Master thesis
The relation between big data and informational privacy in the context of the healthcare.

Student: Robin Sippe (4183800)
Study programme: MSc. Management of Technology

Chair: Prof.dr. Yao-Hua Tan
First supervisor: Dr. ir. G.A. de Reuver
Second supervisor: Dr. Jafar Rezaei
External supervisor: Drs. J. Heisenberg

August 10th, 2015

Public version

10th of August 2015

Abstract

Big data is a broad term that is related to the collection, storage and analysis of large volumes of data. The term big data is often associated with the popular 3V's model, which defined that data is growing significantly in the characteristics volume, variety and velocity. In this research we defined big data as: *the collection, storage and transformation of structured and unstructured data from multiple sources into useful information (or knowledge) to improve decision-making within organizations.*

The significant growth of data is also occurring in the health care sector. A lot of these scattered data sources, possessing large volumes of personal health data of patients, are present in the health care. Big data have shown potential to support health care, by combining and transforming health data. Big data can be used to support medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health (Raghupathi & Raghupathi, 2014). The increasing availability of large data sets from various sources in combination with the development of more advanced analytical tools for big data makes it more and more difficult to ensure privacy.

Big data in its current form is still relatively new, and the knowledge on the implications on the security and privacy issues that it brings is still limited. This study explores the relation between big data and privacy in the health care. **The research objective of this study is to gather knowledge on how big data affects privacy in the health care.** In order to reach this objective, semi-structured interviews have been conducted with eight experts in either big data, health care or privacy in the Netherlands.

In this research, a conceptual model of privacy has been created based on existing theories of privacy (e.g. nonintrusion theory, seclusion theory, control theory and restricted access theory). The conceptual model of privacy defines privacy in the elements: *natural privacy*, *normative privacy*, *control aspect of privacy* and the *condition of privacy* and has been used as a structure to analyze the relation between big data and privacy.

Table of Contents

1. Introduction

Since a few years ago, Big Data has been a promising IT trend. Big data is a very popular term nowadays and experts are even talking about a big data era (Boyd & Crawford, 2012; Manyika, Chui, Brown, & Bughin, 2011). Huge amounts of data are generated from various sources like sensors and location data from mobile phones and other electronic consumer electronic devices, internet data from web searches and social media service like Facebook and Twitter, etc. (Figure 1). In 2012, International Data Corporation (IDC) reported that the volume of digital content will grow to 2.7 zettabyte (ZB, = 10^{21} bytes) and will pass the 8 ZB mark by 2015 (Gens, 2012). This explosion in growth of data in combination with the decreasing price of hardware followed from the commoditization of IT pushed the development of technology that process and analyze big data.

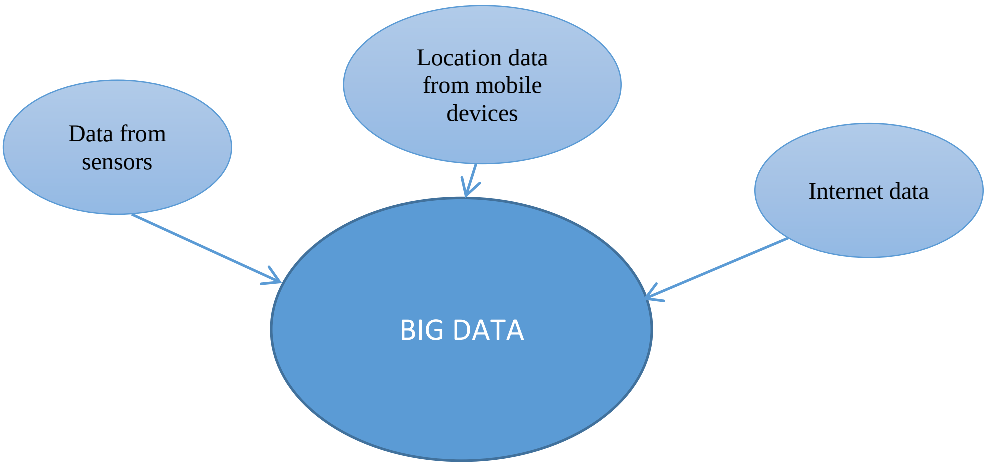


Figure 1: Big data sources

Big data analytics

Analytics is the process of examining data to uncover hidden patterns, unknown correlations and other useful information. Through analytics data can create value for the global economy, driving innovation, productivity, efficiency, and growth (See Figure 2). Big data enhances existing analytics.

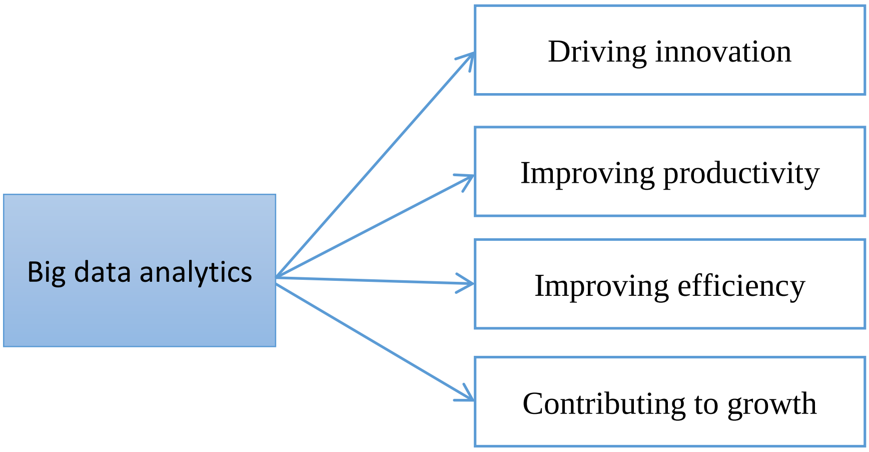


Figure 2: Values generated from Big Data analytics

A classical, real-world example of big data analytics is Wal-Mart's Polaris project, which was launched in 2012 for the purpose to assist consumers to find products of interest. Polaris is a big data platform using predictive analytics, which relies on techniques like: information retrieval, text mining, machine learning and even synonym mining to produce relevant search results on their web shop wal-mart.com. Wal-Mart stated that its new search engine improved online shoppers completing a purchase by 10% to 15% (Walmart, 2012).

Another famous example is Google's Flu Trends. In 2008, Google released Google Flu Trends, which is a web service that "accurately estimate the current level of weekly influenza activity in each region of the United States" (Ginsberg et al., 2008; p1012). Google stated that they found a close relation between how many people search for flu-related topics and how many people actually have flu symptoms. Currently, Google flu trends can estimate influenza activity in 18 countries, by aggregating search queries (Google, 2011).

Big data in the healthcare

A lot of data is generated in the health care sector. Health care providers are storing clinical data in their electronic medical health records (EMR). Besides EMR, there are three other primary sources of data in the health care sector, which include: data of claims and costs of service that was provided, pharmaceutical R&D data and patient behavior and sentiment data that describes patient activities and preferences (Figure 3).

At the same time, the expenditure in the health care is increasing continuously. According to Statistics Netherlands, the health care expenditure in Netherlands in 2011 was 12% of the GDP. The main driver of the expenditure growth is volume growth in demand of health care, which is caused by demographic factors as aging population. These drivers force the healthcare to work more efficiently, while improving quality.

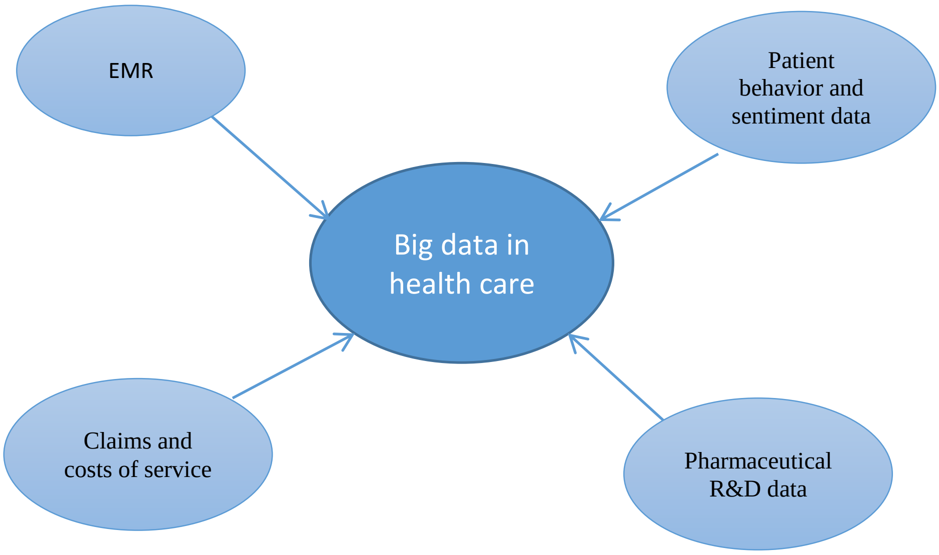

EMR
Patient behavior and sentiment data
Big data in health care
Claims and costs of service
Pharmaceutical R&D data

Figure 3: Big data sources in health care (Groves & Knott, 2013)

Big data also have shown to have potential to support healthcare. Inside the large amount of health data, patterns and trends are hidden, which with the right big data analytics has the potential to improve health care, save lives and lower costs. Big data can be used to support medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health (Raghupathi & Raghupathi, 2014). McKinsey believes that big data can improve efficient and quality in the following areas: Clinical operations, Research & Development, Public Health (Manyika et al., 2011).

Big data security and privacy issues

Security and privacy issues are magnified by several aspects of Big Data. The relatively new big data tools were developed to deal with large sets of data. The main focus in the development was performance and scalability, at the cost of security. With the emergence of cloud technology data storage has become accessible and affordable for more and more parties. It becomes harder to ensure security and protect privacy when information is being multiplied and stored on various servers around the world.

The tools commonly used for Big Data have been called out for their security weaknesses. Many of the big data tools were originally developed by large internet companies to analyze a specific data. High security might not always have been required, since the tool would be part of a larger software system. In those cases the security of those tools relies on external enforcement mechanisms present in the system. Many of these tools shared a common set of security gaps, including weak access control (authentication, authorization, auditing), insecure communications, weak client or API security, no encryption functionality (Cloud Security Alliance, 2013; Hamami, 2011)

NoSQL (commonly known as *Not Only SQL*) databases have been listed as one of the top threats concerning big data. NoSQL were mainly designed to deal with large sets of data and efficiently store unstructured types of data, with a limited emphasis on security. It has been demonstrated that current NoSQL database packages only have a very thin security layer compared to traditional relational database management systems (RDBMS). Many of these packages have the security setting very low or are completely disabled at default. There only a few NoSQL packages that meet the data security requirements (Fidelis Cybersecurity Solutions, 2014).

Big data depends on large quantities of cheap storage and computing resources. In the past, Big Data was limited to very large organizations such as governments and large enterprises that could afford to create and own the infrastructure necessary for hosting and mining large amounts of data. Nowadays, big data storage is affordable and accessible for more organizations through public cloud infrastructure. However, using the cloud as part of a big data solution also introduces security issues on its own. Notable issues with the cloud are: data residency, data encryption, data retention and destruction, and regulatory compliance (Hamami, 2011).

The increasing availability of large data sets from various sources in combination with the development of more advanced analytical tools for big data makes it more and more difficult to ensure privacy. Enforced by law and regulation, organizations use strategies as data encryption and data de-identification to ensure privacy. For example, some privacy laws include standards for de-identification that requires either an expert determination that data cannot be re-identified using common statistical tools and practices or the removal of any data fields that would make it possible to re-identify the individual (Hamami, 2011; Malin, 2012). Organizations are using various de-identification techniques like: anonymization, pseudonymization, encryption, key-coding, data sharding, to distance data from real identities and prepare their data for analytics while maintaining privacy (Tene & Polonetsky, 2012). However, over the years, it has been repeatedly demonstrated that de-identificated data, can often be re-identified (Ohm, 2009).

1.1. Research Problem

In the previous section, it was discussed that big data introduces new security and privacy issues. For the health care sector these issues are even amplified, due to fact that health care data are considered privacy sensitive data. Regulations like the Dutch Privacy law, *Wet bescherming persoonsgegevens* (Wbp) has formulated requirements on how to use process and share our private medical data. However, traditional security and privacy methods to protect privacy health care data seem insufficient or even obsolete.

This is a problem for patients as personal information can unwillingly be derived from these health information systems and end up in wrong hands. Besides that individuals have some rights against intrusion of their personal information, in wrong hands, personal information can potentially harm individuals.

On the other hand, weak security and privacy methods can hinder the adoption of big data in the health care. There can be public resistance from individuals or government against the use of big data in health care, when there is no trust in the protection of their personal information. Hindering in the adoption of big data in the health care could also hinder potential benefits big data could bring to the health care, which are for example improved quality of health.
Therefore, the owners of the problem are the hospitals and other organizations in the healthcare that potentially can benefit from an adoption of big data in health care. These organizations have to deal with hurdles such as privacy legalization and the public perception of privacy before they can successfully adopt big data.

1.2. Research Objective

We can conclude that big data in its current form is still young, and the knowledge on the implications on the security and privacy issues that it brings is still limited. The purpose of this study is to create understanding on how big data impacts privacy issues in the healthcare.

With privacy we mainly refer to ***informational privacy***. According to Parent (1983), informational privacy refers to the "protection against the misuse of personal, sensitive information". This distinct category of privacy includes concerns as access to personal information stored in computer databases (Himma & Tavani, 2008; J. Moor, 1990; Parent, 1983; Tavani, 1999). The challenge in informational privacy is to data for beneficial purposes, while protecting personally identifiable information.

This study will focus on enriching our knowledge about privacy issues of big data within the health care. Big data is still a young topic in the scientific literature. A search in the scientific database of Scopus on the keyword: ***big data***, returns limited results. Scopus shows that the scientific literature started to discuss big data from 2008.

This indicates that privacy issues in the health care are still quite unexplored. Hence, we try to contribute to the scientific by enriching the scientific literature on privacy issues in the health care. We aim to enrich the knowledge of big data, health and privacy literature, by exploring the relation between big data and health care in the context of privacy.

Therefore the following research objective has been formulated:

"To gather knowledge on the relation between big data and privacy in the health care."

We want to gather knowledge on privacy issues that affect health information systems within the healthcare sector. In order to reach this object we will conduct semi-structured interviews with experts.

The healthcare sector is broad and includes actors such as payers, government, medical research institutes, patients, and healthcare providers. In this research we will focus on ***healthcare providers*** that can benefit from big data. More specifically we focus on ***hospitals***. Hospitals are large health care providers that are treating a lot of patients. Hospitals are already collecting large amounts of medical data as they treat large amount of patients. There seem to be opportunities for these large amounts of data. According to the literature these health care services can benefit from these large amounts of data (Groves & Knott, 2013; Manyika et al., 2011; Nambiar, Bhardwaj, Sethi, & Vargheese, 2013; Raghupathi & Raghupathi, 2014). Hence, it is expected that especially hospitals benefit from big data, due to the large pools of data hospitals already possess.

1.3. Research Questions

Based on the research objective four research questions have been formulated. The research questions are presented and described below:

RQ1: What is big data and what are the technologies related to big data and how does it differ from business intelligence?

Research question 1 is related to the exploration of the different views on big data within literature. The different views of big data will provide the basis to the understanding of the different aspects of big data. Additionally, the ***research question 1*** also covers the technology related to big data and the differences between big data and business intelligence. Business intelligence is often referred as a predecessor of big data. We would like to know what the relation is between big data and business intelligence and how big data and business intelligence differ from each other.

RQ2: What are relevant theories to analyze the concept of privacy?

Research question 2 relates to the exploration of the privacy literature, in order to substantiate the theoretical background for this research. We would like to conceptualize privacy by reviewing existing theories of privacy in the academic literature. Answering ***research question 2*** will result in a conceptual model of privacy that can help us analyzing privacy in this research.

RQ3: What is privacy in the health care and what are the regulations related to privacy in the health care?

Research question 3 is focused on privacy in the context of the healthcare. We would like to know what privacy means in the context of the health care. We would like to answer this by providing contextual information about medical data and health information systems that stores medical data. Additionally, we would like to answer ***research question 3*** by reviewing health care laws and privacy laws. Health care and privacy legislation have a significant role on how privacy is experienced within the health care.

RQ4: What is the relation between big data and privacy in the context of the healthcare?

Finally, ***research question 4*** relates to the investigation of the impact of big data on these privacy issues in the healthcare.

1.4. Research methods

The main research method of this research is the semi-structured interview with experts. Our research objective indicates the exploratory character of our research.

A Scopus search showed that only on big data in combination with the health care and privacy. which makes it is interesting to explore to touch this untouched area. Interviews allow us to interact with experts and have a higher chance to find something that has not discovered yet by current research.

We selected experts for our target for this interview since we want to know in depth knowledge instead of the general opinion of the public (e.g. patients in the context of health care). Experts can provide in depth knowledge in their specialized field, while patients can only provide knowledge inclined to be influenced by the media and other people in their surroundings. Experts usually possess much more advanced knowledge in the respective field (e.g. it, health care and privacy) this allows them to form an opinion based on facts from professional experience.

This research doesn't focus on users of the health care system (e.g. doctors and patients) since they are not experts in the above mentioned and therefore mainly provide general opinion rather than a professional opinion based on facts from professional experience.

The semi-structured interview is an appropriate method to use for exploratory research. One of the main strengths of a semi-structured interview is that it allows us to interact with the participants. This allows us to prepare a set of questions that covers the topics of interest, while open for new findings that were initially not covered by the questions. Additionally,

participants are given the space to use their own wording and provide rich information that allows us to get deeper understanding of the relation between big data and privacy in the context of the healthcare.
The initial findings from the literature review will be used to develop an interview protocol that will provide some guidance through the interviews. This interview protocol will provide topics of discussion supported with thematically guiding questions which are not fixed. Non-fixed questions give us the room to explore outside the initial findings.

Expected is that the interviews will bring up discussions that was not covered by the initial findings. Big data revolution is still in its early days (Groves & Knott, 2013; Manyika et al., 2011; Raghupathi & Raghupathi, 2014). And therefore we expect findings that are not covered yet by the literature. A search in the academic literature database Scopus on the keywords big data shows that big data has started to be mentioned since 2008 (Scopus shows ***one*** article from before 2008 which was in 2004).

1.5. Thesis Outline

This thesis is divided into six chapters and the outline is provided below (Figure 4).

Chapter 2 addresses the first research question by discussing the different views on big data and describing the relevant big data technologies. Based on the findings from the literature review, we will conclude this chapter with key characteristics of big data.

Chapter 3 addresses research question two. This question will be answered with an overview of privacy theories in the literature. Based on the main findings from the privacy literature, a conceptual model of privacy has been created and is presented in this chapter.

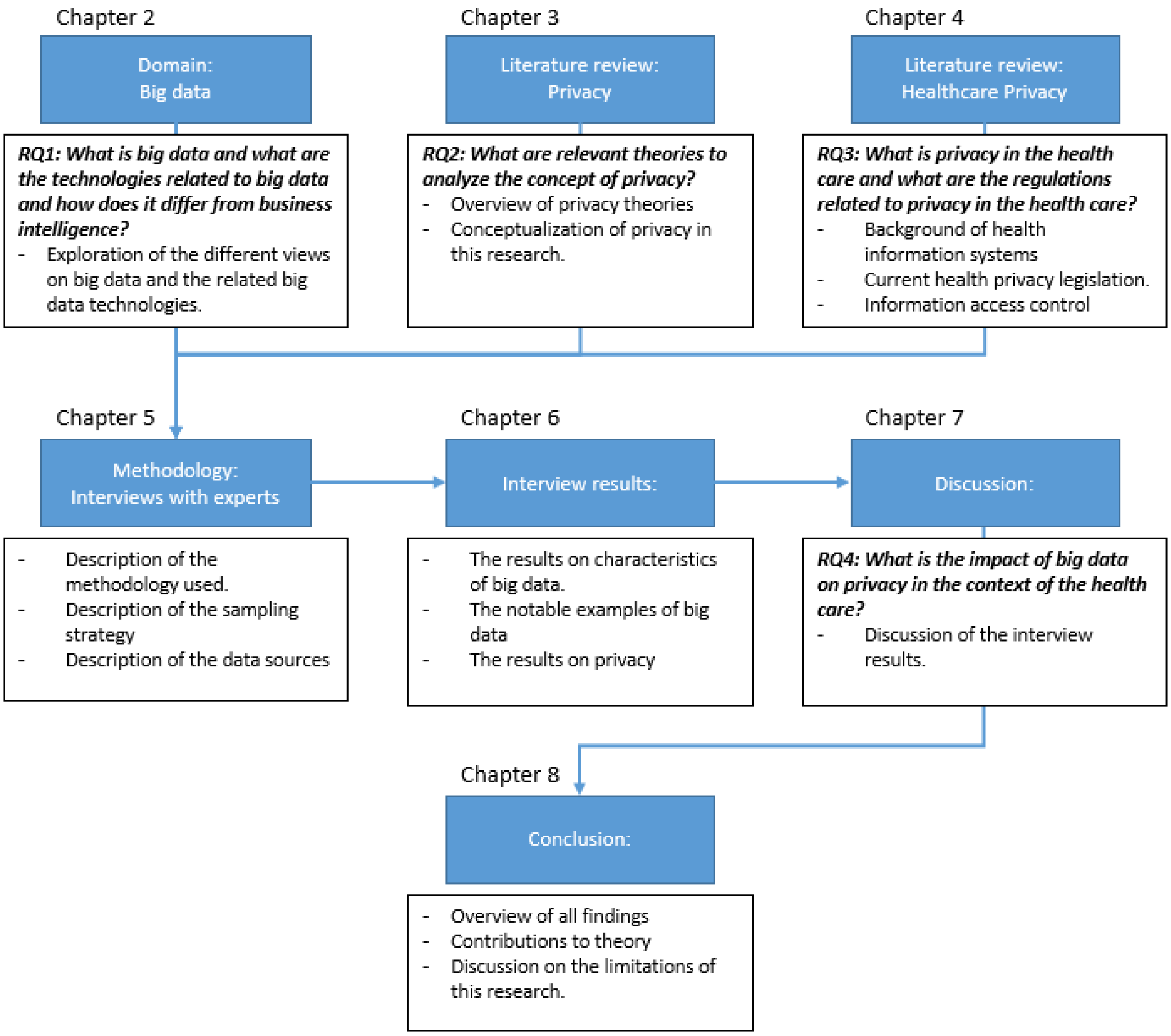
Chapter 4 will zoom in on the health care context of privacy. Health care privacy concerns medical data and the health information systems on which this data is stored on. A background on medical data and health information will be provided as introduction to health care privacy. Finally this part will be concluded with the main findings on the situation of health care privacy.

Chapter 5 provides the methodology that has been used for this research. We used semi-structured interview to collect date for this research. This chapter also discusses the sample strategy and the background of our sample.

Chapter 6 will present the results from the interview and discusses the results.

Chapter 7 will provide a discussion on the results from the interviews.

Chapter 8 will summarize the main findings of this research and will discuss the limitations of this research.



Chapter 2

Domain:
Big data

RQ1: What is big data and what are the technologies related to big data and how does it differ from business intelligence?

- Exploration of the different views on big data and the related big data technologies.

Chapter 3

Literature review:
Privacy

RQ2: What are relevant theories to analyze the concept of privacy?

- Overview of privacy theories
- Conceptualization of privacy in this research.

Chapter 4

Literature review:
Healthcare Privacy

RQ3: What is privacy in the health care and what are the regulations related to privacy in the health care?

- Background of health information systems
- Current health privacy legislation.
- Information access control

Chapter 5

Methodology:
Interviews with experts

- Description of the methodology used.
- Description of the sampling strategy
- Description of the data sources

Chapter 6

Interview results:

- The results on characteristics of big data.
- The notable examples of big data
- The results on privacy

Chapter 7

Discussion:

RQ4: What is the impact of big data on privacy in the context of the health care?

- Discussion of the interview results.

Chapter 8

Conclusion:

- Overview of all findings
- Contributions to theory
- Discussion on the limitations of this research.

Figure 4: Thesis outline

2. Domain: Big data

This chapter will provide a review on business intelligence and big data. Firstly, in Section 2.1, we will review the literature of Business intelligence. Business intelligence is considered as a predecessor of big data and therefore is selected as a starting point for this research. Section 2.2 will review the background and definitions of big data. The popular 3V model of big data will be discussed. Furthermore, other definitions of big data will be reviewed with the aim of exploring the relation between big data and business intelligence. Section 2.3 will describe the technologies that are associated with business intelligence and big data. Section 2.3.1 will review the technological development of the database technologies. New database technologies have been developed in order to process big data. Section 2.3.2 will discuss the Hadoop framework. The Hadoop framework is a popular open-source software framework that is aimed at processing large volumes of data. The Hadoop framework is often associated with big data, as it provides the software tools that could deal with the technological requirements to process big data.

2.1. Business intelligence

Organizations have already used data intelligently since the term Business Intelligence was introduced. The term Business Intelligence (BI) has been defined numerous times and many different definitions exist in literature. In 1989, H. Dresner of the Gartner Group introduced Business Intelligence as an umbrella term to describe ‘concepts and methods to improve business decision making by using fact-based support systems’.

According to Nylund (1999), ‘Using fact rather than intuition was the key to intelligence’. This definition implies that data contains facts that can assist in making intelligent (business) decisions.

In later definitions of business intelligence there seem to be a focus on how to use data to improve ‘intelligent’ decision making in order improve business performance. Azvine, Cui, Nauck (2005) defines BI as “how to capture, access, understand, analyze and turn one of the most valuable assets of an enterprise — raw data — into actionable information in order to improve business performance”. By then it was already accepted that data contain facts or information that can assist ‘intelligent’ decision making and therefore improve business performance. However, the methods to use these data effectively were still being developed or explored.

Jourdan, Rainer, and Marshall (2008) defined Business Intelligence as “both a process and a product. The process is composed of methods that organizations use to develop useful information, or intelligence, that can help organizations survive and thrive in the global economy. The product is information that will allow organizations to predict the behavior of their competitors, suppliers, customers, technologies, acquisitions, markets, products and services, and the general business environment with a degree of certainty”. This definition explicitly separate business intelligence into two elements: an information part and a process part. In relation with the previous definitions we could say that the information part refers to that data possess facts, information. The process part of Jourdan, Rainer, and Marshall (2008) definition, relates to the methods, the actions that need to be taken in order to gather the facts and information from data.

Other definitions of business intelligence explicitly include analytical tools as important part of the process to turn process data. Negash (2004) defines Business Intelligence as “combining data gathering, data storage, and knowledge management with analytical tools to present complex internal and competitive information to planners and decision makers” (p. 178). According to (Elbashir, Collier, & Davern, 2008), “BI systems are defined as specialized tools for data analysis, query, and reporting, (such as OLAP and dashboards) that support organizational decision-making that potentially enhances the performance of a range of business processes”.

Other terms are also used for Business Intelligence, such as Competitive Intelligence, Corporate Intelligence, Competitive Information, or Commercial Intelligence. These terms emphasize the competitive advantage that can be taken by the use of intelligence. Although the terms Competitive Intelligence and Business Intelligence are used interchangeably, there is a difference. McGonagle & Vella (2002) describe the difference by discussing that Business Intelligence is only oriented to the internal processes, while Competitive Intelligence is also oriented to the external processes. According to Negash (2004), Competitive Intelligence is a specialized branch of Business Intelligence.

Many definitions of BI are given in the literature. We learned that business intelligence is used as an umbrella turn. We discussed that business intelligence has an information part that relates to the facts and information that data possesses. Facts and information can be used to improve (business) decision making and therefore improve business performance. Another part is the process part that refers to the methods and actions used to effectively use data to improve decision making. The literature describes this as transforming data into insights or useful information. In the literature we found that analytical tools are important aspect in the process to transform data into insights or useful information. Based on these findings in literature review we adopt the following definition of BI: *The usage of analytical tools to transform data into useful information (or knowledge) to improve decision-making within organizations.*

With this definition we approach business intelligence more as a process and do stress the essence of analytical tools that are used in a business intelligence process. In the next section we will discuss the business intelligence framework, which illustrates how a typical business intelligence process looks like.

2.1.1 Business Intelligence framework

In this section we will use the business intelligence framework to illustrate how a business intelligence process looks like. A typical business intelligence process consists of three parts: data acquisition, data storage and analytics. These three components and their specific parts will be discussed below.

Data capture/acquisition
Before data can be stored, it's needs to be captured from a system which stores data from an operational system (e.g. an SQL server saving data from a production process). To be able to have the data stored, it undergoes an Extraction, Transform and Load (ETL) process. During the ETL process data is extracted from operational systems and it is cleaned to correct missing, inconsistent or invalid values. After extraction, data is transformed into standard formats according to the storage schema. Also business rules can be applied to map the data. The final ETL stage – Load – loads the data into the storage.

Data storage

After data has gone through the ETL process, data is stored in a data warehouse (DW). In a data warehouse, data is specifically structured for query, analysis and decision support. The data's purpose is to support business decisions, not business operations. Data can also be stored in data marts, which are small sized data warehouses created by specific departments to facilitate their own decision support (Khan & Quadri, 2012).

Data access and Analysis

The data access part of a BI system provides business users with an interface while hiding the technical complexity of the data analysis. This interface is provided through the use of BI tools to query the data, to do sophisticated analysis and to visualize the results. The most known discovery techniques are Online Analytical Processing (OLAP) and Data Mining (DM).

OLAP transforms Data Warehouse data into strategic information by using (1) providing multidimensional views of data, by having (2) calculation-intensive capabilities and (3) by having time intelligence (Forsman, 1997). Examples of OLAP are roll-ups (data is summarized with increasing generalization), drill-downs (increasing levels of detail are revealed), slices and dices (performing projection operations on dimensions), and pivoting (cross tabulation) (Khan & Quadri, 2012).

Data mining discovers relationships and patterns in data through comparison, characterization, classification, association and cluster analysis (Han, 1997). The main difference between OLAP and data mining is the fact.

2.2. Big data

The previous section provided a review on the business intelligence. Section 2.2 provides a review on the big data literature. Firstly in section 2.2.1

2.2.1 3V's model

A general accepted description of big data is the 3V's model, which defines big data on its main characteristics: volume, variety, and velocity. The usage of the 3V's model to describe big data is popularized by Gartner. However, before the term big data was even introduced, Laney of Meta Group (now Gartner) already discussed the 3v's of data management in a report from 2001. The report remarks upon the increasing size of data, the increasing rate at which it is produced and the increasing range of formats and representations employed. The model illustrated the growing development of data and data management. Below illustrates how data developed over the years in the three dimensions.

Volume

Volume is the main attribute of big data and refers to the large amounts of data which organizations are willing to use in their big data processes. It is estimated that in 2012 alone, 2.7 zettabyte (ZB) was generated. This amount of generated data will grow by a 40% compound annual rate, passing the 8 ZB mark by 2015 and reaching 45 ZB in 2020 (Gens, 2012). Volume doesn't refer only to the mass amount of bytes of data; it can also be quantified by the amount of records, transactions, tables and files, which a dataset possesses. The scope of big data varies from organization to organization and affects the quantification. Currently, organizations are managing datasets with a volume ranging from terabytes (TB, = 1012 bytes) to petabytes (PB, =1015 bytes). The volume size, including amount of records, transactions, tables and files are growing continuously and the amount that is considered.

Variety

The volume of data grows in variety of forms and comes from various sources. Data can be generated from our activities with internet, sensors, smartphones and other electronic devices and can be in the form of text, video, audio, web data, log files, etc. The variety of data is growing and can come in different structure types. In addition to our use of traditional sources with structured data (e.g. business transactions data), which we already use in traditional business intelligence and analytics, organizations are trying to leverage new data sources that are unstructured (e.g. video, audio, social media data) or semi-structured (e.g. XML, RSS feeds). The amount of unstructured data is growing much faster than the amount of structured data and it is estimated that in 2015 90% of our data will be unstructured (Gens, 2012). These new types of data require new methods to analyze and new technologies to manage, and these are elements which make big data complex.

Velocity

Velocity refers to speed of how fast data is being processed and analyzed. Shorter time between data creation and accessibility results higher levels of velocity. Streams of data can come from sensors of our electronic devices and web data, measuring our activities real-time. Real time data capturing is not really new, and web sites have been collecting clickstream (click activity on web sites) data for years, like search engines and online web stores. However the challenging part is to analyze all these various streams which give us real-time insights that can affect our decision making.

2.2.2 Definitions of big data

In 2012 Gartner formally introduced its definition of big data as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.”. Additional V characteristics to describe big data has been proposed by different sources but aren’t as widely used as the original three V’s. IBM expanded the definition of big data by introducing an additional V for veracity. According to IBM, veracity refers to the level of trust in the information derived from data. Some sources propose value as a characteristic of big data that could be created by using Big Data to enable enhanced decision making, to get insight discovery, and to optimize processes.

Big Data can be described as “massive amounts of stored content (structured or unstructured) that can be easily analyzed in real time (a reasonable amount of time to get a useful answer)” (Arnold, 2012, p. 32). Capgemini views big data in terms of three elements: the data itself, the process for dealing with the data, the holistic view that it can enable.

Boyd and Crawford (2012) observed that big data is often used to refer to the quantity of data, instead they define Big Data as an interplay of the technology which maximizes computational power to gather and compare large data sets, the analysis which identifies patterns to make economic, social, technical or legal claims, and the belief that large data sets offer intelligence with the aura of truth, objectivity and accuracy. According to Gartner: “correlating, analyzing, presenting and embedding insights from structured and unstructured information together enables organizations to better personalize the customer experience and exploit new opportunities for growth, efficiencies, differentiation, innovation and even new business models.”

Big data as an extension to traditional BI

According to Oracle, big data is the inclusion of additional sources to augment existing data analytics operations. Exploiting big data entails dealing with multiple sources and combining structured and unstructured data.

Various sources advices not to dispense existing infrastructures and capabilities of BI, but current capabilities should be integrated with the new requirements of Big Data. Big Data technologies work best in cooperation with the original enterprise data warehouses, as used with Business Intelligence (Nasar & Bomers, 2012). Hence, big data can be seen as an extension to traditional BI.

Big data as a new generation of BI & Analytics

Chen (2012) argues that there are three generations of BI & Analytics (BI&A). Every generation introduces new data types, which require new capabilities to deal with these new data types. The first generation of BI&A applications and research focused on mostly structured data collected by companies through legacy systems and stored in relational database management systems (RDBMS) (H. Chen et al., 2012).

Analytical techniques used in this generation of BI&A are rooted in statistical methods and data mining techniques developed in the 70s and 80s respectively (H. Chen et al., 2012).
The second generation of BI&A is a result of the development of the internet; it encompasses analysis of web-based unstructured content.

The third generation of BI&A is emerging as a result of smartphones, tablets and other sensor-based information supplies and includes analysis on location-based, person-centered and context-relevant analysis (H. Chen et al., 2012). According to Chen, big data refers to the later generations of BI.

The definitions of big data show great similarities with the definition of business intelligence as was defined in section 3.1. Both business and intelligence and big data focus on creating insights or useful information from data. The big data definitions are stressing on the combination of multiple sources of data, combination of structured and unstructured data. These finding resulted in the following definition of big data:

The combination and transformation of structured and unstructured data from multiple sources into useful information (or knowledge) to improve decision-making within organizations.

2.3. Big data technologies

Traditional BI is associated with collection, storage and analysis of structured data. BI stores there structured data in relational database management systems (RDBMS). RDBMS are efficient solutions to store structured data, but are not capable of dealing with the new data types (especially unstructured data).

Big data tries to combine data from different data sources and data types together to provide new insights or useful information. Moreover, since big data deals with large and fast growing amounts of data, redundancy and scalability of storage has become more important.
New technologies like: NoSQL, Hadoop and Mapreduce have emerged, to meet these big data requirements.

Section 2.3.1 will provide a review of the development of database technologies. Database technologies were traditionally focused on storing and processing structured data. Recent development in data base technologies such as NoSQL is focused on storing and processing unstructured data. Section 2.3.2 will provide an overview of the Hadoop framework, which is a popular open-source framework that is focused on storing and processing large volumes of structured and unstructured data. The Hadoop framework uses parallel computing to store and process the large volumes and high variety of data, which is focused on dividing data in smaller parts and processing these smaller parts of data simultaneously.

2.3.1 Database technologies

Early database technologies were optimized to deal with structured forms of data. Big data includes large volumes of unstructured data that were not suitable for traditional database technologies that are focused on structured data. Recent developments in the database technologies, have shown solutions to deal with the requirement of big data, and are focused on dealing with unstructured data as well.

RDBMS

The relational model to store data was originally proposed by Codd in 1970. RDBMS characterize themselves by storing and retrieving data according to the relation data model. Relational databases are often structured and work with a well-defined schema. Data is stored in related tables, in which rows are representing records, and columns representing attributes of those records. The emergence of RDBMS during the 1980s started the revolution in data management.

Due to limitations of the relational model to store and manage large multimedia data objects and to represent complex relationships Object databases were developed. These set of databases provided object support at the database level making it possible to directly store objects using object-oriented programming languages.

NoSQL

NoSQL refers to the databases that don't store data according the relational data model. Over the years we started to generate new types of data that are hard to structure and harmonize. The current RDBMS are not capable of dealing with the new data types and do not provide a convenient way of storing data that is necessary for big data. Therefore, new types of database emerged that dropped the relational model for data storage. Non-relational databases eliminate the need to map data from the database to data structures in a program's memory because an object can be stored and retrieved in any format. As a result, queries from these databases require lower level programming (Pooley et al., 2013). Typical properties of a NoSQL database include scalability, no fixed database schema, a more limited query model compared to traditional databases and optimization for simple read and write operations (Cattell, 2010). The adoption of NoSQL databases was mostly driven by a need to overcome the limitations of relational databases. Three main reasons have been found to be important to the development and adoption of NoSQL databases. The first limitation being high throughput, this used to only be possible by dividing and mirroring the data over multiple expensive servers. Secondly scalability was a key issue, scaling relational databases comes with complex setups and fine-tuning. Also sharing of data has a performance penalty. A third limitation overcome by NoSQL databases is the integration with programming languages.

NoSQL databases can roughly be categorized on four groups based on the storage model that is implemented (Cattell, 2010; Fidelis Cybersecurity Solutions, 2014)

> **Key-Values Databases:** Store un-interpreted arbitrary data values into a system that can be recalled later using a key (hash). This schema less data model allows for easy scaling and very simple APIs for implementations.

Column Databases: Store data in a similar key-value model, except the key is a combination of column, row, and/or timestamp, which points to one or multiple columns (column family). The column family used here is like a table commonly found in a relational database.

Document Databases: Store documents that consist of one or more self-contained named fields in each document, like JSON or BSON format. The structure of documents is dynamic that allows for free modification with the ability to add or remove fields of existing documents. Indexing on the named fields enables fast data retrieval.

Graph Databases: Store data in a flexible graph model that scales across multiple machines. This model is suitable for data with relations that are best represented as a graph (elements interconnected with an undetermined number of relations between them), such as social relations, public transport links, road maps or network topologies.

Google reported running its services on NoSQL database BigTable in 2006 (Chang et al., 2008), Amazon also had most of its services running on Amazon Dynamo in that year (Decandia et al., 2007). Other open source NoSQL databases like Apache Cassandra (Lakshman & Malik, 2010) and HBase where under active development. Although NoSQL databases overcome a number of key limitations related to relational databases, this comes at a cost. Because of the fundamental different design (for easy scalability), NoSQL databases are not ACID (Atomicity, Consistency, Isolation, Durability) compliant, but instead offer BASE (Basically Available, Soft-state Eventual consistency) properties (Pritchett, 2008; Vogels, 2009).

2.3.2 Hadoop framework

Unlike RDBMS and NoSQL, Hadoop is not referring to a type of database, but rather a software platform that allows for massively parallel computing. Hadoop is an open source software framework, which consists of several software modules that are targeted to process big data, large volume and high variety of data. Core modules of the Hadoop ecosystem are Hadoop Distributed File System (HDFS) and Hadoop Mapreduce. Below we describe the most popular modules of the Hadoop framework.

HDFS is the software module that arranges the storage in a Hadoop big data ecosystem. HDFS breaks down data into pieces and distributes these pieces to multiple nodes of physical data storage in a system. The main advantages of HDFS are that it is designed to be scalable and fault tolerant. Additionally, by dividing data into pieces HDFS prepares data for parallel processing. Other modules in the Hadoop framework are designed to take advantage of distributed data over multiple nodes.

Mapreduce is a software framework that provides a programming language that takes full advantage of parallel processing. Tasks that programmed in Mapreduce are divided in smaller tasks, which are sent to the relevant nodes in the system. The Mapreduce framework takes care of the whole process: managing communication between nodes, running tasks in parallel and providing redundancy and fault-tolerance.

HBase is a software module that runs as non-relational database on top of HDFS. HBase is NoSQL database that stores data according a key-value model. As it is a NoSQL type of database it requires low level programming to query. Like other software modules of Hadoop, HBase is open-source and is modeled after Google's BigTable database (Chang et al., 2008).

Hive is essentially a data warehouse that runs on top of HDFS. Hive structures data into concepts like tables, columns, rows and partitions, similar to a relational database. Data in a Hive database can be queried using (limited) SQL like language, named HiveQL.

2.4. Conclusion

This chapter answered the first research question of this thesis:

RQ1: What is big data and what are the technologies related to big data and how does it differ from business intelligence?

Firstly, we discussed the definition of business intelligence. We started the literature reviews with business intelligence, since business intelligence is often considered as a predecessor of big data. Early definitions of business intelligence show the usefulness of data in making decisions. Data possess facts or other useful information, and business intelligence is about using facts rather than intuition to make decision.

Later definitions also approached business intelligence as a process and were focused on how to capture, access, understand, analyze data and transform it into insight or useful information. The business intelligence literature also stressed the usage of analytical tools was a key element to transform data into useful information.

The literature review on business intelligence concluded with the following definition of business intelligence for this research: *The usage of analytical tools to transform data into useful information (or knowledge) to improve decision-making within organizations.*

Furthermore, in Section 2.2 we provided a review of the big data literature. In section 2.2.1 the 3V's model of big data was discussed. The 3V's model is a common definition of big data in the literature. The 3V's model mainly describes how data developed over the years. According to the 3V's model, data is growing in the dimension volume, variety and velocity. The 3V's model highlights that the amount of data that is being generated nowadays is increasing significantly and that the data generated nowadays also appears in various types at various data sources. From the large volumes of data generated nowadays the portion of data that is considered unstructured has grown faster than the portion of structured data. The 3V's model stresses the growing significance of unstructured data.

In section 2.2.2 we discussed other definitions of big data. The definitions discussed in section 2.2.2 have shown great similarities with the business intelligence definition as we defined in this research. Some experts in the literature, view big data as an extension to business intelligence or as new generation of business intelligence. In essential both big data is focused on transforming data into insights or useful information. However, big data differs from business intelligence as big data is focused on transforming data with the 3V's characteristics into insights or useful information.

Moreover, we discussed that the volume of data have increased significantly and that data is nowadays often found in unstructured form. Large volumes of data and unstructured data are difficult to store and process with existing infrastructure and technologies that was used with traditional business intelligence. New technologies such as NOSQL and Hadoop that deal with the new forms (semi-structured and unstructured) of data, has been discussed in this chapter.

Some experts in the literature, view big data as an extension to business intelligence or as new generation of business intelligence. From this view, we learned that there is a need to acquire new types of data and new sources of data to enhance traditional business intelligence. A lot of data is generated nowadays. However, a large percentage of the generated data is not used by any form analytics and therefore not yet transformed to new insights. The term big data is often associated with trying to capture the insights from the still unused data sources. This leads to the first aspect of big data.

1. **Need to capture more insights from data**. The need to acquire, storage and process more data. Combining new sources or new types of data with existing sources and data types. To create new insights.

Moreover, we discussed that the new types of data are now often semi-structured or unstructured and can exist in large size or large volume. Semi-structured and unstructured data are difficult to process with traditional database and analysis technologies. Some technologies are required to store and process unstructured data. Additionally, the large volume of big data makes it difficult to capture with existing infrastructure and technologies that was used with traditional business intelligence, mainly because they are not designed for large volumes. Big data technologies such NoSQL and Hadoop are designed to deal with large volumes of data as they offer better scalability and parallelization   New technologies to deal with the new types of data has been discussed in this chapter and is viewed as an important aspect of big data. These leads to next aspect of big data

2. **Big data technologies**. New infrastructure and technologies that enables the acquisition, storage and processing (3V data), to be done more effectively, efficiently, easily.

Finally, in relation with the third aspect of big data, advanced analytics is used to gain insights from the new types of data. When dealing with larger volumes of big data and data with complex structures (or unstructured data), it becomes more difficult for human beings to read and extract information from data. Big data analytics refers to advanced analytical tools and algorithms that can assist human beings to extract important insights from large volumes of unstructured data.

3. **Big data analytics**. Advanced analytics can assist in extracting insights from large volumes and/or unstructured data. Insights those are difficult or impossible to extract without the right tools.

3. Literature review: Privacy

In this section we will discuss several conceptions of privacy. We will review four theories of privacy and conceptualize a privacy model that can help us to analyze privacy in this study. In section 3.1 we will discuss the nonintrusion theory. Section 3.2 discusses the seclusion theory. Section 3.3 discusses the control theory. Section 3.4 discusses the restricted access theory.

In section 3.5 we try to combine the four privacy theories and explain the conceptualization of privacy in this research, which finally results in the conceptual model of privacy in this research.

3.1. Nonintrusion theory

One theoretical view on privacy is the nonintrusion theory of privacy. Warren & Brandeis (1890) defined privacy as the "right of being let alone" and "being free from intrusion". Warren & Brandeis discussed in their article their concerns about the impact of portable instantaneous photo camera, which was a technological innovation during that time, at the end of the 19th century. According the authors the instantaneous photo camera allows people to intrude one another's personal space, which could be harmful. The authors, defined privacy as a right and they argued for this right to be recognized in legal terms. The nonintrustion theory illustrates that privacy is important in defending liberty. However, this theory fails to distinguish the concept of privacy from liberty (J. Moor, 1990; Tavani, 1999). Privacy and liberty are closely related concepts but should be distinguished. Critics pointed out that it is possible one's liberty can be denied (by not letting him or her alone), but still have privacy (J. H. Moor, 1997; J. Moor, 1990; Tavani, 1999). Moor (1990) supported this criticism with the example of a normal conversation on a public street.

> *If A approaches B on a public street and A asks B what time it is, A has not let B alone but neither has A invaded B's privacy.*

The opposite is also true, as one can be left alone and still has no privacy.

> *If unknown to B and without B's permission, A looks through B's personal files, then A has invaded BS privacy, but, strictly speaking, A has let B alone.*

3.2. Seclusion theory

Another conception of privacy is the seclusion theory, which defines privacy as "being alone" (Tavani, 1999; Alan F Westin, 2003). This theory describes privacy as the voluntary physical withdrawal of a person from the general society in a state of solitude (Tavani, 1999; Alan F Westin, 2003). This theory successfully separates privacy from liberty. However, this theory mixes the concept privacy with solitude, and implicates that one that excludes himself from society more often has more privacy.

3.3. Control theory

The control theory of privacy, defines privacy in terms of control one has over information about oneself (Tavani & Moor, 2001; Tavani, 1999; A F Westin, 1967). According to Fried (1984) "Privacy is not simply an absence of information about us in the minds of others, rather it is the control we have over information about ourselves". Westin (1967) states that privacy is the claim that individuals and groups determine for themselves when, how, and to what extent information about them is communicated to others. Altman (1975) defined general privacy as "the selective control of access to the self". Margulis (2011) elaborated: "Privacy, as a whole or in part, represents the control of transactions between person(s) and other(s), the ultimate aim of which is to enhance autonomy and/or to minimize vulnerability". Early works of control theory defined privacy purely as the control or the ability to control information. Moor (1990) argues that these early definitions emphasizing control are inadequate for there are many situations in which people have no control over the exchange of personal information about themselves but in which there is no loss of privacy. One can autonomously decide to abdicate all informational privacy interests by disclosing all private facts about herself from which she had an interest or right to exclude other people, in this case privacy can be confused with autonomy. Later works of control theory of privacy, however, also argue that control is an important factor to shape privacy (J. Moor, 1990; Smith, Dinev, & Xu, 2011).

3.4. Restricted access theory

According to the restricted access theory, privacy is a matter of the restricted access to persons of information about persons (Allen, 1988; Bok, 1989; Gavison, 1980a). According to Bok (1989) "privacy is the condition of being protected from unwanted access by others, including access to one's personal information." This view treats privacy as a condition, where in the traditional privacy conceptions of nonintrusion privacy was viewed as a right. Gavison (1980b) defines "privacy as a measure of the access others have to you through information, attention, and physical proximity." Gavison's definition suggests that the condition of privacy is not absolute, but different levels of privacy exist according to degree of access others have to one's personal information.

Moor (1990) states that: "an individual or group has privacy in a situation if and only if in that situation the individual or group or information related to the individual or group is protected from intrusion, observation, and surveillance by others". In this definition Moor deliberately selected the vague term situation, which normally would refer to an activity, relation or a location. According to Moor, a situation can be private in two ways: a naturally private situation or normative private situation.

In a naturally private situation, individuals are protected by natural means from access, interference, or intrusion by others. Moor provided an example of hiking alone in the woods in which the natural setting of the woods forms a boundary. If in such a situation, a stranger enters the same woods and sees the hiker, privacy of the hiker can be lost but not is not necessarily violated, since there are no legal or ethical norms that protect one's privacy in such situation (J. H. Moor, 1997; J. Moor, 1990; Tavani & Moor, 2001). Normatively private situations can include "activities" such as voting, "information" such as medical records, and "locations" such as a person's house. In these situations, people's privacy is protected by legal and ethical norms.

3.5. Conception of privacy in this research

Several theories and several definitions of privacy have been discussed in the previous sections (3.1 – 3.4). The literature review has presented different definitions of privacy. In this section, the conception of privacy that will be used on this research will be formulated. This conception is based on the notions that were introduced with the different theories.

3.5.1 Informational privacy

The nonintrusion theory taught us that intrusion to our personal space can be harmful. Privacy was defined as a right that is necessary to protect people against this intrusion. Traditional concepts of privacy were concerned to a more physical view of intrusion. The large impact of the development of informational technology changed traditional concepts of privacy. Modern concepts of privacy are more concerned with information, due to huge impact of information technology on our society. The term "informational privacy" is nowadays used to refer to information-related privacy concerns as distinct category, which includes concerns as access to personal information stored in computer databases (Himma & Tavani, 2008; J. Moor, 1990; Parent, 1983; Tavani, 1999). According to Parent (1983), informational privacy refers to the "protection against the misuse of personal, sensitive information". In this research we will adopt a concept of privacy that concerns personal, sensitive information. The term privacy will be used mostly to refer to what we just described as informational privacy.

3.5.2 Private situations

Moor (J. Moor, 1990) introduced the notion of private situations in which he made the distinction between normative private situations and natural private situations (Figure 5).

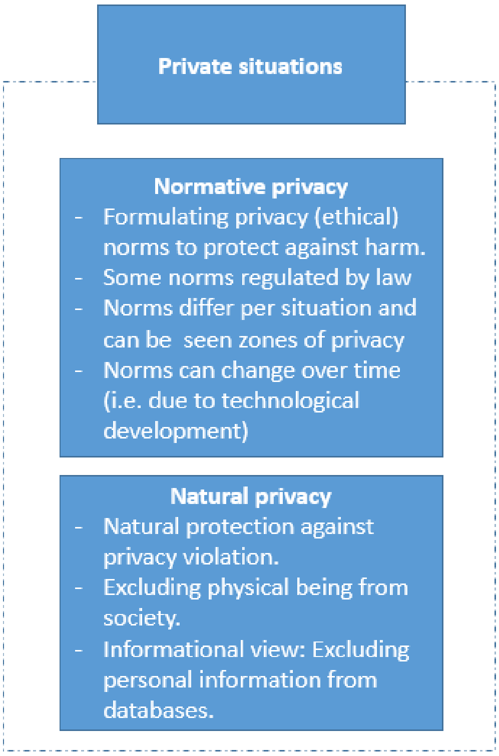

Private situations
Normative privacy
- Formulating privacy (ethical) norms to protect against harm.
- Some norms regulated by law
- Norms differ per situation and can be seen zones of privacy
- Norms can change over time (i.e. due to technological development)
Natural privacy
- Natural protection against privacy violation.
- Excluding physical being from society.
- Informational view: Excluding personal information from databases.

Figure 5: Private Situations: Normative and Natural Privacy

Natural private situation can be linked to the seclusion theory of privacy, the theory which defined privacy as: "being alone". Even though we don't adopt this definition of privacy, it is understandable that this early theory of privacy refers to the notion of "being alone". The relation between privacy and "being alone" can be explained by the fact that seclusion (or exclusion) can contribute to restricting access to one's physical self or his/her personal information. For example, when someone in a public area surrounded by other people is receiving a call, he or she can decide to move away and separated from the others. The person in this example attempt to restrict physical access to his/her voice from others, which will make it more difficult and therefore make it less likely that the others can eavesdrop the conversation. The notion of private situations teaches us that seclusion or exclusion can provide some privacy in the form of natural privacy.

Normative privacy is built upon the rights of privacy. People establish norms in which we define what is potentially harmful and what is acceptable in the context of intrusion. In other words people define privacy norms to define what is considered a violation to privacy. What is considered as violation depends on the situation or context (Barth, Datta, Mitchell, & Nissenbaum, 2006; J. Moor, 1990; Nissenbaum, 2004).

Norms can be defined by a collective of individuals into social norms. In the case of the healthcare sector, social privacy norms can be referred to the privacy and security demands by patients.

Modern privacy laws are implemented as data protection laws. These laws translate privacy norms into organizational and technical requirements, for any situation that deals with. These organizational and technical requirements are aiming at ensure the protection of personal data that is stored and/or processed in information systems.

According to Moor, norms are influenced by culture. Also, Moor stated that privacy norms can change over time. One reason why privacy norms change is due to the impact of ***technological development***. It was the invention of the instantaneous compact cameras during the end of the 19th century, which was the key for the privacy concerns raised by Warren & Brandeis. Nowadays, in a period that is being dominated by rapid information technology development, we try to regulate our privacy norms through data protection laws as we see.

3.5.3 Private situations can be both natural and normative at the same time.

In the previous section (3.5.2) we introduced the notion of private situations as was formulate by Moor. According to Moor private situations are either natural private or normative private. For the conceptualization of privacy in this research, it is argued that private situations can be both natural and normative protected at the same time.

An example from an informational view of privacy could be the exclusion to store personal information on a database. When the database doesn't contain one's personal information it is impossible for others to access his or her personal information through that database. In this example, the fact that the personal information is not present on the database could be seen as the natural border that restricts access to personal information. More information available in databases lowers the natural privacy in such situation. As well, normative privacy is present in the same situation. Modern privacy laws are prescribing data processing norms in order to reduce the chance of unwanted access to personal data that are stored in databases and therefore form a normative border that restricts access to personal information.

The important distinction between normative privacy and natural privacy is that natural privacy naturally contributes to the condition of privacy, in contrast with normative privacy which does not directly contribute to the condition of privacy. To what degree normative privacy contribute to the condition of privacy, depends on the control aspect of privacy.

3.5.4 Control aspect of privacy

Early works of control theory of privacy defined privacy as the control one has over information about itself (Altman, 1975; A F Westin, 1967). This definition is not adopted in the conception of privacy in this research. Instead, in this research, a view is accepted that distinguishes control from privacy. Later works of control theory treated control as separate factor that shapes privacy (Smith et al., 2011). For this research we elaborate on this definition by stating that control is a factor that is involved in the link between privacy norms and the condition of privacy (Figure 6). Control is used to enforce privacy norms in order to restrict unwanted access to personal information. In this research we conceptualize control as a factor that transforms privacy norms into the condition of privacy. Control exists in the form of sanction policy and technology. In this research, we will focus on information access control systems in the healthcare.

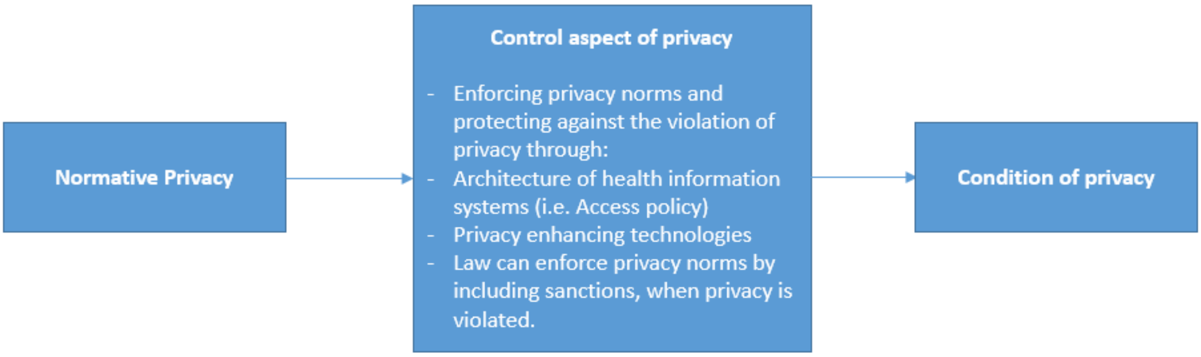

Normative Privacy
Control aspect of privacy
- Enforcing privacy norms and protecting against the violation of privacy through:
- Architecture of health information systems (i.e. Access policy)
- Privacy enhancing technologies
- Law can enforce privacy norms by including sanctions, when privacy is violated.
Condition of privacy

Figure 6: Control aspect of privacy

3.5.5 Condition of privacy

The restricted access theory defined privacy as a condition. Bok (1989) stated “privacy is the condition of being protected from unwanted access by others, including access to one's personal information.”

We accept this definition of the condition of privacy. We elaborate on this definition by stating that the condition of privacy finally leads to the amount of privacy one enjoys. The condition of privacy is not absolute but can exist in various levels of degree. This means that condition of privacy is equal to *the level of degree that restricts unwanted access by others to one’s personal*. The condition of privacy in a certain situation is a combination of the effective amount of normative privacy after control and the natural privacy.

3.5.6 Conceptual model of privacy

Figure 7 presents the conception of privacy in this research. We initially introduced two private situations as either normative privacy or natural privacy. Natural aspect of private situations protects an individual's privacy by natural means. Normative aspect of private situations does not protect an individual from privacy violation by natural means; instead privacy norms protect an individual from privacy violations. We defined the condition of privacy as the condition of being protected from unwanted access by others, including access to one's personal information. Natural privacy directly contributes to the condition of privacy. Normative privacy doesn't directly contribute to the condition of privacy, as norms can be violated. The control aspect of privacy refers to policy and technological systems that enforce privacy norms, the effectiveness of control determine to what degree, normative privacy contribute to the condition of privacy.

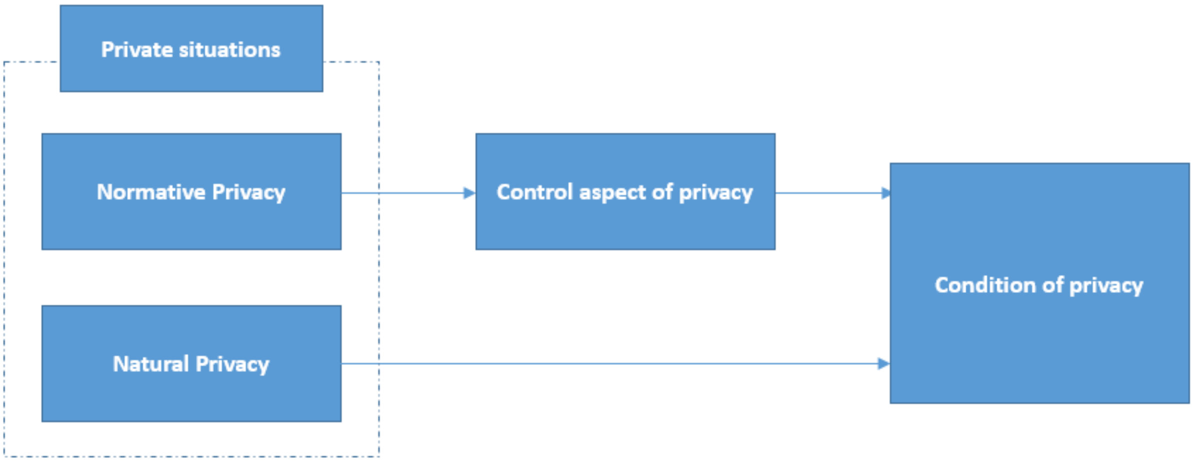


Figure 7: The conception of privacy in this research. Privacy situation consists of a natural aspect that directly contributes to the condition of privacy and a normative aspect of privacy that contributes to the condition of privacy depending control aspect of privacy.

In this research we investigate the impact of big data on normative privacy, natural privacy and the control aspect of privacy. The impact on the different aspect of privacy will be investigated in the context of healthcare.

We found in the literature big data stimulates the need to capture and combine more data types and more data sources. Does this mean that, because more (potential) private information is captured, our natural privacy decreases? Are we then more reliant on privacy norms to protect us in normative private situations?

3.6. Conclusion

Chapter 3 had the purpose to answer the second research question of this thesis:

RQ2: What are relevant theories to analyze the concept of privacy?

In Chapter 3 we discussed four theories and several definitions of privacy. *The nonintrusion theory* defines privacy as the "right of being let alone" (Warren & Brandeis, 1890). *The seclusion theory* defines privacy as "being alone" (Tavani, 1999; Alan F Westin, 2003). *The control theory* defines privacy in terms of control one has over information about oneself (Tavani & Moor, 2001; Tavani, 1999; A F Westin, 1967). According to Fried (1984) "Privacy is not simply an absence of information about us in the minds of others, rather it is the control we have over information about ourselves". *The restricted access theory* defines privacy as "the condition of being protected from unwanted access by others, including access to one's personal information" (Bok, 1989). Additionally, we discussed the notion of private situations, which distinguish normative private situation from natural private situations.

Section 3.5 presented the conceptualization of privacy for this research. The conceptualization for this research, as shown in Figure 8, attempts to combine the different aspects of privacy. The remaining part of this research will mainly focus on natural privacy, normative privacy and control of privacy.

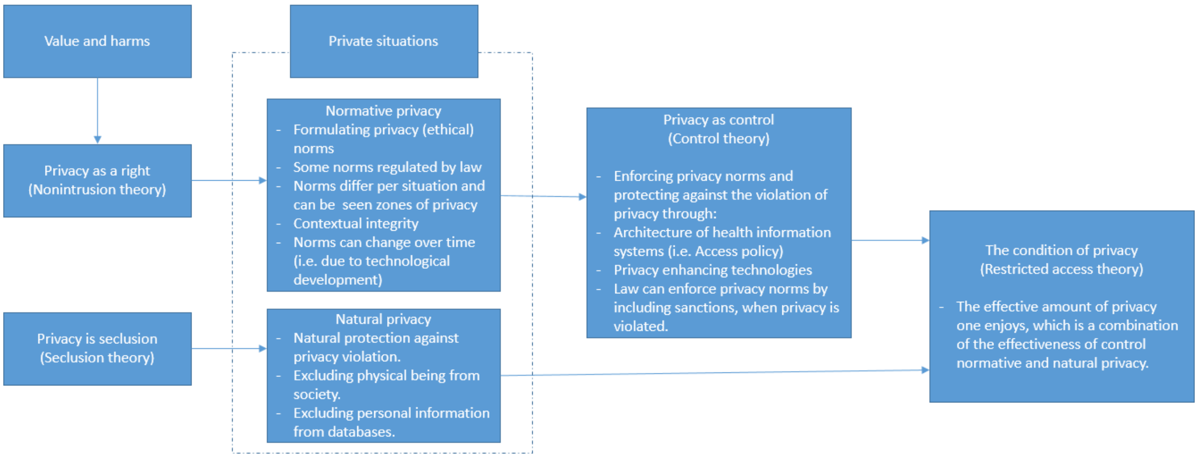

Value and harms
Private situations
Privacy as a right
(Nonintrusion theory)
Normative privacy
- Formulating privacy (ethical) norms
- Some norms regulated by law
- Norms differ per situation and can be seen zones of privacy
- Contextual integrity
- Norms can change over time (i.e. due to technological development)
Privacy as control
(Control theory)
- Enforcing privacy norms and protecting against the violation of privacy through:
- Architecture of health information systems (i.e. Access policy)
- Privacy enhancing technologies
- Law can enforce privacy norms by including sanctions, when privacy is violated.
The condition of privacy
(Restricted access theory)
- The effective amount of privacy one enjoys, which is a combination of the effectiveness of control normative and natural privacy.
Privacy is seclusion
(Seclusion theory)
Natural privacy
- Natural protection against privacy violation.
- Excluding physical being from society.
- Excluding personal information from databases.

Figure 8: Conceptualization of privacy in this research.

4. Literature review: Privacy in the healthcare

Nowadays, these health records are stored electronically in information systems. The use of information systems healthcare is viewed as an important factor to improve the healthcare. Health information systems facilitates the sharing of health information among healthcare employees like physicians and nurses, and thereby improve the efficiency and quality of healthcare (Perera, Holbrook, Thabane, Foster, & Willison, 2011). Several studies, have examined the impact, efficiency and contributions of the use of information systems in the healthcare. For example, information could be used to improve efficiency within the healthcare system, drive public policy development and administration, and in the conduct of medical research (Hodge, 2003).

In this chapter we will provide some background information on privacy in health care. Furthermore, this chapter will review legislation related to privacy in the health care. Privacy is strictly regulated by legislation and therefore it essential study the legislation of privacy.

Privacy in the healthcare concerns the data that are processed on health information systems. Health information systems process the data that are essential for healthcare processes. Modern privacy legislations are mainly regulating what type data is considered private data and are regulating which requirements organization need to fulfill the organization is allowed to process private data.

However, these health information systems include privacy sensitive data which include personal information and medical information, such as Electronic Health Records (EHR). The privacy sensitive characteristics of health information, requires extra attention concerning the storage and the processing of such information. From the perspective to protect privacy, it is important that health information remains confidential and disclosure of such information only happens with the patient's consent.

The type of data that is processed on health information systems contains medical information which are considered privacy sensitive data. Unwanted access to this private data, could lead to usage of private information for purposes other than essential to the health care processes. In section 4.1 we describe the different types of information systems used in the health care. An overview of the different applications of health information and the type of information they process is provided in this section. In section 4.2 we elaborate more on the electronic health record (EHR) systems, which is a type of information system that processes information such a patient’s medical history that is often used for decision support. Section 4.3 discusses the healthcare privacy legislation in the Netherlands. Section 4.4 discusses information access control in the healthcare.

4.1. Types of health information systems

An information system (IS) "is an arrangement of information (data), processes, people, and information technology that interact to collect, process, store, and provide as output the information needed to support the organization" (Whitten & Bentley, 2007). We refer to a health information system, when an information system is used in the healthcare industry to support healthcare organizations. Health organizations in the healthcare include groups such as payers, government, medical research institutes, patients, and healthcare providers. In this research we will focus on healthcare providers as healthcare organization, since essential health-related services are provided by this group of actors. Within this group healthcare providers we mainly refer to hospitals, since the majority of the health-related services are provided by specifically this type of organization.

Healthcare information can be categorized on their purpose and type of information they process.

Table 1 presents an overview of different types of health information systems. We can roughly categorize health information systems into administrative applications and clinical applications. Administrative health information systems processes mainly administrative data and financial to support general operation and management function of the healthcare organizations. These types of systems contain information for the purpose of managing finance, personnel, equipment, etc.

Clinical health information systems on the other hand, processes mainly health-related information. Clinical health information systems assist healthcare providers to diagnose, treat and monitor patients and the patients' healthcare. These clinical systems may be departmental systems and limited to a certain scope of clinical information, such as radiology, pharmacy and laboratory systems. Clinical information systems may also be clinical decision-support, medication administration, or electronic health record systems.

Table 1: Types of Information Systems in Healthcare (Wager, Lee, & Glaser, 2013)

Administrative applications	Clinical Applications
Patient administration systems	***Ancillary information systems***
- **Admission, discharge, transfer (ADT)** tracks the patient's movement of care in an inpatient setting. - **Registration** may be coupled with ADT system; includes patient demographic and insurance information as well as date of visit(s), provider information - **Scheduling** aids in the scheduling of patient visits; includes information on patients, providers, date and time of visit, rooms, equipment, other resources - **Patient billing or accounts receivable** includes all information needed to submit claims and monitor submission and reimbursement status - **Utilization management** tracks use and appropriateness of care	- **Laboratory information** supports collection, verification, and reporting of laboratory tests - **Radiology information** supports digital image generation (picture archiving and communication systems [PACS]), image analysis, image management - **Pharmacy information** supports medication ordering, dispensing, and inventory control; drug compatibility checks; allergy screening; medication administration
Other administrative and financial systems	***Other clinical information systems***
- **Accounts payable** monitors money owed to other organizations for purchased products and services - **General ledger** monitors general financial management and reporting - **Personnel management** manages human resource information for staff, including salaries, benefits, education, and training - **Materials management** monitors ordering and inventory of supplies, equipment needs, and maintenance - **Payroll** manages information about staff salaries, payroll deductions, tax withholding, and pay status - **Staff scheduling** assists in scheduling and monitoring staffing needs - **Staff time and attendance** tracks employee work schedules and attendance - **Revenue cycle management** monitors the entire flow of revenue generation from charge capture to patient collection; generally relies on integration of a host of administrative and financial applications	- **Nursing documentation** facilitates nursing documentation from assessment to evaluation, patient care decision support (care planning, assessment, flow-sheet charting, patient acuity, patient education) - **Electronic health record (EHR)** facilitates electronic capture and reporting of patient's health history, problem lists, treatment and outcomes; allows clinicians to document clinical findings, progress notes, and other patient information; provides decision-support tools and reminders and alerts - **Computerized provider order entry (CPOE)** enables clinicians to directly enter orders electronically and access decision-support tools and clinical care guidelines and protocols - **Telemedicine and tele health** supports remote delivery of care; common features include image capture and transmission, voice and video conferencing, text messaging - **Rehabilitation service documentation** supports the capturing and reporting of occupational therapy, physical therapy, and speech pathology services - **Medication administration** is typically used by nurses to document medication given, dose, and time

4.2. Electronic health records (EHR)

As Table 1 presents, several clinical applications of health information systems exist within the healthcare organization. The core health information system within the healthcare organization is the electronic health record (EHR) system. Patient medical records are used by healthcare organizations for documenting patient care, as a communication tool for those involved in the patient's care, and to support reimbursement and research. Traditionally these health records were stored in paper-based form. These paper-based health records generate large paper trails and studies showed issues with these paper-based records. These records can often be incomplete or unavailable when and where they are needed. In addition, electronically stored health records can improve information sharing and communication. Information in these records can easily be transferred without the requirement of a physical copy. Applications developed to process and analyze EHR, assist healthcare staff to track patient's data over time. Several studies suggest that there is growing evidence that EHR systems help improve the quality and efficiency of care (Perera et al., 2011).

The term electronic health record (EHR) is often used interchangeably, with the terms electronic medical record (EMR) and personal health record (PHR). However, in strict definitions they are not the same. In a report of 2008, the National Alliance for Health Information Technology (NAHIT) defined the terms EMR, EHR and PHR as follows:

EMR: The electronic record of health-related information on an individual that is created, gathered, managed, and consulted by licensed clinicians and staff from a single health care organization who are involved in the individual's health and care.

EHR: The aggregate electronic record of health-related information on an individual that is created and gathered cumulatively across more than one healthcare organization and is managed and consulted by licensed clinicians and staff involved in the individual's health and care.

PHR: An electronic, cumulative record of health-related information on an individual, drawn from multiple sources, that is created, gathered, and managed by the individual. The integrity of the data in the PHR and control of access to that data is the responsibility of the individual.

.

4.3. Healthcare privacy legislation

In this section we will discuss healthcare privacy legislation for the Netherlands. In the Netherlands there are three important laws that regulate privacy in the context of the healthcare. The three laws are the *Wet bescherming persoonsgegevens* (Wbp), *Wet op de geneeskundige behandelingsovereenkomst* (WGBO) and *Zorgverzekeringswet* (Zvw).

4.3.1. Wet bescherming persoonsgegevens (Wbp)

The *wet bescherming persoonsgegevens* (Wbp), ("Privacy law"), is a Dutch that protects privacy of individuals, by regulating how personal information (data) should be processed. This law is the Dutch implementation of the *EU Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data*.

Wbp applies for any situation where personal information is being processed, with the exception of information processed by governmental organizations related to security and intelligence. Information processing in the healthcare is a special situation and has additional regulation in healthcare laws.

The notion **processing** refers to "any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction".

The Wbp defines information as **personal information** if it *(1) consists information about a natural person (a real person)* and *(2) the person is identifiable*.

Information about a natural person can be factual information such as: the IQ of a person, and the medical history of a person. Additionally, information about objects can also be considered personal information, when it can be related to an identifiable person. Examples of such information are: the value of someone's care or the healthcare expenditure of a person.
According to the definition of the Wbp, a person is identifiable when: the identity of the person can be determined without disproportional effort. The information can be directly related (name, date of birth, address), or indirectly (in combination with other information) to a real person.

Some personal information is considered even more sensitive as processing of this information can cause serious privacy violations. The Wbp define this as *bijzondere persoonsgegevens* ("**sensitive personal information**"), which have even more strict regulations. Sensitive personal information includes information concerning: religious beliefs, political opinions, **health**, sexual orientation, race, membership of past organizations.

The Wbp defines the following requirements on data processing:

- There must be a relation between the registered data and the purpose. Other usage is prohibited, unless the person concerned provides permission for to process his/her data.
- It is not allowed to process more information than is required for the purpose for which the data is collected.
- Data cannot be stored longer than strictly necessary. Health care providers are required to store medical data for fifteen years
- Adequate organizational and technical measures should be taken to the data.
- In most cases organizations are required to inform when they are processing private data. In the case of medical data, health care providers are not required to inform patients.
- Specific for medical data, it prohibited to process medical data with the exception for healthcare providers and care worker in the social domain when it is required to provide healthcare.

Furthermore, the Wbp gives individuals certain rights. For example, individuals always have the right to know what is happening with their personal information. Individuals are allowed to view their personal information and can request correction of wrong information or object against the processing of their personal information.

The Wbp designated the College Bescherming Persoonsgegevens (CBP) as the regulatory authority that supervises organizations and exercise the privacy law. All organizations that are processing data have to inform the CBP.

4.3.2. Wet op de geneeskundige behandelingsovereenkomst (WGBO)

The Dutch law *wet op de geneeskundige behandelingsovereenkomst* (WGBO) regulates the healthcare treatment agreement between patient and healthcare provider. WGBO also covers the regulation of health records. Healthcare providers are required to maintain a health record of the patients they are treating (Article 454). The health records should include information about patient's health, the treatments the patient received, and other information which are necessary for healthcare processes. Article 455 states that healthcare providers are required to destroy the health records on patient's request. The patient has the right to view the content of his or her health record (Article 456). Others are not allowed to view the content of the health record, with the exception of people directly involved with the treatment of the patient (Article 457). Another exception allows the content of the health records to be accessed without patient's permission for certain statistical or scientific research purposes (Article 458).

4.3.3. Zorgverzekeringswet (Zvw)

Since 2006, healthcare in the Netherlands is financed by two systems: the *Algemene Wet Bijzondere Ziektekosten* (AWBZ), ("General Law on Exceptional Healthcare Costs"), and the *Zorgverzekeringswet* (Zvw), ("Healthcare insurance law"). The AWBZ is a state-controlled mandatory insurance system that covers long-term healthcare for people with a handicap, chronic disease, and long-term healthcare for elderly people.

The Zvw covers short-term healthcare which is financed by a health insurance system with private health insurance companies. The Zvw requires every individual who enjoys income in the Netherlands to take at least a basic health insurance package with one of the private health insurance companies of the system. The basic health insurance package consists of a set of insured treatments, which are defined by the Dutch government. The basic package covers basic healthcare including: GP, hospital and pharmacy.

In this healthcare system, the healthcare insurance companies are positioned between the patients and the healthcare providers in order to finance healthcare. Hence, health insurance companies are required to process information such as patient treatment and billing.

Chapter 7 of the Zwv regulates the information flow within the Zvw system. Healthcare providers are required to share personal information concerning the health (Article 87). Personal information should be shared with designated healthcare related governmental organizations, when requested (Article 88). The health related governmental organization *college zorgverzekeringen* regulates which personal information should be shared (Article 90). In addition, *college zorgverzekeringen* is also responsible for the electronic infrastructure that facilitates the sharing of this personal information, how personal information is shared and the maintenance of the shared databases that stores this personal information.

4.3.4 Privacy laws are becoming stricter

The current trend is that privacy laws are becoming stricter. An extension to the Dutch privacy law (Wbp) is about to be accepted that introduces two major changes. Firstly, the extension requires organizations that are processing private data to report data breaches. In case of a data breach that violated the required security measures (Wbp article 13) the organization responsible for the processing of private data are required to report that breach to the regulatory authority of privacy (CBP). When a breach has serious harmful consequences to the ones concerning the private data, the responsible organization of data processing are also required to inform the concerned ones. The second major change to Dutch privacy law will grant the Dutch authority body for privacy (CBP) more power and will be able to sanction organizations with fines of an amount up to €810.000.

At a European level, a successor of the current privacy directive (Directive 95/46/EC) is in development under the name General Data Protection Regulation (GDPR). The GDPR will be a regulation rather than a directive, this means that this single regulation will apply to all member states of the EU. The new privacy regulation is expected to be active from 2018. The GDPR will be stricter in several aspects.

Currently, organizations already need consent from the concerned individuals to process private data. The GDPR will require more explicit consent, which means that consent will have to be more explicit on specifying the types of the particular data and the specific purpose they may be used for.

The rights of privacy for individuals will be improved which allows individuals to have more control of what happens with their private data.

The GDPR will also introduce a new supervisory authority at European level that will have to power to sanction organizations with fines up to an amount up to 1.000.000 and in cases of enterprises even up to an amount of 5% of annual turnover of the organization. Table 2 shows an overview of these findings.

Table 2: Main findings on privacy law and regulation

Main findings on privacy law and regulation	
Dutch Privacy Law (Wbp)	Dutch privacy law is becoming stricter. Extension to Dutch privacy law is about to accepted that: - Requires organization (that are processing private data) to report data breaches. - Gives more power to authority body for privacy (CBP). CBP can fine organizations that are violating Dutch privacy law with fine of an amount up to €810.000.
EU directive 95/46/EC and General data protection regulation	General data protection regulation (GDPR) is in development that will succeed the EU privacy directive (Directive 95/46/EC). Recent updates indicate a much stricter privacy regulation. - It is a regulation rather than directive and it will apply to all EU member states. - Creation of EU supervisory authority of privacy that will have more power and have the ability to fine organizations with fines up to an amount up to €100.000.000, for enterprises even up to 5% of annual turnover. - More individual rights of privacy

4.3.5 Conclusion

Three Dutch laws that regulate privacy in the context of the healthcare have been examined. The Wbp provides restrictions and requirements for any organization that is processing personal information. The Wbp defines information as personal information if it *(1) consists information about a natural person (a real person)* and *(2) the person is identifiable*.

Personal information that concerns health is considered sensitive personal information and has additional restrictions and requirements for processing. It is inevitable for health care organization to process health data. The WGBO requires healthcare providers to maintain a health record of patients they are treating. While Zvw regulates the information flow within the Zvw system.

The Dutch privacy law is becoming stricter. Eventually, Dutch privacy law will be replaced by European privacy. The current updates on European privacy laws indicate that European privacy law will become stricter. As part of this research, we would like to explore how the developments in privacy laws relate to big data in the health care.

4.4. Information access control

In the previous sections (4.1 - 4.3) we discussed what privacy is in the health care sector. Privacy in the healthcare concerns the data that are processed on health information systems. Health information systems process the data that are essential for healthcare processes. The type of data that is processed on health information systems contains medical information which is considered private sensitive data. Unwanted access to this private data, could lead to usage of private information for purposes other than essential to the health care processes. At the same time health care staff (e.g. doctors, nurses) might need to have access to medical information stored in the health care information systems. In this section (4.4) we will provide a review of the information access control literature in the context of the health care. The aim of this section is to gather knowledge between information access control and privacy in the health care. In this section we will discuss the most common access control model: the Role Based Access Control (RBAC). Moreover, we will discuss a-priori and a-posteriori access control. In the health care it is standard to have an emergency access control system (a-posteriori) besides a standard access control such as Role Based Access Control (a-priori).

4.4.1 Role based access control (RBAC)

The concept of role based access control (RBAC) has originally been developed for the purpose of managing resources on multi-user and multi-application on-line system. The main idea behind RBAC is that access permissions to resources are based on roles rather than individual users. Based on the various job functions within the organizations, roles are defined, and based on the responsibilities of the role, the required access permission are assigned to the role. RBAC simplifies the management of permissions as RBAC allows representation of the natural organizational view upon access permissions. When a user changes function within the organization, the user can easily be reassigned to other role in within RBAC, and adjustments in permissions for everyone in a role can be adjusted as access requirements for functions in the organizations are changing.

Due to its success, RBAC has found its way into many organizations, including healthcare organizations. However, researchers believe that Traditional RBAC is not suitable for the healthcare, as traditional RBAC systems lacks the flexibility to deal with this dynamic environment. (Appari & Johnson, 2010; Bhatti & Grandison, 2007; Rostad & Edsberg, 2006). Traditional RBAC systems are based on a principle that determine access rights on a prior base, by determining who (the actors) tries to access what information. This principle works fine in a static environment, where the role of the actors involved in an access situation, is the main variable. In the case of a dynamic environment of the healthcare, the purpose of the information access can change contextual integrity of the access situation. Contextual factors other than the actors involved, like the purpose of the information access can justify information access. In the case of the healthcare, timely access to healthcare information can change the outcome between life and death of the patients.

4.4.2 Access control in healthcare

Traditional access control models and policies are based on the assumption that possible access requests that will have to be obeyed are known in advance (a priori) and can therefore be captured by authorizations. The healthcare deals with unplanned and dynamic events like patients being transferred between wards, doctors asking for second opinions from colleagues or simply unplanned patient arrivals like emergencies. Since, the health condition of a patient is prioritized (even) over medical privacy, cases exists that medical information should be accessible, even when standard authorizations prohibits that. As access control shouldn't interfere with the care delivery in the healthcare, access control in the healthcare often include an emergency mechanism that bypasses the standard access policy that in standard conditions protect against unauthorized disclosure. In the literature, this mechanism is often referred to as "break the glass" or short BtG (Appari & Johnson, 2010; Bhatti & Grandison, 2007; Rostad & Edsberg, 2006). While the importance of such mechanism to prioritize patient's health is obvious, this mechanism still introduces weaknesses in the privacy security of the system.

Rostad & Edsberg (2006) studied the usage of break the glass within large Norwegian hospitals, by analyzing audit trails from access logs. Health information systems are logging access to health records and include information to these logs like: time of access, ID of user (physician, nurse, etc.) who tries to access a record, the location of the user, ID of the patient to which the record belongs to, location of the patient, etc. In addition, extra log files are created when break the glass mechanism is used. Upon a break the glass request, the user has to enter the reason for the break the glass. The reason and the time of request are included in the extra break the glass log files. Rostad & Edsberg (2006) found in their study that in the case of the Norway hospitals, that the records of 54% of all the patient registered in the health systems have had their medical information been accessed at least once, by the use of BtG privileges. From the total amount of access requests, 17% of the access requests were by BtG privileges. These number shows that, the use of BtG privileges are not used as an exception, but instead an integral part of the existing access control policy in practice. Dangers of such policy is that it relies too much on a mechanism that allow, unconditional manner which creates room for misuse by healthcare employees (Appari & Johnson, 2010; Ardagna et al., 2010; Bhatti & Grandison, 2007; Rostad & Edsberg, 2006).

4.4.3 A-priori vs a-posteriori

Traditional access control models and policies are based on the assumption that possible access requests that will have to be obeyed are known in advance (a-priori) and can therefore be captured by authorizations. Authorizations are only checked once by a single authority, i.e. at the moment access is requested.

In the healthcare there is a (break-the-glass) mechanism that allows standard access control policy that are based on a-priori to be overridden by a-posteriori access. Emergency access has to be justified using a-posteriori knowledge from log files that registered the access. A-posteriori creates more flexibility, as it allows the medical staff to go ahead with their duties, without worrying about problems like expiration of certificates, passwords or failing network connectivity to some authorization server. Figure 9 presents a graphical illustration of a-priori and a-posteriori access control. Dekker (2007) demonstrated an Audit-Based access control for the EHR setting that minimize a-priori access control, but focus on a-posteriori control based on audit logic.

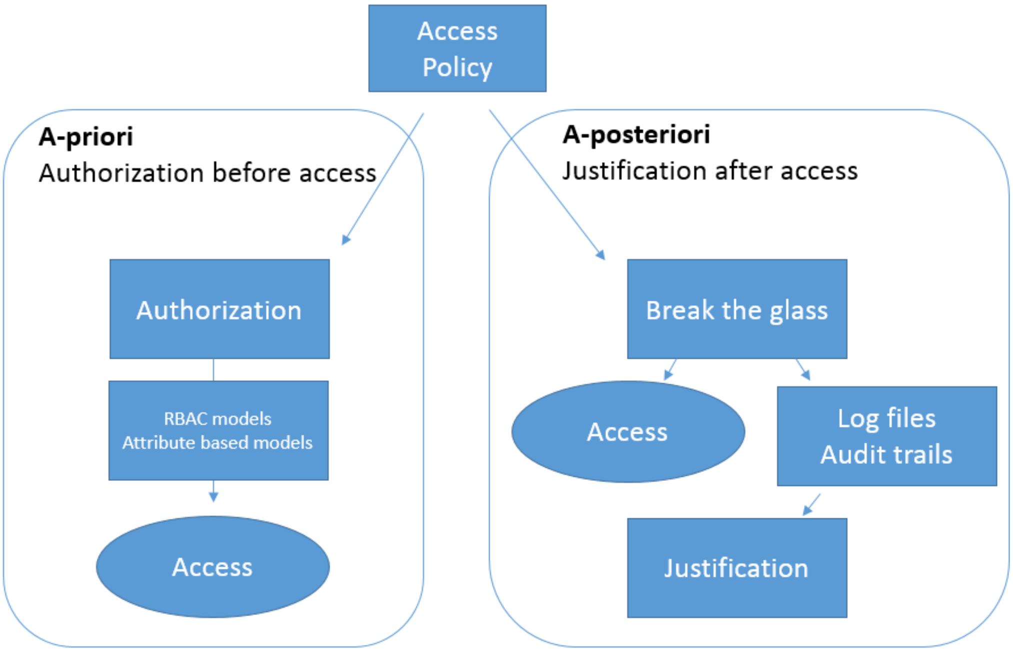


Figure 9: Access control in the healthcare

An important requirement for a-posteriori access control is that there must be some mechanism in place to ensure that users can be held accountable for their actions, i.e. that a user will not vanish after executing his (illegal) actions.

4.4.4 Policy refinement

Bhatti and Grandison (2007) propose the PRIvacy Management Architecture (PRIMA) in which a-priori access control is improved by *policy refinement* (Figure 10). The main idea behind policy refinement is to improve a-priori access control, by analyzing audit trails of emergency access.

Access control requirements are very complex. It is very difficult to capture all situations for healthcare information access in rules. Practical consequence of this is that there are many information access situations that are not explicitly covered by an access policy. This leads to an over-use of BtG in situations which are not covered by the a-priori access control policy.
A more explicit access policy that is more in line with actual clinical workflow can regulate more of information access situations and therefore lower the reliance on BtG.

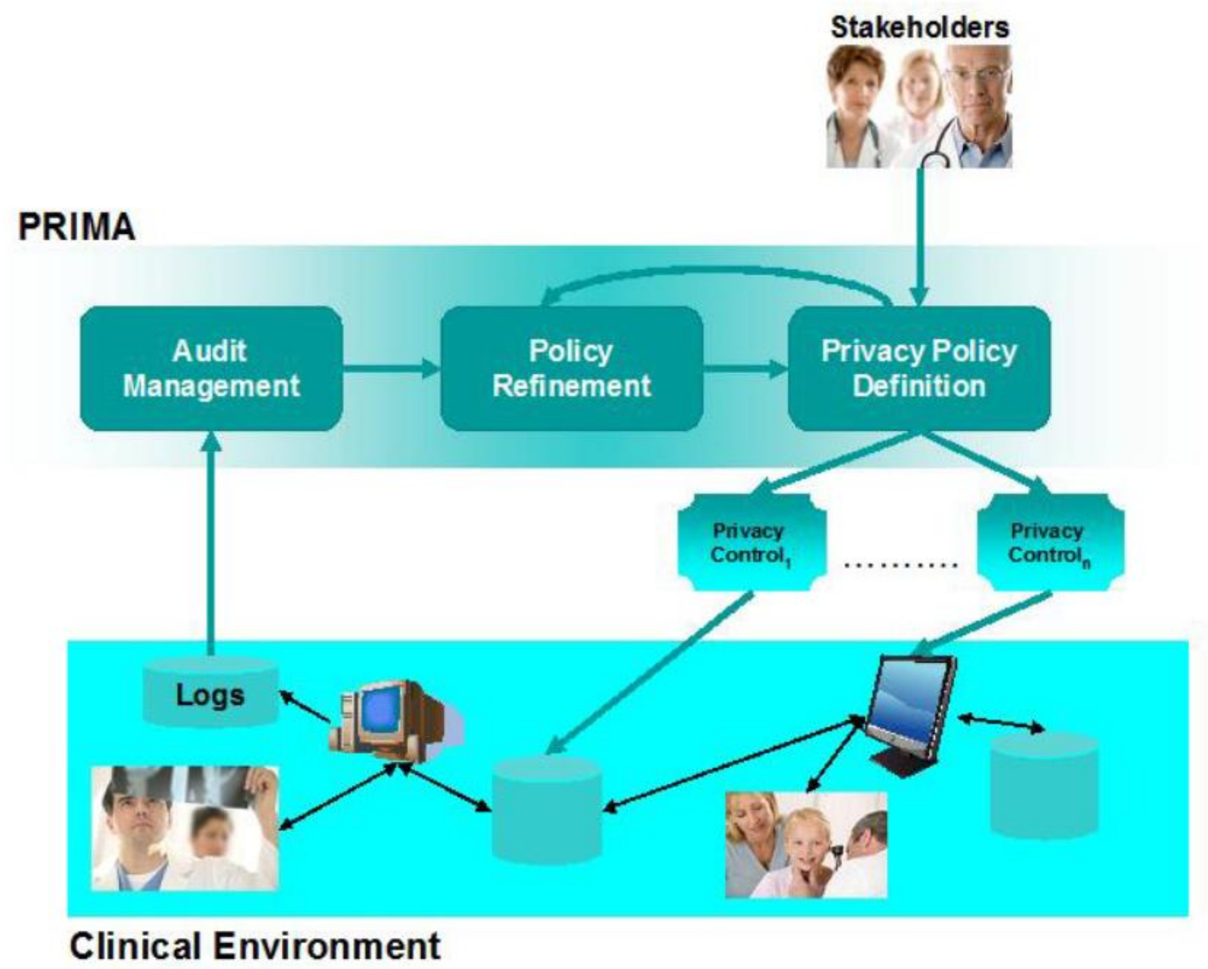


Figure 10: PRIvacy Management Architecture (PRIMA) (Bhatti & Grandison, 2007)

4.4.5 Policy spaces

"A policy space can be defined as a policy repository, whose policies regulate access to resources" (Ardagna et al., 2010). Traditional policy solutions distinguish two policy spaces: authorized access and unplanned exceptions (BtG). Access in those two policy spaces are based on the knowledge before (a-priori) and justification (a-posteriori) respectively. New policy spaces are introduced new policy spaces: denied access and planned exceptions, which allows access control systems to better fit the dynamic needs of the healthcare system.

4.4.6 Conclusion

Rostad & Edsberg (2006) pointed out a serious weakness of the emergency access control system. Several ideas to address the weakness of emergency access control has been proposed in the literature such as *policy refinement* (Bhatti & Grandison, 2007) and *policy spaces* (Ardagna et al., 2010). As part of the control aspect of privacy in our research we would like to explore how access control is currently is organized in the Dutch hospitals and how emergency access affects privacy in Dutch hospitals.

4.5 Conclusion

Chapter 4 had the purpose to answer the third research question of this thesis:

RQ3: What is privacy in the health care and what are the regulations related to privacy in the health care?

We answered this question by providing background medical data and health information systems. As modern conceptualization of privacy concerns information, we firstly describe information and information systems in the healthcare. In section 4.1 we described information systems in the healthcare and provided an overview of the different applications of healthcare information systems and the type of information these systems process. Roughly we can categorize health information system into clinical health information systems and administrative health information systems. In section 4.2 we discussed electronic health records (EHR). EHR includes clinical information about a patient such as medical history, treatment and outcomes.

Furthermore, in section 4.3 we discussed the related Dutch healthcare legislation and privacy legislation.

The *wet bescherming persoonsgegevens* (Wbp), ("Privacy law"), is a Dutch law that protects privacy of individuals, by regulating how personal information should be processed. The Wbp defines information as personal information if it *(1) consists information about a natural person (a real person)* and *(2) the person is identifiable*. Wbp considers some types of personal information even more sensitive. Information concerning health is considered as sensitive personal information. The processing of sensitive personal information has to follow even more strict regulations.

Wet op de geneeskundige behandelingsovereenkomst (WGBO) requires healthcare providers to maintain a health record of their patients. Only the patient and the people involved with the treatment of the patient are allowed to view this record. However, for certain statistical or scientific research, access to the health records is allowed.

Chapter 7 of *Zorgverzekeringswet* (Zvw) regulates the information flow between patients, healthcare providers, healthcare insurance companies and governmental organizations. The governmental organization *college zorgverzekeringen* regulates which personal information should be shared and is also responsible for the electronic infrastructure that facilitates the sharing of this personal information, how the personal information is shared, and the maintenance of the shared databases that stores this information.

The Dutch privacy law is becoming stricter. Eventually, Dutch privacy law will be replaced by European privacy. The current updates on European privacy laws indicate that European privacy law will become stricter. As part of this research, we would like to explore how the developments in privacy laws relates to big data in the health care.

5. Methodology

This chapter will elaborate on the methodology to conduct our research. The chapter is structured as follows:

Section 5.1 contains a motivation for the selection of a semi-structured interview to conduct our research. Section 5.2 discusses the sampling strategy that has been used in this research. Section 5.3 describes how participants for this research have been approached and discusses the background of our sample. Section 5.4 describes the format of the interviews and discusses the steps actions that were taken, including the creation of an interview protocol.

5.1 Semi-structured Interviews

The semi-structured interview is a qualitative data collection method that is used to obtain information on the issues of interest. Participants of an interview are given the time and scope to talk about their opinions on particular subject.

Unlike a structured interview, the questions of a semi-structured interview are not completely fixed. Semi-structured interviews are usually structured around an interview's guide, which include topic and themes that should be covered during interview, rather than a standardized script of fixed questions.

Semi-structured interviews give the participants the freedom to express their views in their own terms. The participant can add detail and depth to an answer providing more rich information. The aim of semi-structured interview is to understand the perception of the participant, rather than making generalizations. Hence, it often credited as a useful method for exploring purposes (Sekaran & Bougie, 2009).

The selection of a semi-structured is appropriate to address our exploratory goals. One of the main strengths of semi-structured interview is that it can provide rich information that allows us to get deeper understanding of the relation between big data and privacy in in the context of the healthcare, while still keeping some structure to guide through the interview.

In this research the findings from the literature study provided the structure of the interview. An interview protocol has been created based on the findings from the literature and was used as guide for the interviews. It is interesting to see if the findings from literature are matching experiences from practice. Hence, the structured part of the interviews provided some validation purposes besides providing only a guiding role.

Besides that, we already demonstrated that research on big data, health care and privacy is very limited and that why we want to explore. Exploration is addressed by the freedom part of the interviews. The freedom allows interviewees to share more in-depth knowledge, which can enrich the knowledge derived from the literature. Enriching knowledge of the literature with knowledge from interviews is aligned with the exploratory purposes of our study.

5.2 Sampling strategy

Since our purpose is to explore, we would like to find study objects that are suitable to contribute to our understanding of the phenomenon. Key topics related to our research objective are: healthcare, privacy and big data.

This will lead to the main criterion of our sampling strategy:

The interviewee should possess the information and knowledge about the relevant topics (healthcare, privacy and big data).

As we don't focus on generalizing, we don't have to follow a random sampling strategy. This leaves room to focus on information rich samples. We should select a sampling strategy that best suits our situation, with main factors time and resources.

Experts possess the information and knowledge about the relevant topics. Hence, we focus on conducting the interviews with experts.

As part of the sampling strategy we would like to categorize experts based on the two following characteristics:

1. The perspective of the expert on the topic of the research.
   (Technology perspective / Health care perspective / Privacy perspective)

Experts will be selected based on the expertise they have with one of the relevant topics and will be characterized on perspective they can provide.

Experts with a *technology perspective* indicate that the participant is mainly selected for his or her knowledge of big data and related technologies. These experts can provide a more technical perspective on big data.

Experts with a *health care perspective* are mainly valued for his or her experiences with big data or IT related projects in the health care. Assumed is that these participants are able to identify the current developments of IT in the health care and more specifically big data development or trends. They are valued for their views on big data and privacy from a health care value perspective.

Experts with a *privacy perspective* are mainly valued for their experiences and expertise with privacy regulation. They have in depth knowledge on data processing regulation and privacy certifications. They are able to identify the main issues of privacy from a legal perspective.

5.3 Participants / Sample

Participants in this research have been approached through personal connections and through LinkedIn. In total eight people participated for an interview in this study. Six of the participants resulted from personal connections in combination with the snowballing effect that occurred after the initial interviews. The other two participants were found through LinkedIn searches. Searches on LinkedIn were performed through the combination of keywords ‘health care’, ‘privacy’ and ‘big data’.

Table 3 presents an overview of the background / occupation of each of participant and the perspective for which they are characterized. Furthermore, each participant has been coded. This code will be used to reference statements to the specific participant. For example, when the principal consultant is quoted, reference code [ITCH1] will be used to refer to him. The last letter of the code indicates the perspective of the participant: T for technology, H for health care and P for privacy.

Table 3: The background of the participants and the main selection criterion to add them to the study.

Code	**Job title**	**Perspective**
DMT	Assistant Professor	Technology
DST	IT Freelancer	Technology
ITC1H	Principal IT Consultant	Health care
ITC2H	Managing Enterprise Architect	Health care
CRTH	Associated Professor	Health care
ITSP	Information Security/ Data Privacy Manager	Privacy
PJP	Owner Legal Advisory Office	Privacy
HISP	Owner Health care privacy software developer	Privacy

5.4 Format of interview and interview protocol

All of the interviews were face to face interviews taken place at the university or at the location of the participant. The interviews followed an interview protocol which included a topic list and guideline questions. The interviews were aiming to last 60 to 90 minutes. Five interviews stayed within the planned time. Three of the interviews lasted longer than 90 minutes.

An interview protocol has been created in advance, which includes a set of questions that addresses the topics of big data, healthcare privacy and information access control. (See Appendix A). The interview protocol provided topics of discussion and initial questions related to the topics. The topics and initial questions provided by the interview protocol provided some structure to our interviews. The interview protocol consists of three sections with each section covering one of the topics. Table 4 provides an overview of the relation between the findings in the literature review and the sections of the interview.

The first section of the interview protocol is focused on the definition of big data. Firstly, we ask participants what their understanding is of the concept big data. Participants might be able share new perspectives of big data that was not covered by the literature review. In Chapter 2 we formulated the three aspects of big data we focus on in this research: *Need to capture more data*, *Big data technologies*, *Big data Analytics*. In order to reach a mutual understanding of the concept of big data, we end the first interview section by introducing our understanding of big data for this research, which is based on the literature review in Chapter 2.

The second section of the interview protocol is focused on privacy. In the second section of the interview protocol, we first ask the participants what their understanding is of the concept privacy. In Chapter 3 we presented a conceptual model of privacy which introduced the following aspects of privacy: natural privacy, normative privacy, control aspect of privacy and condition of privacy. In order to reach a mutual understanding, we end the second interview section by introducing the conceptual model of privacy.

The third section of the interview protocol is focused on the relation between big data and privacy in the health care. The conceptual model of privacy defines privacy in several elements. The third section focuses on the relation between big data and natural privacy, the relation between big data and normative privacy and the relation between big data and control aspect of privacy.

Notes were taken during the interview and after the interviews. Each of the participants was also asked for permission to create an audio record of the interview. All of the participants agreed and gave permission to create an audio record of the interview. Based on the notes and the audio recordings of the interviews, detailed transcripts were created. Unfortunately, the audio recordings for one interview got lost due to technical errors. In this case, the summary and the notes were used for analysis.

Table 4: Relation between the literature review (Chapter 2 and Chapter 3) and the interview protocol.

Concept	*Interview section*
Big data In order to reach a mutual understanding of the concept of big data, we first ask the participants what they understand with the concept of big data. Afterwards, we introduced our definition of big data in this research, which is based on the literature review on business intelligence and big data (chapter 2).	***1. Can you shortly describe what you understand with the concept of big data?*** - *The need to acquire and process more new sources and new types of data.* - *New infrastructure and technologies to process new sources and new types of data.* - *Advanced analytics that can gain more insights of data.*
Privacy In order to reach a mutual understanding of the concept of privacy, we first ask the participants what they understand with the concept of privacy. Consequently, we introduced the conceptual model of privacy (chapter 3), which is the model we use to analyze privacy in this research.	***2. Can you shortly describe what you understand with privacy?*** - ***Nonintrusion theory*** *“right of being let alone”* - ***Seclusion theory*** *“being alone”* - ***Control theory*** *“the control one has over information about itself”* - ***Restricted Access Theory*** *“privacy is the condition of being protected from unwanted access by others, including access to one's personal information”*
Natural privacy The conceptual model of privacy introduced natural privacy as an element of privacy (chapter 3). The purpose of the questions in this section of the interview was to explore how big data affect natural privacy.	***3a. What is the relation between big data and natural privacy?*** - Is this significant for the healthcare? - More private and public information is being stored. - Duplication of data; hard to remove data. - Anonymization of private data doesn’t provide natural privacy anymore. - Does big data decrease our natural privacy?
Normative privacy The conceptual model of privacy introduced normative privacy as an element of privacy (chapter 3). The purpose of the questions in this section of the interview was to explore how big data affect normative privacy.	***3b. What is the relation between big data and normative privacy?*** - Are current norms sufficient to protect against the harms of privacy? - Will we rely more on normative privacy due to the developments of big data? - Do the norms need to be change for big data?

Control aspect of privacy The conceptual model of privacy introduced control aspect privacy as an element of privacy (chapter 3). The purpose of the questions in this section of the interview was to explore how big data affect control aspect privacy.	***3c. What is the relation between big data and control aspect of privacy?*** - How does information access control in the health care currently look like? - Can you tell me more about the use of audit trails to improve information access control? - Can you tell me more about a-posteriori information access control? - How could big data impact (a-posteriori) information access control?

5.5 Data analysis

The audio recording of the interviews were transcribed. The interviews transcripts were then coded. The initials round of coding was done on printed versions of the interview transcripts. The initial round of coding was open coding, in which we read through the transcripts and breakdown transcripts into sections. The hardcopies of the interview transcripts were read carefully and notations were made on the hardcopies to indicate different sections of the transcripts. The main idea is of open coding was to reduce data into sections and associate these sections with the labels. Open coding assisted in improved understanding of the transcripts and helped to filter out the unimportant parts from the data, while highlighting relevant parts. To be sure nothing important was missed out in the first round of coding, this cycle of reading and making notes was repeated at least once more for each interview.

In the following step, the transcripts were transferred to a qualitative data analysis software package. For this study we used the software package *ATLAS.Ti*. The software assisted in structuring the data and refining the codes. With the use of *ATLAS.TI*, we did the next round of coding which was axial coding. During axial coding we search for relations between the sections we highlighted and group different codes together. There was a continuous cycle of refining and restructuring, which included in renaming and merging codes.

The final step of coding was selective coding in which we selected core codes for the story line of our results. We focused to select codes that could provide a story line around our conceptual model. Additionally, we categorized codes based on the background of the interviewee.

6. Interview results

This chapter will discuss the results from the interviews. The chapter is structured as follows:

In section 6.1 we will present and discuss the results related to the characteristics of big data. This section will discuss three categories of big data and will conclude with the characteristics of big data in the health care. Section 6.2 will describe three common examples of big data in the health care that were mentioned during the interviews: *Biobanks*, *National EHR* and the *Integrated Care*. We will elaborate on big data in the health care by discussing the three notable examples. Section 6.3 will present the findings relating to the conceptual model of privacy.

6.1. Characteristics of big data in the health care.

The majority of the participants mentioned the combination of multiple sources and multiple types of data as important characteristics of big data.

> ***"I can't add much to the usual definitions of big data. However, what big data makes special to me is the combination of multiple data sources and as well multiple types of data. [ITC1H]"***

Besides these common characteristics of big data, other characteristic of big data have been given to describe big data. In section 6.1 we will present three categories of big data characteristics that were derived from the interviews: ***big data is difficult to process***, ***big data collecting without purpose*** and ***big data in the health care.***

Participants mentioned big data as "broad" or a "fluid" term. By presenting the three categories of big data, we try to isolate what is meant by big data in the health care. In section 6.1.3 we describe what is mainly meant with big data in the health care. In section 6.2 we elaborate on big data in the health care, by presenting three common examples of big data in the health care mentioned by our participants.

6.1.1 Big data difficult to process

From a technical data management perspective big data is often referred as*: "all the forms of data that were difficult to process by computers, for certain reasons, but which we are now able to process. [DMT]"* The definition can be separated into two parts. First, this definition implies data became more complex or even too complex to process as data is growing in volume, variety and velocity. Relating to the standard definitions this is related to the volume and structure characteristics of data. Second, the developments in big data technologies enable us to store process new complex forms of data. For example, we are now able apply machine-learning on data, we are able to visualize the data or to process the data real-time. These developments were described in terms of advanced analytics and algorithms.

6.1.2 Big data collecting without a clear purpose

One of the participants explicitly distinguished big data from traditional data, and explained why he believed that big data leads to new insights. *“Big data is when you have large sets of data and exploring this large set of data lead to new insights [ITC2H].”* The participant stressed on that in big data, data is collected without a clear plan beforehand. In contrast he gave an example of what he doesn’t consider as big data. *“Research in the medical world, for example, is about a specific topic and about that we want to know as much as possible. However this doesn’t lead to big data [ITC2H].”* With the example the participant was referring to the Parelsnoer Institute, which is collaboration between eight academic hospitals in the Netherlands. According to the website of the Parelsnoer Institute, they are focusing on 14 diseases. For each of the specific diseases, large volumes of medical data is collected and stored in a large database. These large databases that stores medical data genome data, genome interaction and metabolism data and are called biobanks. In the case of Parelsnoer Instituut each of these biobanks are focused on a specific disease.

For these biobanks, a data model and a protocol are defined. This data model and protocol only includes data of which they think is relevant to the specific topic (which is a specific disease in case of Parelsnoer Instituut). *“It is about a lot of data, but it is very structured and focused on a specific topic [ITC2H].”* This leads to large volumes of data, which include many instances of the same data set, although from different people. According to the participant this is different from big data, as this is focused, predefined, and mainly about more observations leads to better conclusions.

> ***“Big data is collecting as much data as possible first, and after that we will see what we conclude [ITC2H].”***

This aspect and related aspects has also been mentioned by other participants [DMT], [DST], [ITC1H]. However, these participants did not mentioned explicitly and did not consider it as a key aspect of big data.

These aspects about big data do cover what can be found in the general literature of big data. However, approaching big data in such a way is difficult when data is considered private data. Dutch privacy law (Wbp) explicitly address that private data must be collected for specified, explicit and legitimate purposes and not further processed in a way that is incompatible with those purposes. Therefore, privacy law limits the collection of as much ***private data*** without a clear purpose beforehand. As the health care sector mainly processes private sensitive data, this form of big data is currently not compatible with the health care sector.

6.1.3 Big data in the health care

Especially the aspect of multiple data sources seems relevant to the discussion of big data within the health care. Data sources in the health care are currently scattered among the health sector, and exist in silos. There is a strong believe that health care can be improved by using and combining these scattered data sources in the health care. Insights from the right data sources could support health care processes, and health care can be aligned with personal needs. There seem to be a vision to create an integral connection between the different sources. Participants described it as *cloud solutions*, *production environments*, *data infrastructure*.

> ***"Big data in the health care is often discussed in relation to the scattered data sources. In that sense we (the health care sector) understand big data differently. Big data is actually like a cloud solution, which makes data more accessible. [CRTH]"***
>
> ***"Big data is not about the technology and is not about the data. At a certain point you have a goal and you want information for this goal… Many sources of data are not integrated or whatsoever. What you want is data environment that integrates these data sources, which allow to data sharing to be faster and more flexible. [ITC1H]"***

It seems that big data in the health care is mainly focused on integrating scattered data sources. The main purposes for this are improved information sharing, combining different information and availability of more complete medical information.

FINDING: Big data in the context of health care is mainly about integrating scattered data sources

6.2. Common examples of big data in health care

In section 6.1 presented the results showing. That big data in health care sector is mainly about integrating multiple sources within the health care. In the series of interviews several examples of big data in the health care have been given. Three notable examples of big data development in the healthcare that tries to integrate multiple sources within the health care. In section 6.2 we discuss three common examples of big data in the health care mentioned by the participants, to elaborate on what big data means in the health care. The examples give more concrete information on how big data is being used within the health care.

The three examples are Biobanks, Landelijk EPD (National EHR), Ketenzorg (Integrated Care). The three examples illustrate three developments that are considered as big data in the health care. A rich amount of available information about these cases came forward in the interviews. The data sources and the stakeholders differ in each example. These three examples seem to be the most relevant to our topic and the rich information was available from the interviews. These examples were used by the participants to explain big data in the health care. In this section we will describe the three examples, as it provides contextual information on big data in the health care. Table 5 provides and an overview of the characteristics of the three common examples. Section 6.2.1 will elaborate on Biobanks, section 6.2.2 will elaborate on National EHR and section 6.2.3 will elaborate on integrated care.

Table 5: The main characteristics of the three common in health care.

Biobanks	National EHR	Integrated Care
- Research central - University, Academic Hospitals - Anonymized data - Standardized data sets, with many instance of the same type of data set collected from different people. - Large volumes of similar data - Data of different people	- More complete medical history of a patient. - Private data - Data scattered among stakeholder of mainly the health care domain.	- Patient central (lifestyle) - Data scattered among very diverse group of stakeholders in the health care and social domain. - Type of data: large variety of data like nurse reports, home situations. - Identifiable data - Regulations are strict

6.2.1 Biobanks

Biobanks are large databases that store a lot of medical data for research purposes. These biobanks are often present among Universities and Hospitals. Medical data are collected of large populations of patients, but also of healthy people. Examples of medical data stored in these databases are genome data, genome interaction and metabolism data. The expectation is that development of diseases can be understood.

> ***"For example, in Delft they work with (medical) imaging systems. At which large volumes of data is being collected about populations of patients as well of healthy people. With the use of this data, researchers try to develop intelligence, which could for example, make predictions about the development of diseases like: brain diseases, Alzheimer, and cardiovascular diseases. [CRTH]"***
>
> ***"Algorithms are being developed that can analyze individual x-ray images other types of imaging-data. They are trying to apply this more and often in hospitals. [ITC1H]"***

Connecting more bio banks with more hospital and universities, results in larger data sets. The current trend is to connect more bio banks and the respective actors with each other. This improves information sharing between the hospitals and the universities. As result, these bio banks will be larger as more complete, as more hospitals and universities store relevant data in these bio banks. Figure 11, provides a schematic overview of the main actors of a biobanks.

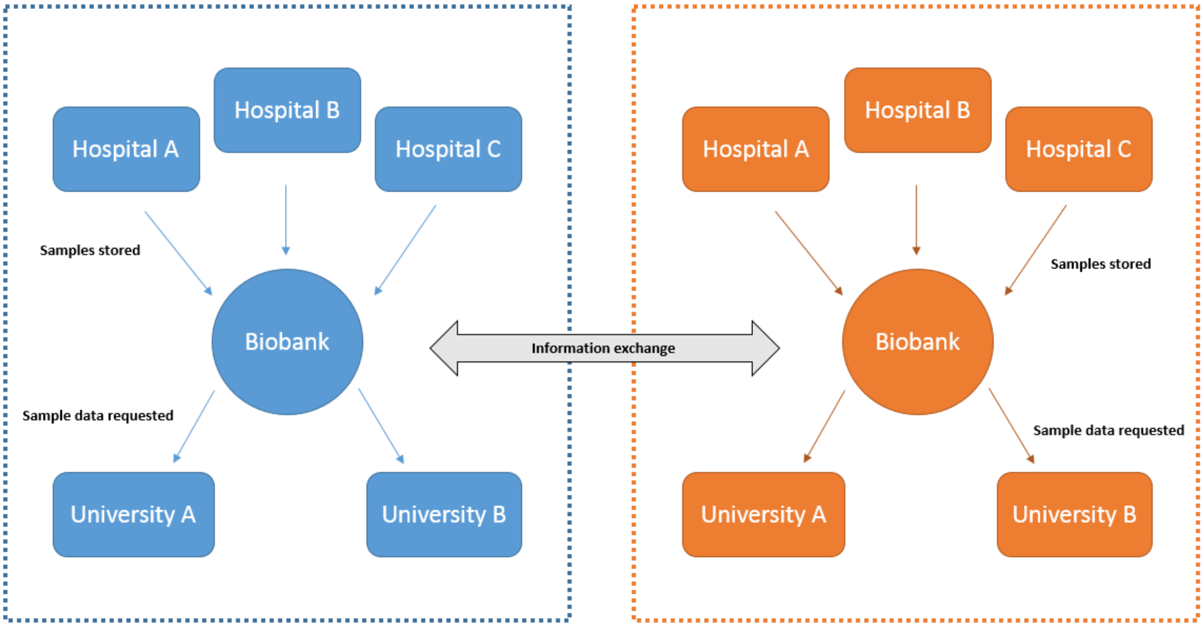


Figure 11: Schematic overview of biobanks.

6.2.2 Landelijk EPD (National EHR)

Landelijk EPD (National EHR) refers to an initiative by the Dutch Government that tried to realize an infrastructure that connects electronic health records data sources with each other at national level. Figure 12, provides a schematic overview of the National EHR.

As discussed, the WGBO law requires healthcare providers to maintain health records of the patients they are treating, which includes information about patient's health, the treatments the patient received, and other information which are necessary for healthcare processes. Nowadays these types of information are stored electronically in so called electronic medical records (EMR) (Dutch: EPD). A patient interacts with various actors when he goes through a healthcare process. For example a patient interacts with his GP for a medical problem. A GP can refer the patient to a specialist for serious and complex medical problems. One time he needs to see specialist in a hospital, for additional lab test or a surgery. Another time he has to visit a physician to revalidate. If there is need for medication for a treatment, the patient needs to interact with the pharmacy. And for each different problem a patient might to see different actors. During his lifetime, a patient may interact with various actors in the health care, and each of the actors maintain pieces of medical information of the patients in medical records. These pieces of information are currently scattered among different data sources at various actors in the health sector. Not all of these data sources are connected with each other. The aim is to connect and integrate these data sources with some sort of infrastructure, which we can refer to *cloud solutions*, *production environments*, *data infrastructure*.

> ***"At a certain moment you are going to the GP with some medical problem, and the GP wants to prescribe your medication. Because of such connection (a connection between data sources containing medical records), the GP can view a complete overview of your medication history, instead of just a small piece of it. The GP might not be aware of this and is only aware of the medication he prescribed himself. The aim (of the landelijk EPD) was to get that information complete. [ITC1H]"***
>
> ***"The communication of this project was a bit unlucky, because the national medical record was never the idea to create a national medical record, but a national switching point (landelijk schakelpunt, LSP). [ITC2H]"***

In the specific case of the Landelijk EPD, the Dutch government tried to realize an infrastructure that keeps track of all these data sources. Such infrastructure would make it possible to determine the location of all the data sources, which possess relevant pieces of medical information of a certain patient. Health care providers can request patient data from data sources at other health care providers through an actor landelijk schakelpunt (LSP). Participants mentioned that name landelijk EPD was maybe not the right name, as it was not the aim to merge all the data sources into a single national database, but rather the aim to realize an infrastructure and a central point which facilitate health information sharing between health care providers [ITC1H], [ITC2H], [PJP], and [HISP].

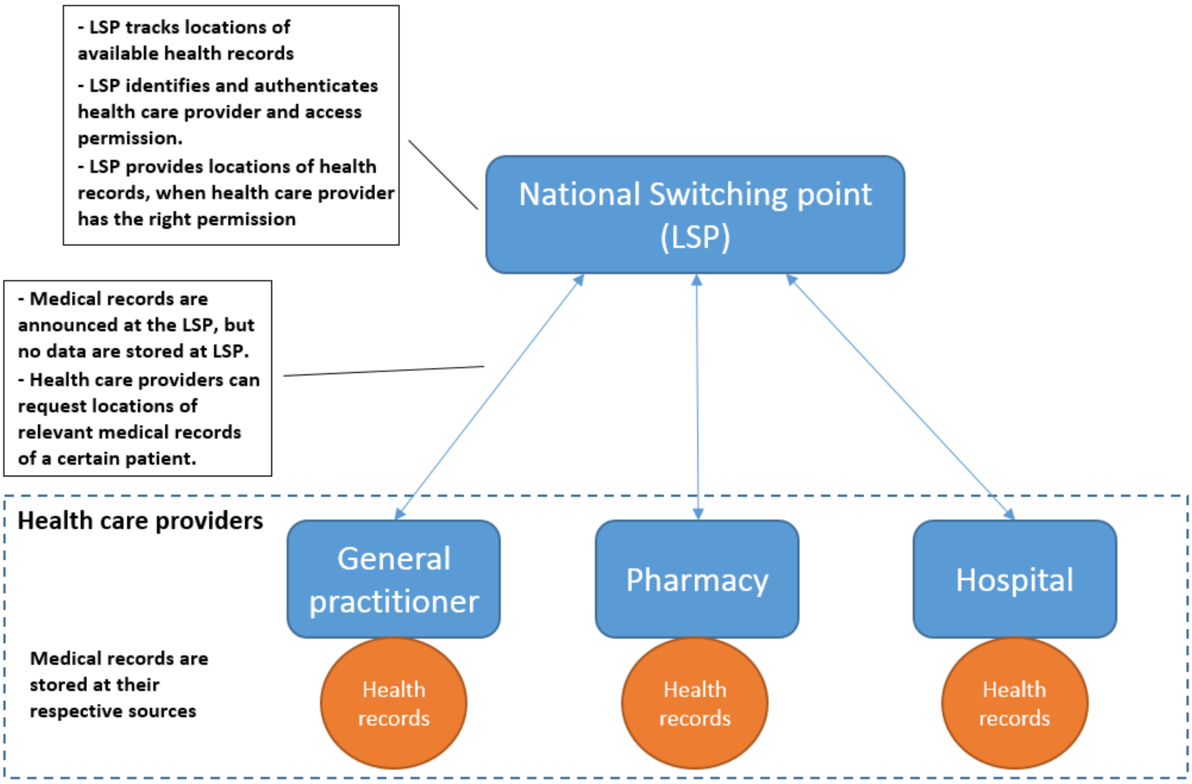

- LSP tracks locations of available health records
- LSP identifies and authenticates health care provider and access permission.
- LSP provides locations of health records, when health care provider has the right permission
National Switching point (LSP)
- Medical records are announced at the LSP, but no data are stored at LSP.
- Health care providers can request locations of relevant medical records of a certain patient.
Health care providers
General practitioner
Pharmacy
Hospital
Medical records are stored at their respective sources
Health records
Health records
Health records

Figure 12: Schematic overview of actors and the health data sources in within the Landelijk EPD

6.2.3 Ketenzorg (Integrated Care)

Various actors in the ketenzorg share pieces of information of a patient to create complete view of a patient's life context. This is at some points similar to the Landelijk EPD, however there are essential differences.

Ketenzorg refers to the alignment of various health care processes and activities that exists in a chain of health care providers, in which the patient is the focus of this alignment.

Health care processes have become more complex and involves various health care actors, from various disciplines. Health care will be organized based on guideline and standards, which allows the right health care to be provided, by the right health care providers, aligned with the required health care needs of the patient. Essential in this collaboration is the sharing of information between the actors. Various actors in the health care possess sets of information of patients and viewed as valuable data sources. Information of these data sources can be combined and may improve the complete view of the context of a patient. With a more complete view on the patient, health care providers can formulate health care treatment plans that are more aligned with the patient needs.

Several health care chains exist for the treatment of chronic diseases, like Diabetes health care chains, cerebrovascular accident (CVA) health care chain.

There seem to be an interest from the health care to extend the scope of the information sharing within health care chain. The aim is to include more actors and more data sources within the scope of a ketenzorg.

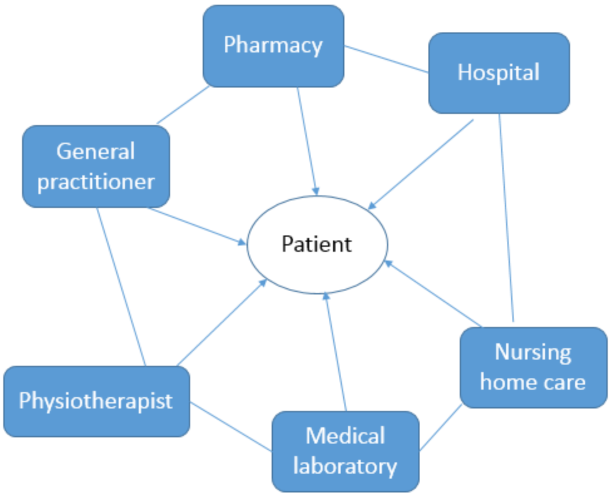


Figure 13: Schematic overview of the actors in integrated care.

Integrated Care at the Social Domain

Due to changes in Dutch health care and social care regulation, some tasks and responsibilities in the social domain have been transferred from Dutch national government to the local municipalities. As a result of this decentralization of these more actors will find each other in the chain and are required to collaborate with each other. Various data sources and information systems that are going to be connected to each other and actors will have access to the connected data and process the data on various ways.

The scope of the actors and data sources in such integrated chain is much broader. The data goes beyond a medical scope and also includes data from a social context. Second the stakeholders involved are from a greater variety: healthcare to social domain. This makes privacy and information security much more complex.

6.3. Privacy

In section 3.1.5, we presented a conceptual model of privacy. The conceptual was used as a structure to reach our research objective: to investigate the relation between big data and privacy. The conceptual model separated privacy I different elements. These elements of privacy were used as a guide through the interviews and provided a structure for our investigation between privacy and big data.

The elements of privacy have been discussed with the participants and the findings from these discussions are presented in this section. Figure 14 presents an overview of these findings.

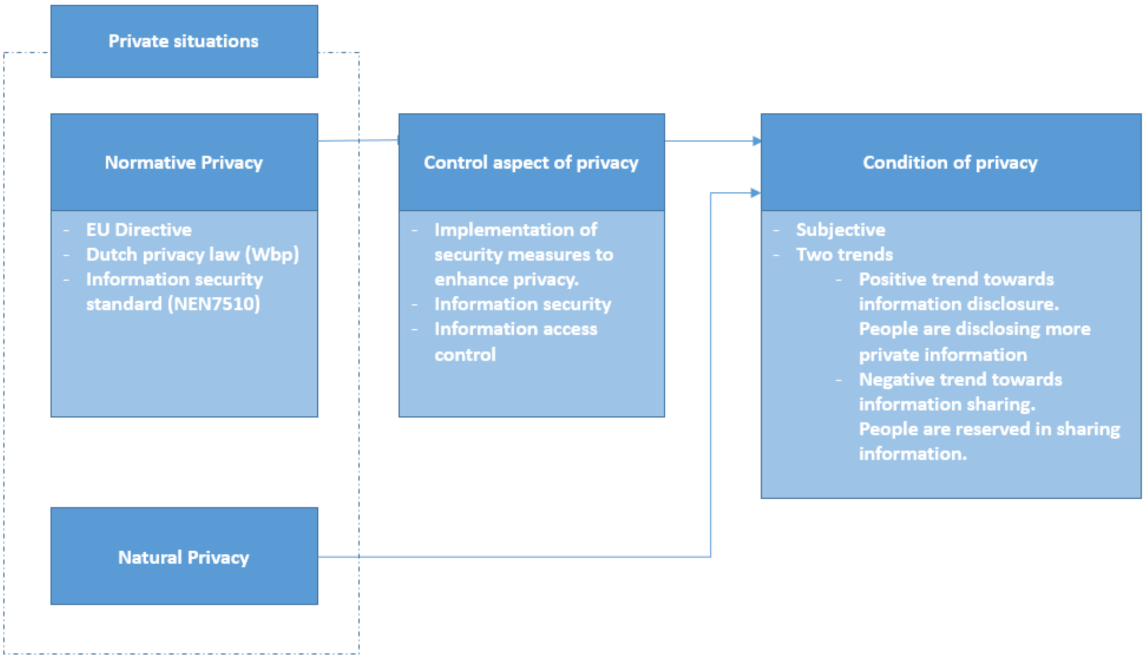

Private situations
Normative Privacy
- EU Directive
- Dutch privacy law (Wbp)
- Information security standard (NEN7510)
Control aspect of privacy
- Implementation of security measures to enhance privacy.
- Information security
- Information access control
Condition of privacy
- Subjective
- Two trends
- Positive trend towards information disclosure. People are disclosing more private information
- Negative trend towards information sharing. People are reserved in sharing information.
Natural Privacy

Figure 14: The main findings related to the conceptual model of privacy

6.3.1 Private situations

All participants agreed that natural privacy is decreasing. Big data is attributed as one of the reasons that contribute to this trend. Factors such as the increased data generation, increased storage of data, and higher data availability are explanations for this. Participants also referred to developments in IT, and technological solutions, such as cloud technology, data base technologies, technologies assisting in data management.

One participant explicitly mentioned that natural privacy doesn't exist anymore, as he believes that modern society cannot function when there is still natural privacy [ITC1H].

> ***If I relate natural privacy to the EHR, then I expect some people to say that they don't want to give permission to be included in the National EHR infrastructure, because some people don't want their private information to be exchanged. This goes beyond the world we are living in right now. At a certain moment you go to the hospital and you want to be medically treated. Medical information will be recorded in different information systems and are exchanged. Your blood sample will be analyzed and the medical specialist can view this information if you then say I want privacy. How do you imagine***

> ***privacy then? Do you think natural privacy still exists? Natural privacy is a phenomenon from a world that doesn’t exist anymore. From the moment a baby is born, medical information is recorded. Natural privacy doesn’t exist anymore, if you want to make use of the current benefits of society and want to operate properly in current society [ITC1H].***

From a health care perspective this is understandable. The health care sector has always been a sector that heavily relied on private data to function properly. This view is reflected in health care law, which requires health care workers to store (private) medical data of their patients. It seems that in the health care sector they are used to the idea to give up natural privacy for better health care.

Even though natural privacy decreases due to big data, developments in IT, and technological solutions, such as cloud technology, data base technologies, technologies assisting in data management. Participants also stated technological improvements do improve our private situation.

> ***Stacks of paper medical records were put at the garbage. I doubt it if that is much safer. I think that the current technology is safer (for privacy), however I think there is a perception among users that current technology is less safe (for privacy) [HISP].***

The electronic storage of medical data and the setup of an infrastructure to share medical data improve our private situation. The main example provided is that medical data used to be recorded on paper. These paper medical records were stored in archives, which took a lot of space. In some occasions these paper records could be found at the garbage. According to some of the participants this paper storage led to much less safe privacy situations.

Normative privacy

All participants agree that we rely more on normative privacy when dealing with privacy situations. The decrease of natural privacy makes us more depended on normative privacy. However, participants questioned if it is a bad thing that we rely more on normative privacy. The decrease of natural privacy might be necessary to operate a modern society. Some benefits at societal level or an individual level can come at the cost of natural privacy. Participants mentioned that we need to give up some of our individual privacy in certain cases, such as: public health, fighting terrorism, and the operation of a hospital.

> ***You have the right to keep private things private. This will be more complicated for data that is normally considered private, but are important for public health and fighting terrorism. I think about it and I think there is a border to what extend private information should keep private. Individuals should give in some privacy for the interest of the society. Some private information is essential for the operation of a hospital and perhaps for hospital staff. In such case the interest of the staff is more important, especially for physically threatening situations [DMT].***

> ***We need to protect against misuse of privacy, but please only in a general way. We have to be careful that we don't throw away all benefits because of privacy [ITC2H].***
>
> ***There should be some balance, but we can be more aggressive (in trading privacy of individuals). Not always requesting permission access to individual private information. But perhaps we should need a council in the Netherlands that determines if privacy is guaranteed [DST].***
>
> ***Health care organizations are trying to follow legislation, since they don't want to take any risk. That is actually also a problem …***
> ***Legislation is fed by fear which makes it more difficult to share information. Instead, we should look at what kind of functionality do we want in relation to big data, and build legislation around that [CRTH].***

The balance between benefits from personal data and privacy is a difficult discussion. Several participants noted that they are afraid that we throw away the benefits because of privacy. Some noted that we could be more aggressive. One of the ideas mentioned is to have an independent council that assesses the tradeoff between benefits and privacy. Each private situation is different and a council could weigh benefits against privacy at certain situations. Participants indicated that privacy laws do limit innovation of big data in the health care. Participants mentioned legislation currently favors privacy over the benefits from big data. One of the participants explicitly mentioned that law makers are led by fear for privacy.

6.3.2 Control aspect of privacy

The discussion about the *control aspect of privacy* shifted during the research. Initially we started this research with the focus on *information access control* to discuss control aspect of privacy. We hoped to gain more insights in the differences between a-priori and a-posteriori information access control systems. After the first few interviews of this research, it became clear that the responses on this topic were limited. Participants noted that they could not provide in-depth information to our questions related to information access control or noted that are noted that the have limited experienced with the information access control. This indicates that we targeted the wrong participants to address the questions related to information access control this specific topic.

Additionally, we believe that the initial questions related to information access control were not suitable to cover the control aspect of privacy in our study. Participants more often discussed information security and NEN7510 to cover the *control aspect of privacy*. Information security seems to be more suitable to address the control aspect of privacy. Hence, later in the research the scope of *control aspect of privacy* shifted towards *information security*, which includes technical and organizational measures that are focused on preventing disruptions in information and information access.

In the section below, we will discuss the results from the first four interviews when the scope of the *control aspect of privacy* was still information access control.

6.3.2.1 Information access control

All four participants who were asked about information access control did mention that role based model is often used for information access control. It is mentioned that there are different variations of a role based model which can vary from a strict and secure implementation to a more flexible implementation. For example, the requirements of information access control system in organization in a context with sensitive military information focuses would often more on security for example. Role based access control can introduce high level in complexity of rules when the situation requires very specific conditions. An example of the complexity of a complexity officer no access to lower ranks.

Questions focused specifically on the requirements of information access control within the health care gave limited responses. The participants could confirm that information access control consist of an a-priori system and a posteriori system. Information access control in health care is designed to work with an a-priori system in normal situations. The a-posteriori system (referred as rode knop procedure) is designed for emergency situations. The participants could add that an a-posteriori system audit trail is created when information is accessed through the a posteriori system. No details could be given on what happens with the audit trails.

FINDING: Limited response on information access control

Participants did not respond positive on the question if they see a posteriori system as serious privacy threat. Some of the participants questioned the conclusion we derived from the Rostad & Edsberg (2006) research which formulated as following: in practice, emergency access is not used as an exception, but instead is used an integral part of the practical access control in the health care. Our conclusion derived from the study by Rostad & Edsberg (2006) did not meet the experiences of the participants. Participants stated that our conclusions from Rostad & Edsberg (2006) are over exaggerated or not applicable to the situation in the Netherlands.

FINDING: Our conclusions that we derived from a study of Rostad & Edsberg (2006) did not meet the experiences of the participants

Participants did agree that the current form of information access control in the health sector has weaknesses. The participants stated that in practice it not unlikely that users authenticate themselves to the access control system using a different account then their own. With users we refer to the employees in the health care who need access to the information systems in order to do their job. It seems to be common that in certain situations users share there authorization accounts. For example, nurses from the same department might share the few computers that are available in that department. According to the participants it is very likely that they will leave one computer open and logged in on one account and share access of this account with a group of nurses. This could be likely, as the nurses in this example don't need to log off and log in every time they need access to the system, which could be seen as a burden to the work process.

Therefore, the main weakness is not solely a technical problem but also has cultural problems. The privacy concerns in the example of account sharing and emergency access are based on the fact that users don't use the system as is designed. This indeed seems to indicate that there are cultural or organizational factors involved in these weaknesses.

FINDING: Current form of information access control in the health care has weakness. The weakness is more a cultural problem than a technical problem.

6.3.2.2 Information security

Later in the research the scope of control aspect of privacy shifted towards *information security* and all the measures to secure the control of privacy. The scope of the *control aspect of privacy* focused on information security measures for information access control.
Participants referred to the NEN7510 to discuss this point. The NEN7510 is a standard for information security for the health care sector in the Netherlands. The relation between NEN7510 and privacy is that the measures as described by NEN7510 standard will improve information security. Improving information security would then lead to more control of privacy, as improved information security would allow an organization to have more control over

The NEN7510 prescribe a set of measures that are focused on preventing disruptions and minimizing the damage from occurred disruptions in information. Disruption in information includes unauthorized access, unauthorized disclosure and unauthorized modification of information. NEN7510 focuses on the following aspects of information security:

- Confidentiality: the protection of information against unauthorized disclosure
- Integrity: securing information from unauthorized modification or destruction.
- Availability: securing availability of information to the users at necessary situations.

These measures include the formulation, execution and evaluation of security policies, risk analysis, and classification of information, and ICT process like user management and access control. An important requirement of secure information sharing is the identification and authentication of the users that are sharing information.

According to the participants there is a misconception that compliance with NEN7510 directly results there is compliance with the Dutch privacy law (Wbp). A NEN7510 certification is used to show that and organization took technical and organizational measures as is required by article 13 of the Dutch privacy law (Wbp). However, it does not address the other requirements specified in other article of the law. Hence, even if the right organization and technical measures has been addressed, article 7 of the Dutch privacy law still requires collection of private date to have a purpose. Therefore, organizations can still not use data to explore usage outside the defined purposes (see section 5.1).

6.3.3 Condition of privacy

Participants mentioned that they find the condition of privacy a difficult topic to discuss about. The participant mentioned that subjective aspect on this element makes is difficult to discuss about. Most of the participants discussed that the level of privacy a person experiences depends on how the individual perceives it. Several examples were given. Participants often referred to an example with social media. There are individuals who disclose all kinds of personal information on social media, while others are more aware and reserved with the personal information they disclose. It is the individual that decides how to deal with the situation when (private) information is disclosed.

> ***"Some people also close colleagues that are active with privacy, are more active and more aware of the possible consequences that happen when certain information is disclosed. Personally, I am more convenient with sharing personal information. I don't think there is a large chance, that someone will misuse my personal information. [ITCH2]"***

This is in an insightful quote as it links two things together. The quote illustrates the differences between how individual approach the condition of privacy. Additionally, it implies that the difference between the individual approaches of the condition of privacy is influenced by a factor *perceived chance of misuse of personal information*.

Participants mentioned two trends relating to the perception on privacy. At one side there is a positive trend, which describes the willingness to disclose personal information. On the other side a group of people who are more critical towards, and more reserved towards the disclosure of private information. Table 6 presents an overview of the characteristics of the positive and negative trend on perception of privacy.

Participants believe that great majority of the people are affected by the positive trend.
More people voluntarily disclose private information on social media, but also other services such as chat services. Besides that, people are willing to trade their personal information for products, money or even only a chance on a prize. Nowadays, it is easier to convince individuals to share their private information. People are willing to fill in forms and share private information for a free sample of a product, to join a contest with a chance to win a car or other big prizes, etc.

It is not new that individuals are trading their personal information for a product or money, though the scale at which this trading happens has grown enormously.

It seems to be normal in modern society to share information as a lot of data is already a public, and the large amount of people are voluntarily disclosing information. This could be explained as a self-enhancing cycle. In which, more people are voluntarily disclosing personal information, which makes it more publically acceptable, which then stimulates more people to voluntarily disclose personal information. Participants mentioned that the majority of the people are not aware of the possible consequences and therefore are more comfortable disclosing information.

The negative trend affects people that are more critical and reserved towards privacy control. They try to avoid as many services that require a lot of private information. They are critical towards services that require certain information without a clear purpose. This small but vocal group of people value private information high. They believe that certain organizations also value this information high and are aware that organizations are willing to pay for private information. There seem to be a market for this information.

Table 6: Overview of the characteristics of the positive trend on perception of privacy and negative perception of privacy.

Positive trend on perception of privacy	**Negative trend on perception privacy**
People are willing to disclose more information On social media voluntarily disclose public data More easy to convince to disclose private information. In trade for products, money or chances on prizes. Many people are not aware of the possible consequences. Values information low. Disclosing personal information becomes more normal and accepted. Since a lot of data is already public. Majority of the people	More critical and reserved towards the disclosure of information. Take in account of all possible consequences, before disclosing information. Values information high. There is market for privacy. Smaller but a vocal group of people.

7. Discussion

In this section we discuss the results from the interview as was presented earlier in this chapter. In section 7.1 we will discuss the findings on the characteristics of big data in the health care. In section 7.2 we will discuss the findings on the common examples of big data in the health care. In this section we will discuss the differences and similarities among the three examples of big data in the health care. In section 7.3 we will discuss the findings related to the conceptual model of privacy and answer the main questions of the interview protocol.

7.1 Characteristics of big data in the health care

Section 6.1.1 presented the set of characteristics that was related to *the processing of big data*. Data has taken forms that became more complex to process, but technologies have been developed and are still in development that can deal with these complexities and allows even complex forms of data to be processed. The results showed that this set of characteristics seem to be mainly relevant to participant with a technological perspective, but participants with a health care perspective also mentioned this.

If we look at the three common examples of big data (presented in section 6.2), these characteristics largely applies to the example of biobanks and not so much to the other two examples. The biobanks have to been described to use advanced analytical techniques and algorithms to process their data. The usage of data within biobanks is mainly focused on processing large volumes of similar data that are aggregated. It seems that these modern algorithms as used within biobanks benefit a lot of this type of data. These big data characteristics: large volume of data and the use of modern algorithms make biobanks an example of big data within the health care. The other two examples: National EHR and Integrated Care are not being associated with these specific big data characteristics.

Section 6.1.2 presented the set of characteristics that were related to collection of data with having a clear purpose beforehand. The general big data literature often does associate big data with the collection without a clear purpose beforehand and the search of patterns that were not expected beforehand. This characteristic is again mainly supported by participants with a technology perspective. Half of the participants with a health care sector background associated this characteristic with big data, while none of the participants with a privacy perspective associated this with big data.

This can be explained with the nature of the data in the health care. Health care data is considered by law as private data. This type of data is treated as private data which requires additional caution. Privacy law put extra requirements on the processing of private data. One important requirement privacy law is that a purpose needs to be formulated beforehand. This makes this set of characteristics of big data in section 6.1.2 incompatible with health care. Participants with a privacy perspective consider privacy law and consider these difficulties.

Section 6.1.3 presented the set of characteristics that seems to be relevant to big data in the health care. Important characteristics of big data in the health care seem to be: integrating scattered data sources, creating a production environment for data, realizing a large scale data infrastructure. These characteristics were mentioned by all experts with a health care perspective. Also experts with a privacy perspective mentioned it often, while experts with a technology perspective hardly discussed this set of characteristics. The common examples of big data in the health care seem indeed to cover these health care characteristics of big data.

7.2 Common examples of big data in the health care

Section 6.2 described the three common examples of big data in the health care. The example of the biobanks has some clear distinguishes from the other examples, while National EHR and Integrated Care showed many similarities. Table 7 presents an overview of the difference and similarities among the three examples of big data in the health care. Below we elaborate

Biobanks are focused on collecting a lot of similar data of a specific type. The data collected is specified for certain purpose from different people. We discussed that biobanks are focused on topic ranging from disease specific biobanks. Biobanks collect data to answers questions related to their topic. Data collected from each person is approached as a sample and more samples enhance the answers for their questions. The developments of connecting biobanks are mainly aimed at improving the amount of samples.

The raw data stored on biobanks does not provide not any insights. The data collected are very specific and has limited context on its own. Algorithms and analytical techniques are essential to process the data. Biobanks are associated with big data mainly because of the large amounts data and algorithms and advanced technologies to process data.

In contrast to Biobanks, both National EHR and Integrated Care are focused on the patient. These developments are focused on combining data sources. Data collected are aimed to provide more contexts to a patient. Individual pieces of data in National EHR and Integrated Care do provide insights. Each piece of data informs part of the context of a patient. Integrating the pieces of information can inform a greater context of a patient. It is believed that availability of more context of a patient will improve health care, as health care providers can make better decisions. National EHR and Integrated Care are both associated with big data, because of the integration of different data sources and the combination of different data to create insights.

Privacy concerns in the examples

The privacy implications for biobanks differ from National EHR and Integrated Care. Biobanks are not allowed to process private data. The essential step for biobanks is therefore to turn private medical data into non-private data. Privacy law doesn’t apply any longer when the data are not considered privacy anymore.

By definition, medical data does posess information about a natural person. The techniques to turn private data in to non-private are focused on make data unidentifiable to a natural person. The main discussion is about the effectiveness of the techniques to turn private data into non-private data. In the literature there are many discussions about the effectiveness of techniques that are used to make data unidentifiable. The aim for biobanks is to make the data they process unidentifiable, without stripping too much information from the data.

The example of National EHR and Integrated care are processing and storing private data. Privacy law requires an organization to take security measures to prevent disruption of information from private data. The main issue here is that is becomes significant more complex to take security measures when dealing with larger amount of data sources and actors. This issue seems to be more significant for Integrated Care as the scope includes a larger variation of actions and data sources, than National EHR.

Table 7: Differences and similarities between the examples.

	Biobanks	National EHR	Integrated Care
Data collected	Aggregated data Focused on specific data Different people	Patient focused Combining scattered health record data sources to enhance information of a single person.	Patient focused Combining data sources to enhance information of a single person.
Data process	Collected data itself does not provide insights. Use of Algorithms to process data Insights from scientific models.	Data is stored for availability Focused on improve the availability of a more complete medical health record Insights from a more complete medical health record.	Data is stored for availability Focused on information sharing between actors of the integrated care to improve context. Insights from greater context.
Privacy	Non-private data Issues with effectiveness of de-identification techniques	Private data Issues with Information security involving large amount of data sources and actors	Private data Issues with Information security involving large amount of data sources and actors

7.3 Conceptual model of privacy

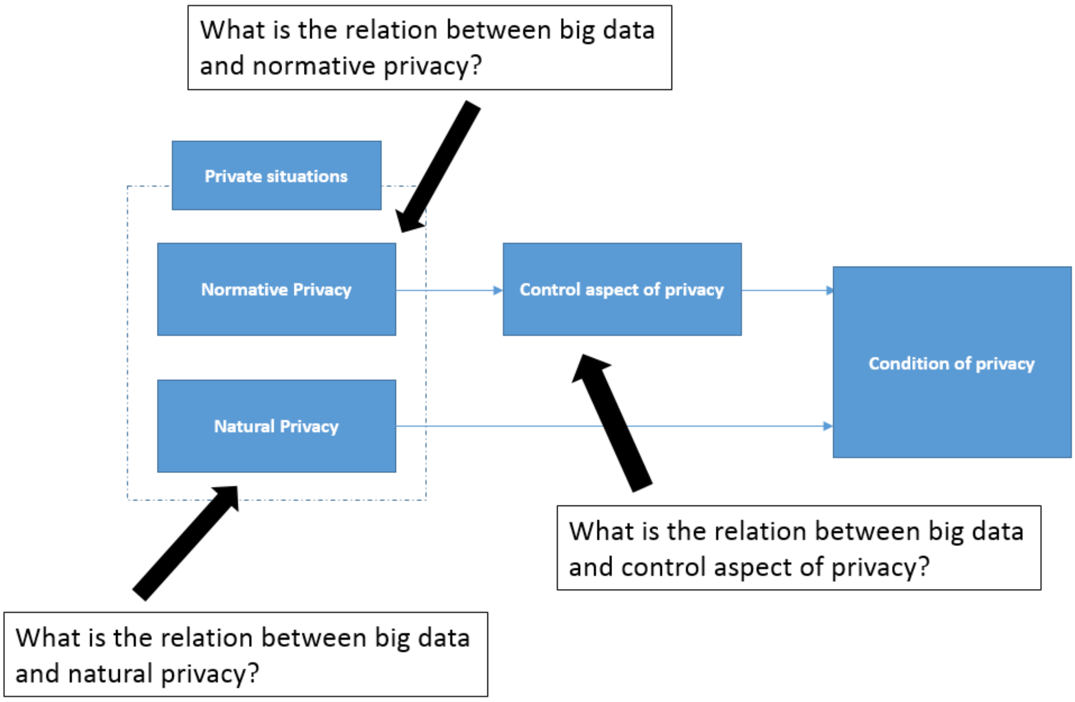
Section 6.3 presented the results of the interviews that are related to the conceptual model of privacy. In this section we will discuss the results related to the conceptual model. The elements of conceptual model are: Private situation, natural privacy, normative privacy, control aspect of privacy and the condition of privacy.

In this section we try to answer the questions that were formulated based on the conceptual model. The purpose of the questions was to investigate the impact of big data on the elements of privacy as we defined with conceptual model.

The following question will be answered:

- What is the relation between big data and natural privacy?
- What is the relation between big data and normative privacy?
- What is the relation between big data and control aspect of privacy?

Table 8: Questions about big data and the conceptual model


What is the relation between big data and normative privacy?
Private situations
Normative Privacy
Control aspect of privacy
Condition of privacy
Natural Privacy
What is the relation between big data and control aspect of privacy?
What is the relation between big data and natural privacy?

7.3.1 The relation between big data and natural privacy

Among the participants there was a general consensus that big data decreases natural privacy. Figure 15, presents the proposed conceptual of the impact of big data on natural privacy. We concluded that big data does impact natural privacy. Big data is often associated with technologies such as: cloud technologies, data storage technologies and developments that are focused on the improvement of availability of large volumes of data. Larger availability of data means that more data are no longer protected by natural means. This is also visible in the health care sector. Within the health care sector it was already common practice to register significant amounts of private data. However technological development allows us to gather more information and store more information. More information can be gathered with modern sensors and modern laboratory tests. Modern genetic testing allows a larger portion of a patient's DNA to be analyzed in a shorter time, which results in more (medical) data. Modern storage technologies and the lower cost of storage allow us to store these data and increase the availability of data and therefore decrease natural privacy.

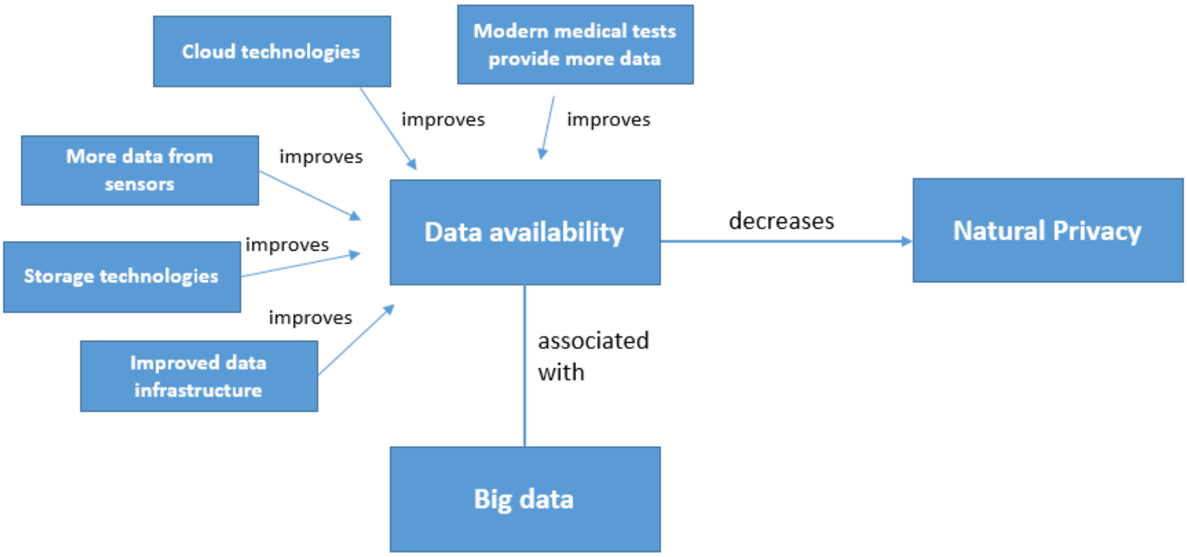


Figure 15: Conceptual model: Impact Big data on Natural privacy

7.3.2 The relation between big data and normative privacy

We observed that normative privacy in the form privacy laws is becoming stricter. At a national level, privacy laws are becoming stricter by introducing the legal requirement to disclose data breaches and by giving the regulatory authority of privacy more power and sanction capabilities.

Figure 16 presents a conceptual model that illustrates how big data impacts normative privacy. We concluded that natural privacy is decreasing due to factors such as big data. As a response on that, governments are trying to compensate the decrease of natural privacy with stricter privacy laws in order to maintain a certain level privacy (see Figure 17).

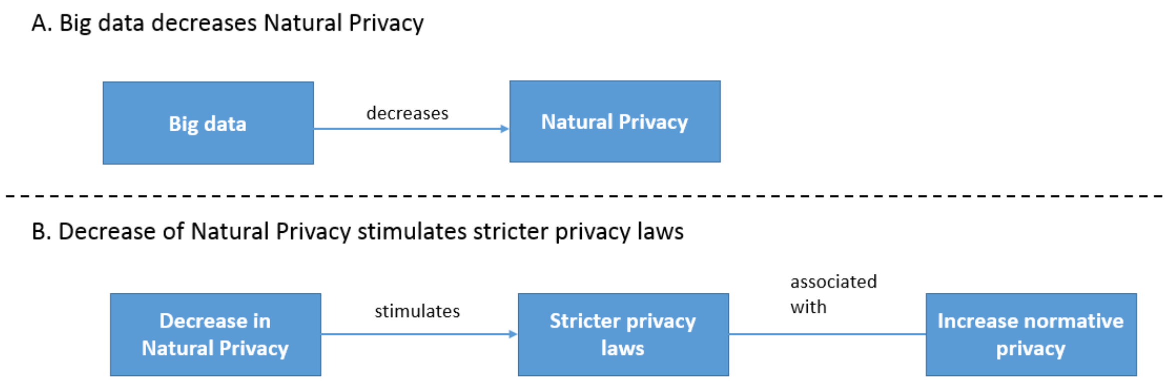

A. Big data decreases Natural Privacy
Big data
decreases
Natural Privacy
B. Decrease of Natural Privacy stimulates stricter privacy laws
Decrease in Natural Privacy
stimulates
Stricter privacy laws
associated with
Increase normative privacy

Figure 16: Proposed conceptual model: Impact Big data on Normative privacy. The relation can be broken down into two parts. First, as we concluded earlier, Big data decreases Natural Privacy. As a reaction governments are trying to compensate the decrease with stricter privacy law and regulation. Stricter privacy law is associated with increase in normative privacy.

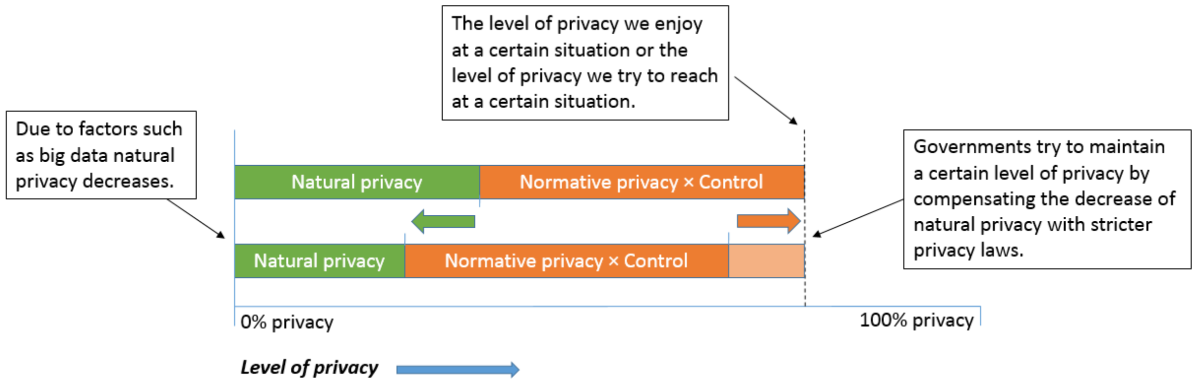

The level of privacy we enjoy at a certain situation or the level of privacy we try to reach at a certain situation.
Due to factors such as big data natural privacy decreases.
Natural privacy
Normative privacy × Control
Natural privacy
Normative privacy × Control
Governments try to maintain a certain level of privacy by compensating the decrease of natural privacy with stricter privacy laws.
0% privacy
100% privacy
Level of privacy

Figure 17: Illustration that shows that when Natural privacy decreases, government tries to compensate by increasing Normative privacy (stricter privacy laws) in order to maintain a certain level of privacy.

7.3.3 The relation between big data and control aspect of privacy

Participants approached the control aspect of privacy as the measures that are taken to secure information against disruption. NEN7510 was used to discuss the norms of these measures as it is currently the most accepted norm for information security measures in the Dutch health care sector. The discussion was mostly about the norms. Hence, the information about the measures itself and the information about implementation of these measures was limited. Participants do believe that big data projects in the health care will deal with increased difficulties to implement security measures that meet the security norms. Those big data projects involve a significant number of users and significant number of data and data sources. Information access control for large size of numbers of users and large number of data sources, introduces complexities.

For example, the developments in the ketenzorg (Integrated Care) and social domain where different actors try to integrate their information systems. The scope of these actors and data sources is really broad. The data goes beyond a medical scope and also includes data from a social context. The stakeholders involved are from a great variety scattered over healthcare and social domain. An access table, with on the rows the various users and in the columns the access rights to the data source can grow enormously in such situation. Setting up the access rights for a large variety of users in combination with the access right to a large number of data sources is a very complex task and will become significant more complex as both more users and more data sources are going to be included.

FINDING: Participants do believe that big data projects in the health care will deal with increased difficulties to implement security measures that meet security norms.

8. Conclusion

The purpose of this study is to create understanding on how big data impacts privacy issues in the healthcare. Our aim in this study was to create understanding by exploring the relation between big data and privacy in the health care. Therefore, the following research objective was formulated:

"To gather knowledge on the relation between big data and privacy in the health care."

In order to achieve this objective, four research questions were formulated and addressed during different steps in the research. Section 8.1 will present the findings and conclusions of the four research steps. Section 8.2 will discuss the contributions of this research to the literature. Finally in Section 8.3, limitation of this research will be discussed and the recommendations will be provided.

8.1 Answers to the research questions

RQ1: What is big data and what are the technologies related to big data and how does it differ from business intelligence?

We discussed the concepts of business intelligence and big data as they form the domain of this research. Based on the literature review we found three aspects of big data.

1. **Need to capture more data**. The need to acquire, store and process more data. Combining new sources or new types of data with existing sources and data types.

2. **Big data technologies**. New infrastructure and technologies that enables the acquisition, storage and processing (3V data), to be done more effectively, efficiently, easily.

3. **Big data analytics**. Advanced analytics that can gain get advanced insights from data.

From the interviews we learned that in the context of health care different aspects of big data matter more. We concluded that in the health care big data is often associated with the ***integration of scattered data sources in the health care***. From the interview we learned three notable examples of big data in the health care: Biobanks, National EHR and Integrated Care (See section 6.2.).

RQ2: What are relevant theories to analyze the concept of privacy?

In Chapter 3 we discussed four theories of privacy:
- Nonintrustion theory
- Seclusion theory
- Control theory
- Restricted access theory

The nonintrusion theory defines privacy as the “right of being let alone” (Warren & Brandeis, 1890). *The seclusion theory* defines privacy as “being alone” (Tavani, 1999; Alan F Westin, 2003). *The control theory* defines privacy in terms of control one has over information about oneself (Tavani & Moor, 2001; Tavani, 1999; A F Westin, 1967). According to Fried (1984) “Privacy is not simply an absence of information about us in the minds of others, rather it is the control we have over information about ourselves”. *The restricted access theory* defines privacy as “the condition of being protected from unwanted access by others, including access to one's personal information” (Bok, 1989). Additionally, we discussed the notion of private situations, which distinguish normative private situation from natural private situations.

Based on the privacy literature we created a conceptual model of privacy. That includes the elements: *private situations, natural privacy, normative privacy, control aspect of privacy and the condition of privacy*. This conceptual model has been used as a guide to analyze the concept of privacy.

RQ3: What is privacy in the health care and what are the regulations related to privacy in the health care?

A literature review resulted in three Dutch laws that describe the normative privacy in the health care (See section 4.3). The three Dutch laws are:

- Wet bescherming persoonsgegevens (Wbp) (referred as Dutch privacy law)
- Wet op de geneeskundige behandelingsovereenkomst (WGBO)
- Zorgverzekeringswet (Zvw)

The Wbp defines information as personal information if it *(1) consists information about a natural person (a real person)* and *(2) the person is identifiable*. Dutch privacy law considers some types of personal information even more sensitive. Information concerning health is considered as sensitive personal information. The processing of sensitive personal information has to follow even more strict regulations.

Privacy law is becoming stricter
We observed that Dutch privacy law is becoming stricter. An extension of the law requires organizations that are processing private data to report data breaches. In case of a data breach that violated the required security measures (Wbp article 13) the organization responsible for the processing of private data are required to report that breach to the regulatory authority of privacy (CBP). When a breach has serious harmful consequences to the ones concerning the private data, the responsible organization of data processing are also required to inform the concerned ones. The second major change to Dutch privacy law will grant the Dutch authority body for privacy (CBP) more power and will be able to sanction organizations with fines of an amount up to €810.000.

RQ4: What is the relation between big data and privacy in the context of the healthcare?

We tried to answer this question by separating privacy into three elements of the conceptual model of privacy (see section 6.3). This resulted into following three sub-questions:

- **What is the relation between big data and natural privacy?**
- **What is the relation between big data and normative privacy?**
- **What is the relation between big data and control aspect of privacy**

The relation between big data and natural privacy

We concluded that big data does impact natural privacy. Big data is often associated with technologies such as: cloud technologies, data storage technologies and developments that are focused on the improvement of availability of large volumes of data. Larger availability of data means that more data are no longer protected by natural means. This is also visible in the health care sector. Within the health care sector it was already common practice to register significant amounts of private data. However a technological development allows us to gather more information and store more information. More information can be gathered with modern sensors and modern laboratory tests. Modern genetic testing allows a larger portion of a patient's DNA to be analyzed in a shorter time, which results in more (medical) data. Modern storage technologies and the lower cost of storage allow us to store these data and increase the availability of data and therefore decrease natural privacy.

The relation between big data and normative privacy

We observed that normative privacy in the form privacy laws is becoming stricter. At a national level, privacy laws are becoming stricter by introducing the legal requirement to disclose data breaches and by giving the regulatory authority of privacy more power and sanction capabilities.

We concluded that natural privacy is decreasing due to factors such as big data. As a response on that, governments are trying to compensate the decrease of natural privacy with stricter privacy laws in order to maintain a certain level privacy.

The relation between big data and control aspect of privacy

Participants do believe that big data projects in the health care will deal with increased difficulties to implement security measures that meet the security norms. Those big data projects involve a significant number of users and significant number of data and data sources. Information access control for large size of numbers of users and large number of data sources, introduces new complexities.

Common example of big data showed that the scope of these actors and data sources is often really broad. In case of the Integrated Care data goes beyond a medical scope and also includes data from a social context.

The actors involved in such system are from great variety scattered over healthcare and social domain. Setting up the access rights for a large variety of users in combination with the access right to a large number of data sources is a very complex task and will become significant more complex as both more users and more data sources are going to be included.

8.2 Contributions to the literature

One of the key contributions of this research is the creation of the conceptual model of privacy, which combined different aspects of privacy in a single conceptual model. The conceptual model was developed by combining different theories from the privacy literature: nonintrusion theory, the seclusion theory, the control theory and the restricted access theory.

The conceptual model of privacy guided this study to investigate the relation between privacy and big data in the health care sector. The conceptual model has shown to be helpful to analyze privacy in this study by including the different elements of privacy. The different elements in the conceptual model of privacy provided focus points of privacy to discuss about within the interviews. Consequently, we focused on natural privacy, normative privacy and the control aspect of privacy in our interviews.

Another contribution of this research is the exploratory efforts on big data and privacy in the health care. The objective of this research was to gather knowledge on security and privacy issues of big data within the healthcare sector. This study contributed to the literature by gathering knowledge by collecting and analyzing data from interviews with experts. The findings from the interviews tried to answer what the relation is between big data and privacy within the health care.

8.3 Limitations and future research

One of the limitations of this research is related to sample size of the interviews. Initially, it was aimed to have at least ten participants for this research. Unfortunately this was not successful and the interviews were only conducted with eight participants. This could influence the validity of the findings of the study.

A recommendation for future research is to increase the sample size and increase the diversification of experiences of the interviewees.

Another limitation of this research is related to finding that participants had difficulties to answer questions related to information access control. Difficulties in getting proper responses to the question related to information access control could indicate that the participants included in the sample were not suitable to answer the questions related to the information access control. Suitable candidates to answer information access control related questions would be information security officers, or information security experts that are responsible for the access control within the health care.

Future research could gather more knowledge about information access control by including information security officers, or information security experts that have experience in the health care.

The scope of this study was to interview only experts. A recommendation for future research is to perform this study from a user perspective (e.g. doctors, nurses, patients etc.) It could be valuable to see to what extend the opinion of users differ from the opinions of the experts.

The questions related to information access control were formulated to cover the control aspect of privacy of our conceptual model. Participants had difficulties to answer the questions related to information access control and instead referred to information security and the NEN7510 to cover the control aspect of privacy of our conceptual model. In hindsight, this could indicate that the questions related to information access control were not very suitable to cover the control aspect of privacy.

Bibliography

Allen, A. L. (1988). *Uneasy access: Privacy for women in a free society*. Rowman & Littlefield.

Altman, I. (1975). The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding.

Appari, A., & Johnson, M. E. (2010). Information security and privacy in healthcare: current state of research. *International Journal of Internet and Enterprise Management*, *6*(4), 279. doi:10.1504/IJIEM.2010.035624

Ardagna, C. a., De Capitani di Vimercati, S., Foresti, S., Grandison, T. W., Jajodia, S., & Samarati, P. (2010). Access control for smarter healthcare using policy spaces. *Computers & Security*, *29*(8), 848–858. doi:10.1016/j.cose.2010.07.001

Azvine, B., Cui, Z., & Nauck, D. (2005). Towards real-time business intelligence. *BT Technology Journal*. Retrieved from http://link.springer.com/article/10.1007/s10550-005-0043-0

Barth, A., Datta, A., Mitchell, J. C., & Nissenbaum, H. (2006). Privacy and contextual integrity: framework and applications. *2006 IEEE Symposium on Security and Privacy (S&P'06)*, 15 pp.–198. doi:10.1109/SP.2006.32

Bhatti, R., & Grandison, T. (2007). Towards Improved Privacy Policy Coverage in Healthcare Using Policy Refinement, 158–173.

Bok, S. (1989). *Secrets: On the ethics of concealment and revelation*. Random House LLC.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, (2012), 662–679. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878

Cattell, R. (2010). Scalable SQL and NoSQL Data Stores, *39*(4).

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., … Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, *26*(2), 4.

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:BUSINESS+INTELLIGENCE+AND+A+NALYTICS:+FROM+BIG+DATA+TO+BIG+IMPACT#0

Cloud Security Alliance. (2013). Expanded Top Ten Big Data Security and Privacy Challenges, (April).

Decandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., … Vogels, W. (2007). Dynamo: amazon's highly available key-value store. *ACM SIGOPS …*, 205–220. Retrieved from http://dl.acm.org/citation.cfm?id=1294281

Dekker, M. a. C., & Etalle, S. (2007). Audit-Based Access Control for Electronic Health Records. *Electronic Notes in Theoretical Computer Science*, *168*(1), 221–236. doi:10.1016/j.entcs.2006.08.028

Elbashir, M., Collier, P., & Davern, M. (2008). Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *International Journal of Accounting Information Systems*, *9*(3), 135–153. doi:10.1016/j.accinf.2008.03.001

Fidelis Cybersecurity Solutions. (2014). *Current Data Security Issues of NoSQL Databases*.

Forsman, S. (1997). OLAP Council white paper. *OLAP Council*.

Fried, C. (1984). *Philosophical dimensions of privacy: An anthology*. (F. D. Schoeman, Ed.). Cambridge University Press.

Gavison, R. (1980a). Privacy and the Limits of Law. *Yale Law Journal*, *89*(3), 421–471. Retrieved from http://www.jstor.org/stable/795891

Gavison, R. (1980b). Privacy and the Limits of Law. *Yale Law Journal*, *89*(3), 421–471.

Gens, F. (2012). TOP 10 PREDICTIONS IDC Predictions 2012 : Competing for 2020, (December 2011).

Google. (2011). *Google Flu Trends*. Retrieved from http://www.google.org/flutrends/about/how.html

Groves, P., & Knott, D. (2013). The " big data " revolution in healthcare, (January).

Hamami, O. (2011). Big Data Security : Understanding the Risks, *19*(2), 20–27.

Han, J. (1997). OLAP mining: An integration of OLAP with data mining. In *Proceedings of the 7th IFIP* (Vol. 2, pp. 1–9).

Himma, K., & Tavani, H. (2008). *The handbook of information and computer ethics*. (K. Himma & H. Tavani, Eds.). Retrieved from http://books.google.com/books?hl=en&lr=&id=ZC7SDyPZUMoC&oi=fnd&pg=PR7&dq=THE+HANDBOOK+OF+INFORMATION+and+computer+ethics&ots=lca-BbB153&sig=I1Rxi-IYvq0ulFs9tm6G5I0KSl0

Khan, R. A., & Quadri, S. M. K. (2012). Business intelligence: an integrated approach. *Business Intelligence Journal*, *5*(1), 64–70.

Malin, B. (2012). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

Manyika, J., Chui, M., Brown, B., & Bughin, J. (2011). Big data: The next frontier for innovation, competition, and productivity, 1–18. Retrieved from http://www.citeulike.org/group/18242/article/9341321

Margulis, S. T. (2011). Privacy Online, 9–17. doi:10.1007/978-3-642-21521-6

Moor, J. (1990). The ethics of privacy protection. *Library Trends*, *39*, 69–82. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Towards+a+Theory+of+Privacy+in+the+Information+Age#0

Moor, J. H. (1997). Towards a Theory of Privacy in the Information Age, (September), 27–32.

Nambiar, R., Bhardwaj, R., Sethi, A., & Vargheese, R. (2013). A look at challenges and opportunities of Big Data analytics in healthcare. *2013 IEEE International Conference on Big Data*, 17–22. doi:10.1109/BigData.2013.6691753

Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev.*, 101–139. Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/washlr79&section=16

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA L. Rev.* Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/uclalr57&section=48

Parent, W. (1983). Privacy, morality, and the law. *Philosophy & Public Affairs*, *12*(4), 269–288. Retrieved from http://www.jstor.org/stable/2265374

Perera, G., Holbrook, A., Thabane, L., Foster, G., & Willison, D. J. (2011). Views on health information sharing and privacy from primary care practices using electronic medical records. *International Journal of Medical Informatics*, *80*(2), 94–101. doi:10.1016/j.ijmedinf.2010.11.005

Pooley, R., Coady, J., Schneider, C., Linger, H., Barry, C., & Lang, M. (2013). *Information Systems Development: Reflections, Challenges and New Directions*. Springer.

Pritchett, D. (2008). Base: An acid alternative. *Queue*, *6*(3), 48–55.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3. doi:10.1186/2047-2501-2-3

Rostad, L., & Edsberg, O. (2006). A study of access control requirements for healthcare systems based on audit trails from access logs. *Computer Security Applications …*, 1–9. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4041165

Sekaran, U., & Bougie, R. (2009). *Research Methods for Business*.

Smith, H., Dinev, T., & Xu, H. (2011). Information privacy research: an interdisciplinary review. *MIS Quarterly*, *35*(4), 989–1015. Retrieved from http://dl.acm.org/citation.cfm?id=2208950

Tavani, H. (1999). KDD, data mining, and the challenge for normative privacy. *Ethics and Information Technology*, (1), 265–273. Retrieved from http://link.springer.com/article/10.1023/A:1010051717305

Tavani, H., & Moor, J. (2001). Privacy protection, control of information, and privacy-enhancing technologies. *ACM SIGCAS Computers and Society*, (March), 6–11. Retrieved from http://dl.acm.org/citation.cfm?id=572278

Tene, O., & Polonetsky, J. (2012). Privacy in the age of big data: A time for big decisions. *Stanford Law Review Online*, 63–69. Retrieved from http://www.stanfordlawreview.org/online/privacy-paradox/big-data

Vogels, W. (2009). Eventually consistent. *Communications of the ACM*. Retrieved from http://dl.acm.org/citation.cfm?id=1435432

Wager, K. A., Lee, F. W., & Glaser, J. P. (2013). Health Care Information Systems : A Practical Approach for Health Care Management , Third Edition Chapter 6 : Federal Efforts to Enhance Quality of Patient Care Through the Use of Health Information Technology, (c).

Walmart. (2012). *Semantic Search Technology Increases Shopper Conversion Rate*. Retrieved from http://news.walmart.com/news-archive/2012/08/30/walmart-announces-new-search-engine-to-power-walmartcom

Warren, D., & Brandeis, D. (1890). The Right to Privacy, *4*(5), 192–220.

Westin, A. F. (1967). *Privacy and freedom*. *Privacy and freedom* (p. 487). Atheneum. Retrieved from http://search.proquest.com/docview/37771098?accountid=27026

Westin, A. F. (2003). Privacy and Freedom, *59*(2), 1–37.

Appendix A: Interview protocol

Interview format and guidelines

Introduction

- Introduce myself
- Discuss topic of the master thesis
- Mention the research objective: to investigate the impact of big data on privacy in the context health care.
- Mention planning: want to interview at least five experts with expertise with health care privacy and big data

Interview guidelines

- Duration: interview will take about an hour
- Focus: Discuss the different concepts of privacy and the impact of big data on privacy
- Mention that in order to ensure quality of the data analysis it would be very beneficial to record the interview
- Ask permission to record conversation
- Discuss confidentiality agreement
    - Only I will access these recordings
    - I will only use these recordings for this research
    - All individual and company names will not be mentioned in the research

Topic list

General information

- Ask details of person being interviewed: name, job title, experience.

Interview section 1: Big data

1. Can you shortly describe what you understand with the concept of big data?

(Key aspects of big data found in the literature review in Chapter 2)

- **Need to capture more insights from data**. The need to acquire storage and process more data. Combining new sources or new types of data with existing sources and data types. To create new insights.
- **Big data technologies**. New infrastructure and technologies that enables the acquisition, storage and processing (3V data), to be done more effectively, efficiently, easily.
- **Big data analytics**. Advanced analytics can assist in extracting insights from large volumes and/or unstructured data. Insights that is difficult or impossible to extract without the right tools.

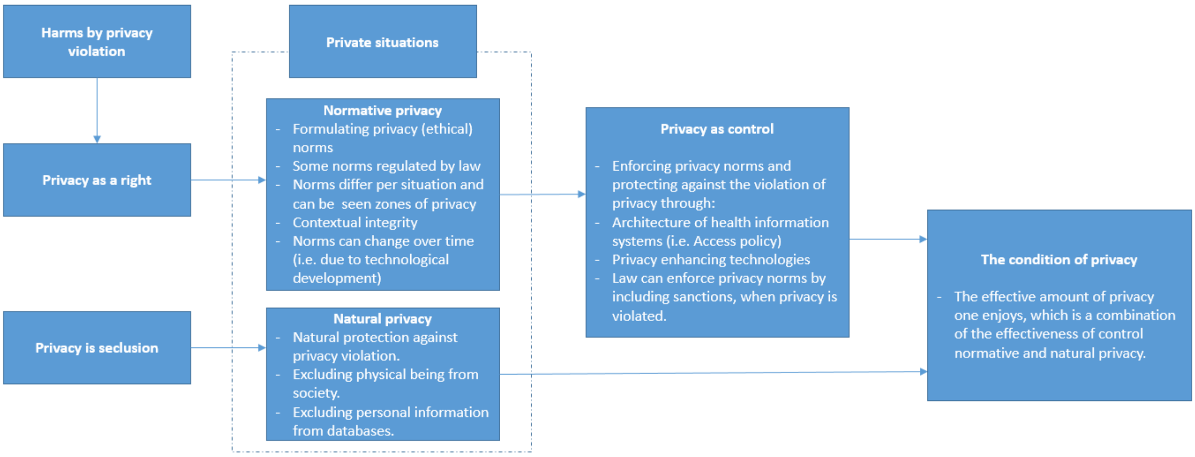
Interview section 2: Privacy

2. Can you shortly describe what you understand with privacy?

(Findings from the literature review on privacy in chapter 3)

- *"right of being let alone"*	***Nonintrusion theory***
- *"being alone"*	***Seclusion theory***
- *"the control one has over information about itself"*	***Control theory***
- *"privacy is the condition of being protected from unwanted access by others, including access to one's personal information"*	***Restricted Access Theory***

Introduce and explain the conception of privacy in this research and that the focus of this research is natural privacy, normative privacy and control of privacy:



Harms by privacy violation
Privacy as a right
Privacy is seclusion
Private situations
Normative privacy
- Formulating privacy (ethical) norms
- Some norms regulated by law
- Norms differ per situation and can be seen zones of privacy
- Contextual integrity
- Norms can change over time (i.e. due to technological development)
Natural privacy
- Natural protection against privacy violation.
- Excluding physical being from society.
- Excluding personal information from databases.
Privacy as control
- Enforcing privacy norms and protecting against the violation of privacy through:
- Architecture of health information systems (i.e. Access policy)
- Privacy enhancing technologies
- Law can enforce privacy norms by including sanctions, when privacy is violated.
The condition of privacy
- The effective amount of privacy one enjoys, which is a combination of the effectiveness of control normative and natural privacy.

Interview section 3a: Relation Big data and natural privacy

3a. What is the relation between big data and natural privacy?

- Is this significant for the healthcare?
- More private and public information is being stored.
- Duplication of data; hard to remove data.
- Anonymization of private data doesn't provide natural privacy anymore.
- Does big data decrease our natural privacy?

Interview section 3b: Relation Big data and normative privacy

3b. What is the relation between big data and normative privacy?

- Are current norms sufficient to protect against the harms of privacy?
- Will we rely more on normative privacy due to the developments of big data?
- Do the norms need to be change for big data?

Interview section 3c: Relation Big data and the control aspect of privacy

3c. what is the relation between big data and control aspect of privacy?

- How does information access control in the health care currently look like?
- Can you tell me more about the use of audit trails to improve information access control?
- Can you tell me more about a-posteriori information access control?
- How could big data impact (a-posteriori) information access control?

Appendix B: Findings on Information access control in the healthcare

A literature review on information access control was performed aiming to explore the situation of information access control.

As we discussed in section 4.4 Information access control is about enforcing rules to ensure that only authorized users get access to informational resources on a system. Medical data on health information systems should only be accessible to the actors that are involved in the treatment of a patient.

According to the literature role based access control (RBAC) is the most common model of access control, this is a model that predefines access rules based on the role of the person that try to access information. In the health care sector, there is besides (standard) information access control, also an emergency access control “break the glass” (“Rode knop procedure” in Dutch) that grants the user temporary full access, but triggers enhanced logging of the access instead. The purpose of this mechanism is to guarantee to health information in the case of emergency. Rostad & Edsberg (2006) found in their study that BtG is not used as an exception mechanism, since more than half of all the patients of Norwegian hospitals has had his health record accessed by BtG. The access control system in the healthcare is too reliant on BtG, which is a source of privacy concerns.

Several studies proposed alternatives for access control systems in the healthcare. Dekker (2007) demonstrated an Audit-Based access control for the EHR setting that minimize a-priori access control, but focus on a-posteriori control based on audit logic. The main benefit of Audit-Based access control is that it offers the flexibility that is required for the healthcare. Bhatti and Grandison (2007) propose the PRIvacy Management Architecture (PRIMA), which built around notion of *policy refinement*. The main idea behind policy refinement is that audit trails of emergency information access, can be used to align the a-priori access policy with the actual clinical workflow, which should lower reliance of emergency access to health information.

The interviews resulted in following findings:

- **Our conclusions that were derived from a study by Rostad & Edsberg (2006) did not meet the experiences of the participants.**

Based on our conclusion from a study by Rostad & Edsberg (2006) we formulated the following statement: “in practice, emergency access is not used as an exception, but instead is used as an integral part of the practical access control in the health care”. The participants responded not to be familiar with this claim. Participants suggested was that these numbers were over exaggerated or not applicable to situation in the Netherlands.

- **There was limited response on the topic of information access control.**

There was a limited response on the topic of information access control. Initially it was the aim to discuss alternative information access control models for the health care and how to adjust the access control models to improve privacy in the health care.

- **Current form of information access control in the health has weaknesses. The weaknesses are a more cultural problem than a technical problem.**

Appendix C: Sampling matrix on background of the participants

Table 9 presents a two dimensional matrix that categorizes the background of the experts. The x-axis presents the perspective of the expert and on the y-axis whether the expert is active inside or outside the health care.

Table 9: Two dimensional matrix of the categorizing the experts on two characteristics: Perspective and Inside or our Health care sector. The sample strategy is to include participants that cover all the six situations.

	Technical Perspective	**Health care Perspective**	**Privacy Perspective**
Inside Hospital			
Outside Hospital			

Table 10 presents the participants in the sampling matrix. The participants sample for this research did not cover the whole spectrum of our characteristics. Only one of the participants has a position in the health sector. This could indicate that the sample is overrepresented by people outside the health care. Additional, our sample covers a technical and privacy perspective by actor who is active at an organization within the health care.

Table 10: The participants placed in the sampling matrix.

	Technical Perspective	**Health care Perspective**	**Privacy Perspective**
Inside Health care	-	[CRTH]	-
Outside Health care	[DMT] [DST]	[ITCH1] [ITCH2]	[ITSP] [PJP] [HISP]

Appendix D: Frequency tables of Big data related codes

Table 11 and Table 12 show the absolute frequency tables of the related codes. These developments are covering very technical aspects of big data. The tables show that these aspects are mainly mentioned by the participants with a more technical background.

Table 11: Frequency table (Absolute): Big data difficult to process.
The table shows how many participants mentioned a certain code.

Code	**Technology (out of 2)**	**Healthcare (out of 3)**	**Privacy (out of 3)**	**Total (out of 8)**
Forms of data that were difficult to process	2	1	1	4
Advanced analytics and algorithms	2	1	0	3

Table 12: Frequency table (Relative): Big data difficult to process.
The table shows relative % of how many participants mentioned a certain code.

Code	**Technology**	**Healthcare**	**Privacy**	**Total**
Forms of data that were difficult to process	100%	33,3%	33,3%	50%
Advanced analytics and algorithms	100%	33,3%	0%	37,5%

Table 13 and Table 14 show the frequency table of the related codes. This characteristic of big data is also mainly supported by participant with a more technical background.

Table 13: Frequency table (Absolute): Collecting without having a clear purpose beforehand.

Code	**Technology (out of 2)**	**Healthcare (out of 3)**	**Privacy (out of 3)**	**Total (out of 8)**
Collecting data without a clear purpose beforehand	2	1	0	4
Exploring new undiscovered patterns in large volumes of data.	2	2	0	4

Table 14: Frequency table (Relative): Collecting without having a clear purpose beforehand.

Code	**Technology**	**Healthcare**	**Privacy**	**Total**
Collecting data without a clear purpose beforehand	100%	66,6%	0%	50%
Exploring new undiscovered patterns in large volumes of data.	100%	66,6%	0%	50%

All participants agreed on the following standard characteristics of big data as often is found the general literature of big data. They refer to multiple sources of data, or data types

Table 15: Frequency table (Absolute): Standard characteristics of big data.

Code	**Technology (out of 2)**	**Healthcare (out of 3)**	**Privacy (out of 3)**	**Total (out of 8)**
Multiple sources data	2	3	3	8
Different types and different structure of data	2	3	3	8
Large volumes of structured and unstructured data	2	3	3	8

Appendix E: Frequency tables of Privacy related codes

All participants agreed that in essence privacy protecting the personal space of individuals and restricting unwanted access by others to personal data. However, the majority of participants were mainly focused in describing the legal aspect of privacy in terms of legal borders that is provided by privacy law and regulation.

Table 16: Frequency table (Absolute): Privacy

Code	**Technology (out of 2)**	**Healthcare (out of 3)**	**Privacy (out of 3)**	**Total (out of 8)**
Privacy is protecting the personal space of individuals	2	3	3	8
Restricting unwanted access by other to personal data	2	3	3	8
Privacy law and regulation	1	3	3	7

Table 17: Frequency table (Relative): Privacy

Code	**Technology**	**Healthcare**	**Privacy**	**Total**
Privacy is protecting the personal space of individuals	100%	100%	100%	100%
Restricting unwanted access by other to personal data	100%	100%	100%	100%
Privacy law and regulation	50%	100%	100%	87,5%

In relation to the conceptual model of privacy, participants described privacy laws in elements that cover normative privacy and control aspect of privacy. Three notable terms were used by the participant to describe legal aspect of privacy: the Dutch privacy law (Wbp), the EU Data protection directive (95/46/EC), and the information security standard used in the Dutch health care (NEN7510).

Table 18: Frequency table (Absolute): Privacy law and regulation

Code	**Technology (out of 2)**	**Healthcare (out of 3)**	**Privacy (out of 3)**	**Total (out of 8)**
Dutch privacy law (Wbp)	1	3	3	7
EU data protection directive (95/46/EC)	1	1	3	5
Information security standard (NEN7510)	0	2	3	5

Table 19: Frequency table (Relative): Privacy law and regulation

Code	**Technology**	**Healthcare**	**Privacy**	**Total**
Dutch privacy law (Wbp)	50%	100%	100%	87,5%
EU data protection directive (95/46/EC)	50%	33,3%	100%	62,5%
Information security standard (NEN7510)	0%	66,6%	100%	62,5%