learning technologies.* If inappropriate sampling was done during the training of the developed algorithms, if the datasets used for training did not include vulnerable minority groups, the results will show irregularities. A study showed that when a model is trained on a semantic text from the internet, it reproduces many stereotypes such as associating domestic terms with the female population or unpleasant terms with African-American minorities. These systems can sometimes make classifications or produce results that are difficult for humans and even their designers to understand their roots. Even though Machine Learning tools learn as described above, they still make unpredictable decisions, which pose the question of whether they should be used in decision-making for saving lives (Schonberger, 2019). All these challenges need to be taken into consideration when using predictive models in the healthcare sector (Chen, Joshi, Ghassemi, & Ranganath, 2020).

1.3.2.3 AI in healthcare

While the usage of these intelligent technologies in the healthcare sector has numerous potential applications and benefits, considerable caution with their implementation is needed. Moreover, there is the concern of how people perceive the AI tools and how do they feel about machines making the decisions about their health, diagnosis, and treatments. The patients are not the only ones that may *resist the implementation* and usage of AI technologies; that can also happen with the healthcare providers even when these

systems will improve and pass the regulatory channels. It is crucial for software developers and authorities to include healthcare providers in the design and testing phase of AI tools, not only to ensure a sense of trust in the applications but also to make sure that they do not add additional work for clinicians. A user-friendly interface and integration with already existing health information technologies are also very important (Reddy, Fox, & Purohit, 2019). There are many technical barriers regarding the *limitations of artificially intelligent tools* when compared to human vision, language processing, or context reasoning. When creating the AI algorithms and their different functions, real data from disease history is taken. This opens a possibility for it to be taken by third parties, which is why there must be reliable *protection from cyber attacks*, especially because in healthcare this is directly connected to human lives. Some extreme possible situations need to be taken into consideration when talking about cyber attacks. If these attacks cause a misdiagnosing or offering a deadly drug or a procedure to a patient or group of patients, they can lead to mass murders (Iliashenko, Bikkulova, & Dubgorn, 2019). This leads to the medico-legal questions that need to be answered. Who will bear the responsibility if and when a medical error occurs, based on the above-mentioned barriers and many more? Even within the very strict medical regulations that exist, there are no specific lines of regulations that state where the responsibility lies in case any medical errors should happen. Because of the mentioned liabilities, these systems and their users need definition and clarification on that matter. This is something that legal representatives must consult with medical providers, health services, software developers, and additional stakeholders (Reddy, Fox, & Purohit, 2019).

2 METHODOLOGY

In this Master's thesis, through data analysis from both primary and secondary sources, the opportunities and barriers of Big Data analytics in the healthcare sector are outlined. In the Literature Review various sources were used to find and explain the practical usage of Big Data Analytics in healthcare, the opportunities and the barriers of this implementation. In addition to the analysis of previous findings, three interviews were conducted with professionals working in this field, more specifically in the three domains that are needed for the implementation of Big Data in healthcare: developers, medical personnel and Medical Informatics experts. The interviewees explained their opinion based on their experience and described practical examples from the field. Additionally, to analyze the opinion of the general public, Rapid Miner was used for Text Mining. As a basis for the Text Mining process, the phrase Big Data Analytics was used, as well as some of the most mentioned words during the interviews. These words are provided in the Results, along with the result interpretation. Data regarding the public opinion of using Big Data analysis in the healthcare sector was extracted from Twitter. The detailed process for text mining, data cleaning, and analysis is explained below.

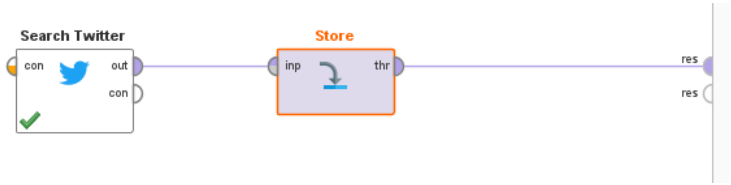
2.1 Interviews

As mentioned before, three teams need to be present for successful Big Data implementation in healthcare: IT developers that develop the solution, a medical team that sets the needs and requirements, and a Medical Informatics team that works as an intermediate between the IT and the medical team. The three conducted interviews for the research are with experts in the before-mentioned three fields. The questions were prepared based on the literature review findings and the analysis that was done for opportunities and barriers of Big Data in healthcare. Each barrier and opportunity was put in a different question, with the possibility of additional sub-questions depending on the interviewee's answer. The interviewees were asked the same questions regarding the barriers and opportunities; however the sub-questions were focused on the field of work and experience of the interviewee. The first's interview duration was 1:45h and it was conducted on 25.9.2021. The first interviewee is an electrical engineer with over 7 years of working experience in the field of developing and implementing Big Data technologies in healthcare. He is currently working as a consultant for a multinational company dedicated to driving healthcare forward by creating intelligent connections and technologies with data usage. The second's interview duration was 1h and it was conducted on 30.9.2021. The second interviewee is a Senior Health Informatics Expert with a Master's degree in Medical Informatics and over 15 years of experience in the field, currently working at an IT company that is developing software and IT solutions for the healthcare sector. The thirds' interview duration was 1:15h and it was conducted on 8.10.2021. The third interview is with a medical doctor, with a Ph.D. in medical neuroscience, a Masters' degree in medical ethics, and a Nanodegree in Data Analytics with practical and research experience in AI implementation in healthcare. Currently, he works for the biggest University Hospital in Europe. For protecting the identity of interviewees I, II, and III, the interviewees will be referred to as A, B, and C, respectively. During the interviews, the usage, opportunities, and barriers of Big Data implementation in the healthcare sector were discussed in detail.

2.2 Text Mining

The process started with connecting the RapidMiner application to Twitter, for which a connection with Twitter and a Twitter account was created. Following, the Search Twitter from the Operators option to mine data was used. As a query for tweets searching "Big Data in healthcare" was used for recent or popular tweets. Additionally, a Twitter search and sentiment analysis was performed for the most mentioned words in the interviews. These words are mentioned in the Results, along with the sentiment analysis results. After the Search Twitter operator, the Store operator was used for storing the dataset as a Big Data repository entry, and the retrieved data was saved as an excel file. The process for extracting the data from Twitter is shown in Figure 1 below:

Figure 1. Extracting data from Twitter

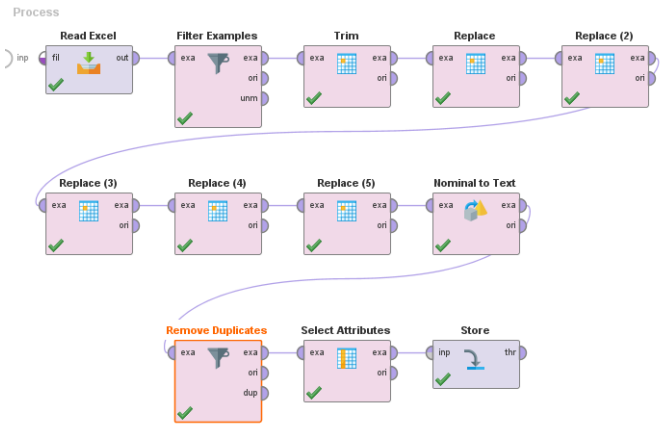


Source: Rapid Miner (2021).

Cleaning the data

After the tweets were extracted from Twitter, the data was cleaned before performing the sentiment analysis. Before the data cleaning, there were 208 mined tweets about Big Data analytics in the healthcare sector. For the data cleaning part, a new process was used and it started with the Read Excel operator to get the previously saved data. Furthermore, the Filter examples operator was used, to filter only the tweets that are in the English language to make the analysis easier and more accurate. Moreover, the Trim operator was used to create new attributes from the text and to get rid of the unnecessary leading and trailing spaces in the nominal values. In the extracted tweets, there was significant unnecessary information such as usernames, URL paths, hashtag signs, and so on, so they were deleted by using the Replace operator. Signs such as RT, @[^]*, #, http[^]*, [-!"#$%&'()*+,./:;<=>?\[\\\]_`{|}~][-!"#$%&'()*+,./:;<=>?@\[\\\]_`{|}~] were deleted. The Nominal to text operator was used to convert all of the nominal values to string values. Next, the Remove duplicates operator was used to remove the duplicate tweets and the Select attributes to get rid of the unnecessary columns. For the Sentiment analysis, several columns were needed: negativity, positivity, retweet count, score, scoring string, text, total tokens, and uncovered tokens. Once the data was clean, the dataset was saved as Big Data cleaned repository. By the end of the data cleaning process, there were 115 tweets left for analysis. The whole process of data cleaning can be seen in Figure 2 below:

Figure 2. Data cleaning

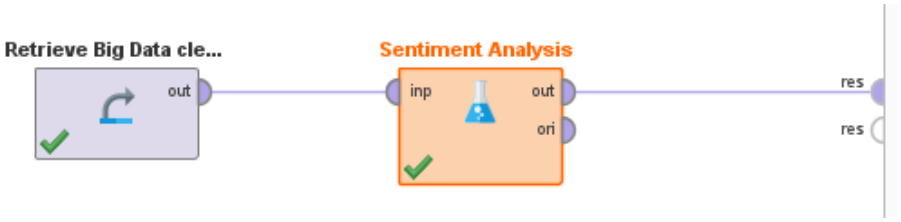


Source: Rapid Miner (2021).

Sentiment analysis

Once the data was extracted from Twitter and cleaned, it was prepared for the sentiment analysis. For the sentiment analysis part, the Sentiment Analysis operator was used, and as an attribute, it was used the text, or in this case the tweets. The text language was limited to English and used the general_en sentiment model. The sentiment analysis process can be seen in Figure 3 below:

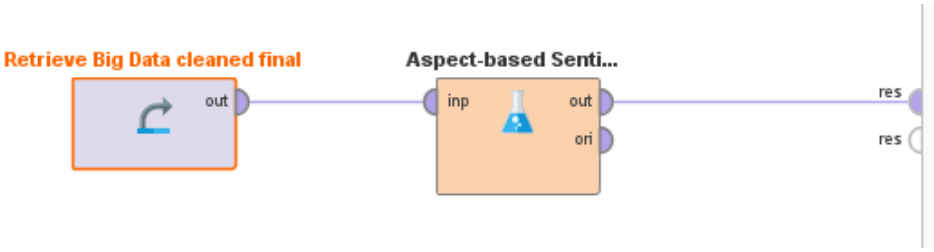
Figure 3. Sentiment Analysis



Source: Rapid Miner (2021).

For more detailed analysis, the Aspect-based sentiment analysis was used, to see exactly which parts of Big Data implementation in healthcare does the public finds positive and negative. The tweets and their content were then analyzed and described. The Aspect-based analysis shows which particular words of each tweet were used in positive, negative, or neutral connotations, and it is shown below in Figure 4.

Figure 4. Aspect-Based Sentiment Analysis



Source: Rapid Miner (2021).

3 RESULTS

In this part of the thesis the results from the interviews and the text mining are explained in detail. The three interviews that were conducted with professionals in the field of IT, medicine, and Medical informatics are described, and each interview is divided into two parts: opportunities and barriers. Moreover, the findings from the Text Mining process with Rapid Miner are presented and the results are explained. The most used words are presented in a table, together with the sentiment analysis results.

3.1 Interview I

The first interviewee, A is an electrical engineer with over 7 years of working experience in the field, based in Bangalore, India. He is currently working as a consultant for a multinational company dedicated to driving healthcare forward by creating intelligent connections and technologies with data usage. A's team works with different areas of Big Data for healthcare analysis, from timing patients' journeys to decision analysis. The focus market of A's team is oncology, currently working with 35 different types of cancer. A and the team are obtaining raw data from various providers such as hospitals, pharmacies, EHR, or ERM, and based on the clients' requirements and needs, they are analyzing the data, searching for answers, and building possible solutions. Many patients are suffering from cancer all around the world. There are billions and billions of records just from the breast cancer market alone, and when processing that data in a traditional database, A mentioned that it takes hours, sometimes up to 30 hours just to run the line of therapy. A and his team require someone to time and monitor the process and in case of failure, to be able to restore the process because they cannot start from the beginning since the timeline is very tight. Another requirement is data storage. The company is using their private cloud, where they run the entire end-to-end Big Data processing. They are providing data to many different healthcare organizations all over the world, and they need to use an optimized way of storing the data because data storage has a huge impact on data processing.

3.1.1 Opportunities

From the implementation point of view, A said that the modeling and implementation itself of Big Data in healthcare is fairly easy. Once Big Data techniques are implemented in healthcare, the efficiency and productivity indicators are higher. The healthcare organization can set up a cloud, where all the Big Data is available and ready for usage for Machine Learning or anything else. Having the data saved into a cloud and being available for processing and usage is much more efficient than saving it traditionally- locally. Not only that saving to a cloud makes the process much more efficient, but also allows the organization to save money on investments. The performance is as well much faster. The company has implemented different Machine Learning techniques, which have helped to drastically reduce the data processing from 30 hours to around three or four hours. With optimized data storage and implementation of Machine Learning techniques, they have saved 80-90% of the time, which has helped significantly to the healthcare organizations they are providing data to, and therefore, the patients. When asked about the benefits that derive from Big Data usage in healthcare, A highlighted that with Big Data usage, the whole healthcare process can be faster and, when diagnosing diseases, the results will have a more accurate ratio. Just with the usage of the smart watch for example, by getting the data from the smart watch of a patient, the healthcare organizations can have much more information to base their diagnosis on, and connect it to the results from other healthcare

devices. This only leads to faster and more accurate disease diagnosis, more effective treatments, or even prevention of certain diseases. According to A, Big Data can be used in the pharmaceutical industry as well, for faster and cheaper medicine development. The process usually starts with a particular market sizing. The pharmaceutical companies want to know the market, how big it is, how common the disease is, and the needs of the patients and to predict how the drug would be moving in the market. For this, they reach out to data-savvy companies, such as the company A works for, which are getting their data from pharmacies or hospitals. The pharmaceutical companies raise a request for analysis and market sizing for a particular product and the results of such analysis can lead to deciding various factors regarding the drug development. Besides development, Big Data can be useful for clinical runs, by running an analysis on how the drug develops and how would affect the patient. This means that there will be less amount of work for the pharmaceutical companies, on which they usually spend billions of dollars. Instead of running clinical runs, the pharmaceutical companies can focus their time on other research for similar or different disease treatments. This is much cheaper and faster than actually performing the trials, which could eventually lead to faster and cheaper drug introduction to the market for general usage.

According to A Big Data can also be used for improving the quality of the healthcare flow and measuring the performance of the whole process. For example, if we take a pharmaceutical company that is coming up with a product launch. Before the product launch, there are many clinical runs and analyses and once the product is ready for the market and general usage, the company may want to know how receptive the market is to the new drug. Before releasing the product the company wants to analyze and to see how well they will perform, and this data can be retrieved from tracking devices, or surveys, or any data that can be found online from potential customers. After releasing the product, the pharmaceutical company and their competitors as well, want analysis on how good this drug is doing. The actual usage, how it is performing, is it helping patients, potential side effects, or any other information connected to the usage of the new drug are very important for pharmaceutical companies. If Big Data is available, it can be tracked in real-time. There are many statistics on patient care and on information from the patients that can be retrieved and used for further research and improvement of the existing drug. This information can help pharmaceutical companies to change the molecular formulations of their products or alter something else that can improve the drug and help the patients. Moreover, A mentioned that Big Data technologies can benefit healthcare by creating personalized medicine. Data-savvy companies can analyze the patient, their medical history, the lifestyle that they lead, their environment, how were they brought up, their genetics, and many other factors that could have an impact on their health, predict diseases, and try to prevent them or start with treatments faster. This can significantly reduce hospitalization and the workload of the medical personnel. For example, by analyzing medical history, it can be found out which are the patients that are usually going through the seasonal flu, and predict whether they are going to get it again. The company A works

for has access to more than 10 years of medical history, which they have collected, and based on that data they can easily see which patients are more likely to get more serious symptoms of the flu. Based on this data, the doctors can recommend and encourage these potential patients to get the seasonal flu and therefore significantly reduce the risk of their hospitalization. A says that this kind of information of specific patients can be used for predicting diseases and disease development for the general population and future generations. Once the analysis is run for a certain number of patients, the data collected can be used on a certain population sharing similar characteristics or living in a similar environment. By analyzing some parameters such as gender, or to which age group they belong, it can be easily predicted which patients are more likely to get certain diseases. While it is very difficult to make these predictions for five or ten years, considering there are so many dependent and independent variables that need to be taken into consideration, such predictions can be made for the next five or seven months with high accuracy. The shorter the period for which the prediction is made, the higher the accuracy is. For example, the accuracy of Big Data predictions for patients for the next year is 60% accuracy but for the next six months is 84% accuracy. Big Data technologies are very useful for disease prediction; however, we have to take into consideration that these are only machines that are making the predictions, and working with the data that is available at the moment. There are many scenarios as well, on how things could work out in the end, and Machine Learning technologies are not always 100% right, but they do provide a huge help with diagnosis and can lead to creating more effective treatments for the whole population, not only certain patients. And even though, the predictions are limited for the next five or six months, and not years, they still have a humongous positive impact on hospital care and patients' health. By preventing certain diseases, the number of hospitalizations decreases significantly, allowing the medical personnel to spend more time with the patients that need help and attention. This is how certain diseases can be controlled, and how somehow the outcome can be manipulated for patients' best interests and, eventually, the entire population.

According to A, Big Data can also be used for more efficient resource usage. With lower patient hospitalization, more hospital beds are being freed for the people that need them. If those beds are not being used, then the hospital costs are lower, since fewer resources such as medications, hospital care, and meals are being used. If the hospital is expecting 2000 visitors each year, preventing certain diseases and avoiding hospitalization is going to reduce the number of expected visits significantly. This means that hospitals end up paying fewer utilities and lowering the workload of the medical personnel and as a result, lower expenses. This could eventually lead to lowering the cost of hospital care for the patients as well. When we take into consideration all of the medical history and the number of increasing patients, we can also say that Big Data can help a great deal with clinical decision support. For example, the Big Data technologies analyze everything, such as the whole medical history and then they suggest to the doctor what might be the cause of the disease or symptoms and what might be an effective treatment for it. Even though in A's

experience this is not yet entirely used and implemented in healthcare, there is a significant amount of data exposure that can be used for this particular instance. It might take more years until Big Data is being fully used and explored, but it is something that the healthcare sector can benefit from. By tracking unusual records of data, some fraud movements can be detected and maybe even prevented. One example is detecting drug abuse, which A had identified in one of their projects. There are some drugs that patients need to refill after a certain amount of time, and if the drug needs to be refilled every 30 or 60 days, and they see that the patient goes and refills the drug more frequently than they should, that is a sign of drug abuse. One example of this is cough syrup and medicine, which some patients keep buying very frequently, and for which they are paying from their pockets, without even using their insurance coverage, because it is still cheaper than buying illegal drugs.

3.1.2 Barriers

Big Data is a huge field, and it is far from being fully explored, A mentioned that the field is still under the testing phase. Because of this, in his opinion, we haven't seen its full potential or all of the barriers. While he considers that he cannot know all of them for sure, with Big Data there are always pros and cons, the size or amount of data. In his experience, he can comment on two barriers: Big Data's storage and processing. The data is being stored in a cloud and being used by different users and everyone is processing the data or reading a database in a particular cloud at the same time. This can lead to performance degradation if the cloud is not optimized. The issue is that generating a report or doing calculations usually takes more time compared to if the cloud is optimized. Some mistakes are happening when using traditional databases that are from the last 20 or 30 or more years which are not optimized. These databases were created when the devices were not as advanced as today and are unable to handle such humongous data, which is why the processing is so slow. Another challenge is that the cloud sometimes has issues and corresponding downtime. While those issues are resolved, the whole system is down and no data is being processed. Sometimes it happens that the server is down up to 30 minutes, which is a lot of time considering how much data is being processed in 30 minutes. In A's experience, the quality of the data has not been an issue so far. There are huge amounts of data coming in, and in his experience, the quality is the main challenge rather than the size.

The data that A and his company are collecting is mainly from pharmacies, EMR, EHR, or hospitals. The hospitals or any other healthcare institution is collecting the data from the patients and then it is in charge of the maintenance. The pharmacies own the data that is being generated in their stores. Some doctors own their clinics, therefore all data that is being generated in their clinics fall under the ownership of the clinic. The organization, as a data-savvy company, collects the data from different vendors such as pharmacies, hospitals, and EHR records. However, the data that is being sold to the customers is not all of the data that is being collected. Healthcare institutions are allowed to pass over, sell or publish only a certain amount of data, and some of that data is well hidden to protect the

patients' privacy. Healthcare organizations can reveal patient ID, or certain doctors' details, however, they are not allowed to reveal patients' private information or demographic information. The data is usually saved in a cloud, and it is completely secure. Only an authorized person who is given access to that particular password can access that information. If someone is not authorized to access that data or is not an employee of the organization that is storing the data, accessing the cloud and retrieving it, is almost impossible, because of all of the security measurements that are set to protect the data. Even someone manages to access the data, which is being encrypted to a great extend and it wouldn't make any sense to the person retrieving it. All of the data is being completely masked, every single detail is encrypted and to someone without a decryption key, the information will not be of any usage.

All of the technologies that are being currently used in healthcare are compatible with Big Data technologies. Accessing data is much easier nowadays because all of the data from the medical devices, and from all of the equipment that is being used in the hospitals is stored in the cloud, from where it can be accessed for Big Data analysis. A highlights that even if the data would be stored in a traditional database, it wouldn't be a big challenge; it is much easier if the data is stored in a cloud. However, Big Data implementation in healthcare is harder in developing countries than it is in developed countries. If we compare a Big Data implementation in healthcare in U.S and India, there are many differences. Executing such a project in a developed country is much easier. Many companies in developed countries are already moving into Big Data because they have the resources and the technological capacity and expertise to do so. In developing countries, this is not the case because they are having a lack of expertise, lack of resources, lack of collected data, and data clarity. The implementation process is much lengthier and complicated due to these factors. In some developing countries, just demanding or collecting the data is already a barrier.

Certain teams need to be involved in the modeling, development, and implementation of Big Data Analytics in healthcare. The development team is in charge of developing the programs and systems. The medical personnel need to be consulted regarding the needs, as well as the correlations between different patterns. The company in charge of developing the programs has its clinical research team, which has to be collaborative with the development team when building the programs. According to A, other teams that need to be involved are the sales, as well as the marketing team. The work starts with the marketing and sales team, which presents and sells a project to healthcare organizations. Once the project is sold, the development team comes into the picture. To make it work, everyone has a specific role to play and everyone needs to be integrated. As for the development of the programs, the medical personnel need to be included. The development team usually doesn't have a background in life sciences, so they need to consult experts in the field or their clinical research team. When talking about using mainly AI and Machine Learning in healthcare, A is quite optimistic and in his opinion, from a technical point of

view, we still have 4-5 years, until we can start using heavily Big Data in healthcare. Nevertheless, we have to take into consideration the marketing and sales point of view, because even if the technology is advanced enough and ready to be implemented, if no one is willing to buy it, we cannot use it. The problems arise when writing a Machine Learning program for example if the collected data is not good enough if there are caveats in the data, which can lead to a disaster. Machine Learning, for instance, cannot be used for all of the needs and steps for creating a full-line therapy. Moreover, mistakes can be made if the variables are not handled properly, and the team in charge must be very careful. Many can benefit from Big Data usage in healthcare. The pharmaceutical companies get more work done in less time and with fewer resources used. Then, the patients will be getting the products and services faster and for a cheaper price. The patients will also benefit from better healthcare services, due to the reduced workload of the healthcare workers and of course, better life quality due to diseases prevention and more effective treatments. The hospitals and other healthcare institutions will benefit to a great extent, due to workload reduction, more efficient resource usage, and lower costs. The processing of the healthcare reports, or CT and MRI scans will take significantly less time and effort and help the medical experts with disease diagnosing. And last, the data-savvy companies, which get the medical data and convert it to information that can be useful for healthcare organizations.

3.2 Interview II

The second interviewee, B is a Senior Health Informatics Expert with over 15 years of experience in the field, currently working at a company that is developing software and IT solutions for the healthcare sector in Hamburg, Germany. B has a Master's degree in Medical Informatics and he and his team are currently working on a couple of AI-based systems, on how they can deploy Big Data on their systems. They are capturing life data on patients' movements and analyze those movements on how the patient walks, and then use the data of the live movements to see if the patient has a particular disease or not. This data analysis can be also used for predicting whether the patient has a chance of developing a disease in the future. The movements of the patient are analyzed, and if their movements are slightly changed, for example, 10% or 20% deviated from normal, the system can predict that the patient might develop a particular disease after one or two years. In case of deviations, the doctors advise certain lifestyle changes and various exercises that the patient can do to prevent the disease if it is detected early in the patient life. In his experience other Big Data implementation is detecting and understanding the effectiveness of a certain treatment, by reading images that are done before and post-treatment. If a patient has knee surgery, upon which he is being analyzed for six months while being led to do exercises by an orthopedist, and after six months the patient is being tested and scanned again to see the results. The patients are being monitored on every doctor's visit and after each appointment, they receive feedback if they are doing the exercises correctly and are they following the doctors' advice. For instance, if the patient is not flexing his arm

correctly and needs to change the angle so they can receive 100% of the benefits from the exercises. Another Big Data usage in healthcare is detecting facial movements in patients who come with facial pains. The doctors collect the information from the patient and we try to visualize what had happened to the patient, whether it is a disease or bad habits, and based on this information the system can predict how the pain will progress and what can be done to treat it or prevent it.

3.2.1 Opportunities

Big Data is largely used in ICU (intensive care units), where there are IoT and BI devices monitoring blood pressure, heart rate, pulse, and respiratory rate. There is a central monitoring station, from where a nurse or a doctor can monitor the patients in real life because any change might be vital. Any change of these rates immediately notifies the physicians so they can attend to the patient. When thinking of Big Data in healthcare, people usually imagine a direct connection to the patients and the treatments; however, Big Data in healthcare has a much bigger role than only that. Currently one of the biggest topics of Big Data usage in healthcare is the COVID-19 vaccines and their storage. The COVID-19 vaccines have very strict rules regarding their storage. The vaccines need to be stored at a particular storage temperature; otherwise, they might lose their effectiveness. Because of Big Data, nowadays there is something called live monitoring of the supply chain, which helps with the COVID-19 vaccine storage. In some countries, some systems are IoT-based and can control the storage unit virtually. There is a virtual control room, which monitors all of the storage units across a particular country and provides live information of temperature, humidity, and other important information, so in case something wrong happens in a storage unit, there is a warning sign that is sent to the staff in charge.

Everything we know is based on data. Everything that we find, analyze or predict is based on data, so the more data we have, the better we can predict and do anything. B says that there are Big Data usages that are quite advanced, and are already developed; however, they are not in the market yet. One example of these usages is intended for diabetic patients. A diabetic patient who shares their activity data, location data, and sleep patterns live. Because of these three parameters, it has been scientifically proven that it is possible to control or alter a patient's blood glucose levels. By monitoring the patients' patterns, the data and the patients' fluctuating blood glucose levels can be mapped and it can be predicted when the blood sugar increases. If at a certain point of the day the patient goes to a coffee bar regularly and their blood sugar is increasing, it can be concluded that he is taking sugar with his coffee, or eating sweets. Based on this information the medical personnel can instruct him to change this habit. Another example is the apple bed, which has an integrated sensor that can identify the quality of sleep, whether the patients sleep well or not and how is it related to their blood glucose levels. Diabetic patients nowadays have a small CGM (continuous glucose monitoring) device installed just below the skin,

monitoring their blood glucose levels. Connecting the blood glucose levels overnight with the sleep pattern allows back-tracking everything. If a patient hasn't slept well, it might mean that their glucose levels went up. Did the coffee, which the patient took with sugar, disturb their sleep, or maybe if the patient didn't do enough exercise. The more data the system gathers, the more predictions we can make. The healthcare flow is quite complex. There are not usually examples where Big Data is produced and analyzed fast. Most of the data is stored and then analyzed. Once the patient walks in, the system collects the data and when the patient does particular examinations, data is being collected as well. There are not huge amounts of patient data and data changes coming in microseconds, so the actual patient journey cannot be that easily affected by Big Data usage. But, by using live Big Data for each patient, related to different sensors the patient's activities and patient's vitals can be monitored. When a patient stays at home 24/7 and is constantly monitored, the data can be used to create personalized treatments. By combining this information with other data, such as medical history, age, demographic information, lifestyle, or the way the patient was brought up. B highlights that Big Data can help increase the chances of more accurate disease detection and treatment. By monitoring the patient daily, the system can also model a disease progression. If the doctors know currently at what stage certain disease is, and they know the patient, his body, and habits, they can try getting ahead of the disease and treat it before it advances, or in some cases, even prevent it. They detect small changes in patients' health or behavior and then connect them. While these might seem like small changes, in the end, the small changes could lead to huge results.

The earlier the system and the doctors can predict the disease in a patient, the less advanced the disease is and the sooner the treatment can start. If they manage to reduce the number of hospitalized patients, by early disease detection, they can reduce the workload of the medical personnel and allow them to spend more time with the patients that need it. Less hospitalization will lead to less bed occupancy and lower utility bills. If a probability of a certain disease is detected earlier, the patient will implement changes to their lifestyle and prevent the disease from ever happening. This means that the patient will not need medical attention and will not use his insurance to go to a hospital, ask for a visit, use human resources to understand what is happening, and in the end start treatment. By using Big Data to avoid hospitalization, Big Data can help save many resources, such as the number of resources a hospital has to spend, how many days has the patient stayed in bed, the patients' expenses, doctors' time, how many nurses, and how many hours of nurse duty has to be added to this patient's management. With good and proper data, Big Data can reduce tremendously the resources used in healthcare and at the same time, improve patient care as well. The healthcare sector is very complex and therefore, generalizing a particular patient or group of patients to a large-scale population is complex as well. When collecting large population data, analyzing it, and then funneling it down to a single patient, the decision support systems work. If a hospital has had hundreds of thousands of patients above the age of 40, that had led a certain lifestyle, medical history, and have developed diabetes, the system can predict quite accurately that the next patient that has the same

parameters, might develop diabetes as well. This is where the Big Data support systems work. But doing it the other way around, this is not the case. B highlighted that the system cannot collect information from a single patient, or group of 10-100 patients, and said that all of the patients in the same group might develop a similar disease. But there are many other benefits, for instance, the pharmaceutical industry can benefit as well since it is mostly focused on clinical research. Whenever the pharmaceutical companies have a new drug and do human trials, the patients are monitored 24h per day to detect any possible side effects and if needed, act upon them. Once a side effect happens, the patient may have a particular or an adverse reaction to the drug and might need immediate medical attention. This might lead to faster and cheaper drug development, which in the end might lead to cheaper drugs in the market. Machine Learning is also used in fraud detection. One example is using Big Data to detect insurance claim frauds. Many hospitals and clinics send claim forms to their patients, so the patients can use those forms to claim insurance from insurance companies; however, many abuse these forms. Big Data can be used to identify these claims and detect points, certain records that do not seem accurate, and prevent fraud from happening. This is one of the most popular usages of Machine Learning in fraud detection in healthcare.

3.2.2 Barriers

The most common barriers to Big Data in healthcare are government regulations, legal regulations, and data sharing. For example, medical data from Germany cannot be shared with another country and some systems can work only within Germany. Because of that, for collecting data the healthcare organizations use anonymized patient data, with which they cover patients' information with a combination of numbers. Getting permission to use a patient's data is a very complex process and it takes time, and by the time the approval comes, the patient's data becomes obsolete. B says that the patient's data ownership is regulated differently in every country. This is still a huge topic that is being discussed and there are still no concrete decisions on that question. Some countries say it belongs to the hospital, some say it belongs to the patient and some say it belongs to both. The ownership of the data also depends on individual norms and regulations within the hospital on how they plan to use the data and for what. If a company is using patients' data for any purposes, means that the legislations of the country and the norms of the hospital allow it, and most likely, that the patient has already consented to share it. According to B, there are still no general standards on this topic; no one with certainty can say that Big Data has to be implemented and run in a specific way. Another barrier is not clean data. He and his team receive large amounts of data that has been manually filled and people start using text. If in the form there is a field for age, some are using free text, which is why we are trying to develop standards of how is Big Data being collected. When talking about Big Data in healthcare, AI, or Machine Learning, B says that we are still very young and there is still a lot of research going on. When we analyze the market, there are only a few projects, which are fully developed, and there are many ongoing projects on how to collect

clean data. And logically, the cleaner the collected data is, the better the prediction rate of AI and Machine Learning will be. If we want at least 90% accuracy, we need very clean data, which is not the case in real life. B says that they have null values in the collected data, for example for age some people enter values like “less than 30” or a range, some enter the date of birth or free text, which is completely unreadable since they have to manually correct it. Moreover, a huge challenge is the mindset of the people and convincing them about the benefits of Big Data usage in healthcare. According to B, the technology is available, as well as the algorithms and everything needed for Big Data usage in healthcare. But for successful implementation, we need clean data, governmental and legal regulations, and patients who understand the benefits from Big Data for them, and the future patients. Professionals need to talk to the patients and allow them to learn and start believing that their data is valuable, that it is important to make the system more transparent on how that data is going to be used and for what. Professionals in the field need to educate people regarding Big Data because most of the issues that they have with Big Data are because of fear of what their data is going to be used for. Once the patients are aware of how their data is used and for what, they become more open to sharing it. The opposition to Big Data usage is happening mostly because of a lack of transparency and lack of education, which we need to change.

Another barrier, according to B is that in healthcare, almost always, we need a human to supervise and if needed to intervene. Since Big Data is built like a machine, we cannot completely rely on the machine alone. He says that Machine Learning algorithms are like black boxes and we do not know what is happening inside those black boxes and how the output is going to be, how it is going to come out of the system. That is why we always need human intervention to cross-check and verify the output. But even with the necessary human supervision, Big Data technologies save time and effort for the medical personnel, since they are not required to do the cognitive process or manually read and input information. Big Data technologies read an image or a graph and based on that they produce an output on what the issue or the disease might be, but a doctor always has to verify that output. The already existing technologies are compatible with Big Data. Big Data technologies for healthcare are almost the same as for any other industry. The essence of the technology is almost the same, especially the way the data is being collected, how is it being processed or analyzed. But B says that the healthcare environment is a completely different domain and the patients’ care flow is completely different from the manufacturing workflow. Even though the same technologies are being used, the flow is completely different, so when implementing Big Data, the developers always have to map the workflow and then implement it onto the existing technology.

The most complex aspect, when implementing Big Data in healthcare, according to B is that a non-medical person, such as the IT team needs to understand the use case. Once they understand the specific use case, then they can map out the particular flow in a hospital environment. When mapping a simple flow in a hospital environment, we imagine a person

that walks in a hospital, goes to the register, then to the doctor, the doctor does a check-up and prescribes a medication. The patient collects this information, gets the medication, and goes out of the hospital. It also might happen that the patient is not eligible for a treatment plan, which means that the patient is discharged without getting treatment. It is very important to understand the use case when building Big Data technologies, so then they can create the design and determine the key points. From the point of view of the medical personnel, a strong IT background or deep understanding of Big Data technologies are not needed. A basic understanding of the processes is enough for their usage. B says that from his experience, the age of the medical personnel is not a barrier. The older population is quite eager to learn and they are learning very fast. He says that the doctors are very interested and are exploring these technologies because they feel it adds value to them and their knowledge, rather than to their workload. However, B says that there is a huge deficit of skilled labor in the Big Data field both in developing and developed countries. The demand is much higher than the supply, every day there are more and new technologies so even the skilled workers need to learn and improve constantly. Big Data is now on the rise, so many companies in every sector want to implement it and use it. The lack of skilled workers in this field is much bigger in developing countries, due to a lack of proper education and lack of resources. Besides that, the infrastructure is a huge problem, since developed countries have already invested in infrastructure and meanwhile developing countries are trying to make up and invest now since they have realized the potential and value of Big Data. Developing countries have started collecting Big Data and they are reframing all legislations concerning the usage of data, but there are the developed countries that have already tried this. The developed countries have existing models from which the developing countries can learn and adopt, but they are still falling far behind and trying to catch up.

When creating Big Data solutions for healthcare, besides the development team, B says that the final user, the doctor needs to be a part of the team. The users, the doctors, talk in medical terms and create the requests, what do they need and what needs to be done. However, they lack IT knowledge and it is difficult for them to explain the requirements to the development team. On the other side, there is the development team that has the IT skills to develop and create the final product but is missing the medical knowledge to understand the requirements and needs of the medical personnel. As a bridge between the two sides, there is a Medical Informatics team, with members that understand both sides, such as B and his team. That team understands the needs and wants of the medical personnel as well as the medical terms, and has the IT knowledge to understand the capacity and to explain the requirements to the development team. For a specific project, for everything to work as it should, all three teams are required for successful execution. Once the product is done, the final users- the doctors test it and give feedback to the Medical Informatics team, which then gives feedback to the development team. According to B, this is the most difficult, time-consuming part of the project. Big Data implementation is a quite costly project, however, in his experience, the hospitals and the

medical personnel see how big the benefits are and that it is worth the money. In his experience, the benefits of Big Data are much bigger than the barriers in general.

Talks and debates are happening on the topic of the ethical challenges of using Machine Learning and AI in healthcare. Some doctors disagree with the usage of Machine Learning and AI because they do not understand and they do not know how those models work. This is the main reason why the doctors do not trust 100% of the output that comes out from the Machine Learning model. This creates ethical doubt in their minds; on whether or not they should do what the machine says is right. These are quite debatable points and they are far from solved at this point. In B's opinion and experience, we are still very young in this field, and that we are still building the foundation. He says that we have covered only 15-20% of what Big Data has to offer and this is not only from a technical point of view but also from the users' side. We still have to develop and test, but also convince people why is Big Data such a good opportunity and how is it benefiting the patients. There are, many undeniable benefits from Big Data implementation in healthcare, and the patients are the ones that should benefit the most. Ideally, the systems are being built to help the patients and their treatments, to help improve their health and therefore the quality of life. The patients are the center of focus, so any technology that is implemented in healthcare should be benefiting the patients. Alongside that core benefit, the medical personnel will benefit from reduced workload and efforts to process information, since Big Data allows this to happen automatically. B says that this is how the Big Data technologies and all of the technologies for the healthcare sector are being developed. The center of the whole system is the patient and everything is revolving around that. Every new system and software is designed in a way to benefit the patient and this is how it should be.

3.3 Interview III

The third interviewee, C is a medical doctor, with a Ph.D. in medical neuroscience and a Master's degree in medical ethics. With over 15 years of experience in the field of Big Data in healthcare, C is currently working as a Senior Healthcare AI Researcher for the biggest University Hospital in Europe and is based in Berlin, Germany. He is also the Chief Scientific Officer of a company focused on medical AI data analytics and personalized AI solutions and a TEDx speaker on the topic of AI in healthcare. Additionally, he has a Nanodegree in Data Analytics and is a visiting professor of Medical Informatics in the UK. In his experience, he says, Big Data analytics have not fulfilled their full potential and they are not used much in healthcare nowadays. The overwhelming majority of applications that he has been involved with, do not use Big Data analytics. Even though Big Data has very promising benefits, it is very difficult to get access to large amounts of data that would refer to as Big Data. The main reason for this is data privacy since healthcare data is very valuable data and has high protection privacy levels. Because of this, it is very difficult to share this data and, at the end, when the actual Big Data technologies are applied, they are usually used with much smaller data sets. There are millions of images and datasets that are freely available for Machine Learning processes,

but for the healthcare field, researchers publish papers with the data of only a couple of hundred patients and they are happy that they obtained that data. It would be amazing if the healthcare sector could leverage all of the available data, but sadly that is not the case. C says that there is a huge gap between what people promote regarding Big Data usage in healthcare and what is done.

3.3.1 Opportunities

C says that the main benefit that we can achieve with Big Data analytics in healthcare is that we can finally have working precision medicine approaches. For example, if someone of a smaller size goes to the doctor with a certain type of infection, the doctor will probably prescribe 500 milligrams of antibiotics, three times per day. And then if someone with 130kg of muscle, a bodybuilder, also goes to the doctor with the same infection, they will get the same medication and dosage. This is how the medicine works and how many studies have been conducted to get a population value that has been tested. All of the people are different and in almost any case that is there, we can probably distinguish different subgroups. In C's experience, Precision Medicine has kind of replaced the term personalized medicine, because personalized medicine has often given the implication that we've tried to tailor individual therapy therapies, which is impossible. The doctors can't create individual therapies; but, they can create individualized therapies for several subgroups and this is the big promise of Big Data analytics in the context of AI. The term personalized medicine is misleading because it lets people think that everyone gets a unique therapy and that is not the case. Big Data can help with making the therapies more tailored to the patients, but that is not a different therapy for everyone. Doctors know that a patient has a certain disease, and they know at what stage is now. When the data is analyzed, the doctors see where the disease is and at what stage is it going, and they can use this information to treat it or prevent it. This is an example of the prediction of disease progression. If we take as an example multiple sclerosis, where a patient has already been diagnosed and from that stage on, the disease can develop in many different ways. In his experience, Big Data cannot help with disease prediction in similar cases to this, since no one can say in which way the disease will develop, but there are cases where Big Data can be very helpful with disease modeling progression.

C says that the moment the heterogeneous data that allows doing precision medicine; the application of it is only a use case. There are almost no differences if it is diagnosing diseases or predicting disease development or anything else. The problem with current Big Data and Machine Learning solutions is that these technologies have been trained on small or narrow data. Usually, these models can work well on this data and the moment they are tested on other data, other hospitals, or different hardware then the performance drops. We also see that in publications and reports from radiologists they say that these tools do not work well on their machines. Ideally, if we do use Big Data for training machines and AI implementation, the heterogeneity of the data will lead to very different results. The

hospitals will have different scanners included or different techniques of data recording. In that case, C says that the developers can create a model that generalizes and adapts well to all machines, all hospitals, and all countries. Potentially this will impact the patient's journey, but in his experience, this is not happening yet, because the data-savvy companies do not have access to these large amounts of data that can make a difference. Even in his organization, as the biggest University Hospital in Europe, they do not have anything centralized that would make the data available. According to C, this is the case with many hospitals everywhere, especially in Germany. Of course, there are some examples, such as Israel, that have all of the data available. Israel is a country where a lot is happening in this area. The insurance system in Israel provides the patients with a GP (general practitioner), specialists, hospitals, and pharmacy services. All of the data generated throughout the process is in the hand of the insurance companies. The insurance companies use this generated data to work with startups that develop new tools for the healthcare system and improve the patients' journey. These are the benefits that Israel has because of the healthcare data digitalization, which is completely different than Germany. C highlights that in Germany all of the medical data that each GP has is usually paper-based and even if it is digitalized, it is difficult for other GP to access it because everyone uses a different system. The hospitals or insurance companies are not connected to the same system as well. This makes it very difficult to implement Big Data analytics in countries where the data is not digitalized and centralized.

Generally, in almost every country the healthcare costs increase annually, and at a higher rate than the GDP increases. This is one of the main challenges of the healthcare systems worldwide, because every year, the healthcare system becomes more and more unsustainable. People complain about the NHS (National Health Service) in the UK, about the long waiting period in Canada, in Germany people are complaining about other things and so on. In the end, everything comes to the fact that people in many countries cannot afford it and it is getting worse and worse by every year and people feel that. Ideally with Big Data and AI implementation, the workload of the medical personnel will decrease and the doctors could spend more time with the patients that need it. However, according to C due to the increasing costs of healthcare, layoffs are more likely to happen. Once the hospitals see that with AI implementation instead of eight radiologists, now six are enough, to cut costs, they will be let go. In that case, the hospitals have saved two salaries. Implementing Big Data in healthcare and becoming more efficient means that the hospitals can save more money, which is still a good goal, but is not what people would expect. Big Data will help the hospitals with more efficient resource usage and with reducing costs and saving money, but this will most likely be done by cutting down personnel. This doesn't mean that AI and robots will replace doctors or radiologists in this example since in healthcare the doctors cannot be replaced by machines. But Big Data will for sure help with decision support and faster disease diagnostics, which will save time, and fewer people can do the workload intended for more personnel. When we are talking about AI and Machine Learning-based solutions, we are always looking at the benefits and how

better would they perform the job compared to the medical personnel. Taking the example of the radiologists, if we have a tool that is better at diagnosing diseases than the average radiologists, how much sense does it make to use this tool as a clinical decision support system, instead of as a lead-in diagnostics? C mentioned that if these tools are more accurate and faster than the average radiologists, the hospitals should use them to diagnose the diseases and the radiologists to check the results for obvious errors. This would be a perfect solution if the AI and Machine Learning solutions are as good as the average and the good radiologists. The problem with using medical personnel is human error. After long shifts, the radiologists get tired, and making a mistake becomes more probable, but with the usage of AI, the radiologist sees the result and realizes what he has missed. However, some study cases indicate that doctors perform worse when using AI clinical decision support. Before implementing Big Data in healthcare many studies need to be done so it can lead to better outcomes. Big Data technologies have to be tailored to a use case and proven that it works in a certain environment.

The availability of other data, that is not medical data, is far less protected and easy to access, says C. Pharmaceutical companies can leverage this data and optimize the distribution of drugs and vaccines. The application of Machine Learning in logistics, in the supply chain, is a very interesting use case according to him. We have already seen disruptions in the global supply chain and we have seen the problems of rolling out the vaccine and their distribution. By using Machine Learning and AI techniques in the supply chain, the pharmaceutical companies can improve their processes and achieve cheaper drug distribution and storage, cut costs and make the drugs cheaper and more accessible to the public. Since the data being used is not from patients, is not as protected and easy to access. Therefore, the development of Big Data technologies is much easier and accurate due to the large amounts of data. Another application of Big Data in the pharmaceutical industry is with drug development. Pharmaceutical companies can use Big Data technologies to predict the potential application of new drugs and chemical compounds. Insurance companies can also benefit from Big Data implementation for anomaly detection and prevent fraud from happening. For example, they can detect unusual procedures, or doctors and practices, that do not seem right compared to historical data. This is also a simpler use case according to C, because the insurance claim data is not as protected as patient data, so it can be used and accessed freely.

3.3.2 Barriers

One of the biggest barriers regarding Big Data in healthcare is data security and regulations. As technology advances, healthcare organizations put more and more layers of protection for the generated data. The cyber securities as well as the laws that are put in place to protect the data make it very hard to access. The generated patient data is being anonymized so it can be used further. There are two solutions on how to make the data more available for usage, says C. The first one is federated learning, which is to train the

machines locally. For this first movement, there is no longer a need to take physical data from someplace else, since the machines are being trained on data that is already available to the organization. The other movement is to use synthetic data, which is as close as possible to the original data, without containing too much private patient information. The second movement is not as easy as it sounds, but it is very promising since the synthetic data is anonymous, and anonymous data has no protection. While anonymous data cannot be traced back to the patient and can be used for further research, it is very difficult to generate anonymous medical data. According to C, in many areas in the healthcare sector, it is even impossible to have anonymized patient data. Each brain is unique, like a fingerprint. As it would be difficult to anonymize a fingerprint, the same goes for the brain. If the system takes the unique characteristics of a fingerprint away, there is nothing left to analyze and the same with the brain. The current method that is being used to anonymize is to take a brain scan and take away the face. But the problem is when you do a brain scan, you scan the face as well and you can simply use a 3D reconstruction and see the face. There are also some examples where the doctors need to see the face, so there is the question of how can the face be anonymized. Another example of the difficulties of anonymized data is a study by the University of Chicago Medical Center. The University wanted to make a study regarding the patients, to see when the patients show up or not and which disease they have so they teamed up with Google. They used anonymized EHRs and they even got the ethics approval for it. For the study, the researchers needed the date, time, and location of the patients' appointments to predict the outcome. However, by working with Google, they are potentially working with someone who can take this patient data, date, and time for each appointment and match it with their localization data and see exactly which patient missed or showed up for the appointment. There are cases like this, where the patients' data can fall into wrong hands. This is a challenge, according to C. As for Big Data acceptance from the patients' point of view, it is a very individual decision. Some differences and factors have an impact on this view, especially in different groups of people or in different countries. For example, C mentioned that the people in Germany are very reserved towards digitalization and new technology implementation, especially when the data is saved somewhere centrally. On the other hand, there are countries like Estonia or Lithuania where many aspects of peoples' everyday life are already digitalized. These countries are an example of successful Big Data usage in healthcare since the complete healthcare data is available for research. The Estonian Biobank is doing genetic research using data. The people living in Estonia, that are already exposed to Big Data usage in healthcare and are open to its implementation, will probably have a more positive view on the matter compared to Germany for example. Even in medicine, some fields are more prone to using Big Data than others. Radiology is a field that has been using technologies for decades, so radiologists are more open to trying new technologies and improving their job. Neurologists, on the other hand, are much less likely to adopt new technologies.

The technological capacity is different from country to country as well, according to C. For example in Germany, the EHR is just being implemented now, and even though they are

becoming available, the doctors are not that eager on using these technologies and not all of the patients' data is there. This is a very bad starting point for Big Data implementation. There might be only one thing in Germany, that is leading the pack in the area of digitalization and that is Giga, a digital health application. Using this application, doctors and therapeutics can prescribe treatments to patients. Currently, there are around 20 apps that are similar to this one, but the first law that was created in Germany for digital prescriptions was created for Giga. France had just implemented the same rule for digital prescriptions. According to C, this is probably the only case where Germany excels and is leading in digitalization. Other countries are much more advanced and with much bigger technological capacity. Big Data is an emerging field and it is emerging very fast. The lack of skilled workers can be felt worldwide; however, there is another issue when we are talking about Big Data in healthcare. The salaries that the skilled workers are offered by the healthcare sector are very low compared to the salaries that the big tech and IT companies offer. So why would the skilled labor work for the healthcare sector, if they can make much more money working in the same field for a tech company? There are exceptions, such as people who want to work in the healthcare sector. According to C, some people accept the position because in healthcare they have a more secure job, but the truth is, they would be paid three times more in a tech company. The systems need to be built where the data is, and in this case that are the hospitals and practices, but these organizations have problems acquiring the funds to hire the much-needed skilled labor to develop them. One solution for this issue might be if the government provides a company that will do this kind of service for hospitals, and the hospital is not bound to pay the salaries. Big Data unfortunately does not equal good data, since we can have large amounts of useless data that is still Big Data. This is the case especially in medicine since we can have a huge dataset that cannot be generalized and connected to all types of the same disease, multiple sclerosis, even though the data has been collected from 100.000 patients. C says that the data quality can certainly be a challenge if it is not the right type of data, but a huge useless dataset.

According to C, the ethical challenges are so many and the topic is so broad that it cannot be explained simply and with a few use cases. One example is data privacy and the ethical issues around balancing data privacy versus the availability of the data. Some value data privacy more over its availability for general usage and vice versa. Then is the ethical question regarding the heterogeneous data and whether the machines have been trained on diverse datasets or the results might be biased. If the data on which the machine has been trained does not include certain groups of people, are they discriminating against certain patients? C says that if we have several German insurance companies and we take their data to build a clinical decision support system to predict the stages of a certain disease, this can be a challenge. If all of the insurance companies are private, chances are, only wealthy people can afford those insurances, and the immigrants or certain minorities are less likely to have them. Based on the data of these insurances companies, how well does this support work for these groups of people? While it might not work for certain groups of

people, it would provide a huge benefit for 80-90% of the citizens, so this is not an easy decision to make. Some companies might even develop a second version of the system, using the money that they made from the first one. This can be one way of improving the decision support systems and training them on broader population so it is useful for everyone. According to C, this is a good start, because otherwise it simply cannot be financed, but other people disagree. In his opinion, it is way too costly to develop a tool that can work with the same accuracy on all minorities and subgroups, considering the time and resources that have to be put in it. The ethical challenges are a very complex field and with these few examples, we have just scratched the surface. C emphasized that currently the most benefits of Big Data implementation in healthcare are seen by people who are working in this field. There is certain hype around it and it is fairly easy to get the investment for project execution. The people that work in the field and are trying to build the solutions benefit the most from Big Data. In C's opinion, for the patient, currently, there is not a strong benefit, especially not on a scale. Some applications can be useful, but Big Data analytics in healthcare should not be about a few examples that have worked and this is not how people should perceive it. Big Data should work on scale and its benefits should be very visible from both the users/doctors and the patients, not only the companies that build the solutions.

3.4 Text Mining Analysis

The tweets regarding Big Data Analytics in Healthcare were posted between 28.9.2021 and 7.10.2021. After the data was extracted, each word and sentiment was separated into a different column, creating more columns with different words and their sentiment. For better analysis, all of the words were organized in one column, next to the tweet for which they were used. Next to the words, in a different column, the sentiment for the corresponding word was included, as shown in Table 1 below.

Table 1. Aspect-Based Sentiment Analysis- results

Text	**Aspect**	**Sentiment**
Digital technologies like faster wireless internet, smart sensors, Big Data, and Artificial Intelligence are transforming healthcare	Big Data	P
Big Data analytics for healthcare supported the rapid development of COVID-19 vaccines. Scientists can draw insight from analyzed data to develop advanced medications very quickly.	Big Data	P
The first big issue was that FDA was slow in getting an exam out, put out a bad test, and didn't share data	exam	N

Source: Twitter (2021).

To analyze the opinion of potential patients and the general population, tweets were extracted with Rapid Miner regarding Big Data Analytics in healthcare, and sentiment analysis was performed on the results. From the 115 tweets, 347 words were analyzed and given a sentiment. Sentiments varied from N+: very negative, N: negative, NEU: neutral, NONE: without sentiment, P: positive, and P+ very positive. Since for this analysis only negative, neutral, and positive opinions were needed, N+ with N were merged, NEU with NONE, and P with P+. The results with the same words were also merged, so the sum of total negative, positive or neutral sentiment for each word could be extracted. From the 347 words that were analyzed for sentiment analysis, 200 values were unique. The sentiment analysis focuses on words that were as well mentioned in the interviews. The values of the top 15 words that were mentioned the most in the mined tweets and were mentioned in the interviews as well are shown in Table 2 below.

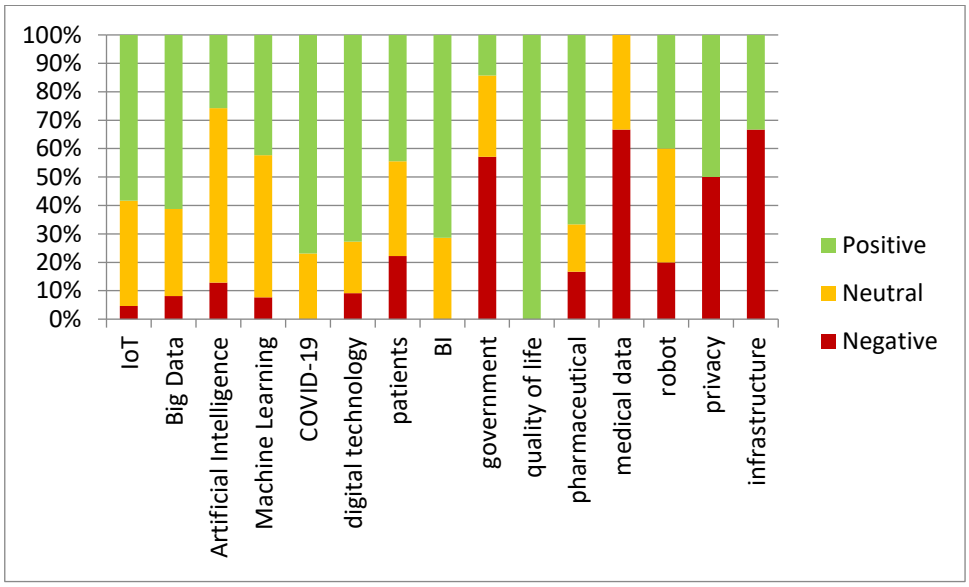
Table 2. Aspect-based sentiment analysis TOP 15 words

Aspects	**Negative**	**Neutral**	**Positive**	**Mentions**
IoT	5	40	63	108
Big Data	4	15	30	49
Artificial Intelligence	4	19	8	31
Machine Learning	2	13	11	26
COVID-19	0	3	10	13
digital technology	1	2	8	11
patients	2	3	4	9
BI	0	2	5	7
government	4	2	1	7
quality of life	0	0	6	6
pharmaceutical	1	1	4	6
medical data	4	2	0	6
robot	1	2	2	5
privacy	1	0	1	2
infrastructure	2	0	1	2

Source: Twitter (2021).

The tweets regarding the words mentioned in the interviews were collected for the period between 01.11.2021 and 10.11.2021. Text Mining and Sentiment analysis was performed on the words in the table and the results were summarized. IoT was mentioned the most in the tweets with 108 mentions, out of which 40 were in neutral connotation, 63 were positive and 5 were negative. The word Big Data was mentioned 49 times, 15 times in neutral, 30 times in positive and 4 times in negative context. The word AI was mentioned 31 times, 19 times in neutral context, 8 times in positive and 4 times in negative. Other words that were mentioned the most in the tweets were Machine Learning, COVID-19, digital technology, patients, BI, government, quality of life, pharmaceutical, medical data, robot, privacy and infrastructure.

Figure 5. Aspect-Based Sentiment Analysis TOP 15



Source: Twitter (2021).

The majority of the mentions, 154 mentions to be precise, were in a positive context. 104 mentions were in neutral context and 31 mentions of the top 15 words were in a negative context. Big Data was mentioned only 4 times with negative context and IoT, digital technology and quality of life were mentioned mostly in positive context. According to the sentiment analysis, there are some concerns regarding almost all of the analyzed words, especially government, pharmaceutical, medical data, privacy and infrastructure.

4 ANALYSIS AND DISCUSSION

According to some sources in the literature review, for successful Big Data implementation, three teams need to be included in the process: developers, medical personnel and Medical Informatics experts. The interviews were conducted with professionals from these three fields. Some of the most important current uses of Big Data in healthcare that the interviewees mentioned are the usage of Big Data in the pharmaceutical industry. With Big Data usage, the pharmaceutical companies can speed up the clinical runs and analyze the market before deciding what kind of medication the patients need most. This usage is very important, since it reduces the costs and time needed for drug development and production, and can eventually lead to faster and cheaper drug accessibility on the market for the patients. Another very important usage is constant monitoring of the vitals of the patients and fast assistance if needed, which can save the lives of the patients. Currently, Big Data is being used for regulating the storage temperature for the COVID vaccines, which is of huge importance, since variations of the temperature (too low or too high) can have an impact on vaccine efficacy. Big Data can be helpful with predicting disease development as well and provide the ability for early

treatments or disease prevention. During the interviews, the opportunities and barriers of Big Data Analytics in healthcare were discussed, and the interviewees provided their opinion based on their experience and described practical examples of Big Data usage. Moreover, public opinion regarding Big Data implementation in healthcare is very important and needs to be considered. Many people state their opinion on social media and interact with other users debating the pros and cons of many things. With Rapid Miner, the tweets regarding Big Data in healthcare were extracted and sentiment analysis was performed to get information about the public opinion on Big Data analytics in the healthcare sector. There were over 100 tweets and over 300 words that went through the sentiment analysis process and focuses on the words used in a positive or negative context and analyzed the tweet to see why people felt positive or negative towards certain aspects. In total, there were 86 unique words used in a positive context in the tweets and 14 words used in a negative context. The rest 100 words on which sentiment analysis was done were written in a neutral context or without any sentiment.

4.1 Opportunities

Most disruptive benefits of Big Data Analytics in healthcare

According to previous findings and literature review, with the help of Big Data, the healthcare organizations would be able to predict epidemics, help cure diseases, improve life quality and reduce the number of deaths that could have been prevented (Kankanhalli, Hahn, Tan, & Gao, 2016). Big Data technologies can be used as clinical decision support systems, designed to assist the medical personnel for easier, more accurate, and faster disease detection and treatment prescription. Because of the overwhelming amount of data and patients the medical personnel have in a single day, human error is highly probable to happen during disease identification or treatment prescription. With the help of clinical decision support, the medical personnel will have access to the output that these systems create which is optimized and contains all the relevant data. These systems might be able to highlight clinical patterns that were previously not considered and effectively analyze routinely collected data for insights about patients' health (Dagliati et al, 2018). The results regarding the most disruptive opportunities of Big Data in healthcare are similar to the results gathered from the literature review analysis. According to the interviewees, the most disruptive opportunities of Big Data analytics in healthcare are its usage as clinical decision support, predictive disease diagnostics, and personalized medicine. The huge data generated in the healthcare sector can be used to make disease diagnostics easier and more accurate and therefore, help the patients. Another potential is using Big Data to improve the life of patients with chronic illnesses, such as diabetes. By using real-time data regarding glucose levels and other related information, the doctors can make changes to the patient's diet and lifestyle that can lead to fewer hospitalizations, less serious disease symptoms, and therefore, life quality improvement. Big Data can be used for predictions of diseases and their early treatments or even prevention. The more data is gathered, the more

predictions can be made. Another disruptive opportunity of Big Data in healthcare is personalized or precision medicine approaches for the patients. Even though the treatments for certain diseases are almost the same for all the patients, with the help of Big Data, doctors can tailor the therapy to be better suitable for each patient and therefore more effective. Moreover, Big Data can be used for improving performance and increasing resource usage efficiency. These technologies have not fulfilled their full potential yet and there are still many opportunities for their implementation in healthcare that we are not even aware of. In the tweets and sentiment analysis, *Big Data* is used in a positive context in a tweet saying that healthcare is making significant progress due to Big Data technologies, systems biology, and blood analysis. Big Data is used in a positive context several times, in tweets saying that Big Data contributed a great deal in COVID-19 vaccine development, the growth of Big Data usage in healthcare in North America but also its impact of growth on other companies and industries. Moreover, positive tweets about Big Data regarding its ability to address even the most challenging aspects of healthcare, how the pandemic and COVID-19 had impacted the growth of Big Data and AI in healthcare, the transformation in healthcare that is happening because of Big Data technologies, and the many usages of Big Data in healthcare.

Improving the patient's journey

According to previously done studies, Big Data plays a huge role in optimizing health processes and providing guidance for better and more effective resource usage. By removing unnecessary tests and examinations, both patients and medical personnel save time and money (Shilo, Rossman, & Segal, 2020). The findings of the interviews are confirming the results from the literature review analysis. With Big Data usage in healthcare, the whole healthcare process can be faster and, when diagnosing diseases, the results will have a more accurate ratio. Just with the usage of smart watches for example, by extracting the data from a device of patient, the healthcare organizations can have much more information to base their diagnosis on, and connect it to the results from other healthcare devices. This only leads to faster and more accurate disease diagnosis, more effective treatments, or even prevention of certain diseases. However, according to one interviewee, the healthcare flow is very complex and the data is not produced that fast. When talking about patients' journeys, we are talking about all procedures, check-ups, and examinations or treatments that take place in a healthcare organization from the moment the patient has been admitted. Patient data is generated once the patient checks in or when certain examinations are being done. This data can be used to improve the disease diagnosing or treatments, but implementing it so it can improve the patient's journey for day-to-day check-ups is difficult. In the tweets and the sentiment analysis, *Biomedical, Machine Learning,* and *BI* are used in a positive context to describe the possibilities of Big Data technologies. *Internet of Things* and *digital technology* is used in a positive context as well, about working together with Big Data in improving healthcare, and how much it has advanced in the past years due to the pandemic. *Data mining* is used in a positive context

in a tweet that talks about the advantages of Data Mining in healthcare. One tweet praises *innovators* saying in a positive context that the innovators are the ones driving Big Data analytics and AI in healthcare.

Personalized medicine and modeling disease progression

With Big Data, healthcare providers can get deep knowledge about their patients through similarities and connections. This provides an opportunity to develop personalized healthcare for each patient, and proactively manage and treat diseases (Chawla & Davis, 2013). By knowing the patients and their medical data history, the medical history of their families, and combining this information about the patient's lifestyle and environment, many diseases can be prevented and the quality of life can be improved (Kim, 2018). Big Data can especially help patients with chronic diseases such as heart disease, obesity, cancer, or diabetes. These diseases are the number one reason for early death in many countries and are known for slowly progressing over a long period. If the risk factors of these diseases are monitored, the diseases can be controlled at an early stage, and mortality or severe symptoms can be avoided. With the help of Big Data and modeling disease progression early detection of the above chronic diseases is possible and many patients' lives can be saved (Zhoua, et al., 2020). The interview results regarding personalized medicine were similar to the results acquired through literature analysis. Personalized medicine is one of the best advantages of Big Data usage in healthcare according to the interviews. By collecting the patients' data, medical data, their lifestyle, and environment as well as their genetics, we can combine them with live data from patients' medical devices and analyze multiple factors to predict their impact on the patients' health. These analyses can help predict diseases and start early treatments or prevent them together. Early disease treatments or their prevention can lead to reduced hospitalization, lower costs for the patient as well for the hospital. By monitoring the patients and knowing their backgrounds and medical history, the doctors can get ahead of the disease and treat it before it progresses. Even with small changes in a patient's diet or lifestyle, big results can be achieved. Once these analyses are run on enough patients and enough data is being collected, Big Data can be used on a certain population sharing similar characteristics or living in a similar environment. Even though these predictions cannot be made for the far future, it can be predicted which patients are more likely to get the flu and advise to get the shot. Certain patients, such as seniors or chronically ill might react to the flu more seriously, so getting the flu shot early will reduce the risk of their hospitalization. Another possibility is an analysis of a group of patients that had led similar lifestyles, which can lead to predicting diabetes and creating personalized treatments for that specific subgroup. In the tweets and the sentiment analysis, *Quality of life* was used in a positive way to explain how Big Data can improve the quality of life eventually. Words like *strategy, plan, project, design, and market* were used in a positive context as well about the plans and opportunities of Big Data analytics in the healthcare sector.

Efficient resource usage

According to the results of the literature review, Big Data can be used for more effective resource usage, especially in the pharmaceutical industry. The pharmaceutical companies spend a huge number of resources on drug development and testing, and yet up to 90% of the tested new chemicals do not make it to the market. With Big Data usage, these costly clinical runs can be done automatically, which will save time and money for pharmaceutical companies (Shilo, Rossman, & Segal, 2020). The results from the interviews were similar to the literature review analysis and showed that Big Data can be used for more efficient resource usage. With fewer hospitalizations, the hospital beds are being freed for patients that need them or need to be monitored for a longer time. If the probability of certain disease such as diabetes is detected earlier, the patient can implement lifestyle changes and prevent the disease from ever happening. If some of the beds remain free, that means lower costs for the hospital and less usage of resources such as medication, hospital care, or meals. For the patients, this means fewer expenses for medical care. Preventing certain diseases or their early treatment can lead to more efficient resource usage with the medical personnel as well. This can eventually lead to lowering the hospital costs for the patients as well. With good and proper data, we can reduce tremendously the resources used in healthcare and at the same time, improve patient care as well. However, more efficient usage of resources for some healthcare organizations might mean letting some medical professionals go. If the workload of the medical personnel is lowered, this will not necessarily mean more time with the patients, but fewer doctors or nurses. Implementing clinical decision support in healthcare might lead to the same outcome of cutting off medical personnel. This is a way of the hospital's way of lowering costs, which is still efficient resource usage, just it might be not as people would expect.

Big Data Analytics in the pharmaceutical industry

Big Data plays a huge role in the efficient usage of resources in the pharmaceutical industry, according to previous findings, and it can help with easier target identification for drug testing as well. Target identification for patients suitable for a new drug testing is a very big part of the clinical runs. The traditional screening for suitable patients is a very long and costly process with many possible errors. This method is known to have thousands of failures per one successful drug candidate, which is why animals are used for drug testing. However, with the help of Big Data, pharmaceutical companies can find the right candidates with very high certainty of success. This may lead to faster and better drug development and cheaper drug availability for the patients (Shilo, Rossman, & Segal, 2020). According to the interviews, there are several other ways on how the pharmaceutical companies can benefit from Big Data implementation in the healthcare sector. Pharmaceutical companies can use Big Data to analyze certain markets, how big is it, how present the disease is, which are the potential customers, their needs, or the competitors. Based on these analyses, the pharmaceutical companies decide in which market to focus for developing medications. Moreover, Big Data can be used for clinical

runs by running an analysis on how the drug develops and how it would affect the patient. Big Data can be used for monitoring the patient 24 hours per day to detect any possible side effects and provide medical assistance if needed. The information of the patient monitoring and clinical runs can be used for drug improvement and faster availability on the market. These trials take a tremendous amount of time and resources for the pharmaceutical companies, and with Big Data usage the time needed and the costs can be significantly reduced. This means that the pharmaceutical companies will have more money and time to spend on the development and improvement of other researches and drugs. Big Data can lead to cheaper and faster clinical runs, which could eventually lead to faster drug introduction to the market for general usage. The distribution of the drugs and vaccines can be also optimized with Big Data. The usage of Machine Learning in logistics and supply chain is already being implemented and it provides cheaper drug distribution and storage. An example of this is using Machine Learning for maintaining the correct temperature for COVID-19 vaccines since variations in the temperature can have an impact on the vaccine efficacy. Cutting costs for research, development, testing, distribution, and storage can make the medications cheaper and more accessible to the public. In the tweets and sentiment analysis, *Artificial Intelligence* is used several times in a positive context regarding its usage in healthcare, how it changed since COVID-19, and how Big Data technologies and AI transforming healthcare and other sectors are. *COVID-19* and *vaccine* are used in a positive context as well, for reacting as a driver for Big Data improvements and implementation. These results show that the people are familiar with the benefits of Big Data and how it helped the pharmaceutical industry during the pandemic.

Fraud detection and prevention

According to previous findings, Big Data can be used in healthcare to detect suspicious records, data movements, or detecting insurance fraud. With Big Data usage in healthcare, the risk of scams happening can be significantly lowered and many frauds prevented (Shilo, Rossman, & Segal, 2020). Big Data can be used for fraud detection and prevention, according to the interviews as well. One example is detecting and preventing drug abuse. By tracking the refills on certain drugs, and the time passed between each refill, we can see which patients are refilling their medication more often than needed and suspect drug abuse. Many people abuse medications without even using their health insurance to cover them, because it is cheaper than buying illegal drugs. Another example is detecting insurance claim fraud. Many hospitals and clinics send claim forms to their patients, so the patients can use those forms to claim insurance from insurance companies; however, many abuse these forms. With Big Data usage it can be detected if certain records do not seem accurate and prevent insurance claim frauds from happening. By comparing historical data with the data of the insurance claim, such as procedures, doctors, or practices, it can be seen if there are some inaccuracies and suspicious claims.

4.2 Barriers

Most common barriers to Big Data in healthcare

According to the literature review analysis, the most common barriers to Big Data implementation in healthcare are its bad quality, inconsistency, and instability of the collected data. Moreover, there are barriers regarding laws and legislation of data ownership and accessibility as well as ethical issues for data privacy and protection. Some technological challenges need to be considered when implementing Big Data in healthcare as well (Lee & Yoon, 2017). The results from the interviews show that these barriers are indeed present in the field. However, Big Data is a huge and not yet fully explored field, it is still under the testing phase, so not all of the barriers are known or explored. The most common barriers when it comes to data are its storage and processing. The data is being stored in a cloud and being used by different users. Everyone is processing the data or reading a database in a particular cloud at the same time. This can lead to performance degradation if the cloud is not optimized. The processing of the data is significantly slower if the data is being stored in a traditional database or if the cloud storage is not optimized. The quality of the data is an issue as well since there are cases where the data has been entered manually and needs to be re-entered in the system digitally. This is where many errors can happen. Other most common barriers to Big Data in healthcare are governmental regulations, legal regulations, and data sharing. These regulations vary from country to country and the patient's data ownership is regulated differently everywhere. In some countries, it is next to impossible to access the patient's data due to regulations, while in others it is fairly simple. Moreover, a huge challenge is the mindset of the people and convincing them about the benefits of Big Data usage in healthcare. Many patients do not understand the opportunities of Big Data implementation in healthcare and oppose its implementation. Even in medicine, some fields are more open to using Big Data than others, and the differences are even bigger when comparing different countries. In the tweets and the sentiment analysis, *Big Data* was used in a negative context only a couple of times, regarding the expenses connected to the implementation process, the implementation speed with which is being implemented in healthcare, and how are these tools addressing the inefficiencies of the global healthcare system.

Patient's acceptance of Big Data in healthcare

Studies show that the patients might feel reluctant regarding Big Data implementation in healthcare and Machine Learning tools making decisions about their health. The lack of knowledge on these topics makes it even more difficult for people to adopt the idea of AI and Machine Learning usage for healthcare processes. Data usage and transparency are other reasons why patients are not eager for their implementation (Janke, Overbeek, Kocher, & Levy, 2016). Patients are not usually aware of how their data has been stored and used, and once they are familiar, they express alarm. This affects the doctor-patient relationship and the trust that the patients have for their health providers and makes it

harder for them to trust their lives to the medical personnel (Brill, Moss, & Prater, 2019). According to the interviews, the acceptance of Big Data in healthcare is a very subjective decision. Some differences and factors have an impact on acceptance, especially in different groups of people or in different countries. There are countries where people are very reserved towards digitalization and new technology implementation, especially when the data is saved somewhere centrally. On the other hand, there are countries where already everything is digitalized. These countries are an example of successful Big Data usage in healthcare since the complete healthcare data is available for research. Some fields in medicine that have been using technology more are also more prone to Big Data implementation compared to some medical fields, where the technology has not been implemented and used that much. In the tweets and the sentiment analysis, *disaster,* and *media* were used in a negative context in a tweet regarding journalism, saying that it is a disaster that the journalists are siding with healthcare instead of asking questions and looking at the data and stories. This shows that the people are concerned about Big Data implementation in healthcare. The second interviewee mentioned that the patients are usually opposing Big Data implementation because of a lack of transparency regarding how their data is being used: "*Once the patients are aware of how their data is used and for what they become more open to sharing it. The opposition to big data usage is happening mostly because of lack of transparency and lack of education, which we need to change."* To overcome this barrier, professionals need to talk to the patients and allow them to learn and convince them that their data is valuable. It is important to make the system more transparent and educate patients on how their data is going to be used and for what purpose. The experts need to educate people regarding Big Data because most of the issues that they have with Big Data are because of fear of how their data is being used. Once the patients are aware and have more knowledge in this subject, they might be more open to its implementation in healthcare and realize the potential benefits from its usage.

Data ownership and security

According to the literature review analysis, there are different rules regarding data ownership in different countries. Sometimes it is not clear who the owner of the data is, whether the patient or the medical institution, and this might lead to legal problems. Another barrier is the security of the data and the protection of cyber-attacks. Since the data is being distributed to different places, the vulnerability of it is much higher (Raguseo, 2018). Data ownership is regulated differently in every country and the ownership regulations are different everywhere, according to the interviews as well. In some countries, such as India, getting the data is very simple since the data belongs to the hospitals or the pharmacy where is being generated. The governmental regulations still protect the patient's medical data, since the healthcare organizations are not allowed to share the personal details of the data with which the patient might be identified. Healthcare organizations are allowed to reveal patients' IDs or some doctors' information, but they cannot share demographic or private patients' information. In some countries, it is almost

impossible to access the patient's data due to governmental regulations and privacy. This makes it very difficult for Big Data implementation in healthcare since the usage of Big Data depends on the access of patient data. In some countries, this topic is not even regulated yet. Some professionals say that the data belongs to the patient, some say that it belongs to the hospital where it is being generated and there are still no concrete decisions on that question. In the tweets and the sentiment analysis, he words *device* and *researches* are used in a negative context about collecting data from wearable devices from the people and giving it to the researchers. *Blockchain* is also used in a negative context in a tweet saying that blockchain can't help with the addition of security to data. However, there were also positive tweets on the topic of data security and usage. *Medical data* is used in a positive context in a tweet regarding AI and anonymous data usage of medical data. P*rivacy* was also used in a positive context because of the usage of patients' medical data to help others. Another tweet about data security uses the words *End2End-encryption* and *FHIR* in a positive context saying that FHIR (Fast Healthcare Interoperability Resources) now allows secured patient data downloads and migration with end-to-end encryption. The owner of the data should be the patient, and this is a challenge regarding data usage. Anonymized data might be a good solution for a big part of this problem, for the patients' data that does not need to reveal any personal information. If the data is anonymized enough, and the patient cannot be identified through it, then the data ownership and security is not that big of a challenge, since the confidentiality is not broken and personal information from the patients is not used. However, for the data that cannot be anonymized, the patients should be asked to give their consent on data sharing since medical data is private information and if it is in the wrong hands it can be used against the patients.

Data quality

According to previously done studies, the data quality varies depending on the source. The data collected from EHR does not have the same quality as the data that is being continuously collected from medical or wearable devices. The collected data from different devices often contain missing values or errors, that might lead to Big Data techniques and create patterns where there are not (Janke, Overbeek, Kocher, & Levy, 2016). Big Data unfortunately does not equate to good data, since there can be large amounts of useless data that is still Big Data, according to the interviews. This is the case especially in medicine because there can be a huge dataset that cannot be generalized and connected to all types of the same disease, such as multiple sclerosis, although the data has been collected from 100.000 patients. The data quality can certainly be a challenge if it is not the right type of data, but a big useless dataset. There are many because the data sometimes is being entered manually and as free text. For the field age, there might be different values, such as date of birth, a range, or the age in text. There are some empty fields and missing values of the gathered data as well. These need to be corrected or calculated manually, incurring many mistakes that can happen. If Machine Learning solutions have been trained

on small or narrow data, it can affect the overall effectiveness and accuracy. Once these models are implemented on a different hospital or different hardware, their performance plummets. The lack of data heterogeneity can be an issue with the outcome as well. If the Machine Learning solutions were not trained on enough data or different data, it might affect the result accuracy. The problem with lack of data heterogeneity arises with the difficulty to access the patient's data, especially in countries where the data is not being digitalized, centralized, and free for usage. Sometimes it is difficult to access the data because every healthcare organization uses a different system. The implementation of Big Data itself becomes an issue in these countries, whilst the countries, which have the majority of the data digitalized, centralized, and available for usage, do not face this problem as much. In the tweets and the sentiment analysis, one word which was used in a negative context was *Italian* in a tweet that talks about the inaccuracy of medical data of health patients and their immune systems. The world is slowly moving towards digitalization of almost every aspect of our lives, and the healthcare sector is not an exception. The sooner countries start implementing EHR and train medical personnel on its usage; the sooner they can adapt to this global digitalization and start seeing the benefits from it. While the culture in some countries is a big barrier regarding digitalization, this is, inevitable and will change at a certain point in time. The governments should proactively promote it and educate people as much as possible. With the digitalization of the healthcare system, data quality will improve significantly. Entering free text or wrong values will be more difficult in a digital form, which will lead to fewer missing values and errors compared to now

Technological capacity and compatibility

The technological capacity of the healthcare sector is an important factor when deciding whether or not to adopt Big Data. The already existing technologies need to be compatible with the new ones for smooth operations. If the new technologies are too advanced compared to the engineering capacity, their acceptance might take longer and cost more (Hall & Khan, 2003). Based on the experience and opinion of the interviewees, the compatibility of Big Data with the technology that is already being used in healthcare currently is not an issue. As for in any other industry, the essence of the Big Data technologies is almost the same, for example, how data is collected, processed, and analyzed. All of the technologies such as CT scan, MRI, x-ray scanner, and ECG machines are already compatible with Big Data tools and it is made that way on purpose. The data that is collected from the medical devices are usually being stored in a cloud, from where it can be accessed for Big Data analysis. Even in examples where Big Data is saved in a traditional database, it is not a big challenge but is stored in a cloud and this is much more convenient. However, the technological capacity is different from country to country. In Germany, EHRs are just being implemented now and the medical personnel are not very eager to use them. This is a huge barrier for Big Data technologies since digitalized data is crucial for their implementation. On the other hand, countries such as Israel or Baltic

countries, that have most of their data digitalized and centralized, so technological capacity is not an issue. In the tweets and the sentiment analysis, *Collaboration* was used in a negative context addressing the inefficiencies of the healthcare system.

IT skills of the medical personnel

According to the results of the literature review, one barrier to Big Data implementation in healthcare is the insufficient IT knowledge of the medical personnel. The medical personnel usually do not have any IT background and their IT knowledge is minimal. This training will take a toll on the financial resources and the time of the medical personnel. Another issue regarding the IT skills of the doctors is their reluctance to change and adapt to new technologies (Raguseo, 2018). However, the results of the interviews do not indicate this example as a barrier. The second interviewee mentioned: *"From the point of view of the medical personnel, they do not need a strong IT background or deep understanding of the big data technologies. Basic understanding of the processes is enough for their usage.*" While the medical personnel need some basic IT knowledge and understanding of the processes, they do not need a strong IT background or profound knowledge of Big Data, so their IT skills do not present a barrier. The willingness to accept the Big Data technologies in healthcare and the ability to use them does not depend on the age of the medical personnel. The older and more experienced medical professionals are eager to learn and adapt to new technologies because they feel that it adds value to them and their knowledge. The medical personnel might not need extensive knowledge in IT, but some training and education should still take place when implementing Big Data. This will not only help the healthcare providers understand the Big Data technologies but also allow them to try using them under the supervision and gain confidence. Some medical professionals are concerned about Big Data implementation in healthcare. In the tweets and the sentiment analysis, the word *radiologist* is used in a negative context, in a tweet from a radiologist's point of view saying that today's data flood is toxic to the patient.

Lack of skilled workers

The availability of skilled workers in the market is crucial for Big Data analytics in the healthcare sector. If the available workers do not possess the needed skills, it will take them some time to acquire them, which will incur expenses. If the availability of skilled labor is limited, the adoption of Big Data in healthcare will last longer and this can be considered a barrier (Hall & Khan, 2003). According to the interviews, Big Data is an emerging field and it is emerging very fast and the lack of skilled workers can be felt worldwide. There is a huge deficit of skilled workers in the Big Data field, and the demand for this kind of labor is much higher than the supply. Every day, there are more and more new technologies, so even skilled and experienced workers need to learn constantly. Many companies want to implement Big Data in their processes, so the demand increases constantly. There is another issue regarding the skilled workers in Big Data when talking about healthcare. The salaries that these workers are being offered in the healthcare sector

are three times lower compared to the salaries that the big tech and IT companies offer. While some people want to work in the healthcare sector or like the idea of having a more secure job, this is not the case with everyone. Taking into consideration the low supply of skilled workers worldwide and the lower salaries in the healthcare sector, the lack of skilled labor is a big challenge when implementing Big Data. The systems need to be built where the data is, and in this case that are the hospitals and practices, but these organizations have problems acquiring the funds to hire the much-needed skilled labor to develop them. Big Data is slowly starting to take place in many companies and it won't be long until it starts taking place in our daily lives. This field needs to be promoted, so more and more new students enroll in programs that teach it. The government could help promote the field by offering scholarships to the students and the companies could offer internships. To achieve a broader audience and target the potential future experts in this field, Big Data technologies should be taught in high school programs. Although not as extensively, even just learning about their usage and importance, might lead students to get interested and consider studying and later working in this field. This might lead to more students and later professionals in the field of Big Data, improving and implementing these technologies.

Required skills for Big Data implementation in healthcare

Big Data technologies are usually written in programming languages such as Python and R, therefore a good knowledge of IT and coding is very important for the development of these solutions. However, most IT experts and coders do not have a background in medicine and do not understand the terms and correlations of different factors. Some fields of study do offer knowledge in both IT and medicine, such as Bioinformatics or Medical Informatics. For a successful implementation, several skills need to be present in the team. Knowledge in IT and coding skills are needed, to develop the Big Data technologies, knowledge in medicine and medical terms, and if needed someone in the field of Bioinformatics or Medical Informatics that knows both, to work as a bridge between the first two (Viceconti, Hunter, & Hose, 2015). As mentioned before, for the development of Big Data technologies, three teams need to be included. The development team needs to be included actively because they understand the IT capacity and have the knowledge to develop the solutions. The doctors need to be consulted regarding the needs of the patients and need to set the requirements that they need from the Big Data technologies. Later on, when the product is developed, they are testing it and report back if there are any needed changes and corrections. A Medical Informatics team needs to be included as well since they understand the developers, the IT possibilities and capacities as well as the medical personnel, medical terms and patients need and correlations. The Medical Informatics team works as an intermediate between the developers and the medical personnel and makes sure that they understand each other and what needs to be done. These three teams need to be included for a successful Big Data implementation for healthcare, but of course, other teams and departments play an important role as well. The sales team and the marketing

team from an IT company are important for finding projects, presenting them to the healthcare organization, and selling them. Once the project is sold, the development team, the Medical Informatics team, and the team of medical personnel from the healthcare institution are relevant in the process of Big Data implementation. To make it work, everyone has a specific role to play and everyone needs to be integrated.

Requirements for Big Data implementation in healthcare

The most important requirement for Big Data implementation in healthcare is to have the right skilled workers in the team. Needed skills are IT knowledge and coding skills, as well as medical knowledge and patients' knowledge (Viceconti, Hunter, & Hose, 2015). Infrastructure and technological capacity are also important for Big Data implementation (Hall & Khan, 2003). The ethical requirements include technical precision of data and accurate statistical performance. Data transparency is necessary for using the data and analyzing it, however, too much transparency might threaten data privacy. Moreover, enough heterogeneous data needs to be included in the training, so the Machine Learning solutions can make accurate predictions about the minorities as well. Laws and regulations need to be standardized regarding data usage and ownership and patients should be asked about data access (Leon-Sanz, 2019). According to the interviews, many requirements and preparations need to be done before Big Data implementation in healthcare. It needs to be decided where the data will be stored, and that is usually a cloud. To access the data, there are some protections layers that the person needs to go through. Only an authorized person who is given the key can access the data. These protective layers make sure that anyone who is not an employee and not authorized to use the data cannot access it and, in that way prevent possible data misusage. Another requirement is a complete data encryption and pseudonymization. These encryptions are set in place as an additional protective layer if someone somehow manages to get access to the data. Without the decryption key, the data doesn't make sense to the person who retrieved it and it cannot be used. As for the development of Big Data technologies, the most important requirement is to understand the use case. The development team needs to understand each use case before they start developing, and map the entire patients' journey and every possible outcome. This is a very important part of the development, so the developers can understand the entire process and create the design and determine the key points.

Expensive implementation

According to literature review results, healthcare organizations need to be ensured that the costs of the implementation will be covered by future income made from these technologies. The bigger and more likely the probability is for future demand and stability, the more likely the hospitals are to implement new technologies and inventions such as Big Data analytics (Hall & Khan, 2003). While the implementation of Big Data is quite an expensive project, the opinion regarding this possible barrier was divided. One interviewee stated that healthcare organizations are aware of the benefits that these technologies bring

to the hospital and patients, and when these benefits are compared to the costs, they see that it is completely worth the money and funding. Another opinion in the interviews was that the healthcare system itself is not sustainable and the hospitals are cutting costs so they can continue operating and serving the patients. Even with Big Data implementation, the most probable outcome for the hospitals would be to cut down personnel instead of using Big Data for spending more time with each patient for example. The government should be more invested in funding such projects and helping the hospitals set up Big Data Analytics and start using it. This will take-off of the financial burden of the implementation, so the healthcare organizations can still keep their employees and lower the risk of letting medical personnel go. In many hospitals being understaffed is also an issue, so Big Data implementation will help solve it. Since the usage of Big Data will help healthcare organizations to use their resources more efficiently and save money, the Big Data implementation will pay off. For the initial investment, besides the governmental help, the hospitals could use bank loans or investments, which in the long run will pay off.

Relying only on Big Data vs Human supervision

The complexity of the healthcare flow and the other factors contributing to the health of patients makes it difficult for technologies to diagnose certain diseases. Big Data technologies cannot yet be implemented fully and in every process in healthcare, due to the complexity of the care processes. While the healthcare sector can benefit greatly from using Big Data, there will always be a need for human supervision and consultation (Wang, 2019). The results from the interviews indicated similar results that the healthcare flow is very complex and in most cases, we need a human to supervise and intervene. Big Data is built like a machine and we cannot completely rely on a machine alone. The algorithms used for building Machine Learning are similar to black-box processes and the medical personnel never really know what is happening inside and how the output is going to be. This is why human intervention is always needed to check and verify the output of the system. Even though Big Data technologies require human supervision, they still save time and effort for the medical personnel, since they do not have to do the process by themselves by manually entering and reading information. While Big Data technologies read an image or a graph and produce an output on what the disease might be, a doctor needs to verify that. A mix of medical personnel and AI solutions would be the best solution and probably the most accepted one. While machines can make errors, due to small or narrow data training, this is the case with human resources as well. The healthcare providers often spend many hours and work different shifts during the day, which can lead to sleep deprivation, stress and that can lead to human error. With the help of AI and Machine Learning, the medical personnel will get the support that they need for easier data analysis and faster and more accurate disease diagnostics. After each AI and Machine Learning outcome, the results need to be checked by a doctor to see if any errors might have happened. Using these machines as clinical decision support would improve the flow,

accuracy, and quality of healthcare. Big Data could decrease the workload of the doctors, and the working hours which will prevent sleep deprivation.

Big Data in developing countries

The implementation of new technologies, such as Big Data is especially difficult in developing countries. These economies struggle with poor policies and a lack of infrastructure investment, which makes it difficult for healthcare organizations to adopt Big Data. Moreover, the lack of skilled workers is a much bigger barrier compared to the developed countries. Not only that the skilled workers are underpaid, but also the supply is very limited due to lack of education. Developing countries do not have many graduates from the IT field, despite the awareness of the need and importance of such skills. Many universities have limited resources, such as access to the internet and IT technology. An even bigger issue is the lack of human resources that can pass over the knowledge in this field (Yu, Wu, Yu, & Xiao, 2006). There are many differences between Big Data implementation in developed economies compared to developing countries, according to the interviews. The execution of such a project is much easier in developed countries due to many reasons. In developed countries, many companies are moving into Big Data and implementing it in their processes, since they have the resources and technological capacity to do so. In developing countries, this is not the case because they lack the expertise, resources, not enough collected data and data clarity. The lack of skilled workers in this field is much bigger in developing countries, due to a lack of proper education and lack of resources. The implementation process is much lengthier and complicated due to these factors. In some developing countries, demanding or collecting data is already a barrier. The infrastructure is another significant barrier in developing countries. The developed countries have already invested in infrastructure and on the other hand, developing countries are trying to accelerate the process and invest now when they have realized the potential and value of Big Data. In the tweets and the sentiment analysis, the word *Africa* is used in a positive context saying that the countries advocate health data as a critical element to meet sustainable development goals and the companies that are servicing healthcare services must come up with strategies of how to utilize the data. Although developing countries are trying to catch up with developed countries by collecting Big Data, investing in infrastructure, and reframing the legislation about data usage, they are falling far behind and there is much work to be done. Developed countries need to try and help the developing countries with IT infrastructure, through financial aid as well as human resources and expertise. Many programs for studying abroad and helping foreign students are already happening, but they need to be promoted more, and students are encouraged to take a chance of studying abroad especially in the IT and Big Data fields. These young graduates can help with the education of new students that will continue working and improving the conditions of Big Data implementation in healthcare in developing countries.

Ethical challenges

Some of the most common ethical challenges revolve around data quality, ownership and security. The data needs to be cleaned before usage otherwise it can lead to unwanted outcomes such as wrong results from disease prediction or disease progress modeling. The data needs to be transparent for Big Data tools to use for analysis; however, transparent data threatens the privacy of patients' medical data and the right to confidentiality. Additionally, cyber security attacks have a significant impact on patient health due to data misuse (Leon-Sanz, 2019). The interviewees say that ethical challenges are the focus of current debates. Some doctors disagree with Big Data usage because they do not understand how these models work or how they are producing the outcome. The doctors do not know what happens inside the Machine Learning solutions or how they derive solutions, and they are conflicted about whether or not to listen to them. Data privacy and the ethical challenges on this topic are other barriers. Many governments are struggling to balance data privacy to protect the patient versus the data available for further researches and advancement. Some cultures value data privacy more over its availability for general usage and others do not. Then is the ethical question regarding the heterogeneous data and whether the machines have been trained on diverse datasets or the results might be biased. While training Machine Learning solutions, it is very difficult to develop the algorithms for every subgroup and minority of the population, therefore increasing inaccurate disease diagnoses. On the other hand, they can be helpful for hundreds of thousands of other individuals. A significant challenge is how can these solutions be trained and improved if they are not implemented and able to collect more data from a more diverse population. The ethical challenges are many and it is a very complex field that needs to be discussed among professionals in the field, country leaders, including the general public as well. In the tweets and the sentiment analysis, the people expressed their concerns regarding the patients' data collection and using it for profit. The words *Facebook* and *America* are mentioned in a negative context in a tweet about the collection of data from Facebook in America and using it against the population in everything from business to healthcare.

Using AI and Machine Learning in healthcare

As for the usage of AI and Machine Learning in healthcare, the opinions in the interviews were different, considering their different backgrounds and professions. From a developers' point of view, Big Data technologies are quite advanced and from that point of view, we will be able to implement AI and Machine Learning solutions in almost any aspect of healthcare in 4-5 years. Before the technology is advanced enough, it will be difficult to do so due to other factors. The problems arise when the collected data is not clean enough and the data quality is bad. This happens due to poor data digitalization, centralization, and availability. If the Machine Learning solutions are made on bad quality data, the outcomes can lead to disasters. Another opinion is that this field is very young in its implementation and that we have only covered 15-20% of the Big Data potentials with a long way to heavy usage of Big Data in healthcare. This is not only from a technological point of view but also considering other factors, such as public acceptance, laws and

regulations, and ethical challenges. These factors make Big Data implementation more difficult, and because of these barriers, this field is still very young and we are still building the foundation.

Who benefits most?

There are many benefits from the implementation of Big Data in the healthcare sector. According to the interviews, several parties benefit the most from their usage. Pharmaceutical companies can obtain various benefits from Big Data analytics. Pharmaceuticals get more work done in less time and with fewer resources used. Then, the patients will be getting the products and services faster and for less money. The patients will also benefit from better healthcare services, due to the reduced workload of the healthcare workers and of course, better life quality due to diseases prevention and more effective treatments. The hospitals and other healthcare institutions will benefit, due to workload reduction, more efficient resource usage, and lower costs. And last, the data-savvy companies and the professionals that work in the Big Data field, which retrieve the medical data and convert it to information that can be useful for healthcare organizations. However, while the patient's welfare should be at the core of Big Data implementation and the patients should be the ones benefiting the most, this is not happening on a full scale. The interviewees mentioned that the pharmaceutical companies and the people working in the field are currently benefiting the most from Big Data analytics implementation and usage in healthcare.

In the literature review, the adoption factors for IT are analyzed and information is provided about what people find most important when making decisions on whether or not to start using new technologies. Moreover, the broad topic of Medical Informatics, its history, and its importance are explained. Furthermore, information about how information technologies are used in the medical field nowadays is provided. Big Data, enabling technologies and skills are trending topics overall, but especially in the healthcare industry. There are many examples of how can AI, Machine Learning, and Data Mining techniques be implemented in healthcare and the benefits they entail. The literature review served as a base for the research, for structuring the interviews, and the Text Mining analysis. To gather opinions on Big Data analytics in the healthcare sector from experienced professionals in the field, interviews were conducted with professionals from the three teams required for Big Data implementation in healthcare: a developer and consultant, a Medical Informatics expert, and a medical doctor and researcher in the field. The results of the interview explain in detail how are Big Data technologies used in healthcare, what are the most disruptive opportunities that these technologies can offer, and the barriers that we are facing with their implementations and usage. The main and most important contribution of the thesis is defining the most important opportunities and barriers of Big Data implementation in healthcare through practical examples and different experiences in the field. With the comparison of the literature review and the interviews, the differences and similarities of the results were pointed out, showing an accurate estimate of the current

state in this field. The findings and the contribution of the thesis, the opportunities and barriers are presented in the Tables 3 and 4 below, respectively.

Table 3. Thesis contribution- Big Data analytics in healthcare: Opportunities

Big Data analytics in healthcare: Opportunities	**Literature review**	**Interview analysis**	**Text Mining analysis**
Disease prediction and diagnostics	Disease prediction, diagnostics, modelling progression and cataloguing	Disease prediction, disease diagnostics, disease modelling progression	Helping healthcare with systems biology and blood analysis
Clinical decision support	Assisting medical personnel for faster and more accurate disease detection and treatment prescription	Assisting medical personnel for faster and more accurate disease detection	Helping the healthcare sector and the medical personnel during the pandemic
Big Data Analytics in the pharmaceutical industry	Finding suitable patients for drug testing	Monitoring test subjects during clinical trials, maintaining storage conditions, analyzing markets and potential patients	COVID-19 acting as a driver for improving AI implementation in healthcare and helping with vaccine development and distribution
Fraud detection and prevention	Detecting strange movement of records	Detecting and preventing insurance fraud and drug abuse	/
Improving patients quality of life	Early possible disease diagnosing and altering patients' habits for disease and hospitalization prevention	Early possible disease diagnosing and altering patients' habits for disease and hospitalization prevention	Early disease diagnosing and eventually improving patients' life
Personalized medicine	Developing personalized medicine tailored to each patient	Developing personalized medicine tailored to groups of patients	Many opportunities lying ahead for Big Data in healthcare and helping patients
Healthcare process improvement	Improving patients' journey, efficient resource usage	Efficient resource usage	Improving healthcare processes

Source: Own work.

Table 4. Thesis contribution- Big Data analytics in healthcare: Barriers

Big Data analytics in healthcare: Barriers	Literature review	Interview analysis	Text Mining analysis
Laws and regulations	Not regulated in many countries, not standardized laws	Not regulated in many countries, not standardized laws	Concerns regarding data collections from patients' devices and giving it to researchers
Data related challenges	Data ownership, data availability, data security and data quality issues	Data ownership, data availability and data quality issues	Inaccuracy of medical data of patients' health and their immune systems
Skilled labour	Lack of skilled workers in the Big Data field, lack of IT skills of medical personnel	Lack of skilled workers in the Big Data field	Concerns regarding the doctor's negative opinion regarding Big Data usage in healthcare
Technological capacity and compatibility	Expensive implementation, lack of compatibility with already existing devices	Lack of technological capacity, slow healthcare sector digitalization, expensive implementation	Public concern regarding expenses of Big Data implementation, and the inefficiencies of the healthcare system
Ethical challenges	Relying only on Big Data, questionable data accuracy, threatening patient's privacy	Questionable data accuracy, threatening patients' privacy, doctors questioning results	Concerns regarding the data collection and selling to businesses
Patient's acceptance of Big Data in healthcare	Patients not familiar with how their data is used, lack of knowledge in the topic, data transparency and miss usage	Patients reserved towards digitalization, not familiar with how their data is being used	Concerns regarding social media interpretations of Big Data implementation in healthcare
Big Data in developing countries	Poor policies, poor infrastructure, lack of skilled labour, lack of education, limited resources	Lack of technological capacity, lack of skilled workers, availability of the data, poor infrastructure	Countries in are Africa trying to catch up with developed countries and meet sustainable goals

Source: Own work.

When compared to the literature review, the interviews highlighted the most important positive and negative sides of Big Data analytics and whether or not certain topics present a challenge. Moreover, through the Text Mining analysis, information on how people feel

about Big Data analytics in healthcare was collected and what exactly is the problem when they are opposing its implementation. The sentiment analysis showed that words such as Big Data, AI, Medical Data, Biomedical, innovations, digital technology, Internet of Things or Quality of life were used in a positive context, showing that the people that posted these tweets were familiar with the benefits of Big Data usage in healthcare. The most important opportunities of Big Data in healthcare are their usage for *predicting and diagnosing diseases, clinical decision support, several different uses in the pharmaceutical industry improving, fraud detection and prevention, improving patients' quality of life, personalized medicine and improvement of the healthcare processes*. According to the interviews, there are many opportunities and barriers that we haven't even faced yet, because of the low Big Data usage in healthcare nowadays. Some people are worried about their data privacy, availability, and misusage, and this is shown from the words that were used in a negative context such as Facebook, device, researches, and collaboration. The analysis on the words used in a negative context confirms the findings from the literature review and the results from the interviews, which stated that people oppose Big Data implementation in healthcare because they do not know how their data is being used and who has access to it. The main barriers of Big Data Analytics in healthcare are *the lack of standardized laws and regulations, data related barriers, such as quality, security and ownership, the lacked of skilled labor, technological capacity in the healthcare sector, ethical challenges, patients' and medical personnel acceptance of Big Data in the healthcare sector, Big Data implementation in developing countries.* These findings answer the research question which was defined in the thesis: What are the opportunities and barriers of Big Data Analytics in the healthcare?

There are some limitations regarding the research results of the thesis. The interviews were conducted with one representative per team, one developer, one Medical Informatics expert, and one doctor. While the interviewees provided extensive answers to the questions and connected them with practical examples, there might be professionals working in those fields in different countries or organizations that have different opinions. Even though most of the tweets and words with Big Data were used in a positive context, these results came only from one social media site, and only in the English language. Many other opinions in different languages or opinions that were not expressed on Twitter were not considered. Moreover, there might be other results for different search phrases and words or tweets that have been posted at a different time, other than the one that was analyzed. The Text Mining analysis just scratched the surface on what is being discussed on a social platform about Big Data analytics in the healthcare sector. To discover all the possibilities that these technologies have to offer, the healthcare sector needs to promote and implement Big Data in many more aspects, test it and report back to the development team and the IT companies for further improvements. While professionals always have to be careful and consider all of the possible barriers that Big Data Analytics faces, embracing the future and the new technologies will benefit the patients, healthcare organizations, and pharmaceutical companies as well as the companies working with Big Data. The countries

and their leaders must find a way to digitalize patient data and use it for further analysis. While this might not be easy, there are some countries where this has been done and where Big Data technologies are thriving. The experts need to educate the medical personnel as well as the patients on how their data is being used and that these analyses are in everyone's interests. By promoting this field, and providing the option for gaining knowledge and practice, the governments can solve the issue with the lack of skilled workers in Big Data. There are still many ethical challenges that need to be revised and many studies to be done before everyone accepts Big Data in healthcare, further research can be done on the topic of ethical challenges of Big Data implementation in healthcare and how this barrier can be overcome. Another suggestion for further research is Big Data implementation in healthcare in developing countries, about the issues that developing countries face and how can they be solved.

CONCLUSION

There are many potential opportunities and barriers of Big Data implementation in healthcare that are currently known of and professionals are working on solutions. This thesis analyses the application of Big Data analytics in the healthcare sector, and the barriers and the opportunities of this implementation. With consultation of different literature sources, the opinion of experts working in the field of Big Data in the healthcare sector and the public opinion of current and future patients, the thesis answers the question: which are the barriers and opportunities of Big Data Analytics in the healthcare sector? Through detailed literature analysis of Big Data analytics in the industry and their practical usage, the most important opportunities and the barriers are defined. The goals of the thesis were fulfilled; data from primary and secondary sources were collected and analyzed. Various literature sources were analyzed in order to find and define the barriers and opportunities of Big Data analytics in healthcare. The data collected from literature review analysis, was compared to the interview analysis and the sentiment analysis of the mined text. The interviews' results provided extensive information on the topic, from a point of view and experience of experts in the field. The public opinion on the topic of Big Data analytics in healthcare was analyzed through sentiment analysis which was performed on a mined data from a social media platform. The findings of literature review, the results of the interview analysis and the sentiment analysis were analyzed and compared to each other, which allowed for the main opportunities and barriers of Big Data analytics to be defined and described.

While the interviewees provided extensive information based on personal experiences during the interviews, there might be experts in the field who disagree or have different experiences. Another limitation of the research for the Text Mining analysis is the number of tweets analyzed and only in the English language. Moreover, there might be people who haven't posted their opinion on the topic of Big Data in healthcare, and those opinions were not taken into consideration in the sentiment analysis. There might have been tweets

for the topic that used different terms or phrases, which also weren't taken into consideration. Future research can be done on the topic of ethical challenges of Big Data implementation in healthcare, considering that this is still an ongoing topic of discussion. Another suggestion for future research is Big Data implementation in developing countries and the barriers they face. The implementation of Big Data can lead to many benefits for the patients, which is the center of the healthcare system, however, these technologies should always be supervised by humans and the results should be revised. With constant usage and improvement, Big Data Analytics can lead to great changes to the healthcare sector as we know it, change it for the better and improve the healthcare service for the patients, as well as their quality of life.

REFERENCES

1. Aboudi, N. E., & Benhlima, L. (2018). Big Data Management for Healthcare Systems: Architecture, Requirements, and Implementation. *Advances in Bioinformatics* , *2018* , 1-11.

2. Ahmad, P., Qamar, S., & Rizvi, S. Q. (2015). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications* , *120* , 38-50.

3. Alharthi, A., Krotov, V., & Bowman, M. (2017). Addressing barriers to big data. *Business Horizons, 60 , 3* , 285-292.

4. Bezemer, T., Groot, M. C., Blasse, E., Berg, M. J., Kappen, T. H., Bredenoord, A. L., Solinge, E. W. V., Hoefer, I. E. & Haitjema, S. (2019). A Human(e) Factor in Clinical Decision Support Systems. *Journal of Medical Internet Research, 3* , 1-9.

5. Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017). A Study of Machine Learning in Healthcare . *Annual Computer Software and Applications Conference* (pp. 236-241). San Jose, CA, USA: IEEE.

6. Bora, D. J. (2019). Big Data Analysis in Healthcare: A critical analysis. In N. Dey, H. Das, B. Naik, & H. S. Behera, *Big Data Analytics for Intelligent Healthcare Management* (pp. 43-57). London, United Kingdom: Academic Press.

7. Brill, S. B., Moss, K. O., & Prater, L. (2019). Transformation of the Doctor–Patient Relationship: Big Data, Accountable Care, and Predictive Health Analytics. *HEC Forum, 4* , 261–282.

8. Chawla, N. V., & Davis, D. A. (2013). Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. *Journal of General Internal Medicine* 28, 660–665.

9. Chen, I. Y., Joshi, S., Ghassemi, M., & Ranganath, R. (2020). Probabilistic Machine Learning for Healthcare. *Annual Reviews of Biomedical Data Science 2021* , 4, 393-415.

10. Dagliati, A., Tibollo, V., Sacchi, L., Malovini, A., Limongelli, I., Gabetta, M., Napolitano,C., Mazzanti, A., Cata, P. D., Chiovato, L. & Bellazzi, S. P. (2018). Big Data as a Driver for Clinical Decision Support Systems: A Learning Health Systems Perspective. *Frontiers in Digital Humanities* , 5, 2297-2668.

11. Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data* , 56 , 1-25.

12. Software Testing Help. (2021). *Data Mining Vs Machine Learning Vs Artificial Intelligence Vs Deep Learning*. (2021, 5 30). Retrieved 6 2, 2021 from https://www.softwaretestinghelp.com/data-mining-vs-machine-learning-vs-ai/

13. Dey, N., Das, H., Naik, B., & Behera, H. S. (2019). *Big Data Analytics for Intelligent Healthcare Management.* Academic Press.

14. Ejiaku, S. A. (2014). Technology Adoption: Issues and Challenges in Information Technology Adoption in Emerging Economies . *Journal of International Technology and Information Management*, 23(2) , 58-68.

15. Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review* 1 (2), 293–314.

16. Fu, Y. (1997). Data mining: Tasks, techniques and applications. *IEEE Potentials, 29* , 18-20.

17. Gamache, R., Kharrazi, H., & Weiner, J. P. (2018). Public and Population Health Informatics: The Bridging of Big Data to Benefit Communities. *Yearbook of Medical Informatics* , 27 (1), 199–206.

18. Ghobakhlo, M., Sabouri, M. S., Hong, T. S., & Zulkifli, N. (2011). Information Technology Adoption in Small and Medium-sized Enterprises; An Appraisal of Two Decades Literature. *Interdisciplinary Journal of Research in Business* , 1 (7), 53-80.

19. Hall, B. H., & Khan, B. (2003). Adoption of New Technology. *NBER Working Paper Series* , 1-19.

20. Hand, D. J., & Adams, N. M. (2015, June 22). Data Mining. *Wiley StatsRef: Statistics Reference Online* , 15(2) , 1-7.

21. Haux, R. (2010). Medical informatics: Past, present, future. *International journal of medical informatics* , 79(9) , 599-610.

22. Hersh, W. R. (2002). Medical Informatics: Improving Health Care Through Information. *Contempo Updates* , 16 , 23-30.

23. Hoerbst, A., & Ammenwerth, E. (2010). Electronic Health Records: A Systematic Review on Quality Requirements. *Methods of Information in Medicine ,* 49(4) , 1-17.

24. Hüllermeier, E. (2005). Fuzzy methods in machine learning and data mining:Status and prospects. *Fuzzy Sets and Systems* , 156(3) , 387–406.

25. Iliashenko, O., Bikkulova, Z., & Dubgorn, A. (2019). Opportunities and challenges of artificial intelligence in healthcare. *E3S Web of Conferences* (pp. 1-8). St.Petersburg, Russia: Peter the Great St.Petersburg Polytechnic University.

26. Janke, A. T., Overbeek, D. L., Kocher, K. E., & Levy, P. D. (2016). Exploring the Potential of Predictive Analytics and Big Data in Emergency Care. *Annals of Emergency Medicine* , 67(2) , 227-236.

27. Jothia, N., Rashidb, N. A., & Husainc, W. (2015). Data Mining in Healthcare – A Review. *The Third Information Systems International Conference* (pp. 306 – 313). Minden, Penang Malaysia: School of Computer Sciences, Universiti Sains Malaysia,.

28. Kankanhalli, A., Hahn, J., Tan, S., & Gao, G. (2016). Big data and analytics in healthcare: Introduction to the special section. *Information Systems Frontiers* , 18 , 233-235.

29. Karahanna, E., Straub, D. W., & Chervany, N. L. (1999). Information Technology Adoption Across Time: A Cross-Sectional Comparison of PreAdoption and Post-Adoption Beliefs. *Management Information Systems Research Center, University of Minnesota* , 23(2) , 183-213.

30. Kim, W.-J. (2018). Knowledge-based diagnosis and prediction using big data and deep learning in precision medicine. *Investigative Clinical Urology* , 59(2) , 69-71.

31. Koh, H. C., & Tan, G. (2011). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management* , 19(2) , 64-72.

32. Konduforov, O. (2021, January 26). *Data Science vs Machine Learning vs AI vs Deep Learning vs Data Mining: Know the Differences*. Retrieved July 7, 2021, https://www.altexsoft.com/blog/data-science-artificial-intelligence-machine-learning-deep-learning-data-mining/

33. Kumar, S., & Singh, M. (2019). Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools. *Big Data Mining and Analytics* , 6(54) , 48-57.

34. Lee, C. H., & Yoon, H.-J. (2017). Medical big data: promise and challenges. *Kidney Research and Clinical Practice* , 36(1) , 3-11.

35. Leon-Sanz, P. (2019). Key Points for an Ethical Evaluation of Healthcare Big Data. *Processes* , 7(8) , 1-12.

36. Maity, N. G., & Das, D. S. (2017). Machine Learning for Improved Diagnosis and Prognosis in Healthcare. *Aerospace Conference* (pp. 4-11). Montana, USA: IEEE.

37. Murdoch, T. B., & Detsky, A. S. (2014). The Inevitable Application of Big Data to Health Care. *JAMA* , 309(13) , 1351-1352.

38. Pramanik, I., Lau, R. Y., Azad, A. K., Hossain, S., Hossain, K., & Karmaker, B. (2020). Healthcare Informatics and Analytics in Big Data. *Expert Systems with Applications* , 152.

39. Prokosch, H. U., & Ganslandt, T. (2009). Perspectives for Medical Informatics: Reusing the Electronic Medical Record for Clinical Research. *Methods of Informatics in Medicine* , 48(1) , 38-44.

40. Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, a. A. (2020). Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Reviews in Biomedical Engineering* , 14 , 156 - 180.

41. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems* , 2(3) , 1-10.

42. Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management* , 38(1) , 187–195.

43. Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine* , 112(1) , 22–28.

44. Rong, G., Mendez, A., Assi, E. B., Zhao, B., & Sawan, M. (2020). Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering 6* , 6(3) , 291–301.

45. Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42-47). San Diego, CA, US: IEEE.

46. Schonberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology* , 27(2) , 171–203.

47. Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018). Machine Learning in Healthcare: A Review. *International conference on Electronics, Communication and Aerospace Technology* (pp. 910-914). Hyderabad, Telangana : IEEE .

48. Shilo, S., Rossman, H., & Segal, E. (2020). Axes of a revolution: challenges and promises of big data in healthcare. *Nature Medicine* , 26 , 29-38.

49. Tekieh, M. H., & Raahemi, B. (2015). Importance of Data Mining in Healthcare: A Survey. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1057-1062). Ottawa, Canada: Telfer School of Management, University of Ottawa.

50. Vayena, E., Blasimme, A., & Cohen, I. G. (2016). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine* , 15(11) , 1-4.

51. Viceconti, M., Hunter, P., & Hose, R. (2015). Big Data, Big Knowledge: Big Data for Personalized Healthcare. *Journal of Biomedical and Health Informatics* , 19(4) , 1209-1215.

52. Wang, C. (2019). The Strengths, Weaknesses, Opportunities, and Threats Analysis of Big Data Analytics in Healthcare. *International Journal of Big Data and Analytics in Healthcare* , 4(1) , 1-14.

53. Wang, Y., Kung, L. A., & Byrd, T. A. (2016). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting & Social Change* , 126 , 1-11.

54. Wiens, J., & Shenoy, E. S. (2018). Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Healthcare Epodemiology* , 66(1) , 149-153.

55. Wyatt, J. C., & Liu, J. L. (2002). Basic concepts in medical informatics. *Glossary* , 56 , 808–812.

56. Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., & Wang, F. (2020). Federated Learning for Healthcare Informatics. *Journal of Healthcare Informatics Research*, 5 , 1-12 .

57. Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. & Hua, L (2011). Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *Springer Science+Business Media* , 36 , 2431–2448.

58. Yu, P., Wu, M. X., Yu, H., & Xiao, G. Q. (2006). The Challenges for The Adoption of M-Health. *IEEE* , 12 , 181-186.

59. Zhang, X.-D. (2020). *A Matrix Algebra Approach to Artificial Intelligence.* Beijing, China: Springer, Singapore.

60. Zhoua, C., Li, A., Hou, A., Zhang, Z., Zhang, Z., Dai, P. & Wang, F. (2020). Modeling methodology for early warning of chronic heart failure based on real medical big data. *Expert Systems With Applications* , 151(5) , 1-13.

APPENDICES

Appendix 1

Appendix 1. Summary in the Slovenian language

Tehnologija je postala ključni del našega vsakdana in igra pomembno vlogo v poslovanju, na primer s povečanjem učinkovitosti številnih procesov in večjim zadovoljstvom strank. Številne dejavnosti imajo koristi od digitalizacije svojih procesov in zdravstveni sektor ni izjema. Tako zasebni izvajalci zdravstvenega varstva kot javne zdravstvene ustanove uvajajo informacijske sisteme ter ustvarjajo in delijo digitalne informacije. Namen te digitalne preobrazbeje izboljšati proces prepoznavanja, spremljanja, zdravljenja in, če je mogoče, preprečevanja negativnih zdravstvenih posledic. Različni koncepti, kot so elektronski zdravstveni kartoni, komunikacijska tehnologija, pa tudi tehnike rudarjenja podatkov, strojno učenje in umetna inteligenca, so že del sektorja zdravstvenega varstva. Ta implementacija tehnologije v medicini prinaša številne prednosti, na primer zbiranje podatkov iz nosljivih naprav in pomoč ljudem pri izboljšanju svojega zdravja. Vendar pa obstajajo tudi izzivi, na primer kakovost velikih količin podatkov, ki so zbrani iz več virov in so neenakomerni in običajno nestrukturirani.

Ko govorimo o veliki količini strukturiranih, polstrukturiranih ali nestrukturiranih zbranih podatkov, ki so jih ustvarili različni viri, govorimo o masovnih podatkih (angl. big data). Številna globalna podjetja zbirajo te masovne podatke od svojih uporabnikov in iz njih pridobivajo dragocene informacije. Tehnologije masovnih podatkov prinašajo velik potencial zadružbo, kot je odkrivanje subtilnih vzorcev človeškega vedenja, ki jih ni mogoče narediti z majhnimi količinami podatkov, in razvoj napovednih modelov za prihodnje vedenje. S temi podatki lahko podjetja predvidijo potrebe in preference svojih strank ter ponudijo prilagojene storitve, kar vodi do večjega zadovoljstva. Kljub temu te tehnologije niso brez izzivov. Te ogromne količine podatkov ni tako enostavno upravljati. Poleg tega ta vrsta podatkov praviloma vsebuje veliko napak ali manjkajočih vrednosti. V mnogih primerih je treba nadgraditi informacijsko infrastrukturo in namestiti nove prakse upravljanja v celotni organizaciji, ki želi sprejeti velike podatke. Drug izziv je, da primanjkuje zaposlenih, ki imajo potrebne veščine za ravnanje z velikimi podatki. Številne organizacije ne razumejo, kako lahko veliki podatki pomagajo organizaciji in poslovnim procesom, kar vodi v odpor do njihovega sprejemanja v organizaciji. Zato bom v svoji magistrski nalogi analizirala gonilne sile in ovire analitike velikih podatkov v sektorju zdravstvenega varstva. Namen te magistrske naloge je prispevati k razumevanju vpliva tehnologije, natančneje, masovnih podatkov v zdravstvenem sektorju. Raziskovalno vprašanje magistrske naloge je: Kakšne so priložnosti in ovire analitike masovnih podatkov v zdravstvenem sektorju?

Najprej so bile identificirane in predstavljene ovire in priložnosti za analitiko masovnih podatkov v zdravstvu, ki izhajajo iz dosedanjih spoznanj v literaturi. Za pridobitev mnenj na to temo od izkušenih strokovnjakov so bili opravljeni intervjuji s strokovnjaki iz treh skupin, potrebnih za implementacijo masovnih podatkov v zdravstvu: razvijalec in

svetovalec, strokovnjak za medicinsko informatiko ter zdravnik in raziskovalec na tem področju. Rezultati intervjujev podrobno pojasnjujejo, kako se tehnologije Big Data uporabljajo v zdravstvu, katere so najbolj zanimive priložnosti te tehnologije in ovire, s katerimi se soočamo pri njihovi implementaciji in uporabi. Poleg tega je bilo opravljeno rudarjenje po besedilih oz.analiza razpoloženja na tvitih v zvezi z masovnimi podatki v zdravstvu, da bi razumeli mnenje splošne javnosti o tej temi. Pregled literature je služil kot osnova za raziskavo, za strukturiranje intervjujev in za rudarjenje po besedilih.

Glavni in najpomembnejši prispevek magistrskega dela je opredelitev najpomembnejših priložnosti in ovir pri implementaciji velikih podatkov v zdravstvu s praktičnimi primeri in različnimi izkušnjami na tem področju. S primerjavo pregleda literature in intervjujev so bile izpostavljene razlike in skladnost med rezultati, ki kažejo podrobno oceno trenutnega stanja na tem področju. Raziskava je pokazala, da so najpomembnejše priložnosti masovnih podatkov v zdravstvu so njihova uporaba za *napovedovanje in diagnosticiranje bolezni, podpora pri kliničnem odločanju, izboljšanje več različnih uporab v farmacevtski panogi, odkrivanje in preprečevanje goljufij, izboljšanje kakovosti življenja bolnikov, personalizirana medicina in izboljšanje zdravstvenih procesov*. Glavne ovire analitike masovnih podatkov v zdravstvu so *pomanjkanje standardiziranih zakonov in predpisov, ovire, povezane s podatki, kot so kakovost, varnost in lastništvo, pomanjkanje strokovnjakov, tehnološke zmogljivosti v zdravstvenem sektorju, etični izzivi, sprejemanje masovnih podatkov zaposlenih v zdravstvenem sektorju, implementacija masovnih podatkov v državah v razvoju*. Čeprav je bila za magistrsko delo opravljena poglobljena raziskava, obstajajo nekatere omejitve. Medtem ko so intervjuvanci na vprašanja posredovali obširne odgovore in jih povezali s praktičnimi primeri, so morda strokovnjaki, ki delajo na teh področjih v različnih državah ali organizacijah, ki imajo drugačna mnenja. Čeprav je bila večina tvitov in besed, povezanih z uporabo masovnih podatkov v zdravstvu,uporabljenih v pozitivnem kontekstu, so ti rezultati prišli le z enega družbenega medija in to samo v angleškem jeziku. Številna druga mnenja v različnih jezikih ali mnenja, ki niso bila izražena na Twitterju, niso bila upoštevana. Rudarjenje po besedilih je razkrilo samo površino tega, o čemer se na družbeni platformi razpravlja o analitiki masovnih podatkov v zdravstvenem sektorju. Še vedno je veliko etičnih izzivov, ki jih je treba razumeti, in veliko študij, ki jih je treba narediti, preden bo analitika masovnih podatkov široko sprejeta v zdravstvu. Nadaljnje raziskave je mogoče narediti na temo etičnih izzivov implementacije masovnih podatkov v zdravstvu in kako je to oviro mogoče premagati. Drug predlog za nadaljnje raziskave je implementacija masovnih podatkov v zdravstvu v državah v razvoju in reševanje specifičnih vprašanj, s katerimi se pri tem soočajo države v razvoju. S širšo uporabo in stalnim izboljševanjem lahko analitika masovnih podatkov vodi do velikih sprememb v zdravstvenem sektorju, kot ga poznamo, ga spremeni na bolje in izboljša zdravstveno storitev za paciente ter kakovost njihovega življenja.

Appendix 2

Appendix 2. Interview Questions

General questions for introduction to the interview

- Can you tell me about your experience with Big Data analytics in healthcare?
- How can Big Data be successfully implemented in healthcare? What are the requirements?

Opportunities

1) How are Big Data tools being used in healthcare nowadays?
2) What are the most disruptive potentials of Big Data Analytics in the Healthcare sector and what can we achieve with its implementation on long run?
    a) Are the benefits of Big Data implementation likely to significantly improve the healthcare as we know it? Is this a factor that the healthcare organizations consider before they decide on its usage?
3) Can Big Data technologies help with health process improvement? (from admission to treatment to dismissing the patients) How?
    a) Can Big Data analytics improve the quality of care and performance measurement? How? What about the healthcare management?
4) Can the Big data technologies be used for creating personalized medicine and therefore improving the healthcare? How?
5) Can the healthcare benefit from Big Data for modeling disease progression? How?
    a) Can these disease models be later used on other patients and help with creating better and more effective treatments?
    b) Is a bigger treatment effectiveness one of the opportunities of Big Data analytics usage in healthcare?
6) Is one of the benefits of Big Data in Healthcare faster and more accurate disease diagnostics?
    a) Can these technologies help with early disease diagnosis and eventually lead to less hospitalized patients because of it?
    b) Will this lead to lowering the workload that the medical personnel has, and allow them to use more time with each patient?
7) Can Big Data be used for predictive modeling for risk and resource use, and ensure bigger effectiveness? How?
8) Will Big Data allow better disease cataloging (in sense of analyzing, determining, and predicting)?
    a) Can the disease cataloging help with future cases and lead to faster diagnosis, treatments or preventions?
9) Can the Big Data help with target identification of suitable patients/volunteers when pharmaceutical companies are doing human trials for certain drug?
    a) Can this lead to faster and more accurate drug tests and approvals, and their faster and cheaper arrival to the market for general public usage?

10) Can Big Data be used for clinical decision support systems and how?
    a) Can it be a clinical decision support by its own, or its purpose is to help the medical personnel with the decision making?
11) With the help of Big Data analytics, is it possible to reduce the costs for diagnosing diseases?
12) With Big-data analytics usage, is it possible to reveal certain patterns for disease development in a population and therefore create a better treatment?
    a) Will this eventually lead to prolonging the human life and improving its quality?
13) Can Big Data analytics in the healthcare help with fraud detection and prevention, by detecting suspicious records of health data?
14) Would the ability to try out the technologies by the medical personnel before their implementation lead to a more positive outcome and their adoption?

Barriers

1) What are the barriers of Big Data implementation in healthcare? Elaborate why and how.
2) Is relying only on Predictive Models in healthcare a challenge or an opportunity? Why?
    a) Is the human supervision necessary? Is this a challenge? Why?
    b) Is the doctor-patient relationship in danger when implementing big data in healthcare and why?
3) Is the Data Quality a challenge when talking about Big Data in Healthcare? Why and how?
    a) What are the biggest challenges with the data which is being collected from medical devices and smart devices?
4) Is regulating the data ownership and security a challenge, considering its sensitive nature? Why and how can be solved?
    a) When collecting the data, does the patient need to approve that his data will be digitized and used in further researches and analysis?
    b) Does that data need to go through pseudonymization process in order to protect the patients?
    c) Where the collected medical data is usually saved and who can access it?
    d) Is fraud and abuse a challenge when talking about the data?
    e) Is the setting measure for protection from cyber attacks challenging?
5) Is the compatibility with already excising technology an important factor for Big Data adoption?
    a) Are the healthcare organizations willing to change the already existing technologies in order to be able get the potential benefits from the new technologies, despite the expenses?
    b) What needs to be done before this technological adoption, in order to insure Big Data's successful implementation?
6) Is the technical capacity of the industry an important factor for Big Data usage in healthcare?

a) In your experience, are the new technologies too advanced compared to the already existing systems in the healthcare industry?

7) Is the complexity of the technology a factor that can lead to adoption or rejection of Big Data or is the medical personnel more willing to learn and adapt?
8) Are the limitations of artificially intelligent tools a challenge in today's usage of Big Data in Healthcare? What are they, why and how?
9) Is the Implementation of Big Data technology in developing countries more challenging than implementation of Big Data in developed countries? How and why?
   a) Poor policies, insufficient infrastructure investment, depend on foreign aid, Poor IT, lack skilled IT workers, limited access to internet and modern computing?
10) Is the level of skilled workers in the healthcare industry for Big Data analytics a factor for its successful implementation and usage? What are the needed skills, education, and experience?
   a) Can this be resolved easily or is it a challenge?
11) Is setting the Information System structure support for Big Data in Healthcare a challenge and why?
   a) When creating IT solutions for the healthcare sector, beside the IT team, are there other skills from the medical field required? Is this a challenge?
   b) Do the companies in charge for creating and implementing Big Data in healthcare need to consult professionals regarding the software or for its testing? Is that challenging?
12) Are the costs a factor when a healthcare organization decides whether or not to adopt new technologies?
   a) How important is it for the healthcare industry to be assured that the usage of Big Data will help cover the costs of its implementation?
13) Which are the ethical challenges that need to be considered and why?
   a) In your experience, how do people feel regarding healthcare institutions using AI and Machine Learning for making decisions?
   b) Is the general opinion on Big Data Analytics in healthcare an important factor that the organizations consider before implementing it?
   c) Would they oppose its implementation if the potential patients don't agree with it, or would they go ahead and adopt it because of its potential?
   d) What are the patients' biggest concerns for using these technologies?
   e) How can professionals/experts help the public to accept big data in healthcare and insure a smooth transition?
14) Is a challenge to have big data technologies effectively regulated by the government and the law? Why?
   a) Are the government regulations regarding Big Data analytics in healthcare a significant factor, when it comes for their implementation and adoption in healthcare?
   b) Is it a challenge, on one side achieving transparency regarding big data usage in healthcare and on the other side respecting patient's privacy? Why and how can it be done?

General review of the interview

1) Who in your opinion benefits the most of Big Data implementation in healthcare and why?
2) In your opinion, how far are we from using mainly Machine Learning and AI technologies in the healthcare sector?
3) Anything else you would like to add/crossed your mind on the subject?