Aalto University
School of Science
Master's Programme in Industrial Engineering & Management

Mikko Rajala

The Importance of Finding Right Data: Case Healthcare Operational Improvement Projects

Master's Thesis

Espoo, July 27, 2018

Supervisor: Timo Seppälä, Professor

Thesis advisor(s): Niki Kotilainen, MS.c. (Tech.)

Author Mikko Rajala

Title of thesis The Importance of Finding Right Data: Case Healthcare Operational Improvement Projects

Master's programme in Industrial Engineering & Management

Thesis supervisor Timo Seppälä

Major or Minor/Code Strategy & Venturing

Department Industrial Engineering & Management

Thesis advisor(s) Niki Kotilainen

Date 27.7.2018 **Number of pages** 76+3 **Language** English

Abstract

The utilization of data in healthcare improvement projects is currently a very topical subject. Several public and private companies have shown the value of utilizing data to improve operational efficiency. Not all datasets are, however, equally useful – thus, understanding of the data quality is required to ensure correct decision-making. Currently, two streams of literature exist to guide the improvement teams: the literature on operational improvement, e.g. through methods such as Total Quality Management, Lean, and Six Sigma, and the literature on data quality. From the point-of-view of an improvement project team, a linkage between these two streams of literature is missing. This paper aims to bridge the gap between the two streams of literature by helping healthcare improvement teams to assess whether the data quality is sufficient to support decision-making.

The academic framework illustrates, how the viewpoint of data quality has transformed from an intrinsic focus on the 1970s, to fitness for use on the 1990s, finally to describing the specifics of the new trends, such as big data or unstructured data, in the 2010 onwards.

Using the case study method, the findings were expanded by observing an improvement project in a private Finnish healthcare company. Together with the project team, I went through an iterative process with five steps: each of which was guided by a distinctive, new set of data. Finally, the actual improvement was gained by gathering the data manually: a dataset which was highly relevant for the end users, but likely to be intrinsically less robust as the previous datasets.

As a conclusion, the current data quality literature can bring only modest guidance for the improvement teams in terms of choosing the right dataset. Rather, a new model for the data quality in healthcare operational improvement was created. The model suggests that the teams should first consider whether the dataset is relevant for the goal of the improvement project. After that, the improvement team should consider if the dataset can add value to reaching the goal of the project. After these two steps, the other key data quality attributes linking to the following four dimensions come to play: accessibility, intrinsic, representational, and contextual quality.

Keywords Data quality, improvement projects, healthcare improvement

Tekijä Mikko Rajala		
Työn nimi Oikean datan löytämisen tärkeys: case terveydenhuollon operaatioiden kehittämisprojektit		
Koulutusohjelma Tuotantotalous		
Valvoja Timo Seppälä		
Pää tai sivuaine/koodi Strategy & Venturing		
Työn ohjaaja(t) Niki Kotilainen		
Päivämäärä 27.7.2018	**Sivumäärä** 76+3	**Kieli** englanti

Tiivistelmä

Datan käyttäminen terveydenhuollon prosessikehityksessä on laajaa kiinnostusta herättävä aihe. Kaksi pää kirjallisuussuuntaa on kehittynyt datan laadun tutkimiseksi: kirjallisuus operaatiokehityksestä eli aiheista, kuten TQM, Lean ja Six Sigma, ja kirjallisuus datan laadusta. Nämä kaksi suuntausta ovat kuitenkin usein riittämättömiä kehitystiimien päätöksenteon tueksi. Tämän diplomityön tarkoitus on yhdistää nämä kaksi kirjallisuussuuntausta frameworkiksi, joka auttaa tiimejä arvioimaan datan soveltuvuutta omaan kehitysprojektiinsa.

Työn kirjallisuuskatsaus kuvaa, miten käsitys datan laadusta on muuttunut 1970-luvulta nykypäivään. 1970-luvulla datalaadun kirjallisuuden fokus oli sisäisessä laadussa (intrinsic quality). 1990-luvulle siirtyessä painopiste siirtyi kuvailemaan datan laatua sen soveltuvuuden kautta (fitness for use), ja 2010-luvulle siirryttäessä kirjallisuuteen tuli mukaan uusia trendejä, kuten big data tai strukturoimaton data.

Tuloksien tueksi seurattiin kehitysprojektia, joka toteutettiin suomalaisessa yksityisessä terveydenhuollon yrityksessä. Yhdessä projektitiimin kanssa, kirjoittajan matka projektin edetessä voidaan tiivistää viiteen vaiheeseen, joista jokaisessa uusi datasetti näytteli tärkeää roolia. Lopulta suurin edistysaskel projektissa saatiin keräämällä data manuaalisesti. Manuaalisesti kerätty data oli erittäin relevantti projektille, mutta sisäisiltä ominaisuuksiltaan huonompi.

Tulosten pohjalta voidaan päätellä, että nykyinen kirjallisuus datan laadusta voi tuoda enintään keskinkertaista tukea kehitystiimien datan laadun arvioinnille. Tästä syystä uusi malli data laadun tutkimiselle terveydenhuollossa luotiin työn tuloksena. Malli ehdottaa, että projekti tiimien pitäisi ensimmäisenä arvioida datasetin relevanttiutta käyttötarkoitukselle. Toisena askeleena tiimin kannattaa miettiä onko data arvokasta vastaamaan projektin senhetkisiin haasteisiin. Näiden kahden askeleen jälkeen, tiimin kannattaa käyttää kirjallisuudessa laajasti tunnistettuja datalaadun tekijöitä oman datasetin laatunsa arviointiin.

Avainsanat Kehitysprojekti, datan laatu, terveydenhuolto

Table of Contents

Table of Figures

1 Introduction

The utilization of data in guiding quality improvement projects has been a hot topic in the healthcare industry. A lot of articles have focused on the usage of e.g. Lean, TQM or Six Sigma: how to successfully conduct and organize the improvement efforts inside the organization. On the other hand, another established stream of literature has been created around the data quality: a stream focused more on defining data quality to help information system providers to improve the data quality in organizations. This paper aims to bridge the gap between the two streams of literature, i.e. how the data quality dimensions could be used to help improvement teams in healthcare.

The academic framework of the study starts with defining how the understanding of data quality has evolved from the 1970s to the modern days. In the 1970s and 1980s, the data quality literature was focused mostly on the intrinsic qualities of data, e.g. whether the dataset is accurate, timely and consistent. With the rise of the internet, 1990s brought up a more holistic view of the data quality: e.g. Wang & Strong (1996) presented data quality as "fitness for use" comprising combination of intrinsic, contextual, representational and accessibility data qualities. From the 2010s onwards, several studies assessed the data quality in specific domain, e.g. unstructured data or data-as-a-service. Based on the literature review, the findings were mapped on the data quality framework proposed by Wang & Strong (1996).

The literature review was supported by findings from the case study in a Finnish healthcare company. The case study method was used, as the effects of the data quality attributes to improvement projects would be hard to assess through a laboratory experiment. Furthermore, a descriptive approach for the case study was chosen, as the basic data quality research was well established, but not yet focused on the use of improvement projects inside organizations. In order to conduct the case study method, the data was collected by using the action research method, where the author was actually part of the improvement team himself.

The case study focused on assessing the impacts of data quality in an operational improvement project. The improvement project was conducted in a private Finnish healthcare company, and focused on reducing invoicing claims. The project comprised five distinct phases: starting from utilizing high level summary data to finally manually gathering data on the topic. Between the different phases, the project team faced decisions on whether to continue search for better data or to continue with the decision made with the current knowledge. Ultimately, the best results were gotten when the data was gathered manually on the topic, as the data provided actionable insights on the root causes, and showed clear direction on where to focus.

Based on the literature review and the case study findings, the data quality attributes were mapped against the case study findings. As a conclusion, the current data quality attributes are unlikely to provide adequate guidance for the improvement project teams. The findings suggest that the improvement project teams should first focus on assessing whether the data is relevant for the purpose: that the topic of the data relevant for the end goal of the project team. If the dataset is relevant, the project teams should focus on the value-added qualities of the data: if the dataset can drive action both inside the team, and among the important stakeholders. In case the dataset is relevant, and value-adding, it is important to assess whether the other data quality attributes are in an adequate level: i.e. the intrinsic, conceptual, accessibility and representational data qualities. The suggested model should, however, be supported with more research, as research done with one case study hardly supports any generalizable conclusions. The evidence is clear, on the other hand, that the data quality attributes should be assessed again in order to be useful for the improvement teams: and preferably linked to the improvement project literature.

This thesis starts by describing the analytical framework in the section 2. After the analytical framework, the methods used in this thesis are depicted in detail in the section 3. The section 4 will, then, focus on the findings from the case study followed by the section 5 about the results of the case study. Combining the results from the case study and the analytical framework, the rest of the sections will focus on the conclusions, the implications and the discussion about the obtained results. Before going to the analytical

framework, the following chapter will shed more light on the industrial setting in the Finnish healthcare industry.

1.1 Industrial setting

The case company operates in the Finnish private healthcare industry. The Finnish healthcare segment is divided into two main segments: the public and private organizations. According to THL (Terveyden ja hyvinvoinnin laitos), the total healthcare spending in Finland in 2014 was 19,5 billion euros, i.e. 9,5% of the GDP. Out of the total sum, 76% was publicly funded with rest being funded by the private sector. The biggest two groups of expenditures were specialized healthcare, and primary care: the combined share of the two was slightly over half of the total costs. The healthcare industry is highly segmented with several small organizations providing specialized services for the customers. The case company can be classified, however, as one of the major players in the Finnish private healthcare industry with revenues over 100 million euros per annum, and with employing over 1 000 people in Finland. Based on the description given by Orbis, the case company works in comprehensive healthcare, occupational healthcare, and hospital services.

2 Analytical framework

Data quality and its implications have been present in academic literature for decades (e.g. Neter & Yu, 1973, Cushing, 1974, Laudon, 1986, Johnson, Leicht & Neter, 1981, Knechel, 1985). The following analytical framework will first shed a light on how the concept of data quality has changed over time. In the second chapter, the data quality attributes are linked to the model proposed by Wang & Strong (1996), and the findings of the analytical framework are summarized.

2.1 The development of data quality literature

While understanding the data quality has been a prevalent topic in the academic literature, the scope of study has changed over time. In the 1970's and 1980's, the data quality literature focused more on the intrinsic qualities of data, e.g. accuracy and reliability (e.g. Neter & Yu, 1973, Knechel, 1985, Cushing, 1974). In the 1990's with the rise of Internet, data quality was started to be considered as "fitness for use", i.e. the data quality is relative to the context (e.g. Tayi & Ballou, 1998). The big data and platform-driven businesses have created new problems with the data quality: new areas of data quality research have, thus, arisen e.g. around unstructured and semistructured data (Madnick et al., 2009) and the cognitive effects of data quality (Watts, Shankaranarayanan & Even, 2009).

2.1.1 Before the internet age - 1970's and 1980's

The 1970's and 1980's were an era of rapid technological development: e.g. personal computers, mobile phones and CD's were developed during the time. With the Internet still making its way to the mass market, the amount of available data was limited and processing power expensive. Due to the expensive processing power, many companies lacked proper information systems capable of providing quality data for the end-users. A lot of data was still presented in a hard-copy paper format rather than in the information systems (Wang & Strong 1995). Thus, the focus of the era was more on the intrinsic

qualities of data, i.e. that the data itself was accurate and precise (e.g. Neter & Yu, 1973, Knechel, 1985, Cushing, 1974).

The information systems in the pre-internet era were often plagued with problems in the data accuracy: in other words, the data in the information system did not match the real-world system it tried to depict (e.g. Knechel, 1985, Miller & Doyle, 1985). The 1970's and 1980's brought up several studies which illustrated the magnitude of the data accuracy problems in the information systems (e.g. Laudon, 1986). Previously, the data accuracy was identified as one of the key drivers for data quality, but the scale of the phenomena was yet to be understood. The studies of the era focused on data problems in several different industries: e.g. accounting (Knechel, 1985), financial services (Miller & Doyle, 1985) and criminal records (Laudon, 1986).

In the 1970's and 1980's, being error-free and complete were deemed as one of the most important aspects of data quality (e.g. Laudon, 1986, Bailey & Pearson, 1983, Johnson, Leitch & Neter 1981). The data completeness depicts to which extent all relevant data points are presented in the information system, i.e. that no information is missing from the system (Ballou & Pazer 1985). As data completeness describes whether the data takes into account all data points in the real world, it can be considered as the first step towards contextual quality: that all real-world stages are correctly and exhaustively mapped in the information system (Wand & Wang 1996). The completeness attribute was, however, interpreted more as an intrinsic attribute in the 1980s, rather than a quality related to the representation of the data to the user.

To be considered accurate, the data should consist of values, whose presentation in the information system does not differ from the real values (Laudon 1986). According to the data users, the accuracy of the data was the most important attribute of the data quality (Bailey & Pearson 1983). Several studies estimated the number of errors through mathematical models: e.g. Morey (1982) built a mathematical model for the lower bound error-rate estimates. The problems with high error-rates were also identified in different studies focusing on individual industries (e.g. Laudon, 1986, Johnson, Leitch & Neter, 1981). In the study of US criminal record databases, Laudon (1986) found out

that approximately 74% of ident records were not “complete, accurate or unambiguous” - summing to a total of 1,75 million disseminations with quality problems.

In addition to being error-free, the timeliness of data was considered as an important attribute for the data quality (e.g. Ballou & Pazer 1985, Laudon 1986). The timeliness of data describes how topical the data is, i.e. that the data values are not out-of-date. Ballou and Pazer (1985) describe the linkage between timeliness and error-free data; if a data value is outdated, it differs from the current real value, and thus can be considered as erroneous. As the effect of not being timely is linked to erroneous data values, majority of the studies reviewed for this thesis did not make a difference between timeliness characteristics and the error-free qualities of data.

Several studies emphasized the importance of reliability and consistency in the information systems: i.e. that the information system provides the correct output each time the system is used (e.g. Bodnar 1975, Knechel 1985). Ballou & Pazer (1985) extend the description of consistency: “the representation of the data value is the same in all cases”. Accessing the data from the information system was not always easy due to the problems in the information system reliability in the 1970’s and 1980’s. Thus, the reliability attribute of data quality was deemed to be the second most important factor affecting the data quality by Bailey & Pearson (1983), and one of the top four dimensions of data quality by Ballou & Pazer (1985). Models were created to simulate the information system reliability as a whole: e.g. Bodnar (1975) used a mathematical model to improve the information system reliability. Several studies also aimed to improve the information system reliability in a certain industry (e.g. Agmon & Ahituv, 1987, Knechel, 1985). In the context of accounting systems, Knechel (1985) created a simulation model which could be used to assess the reliability of different accounting systems. The problems with the information system reliability can be linked to the data accessibility issues - a prevalent topic in the data quality literature.

Already in the 1960s, the researchers argued that both perceived technical data quality, and data accessibility were key determinants of the overall data usage (e.g. Allen, 1966, Gerstberger & Allen, 1968). Studies conducted that the perceived ease of access was the

most influential driver of the data usage (e.g. Rosenberg, 1966, Gerstberger & Allen, 1968). The accessibility of the data cannot be, though, assessed independently, as the experience of the data user affects the perceived accessibility - thus, the level of data accessibility is unique for each user (Gerstberger & Allen, 1968). The data accessibility was, however, considered as independent from the actual data quality: a distinction likely to be based on the common usage of hard-copy reports rather than digital data (Wang & Strong, 1996).

Studies started to include more contextual attributes to the data quality research in the 1980's: in other words, the data quality was considered to be relative to the context of data usage (e.g. Bailey & Pearson, 1983, Miller & Doyle, 1987). For example, attribute data can be interpreted in very different ways in different contexts: e.g. warm weather can mean totally different things for inuit and Israeli person (Agmon & Ahituv 1987). Several contextual attributes were identified in the literature: e.g. responsiveness, relevance and timeliness. Due to the relative nature of the data quality, information systems were required to be responsive to changing user needs - including the possible future needs (Miller & Doyle, 1987). The relevancy of the data was also considered as one of the top five criteria used by users to assess the data quality - an attribute dependable on the context of the data usage (Bailey & Pearson, 1983). In later decades, the timeliness was also linked to contextual characteristics, as users need to make the tradeoff between timely, inaccurate data and historical but accurate data (Ballou & Pazer 1995). While most of the individual characteristics of contextual quality were identified already in the 1970s and 1980s, they were not linked comprehensively together before the 1990s.

In addition to the contextual qualities, the 1980s brought up studies about the representational attributes of data quality, i.e. that the interpretation of the data is relative to the graphical presentation (e.g. Benbasat & Dexter 1985, Doll & Torkzadeh 1988). Several studies focused on describing how the data in information systems should be presented in order to support decision-making: e.g. Järvenpää (1989) took a cognitive approach to understand how graphical format affects processing information. The studies in the 1980s did not, however, link the representational characteristics explicitly

to the data quality attributes, but focused on understanding how the graphical representation should be built (e.g. Santos & Bariff 1988, Doll & Torkzadeh 1988), how the representation affects decision-making (e.g. Järvenpää 1989, Benbasat & Dexter 1985), and building taxonomy to describe the representational attributes (e.g. Tan & Benbasat 1990). These studies built the groundwork for the more comprehensive data quality models created in the 1990s.

The work to improve the quality of data in the information systems started with understanding how mathematical models could help improving the data quality (e.g. Cushing, 1974, Bodnar, 1975). The mathematical models were created to provide an objective, quantitative understanding of the data accuracy: the objective understanding of the problem then served as a starting point for further improvement (Neter & Yu, 1973). The mathematical models were also used to improve the quality of data used in the internal control systems (Cushing, 1974). Building on the work of Cushing (1974), Bodnar (1975) suggested that reliability modeling could improve the efficiency of internal control systems by helping to design better control procedures.

In the 1970's and 1980's, understanding the systemic qualities of data were deemed important (e.g. Ballou & Pazer 1985, Bodnar 1975, Knechel 1985). The studies focused on the information system itself: how the information system was affecting the data quality. Ballou and Pazer (1985) described a model for understanding the multidimensional data quality in different system nodes: what was the data quality of the input and output of the stage in hand. Furthermore, the data quality attributes were not seen as independent on each other, e.g. perceived data quality was suggested to affect the data accessibility (Gerstberger & Allen, 1968). Thus, looking at the data quality one attribute at a time can lead to wrong conclusions: e.g. if the information system does not provide proper access to the data, there is no value of perfectly precise and contextually appropriate data as it is not used.

Overall, the 1970's and 1980's brought up a rapid increase in the number of articles about the data quality. During the period, a lot of the data quality attributes were given formal definitions, which were used also in the later decades: e.g. data completeness,

timeliness or accuracy. The focus of the era was more on the intrinsic qualities of data – whether the data itself was correct. A lot of other attributes, which were added to the list of data quality attributes in the 1990's, were identified already during the era: e.g. the first attributes of representational, contextual and accessibility data quality. The next section will dig deeper on how the 1990s and 2000s changed the understanding of the data quality attributes to a more holistic view with interactions between the attributes itself.

2.1.2 Rise of the internet - 1990s and early 2000s

The 1990s brought an Internet access for the majority of companies and individuals in the Western world. With the internet access, people were able to get their hands on ever growing body of knowledge and data. In addition to the Internet, the rise of the computing power followed the Moore's law, thus seeing an exponential decrease in the cost of computing. At the same time as computing power became more affordable, the information systems itself were progressing in a fast pace. Companies started to turn their hard-copy reports to digital data on the information systems. Overall, the expansion of the amount of available data was rapid, and the decreasing cost of computing power enabled better analysis of the growing data masses.

While the definition of data quality was starting to expand, the articles still deemed intrinsic attributes as the most important factors of data quality (e.g. Wang, Reddy & Kon 1995, Kahn et al. 2002). Whereas the intrinsic qualities were mostly described as error-free and accurate data in the 1980s, more and more studies brought up different point-of-views for the intrinsic data qualities. For example, several studies described the intrinsic quality through a combination of accuracy, timeliness and completeness (e.g. Miller 1996, Wang, Kon & Madnick 1993). While most of the attributes were identified and described already in the 1970s and 1980s, the articles from 1990s onwards used the different intrinsic attributes as a combination - in other words, the overall understanding of the intrinsic quality as a sum of many different attributes started to be common in the articles of 1990s.

The development of the intrinsic qualities was not limited to summarizing the already identified attributes: rather, several articles described new attributes for intrinsic data quality (e.g. Miller 1996, Pipino, Lee & Wang 2002, Strong, Lee & Wang 1997). The believability of the data, i.e. how truthful and credible the user regards the data to be, became a new data quality attribute (Wang & Strong 1996, Strong, Lee & Wang 1997, Wang, Reddy & Kon 1995). According to Wang, Reddy and Kon (1995), the data believability in fact is a higher-level definition of timeliness, data source credibility and accuracy. On the other, the data objectivity, i.e. the unbiased and unprejudiced data, was defined as a factor of intrinsic data quality (Pipino, Lee & Wang 2002, Klein 2001). The objectivity differs from accuracy and being error-free, as the data can be precise, but at the same time biased towards indicating only one side of the truth. Linking to the other intrinsic factors, also the reputation of the data source was considered as an intrinsic attribute (Wang & Strong 1996). If the reputation of the data or data source is low, the data usage is reduced, and thus the value-added to the data consumer is lower than it otherwise would be (Wang, Strong & Lee 1997).

Rather than only as a combination of intrinsic attributes, the data quality was defined more as a "fitness for use" for the data consumer from the 1990s onwards (Tayi & Ballou 1998). Increasing number of studies included contextual attributes as data quality dimensions - in other words, the studies recognized the role of the context the data is used at (e.g. Wang & Strong 1996, Miller 1996). Bovee, Srivastava & Mak (2003) wrote that the data must be relevant to our purpose and context of the data usage. Each data user has different assumptions of the meaning and quality of the data, i.e. different context in which they interpret the data. Thus, error-free information can be misinterpreted when it is transferred from one context to another (Madnick 1995). Madnick (1995) provides an example of global currencies: given that a French person sends data about the prices in Euros, it is possible that a US recipient assumes the data to be in US dollars, thus misinterpreting the data.

While a lot of the contextual attributes were identified in the 1980's, the understanding of the individual attributes expanded on the 1990's. For example, most of the studies in the 1980s understood data completeness as an intrinsic attribute, i.e. having all relevant

data points present in the data (e.g. Laudon 1986, Ballou & Pazer 1985). From the 1990s onwards, the completeness was understood as also having the right variables present in the data for context of the data user (Wang & Strong 1996, Nelson, Todd & Wixom 2005). Nelson, Todd and Wixom (2005) described the data completeness as representing all the relevant states for the context of the user. However, as the amount of accessible data was growing very rapidly, also the appropriate amount of data was seen as an important attribute of data quality (Wang & Strong 1996, Pipino, Lee & Wang 2002). Especially in areas where the amount of raw data was huge, having an analyzable amount of data with not too many variables, was an important attribute for data quality (Pipino, Lee & Wang 2002).

The studies of 1990's also recognized the value-add for the data consumer as an important attribute of data quality (e.g. Wang & Strong 1996, Wang, Strong & Lee 1997, Wang, Strong & Kahn 2002). In order that the data is considered value-adding, it must have relevance for the task the user is trying to perform. While the added value is considered important for the data quality, it is also intangible and inherently difficult to measure (Wang, Strong & Kahn 2002). Thus, the added value is often embedded into the definition of other attributes: e.g. relevance defined as relevance for the task the user is performing (Bovee, Srivastava & Mak 2003). The added value can be, however, interpreted as the glue between different contextual attributes: that the data quality attributes must aim to help the data consumer to perform the task in hand as well as possible.

In the 1990's, also the representational attributes of data quality started to be more common: i.e. that the data must be interpretable and easy to understand (e.g. Wang & Strong 1996, Lee, Strong & Kahn 2001). Whereas in the 1980's many of the representational attributes were already described, the linkage to the models of data quality came only in the 1990's: e.g. the models of Wang and Strong (1996) and Bowee, Srivastava and Mak (2003). Linking the representational attributes to the holistic data quality models enabled studies to find interactions between data quality attributes: e.g. Wang, Lee and Strong (1997) suggested that representational and accessibility attributes of data quality are linked. When the data representation is not unambiguous and easy to

understand, the bad representation can become also a barrier to the data accessibility (Wang, Lee & Strong 1997).

The one of the most important representational attributes in the 1990's studies is the interpretability of the data, i.e. the ability to make the same, correct interpretations of the meaning of the data regardless of the context of the user (e.g. Wang & Strong 1996, Wang, Kon & Madnick 1993). The interpretability can be defined as a subjective quality indicator: e.g. is the data presented in a correct language, symbols, and units (Wang, Reddy & Kon 1995). In order to have a good interpretability, the names, codes et cetera in the data must have a clear meaning for the user: e.g. in medical setting, understanding the diagnosis can be hard for outsiders, thus leading to bad data interpretability (Wang, Lee & Strong 1997, Nelson, Todd & Wixom 2005). Close to the interpretability attribute, Wang & Strong (1996) also identified the ease of understanding as one of the key criteria for good representational data quality. As an ingredient of the ease of understanding, the data in the information systems must have a mapping to the real-world systems (Wand & Wang 1996). Wand & Wang (1996) suggested that firstly, the mapping to the real-world system should be unambiguous, i.e. no two states in the real-world should be together in the information system; secondly, all states should be meaningful, i.e. all states should be traceable back to the real-world system; lastly, the representation should be complete, i.e. all real-world states should be mapped in the information system.

While interpretability and ease of understanding are linked to correctly understanding the meaning of data, several studies also identified the need for the correct format of data presentation (e.g. Wang & Strong 1996, Miller 1996, Nelson, Todd & Wixom 2005). The formatting attribute of data quality can be divided into two separate sub-attributes: the graphical format and the context provided for understanding the information (Miller 1996). In order to be of a good quality, the data must be represented in a concise manner, i.e. to avoid overwhelming the data user with too much information at once (Wang, Strong & Kahn 2002). In addition to being concise, the representation of the data must be consistent: in other words, the format of the data must be the same independent on the place or time of the data usage (Wang & Strong 1996, Pipino, Lee &

Wang 2002). Furthermore, the graphical representation must be clear, i.e. the resolution of the graphical representation needs to be high enough for the data user (Wang, Kon & Madnick 1993).

In the 1990s, data accessibility was started to be considered as an attribute of the data quality itself, rather than a separate feature (e.g. Wang & Strong 1996, Lee et al. 2003, Nelson, Todd & Wixom 2005). The data accessibility was the starting point to consider the data quality: if the data is inaccessible, all the other data quality attributes are meaningless (Bovee, Srivastava & Mak 2003). On the other hand, problems with other data quality attributes can lead to poor accessibility, and thus, unusable data. For example, using definitions or terms which are incomprehensible for the user can become a barrier for accessibility (Strong, Lee & Wang 1997). Additionally, Miller (1996) suggested that timeliness and accessibility should be tightly coupled: if accessing data takes a lot of time, the data might turn out to be unusable for the data consumer.

In the 1990s, several new viewpoints to the data accessibility emerged. With the expansion of the available data, integration of different datasets possibly even in different information systems, became important for data accessibility: all information can be accessible separately, but combining them to get the insights can turn out to be impossible (Nelson, Todd & Wixom 2005). The increase in data also emphasized having appropriate amount of data to analyze (Strong, Lee & Wang 1997): large datasets could take a lot of time to analyze and would potentially incur significant costs during the process (Bovee, Srivastava & Mak 2003). Thus, response time can be considered as an important attribute for the data accessibility (Miller 1996). Strong, Lee & Wang (1997) combined these findings about the data accessibility to three distinct categories: technical accessibility, data-representation issues leading to poor accessibility and the accessibility problems caused by large volumes of data.

In addition to the traditional accessibility concerns, the studies of 1990s also included access security as one of the data quality attributes (e.g. Wang & Strong 1996, Kahn, Strong & Wang 2002, Miller 1996). The access security can be divided into two different categories: securing the data from the humans, i.e. attacks or security breaches

from the outsiders, or securing the data from the natural disasters (Miller 1996). The concerns about the access security can undermine the trust of the data consumers, thus leading to reduced use of the data: a problem which can considered as a hurdle to the data accessibility (Strong, Lee & Wang 1997, Miller 1996). In order the data to be seen as secure and truthful, the data consumers should be able to verify the information accuracy, timely and secure: a data quality attribute also known as the data validity (Miller 1996).

Overall the 1990s and early 2000s were an era of huge advancements in the data quality literature. The data quality attributes were linked to comprehensive models which aimed to create a more holistic and systematic view of the data quality attributes. The focus also shifted more towards understand the data quality as a "fitness for use": a combination of several different types of data quality attributes rather than purely intrinsic ones. The next section will continue to describe the advances of the data quality literature in the era of the big data, now described as the time from 2007 onwards.

2.1.3 Platform-driven businesses and big data - 2007 onwards

The amount of data has been continuing its exponential growth in the 2000s. As the available data masses grew to huge sizes, the scientists and the industry started to call the phenomena "the big data". The big data has posed new problems for the data quality: e.g. accessing huge data masses takes significantly more time, and simplified graphical illustrations become increasingly important as the sheer amount of possible features to analyze easily becomes too big to comprehend with just one look. The era also brought up several new types of data, which have previously been out of focus of the research: e.g. unstructured data (e.g. Batini et al. 2009).

While the environment of the data usage changed significantly from the 1990s until late-2000s, the underlying data quality attributes remained largely the same (e.g. Tee et al. 2007, Batini et al. 2009, Peralta 2008, Wang et al. 2008). While the data quality literature often references a broad scope of data quality attributes, they tend to base their studies on intrinsic qualities of data identified in the earlier decades (e.g. Alizamini et al. 2010, Peralta 2009, Batini et al. 2009, Batini & Scannapieco 2016). Peralta (2009)

focused on the two major data quality dimensions: data freshness, i.e. currency and timeliness of the data, and data accuracy, i.e. correctness, validity, and precision. Similarly, Batini et al. (2009) used accuracy, completeness, consistency, timeliness as a basis for their comparison study on data quality assessment and improvement studies. Alizamini et al. (2010) created, on the other hand, a new model on how to quantify data accuracy through fuzzy association rules. All in all, several articles built on top of existing research on data quality dimensions, and used them to create better implications for data quality management (e.g. Batini et al. 2009, Wang et al. 2008), or to better quantify the used data quality dimensions (e.g. Peralta 2009, Alizamini et al. 2010, Batini & Scannapieco 2016).

While a lot of articles were focused on the intrinsic qualities, some expanded also the contextual attributes of data quality. Watts et al. (2009) suggested that data quality had not been studied before as "fitness for use in context", i.e. how different the user's cognitive processes affect the interpretation of data. Understanding the cognitive processes can enhance the user's ability to interpret the data correctly. As a practical example, data consumers who assumed the data analysis task to be ambiguous took a more structured approach, and achieved better results compared to the ones who perceived the task to be less ambiguous. Similarly, having more relevant expertise on the topic was likely to increase the chances of taking a structured approach to data analysis, thus leading to more correct interpretation of data. (Watts et al. 2009)

While interactions between data quality dimensions had been identified before (e.g. Strong et al. 1997, Ballou & Pazer 1995), more work was done after 2010 to structure the dependencies to a comprehensive framework (e.g. Panayi et al. 2013, Barone et al. 2010). While earlier studies identified specific two-way interactions, e.g. timeliness/accuracy tradeoff, the focus of studies after 2010 was to build generalized models on how to assess multidimensional interactions between data quality dimensions. The models were built on top of mathematical analysis of the different dimensions: e.g. Barone et al. (2010) used Bayesian networks. The goal of the studies was to give tools for database analysts to help them with creating alternative data quality improvement strategies (e.g. Barone et al. 2010).

The rapid increase of data also brought up a need to increasingly distinguish between three types of data: structured, semi-structured and unstructured data (e.g. Batini et al. 2009, Batini & Scannapieco 2016). In the structured data, the data values are defined in different domains, i.e. the range of possible values; relational data tables are a common example of structured data (e.g. Li et al. 2008, Batini et al. 2008). Semi-structured data has a flexible and often self-describing structure, as is the case e.g. in the XML documents (e.g. Li et al. 2008, Batini et al. 2008). Most of the literature in the early 2000s and before focused on the semistructured and structured data (e.g. Batini et al. 2009, Batini & Scannapieco 2016). The unstructured data is, on the other hand, a set of symbols representing often natural language; common examples of unstructured include emails and text documents (e.g. Li et al. 2008, Batini et al. 2008). As unstructured data provides a wealth of opportunities for future data analysis, several calls for future research have been raised for understanding data quality in unstructured data (e.g. Madnick et al. 2009, Batini et al. 2008).

2007 and beyond brought up several articles which focused on identifying data quality problems in specific domains which were not perfectly explained by the general data quality models. One of these specific topics was crowdsourced data: how data quality should be understood in situations where data is based on large number of individual responses (e.g. Buhrmester et al. 2011, Hsueh et al. 2009). In the case of crowdsourcing, individual respondents may have inaccurate and highly variable data which is in big masses very close to the actual values (Buhrmester et al. 2011, Hsueh et al. 2009). Assessing the data quality as individual data items might, thus, lead to wrong conclusions about the underlying data quality in the crowdsourced services. Data quality in crowdsourcing services must, in other words, be assessed as a holistic system rather than assessing whether individual data points are correct or complete.

The traditional data quality dimensions are also incapable of fully capturing the dimensions needed to describe the cloud-based data storages, or “data-as-a-service” (e.g. Truong & Dustdar 2009). While the individual sources of data might be of a perfect quality, combining several different data sources in a cloud-based environment poses a new threat of mixing different abstraction levels (Curry et al. 2013). As a practical

example, the user might be searching for a comprehensive data set about a set of customer which links both external and internal sources. Some of those datasets might, however, contain the estimated purchases of an individual customer whereas similar looking datasets might hold the estimated purchases of the household. Both datasets being of a good quality based on the traditional data quality terms, the linked dataset might contain values which are non-comparable - even if presented as being the same.

Overall, majority of the data quality literature after 2007 has transformed from perfecting the data quality dimensions to understanding the meaning of data quality in specific contexts. The many of the hallmarks of the 21st century - the big data, and rise of the unstructured data - are considered in the data quality literature. Though, as the importance of unstructured data has exploded in a short period of time, a lot of research is still lacking in this realm of research.

2.2 The model on data quality

Based on the literature review described in the previous section, the found data quality attributes are now described in more detail. In several studies, the data quality attributes might have come with different names, but similar meanings. To simplify the model, the attributes with close to similar meanings have been grouped, e.g. "correctness" and "accuracy". To structure the findings from the literature study, the data quality attributes were then grouped by a fourfold model proposed by Wang & Strong (1996)[1].

2.2.1 Building on the model of Wang & Strong

The fourfold model of Wang & Strong (1996) groups the data quality attributes into four categories: intrinsic, representational, contextual and accessibility data qualities. Numerous studies have referenced the article by Wang & Strong (1996), and several

[1] Wang & Strong (1996) model was chosen as it is one of the most cited frameworks in the data quality literature. The Wang & Strong (1996) model combines the thoughts of "fitness for use" to data quality attributes studied in the previous studies (e.g. Ballou & Pazer 1985, Gerstberger & Allen 1968). The model was chosen, as it is more comprehensive compared to the previous data quality models, such as Doll & Torkzadeh (1988). On the other hand, the newer models tended base their findings on very similar dimensions as Wang & Strong (1996), while bringing a novel viewpoint to the table (e.g. Nelson et al. 2005).

have used their fourfold model to structure their own studies: e.g. Sonntag (2004) used the model of Wang & Strong (1996) in their study about the quality of natural language text data. This section is divided into 4 subcategories, each of which describes the data quality attributes belonging to one of the categories.

2.2.1.1 Intrinsic

According to Wang & Strong (1996), intrinsic data qualities mean that the data has quality in itself: in other words, the intrinsic qualities focus on the data quality assessed separately from the context of use. Strong et al. (1997) describe that the intrinsic qualities are the most common studied attributes in the data quality literature: yet, they are inadequate itself as they do not take into account the consumer who uses the data in the end. On the other hand, the lack of focus on the context of use enables more objective assessment of the intrinsic data qualities, which make them a basic starting point of quantitative models on data quality (Stvilia et al. 2007).

Accuracy of the data refers to the extent that the data is error-free, correct and flawless (e.g. Wang & Strong 1996, Klein 2001). Whereas Ballou & Pazer (1985) take a more straightforward view on data accuracy, i.e. whether a difference exist between the recorded and actual value, Peralta (2008) describes data accuracy through 3 different categories. According to Peralta (2008), data accuracy can be understood through semantic correctness, i.e. the accuracy and validity of the data, syntactic correctness, e.g. the amount of misspellings in the data set, or precision factor, e.g. has the data been stored with the precision of 3 decimals or 15 decimals.

Believability of the data refers to how credible or true the data consumers deem the data (e.g. Wang & Strong 1996, Kahn et al. 2002). As the believability attribute links more to how data consumers perceive the data than the actual correctness of the data, Kahn et al. (2002) label believability as one of the service quality attributes of data. In other words, the believability of the data can be seen as a higher level construct of other intrinsic quality attributes, namely timeliness, credibility and accuracy (Wang et al. 1995).

Objectivity of the data means that the data should not be biased or contain prejudices (e.g. Wang & Strong 1996, Kahn et al. 2002). Especially in cases when the data recording is based on human judgement, the data objectivity can be compromised (Strong et al. 1997). Low objectivity of the data might also be caused by impartial data: the data might be accurate but portraying a biased picture as all values are not recorded (Kahn et al. 2002). According to Kahn et al. (2002), data objectivity contains aspects from both sound information attributes, i.e. conform to specifications such as accuracy and completeness, and useful information attributes, i.e. increase the usefulness and relevancy of the data for the consumer.

Reliability of the data refers to how the data accuracy is sustained over time (e.g. Nelson et al. 2005, Ives et al. 1983). Measuring data reliability can be divided into two different parts: "test-retest" reliability or the gross amount of error in the measurement (Ives et al. 1983). Wang & Strong (1996) adopted the same viewpoint for their study, as they linked reliability as a sub-dimension of accuracy. On the other hand, reliability can be also seen through the lens of an information system: is the data available for the end-user at the times it is needed (Nelson et al. 2005).

Reputation of the data relates to whether the data is trusted and kept in high regard by the data users (e.g. Wang & Strong 1996, Stvilia et al. 2007). Reputational data quality defines the place of data in the cultural and activity hierarchy of the organization: in practice, whether the users see the data as a credible source of information (Stvilia et al. 2007). On the other hand, if the data consumers perceive the data as untrustworthy, the low data reputation might become a barrier for data accessibility: the data might be accurate and value-adding, but it is never used based on the user perceptions (Strong et al. 1997).

2.2.1.2 Contextual

Several studies have observed that the quality of data is related to the context of use (e.g. Wang & Strong 1996, Tayi & Ballou 1998). Rather than a stable measurement, the data quality can, thus, change from a user-to-user basis. The differences stem from the finding that data quality is linked to whether or not the data adds value and is relevant to

the end user (e.g. Bovee, Srirastava & Mak 2003, Wang & Strong 1996). Along these lines, Tayi & Ballou (1998) defined data quality as the "fitness for use" - a definition where the contextual attributes are in the core of all data quality attributes. Watts et al. (2009) took the definition a step further by defining data quality as a "fitness for use in context", i.e. how different the user's cognitive processes affect the interpretation of data.

Appropriate amount of data for the context is an important data quality feature (e.g. Wang & Strong 1996, Strong et al. 1997). Especially in the modern era of big data, users can be easily overwhelmed in case there are dozens of different data attributes of which the user needs to find the appropriate ones. On the other hand, the problem with having too much data can also lead to timeliness problems: getting a huge data set out of the system might take a long time, which can lead to timeliness problems where the data is inaccessible at the moment it is needed (Strong et al. 1997). Similarly, too large data masses might take a long time or incur other costs for the user; in these cases, the data can become inaccessible making other data quality attributes irrelevant (Bovee, Srirastava & Mak 2003).

Completeness of the data refers to whether all relevant values for different variables are stored in the data (e.g. Ballou & Pazer 1985, Wang & Strong 1996). Completeness can be understood from an intrinsic point-of-view: whether there are missing values from the stored variables (e.g. Ballou & Pazer 1985). On the other hand, also a contextual viewpoint of often used: are all the states mapped which are relevant for the user (e.g. Nelson et al. 2005, Wang & Strong 1996). As many of the more modern articles are inclining towards the contextual point-of-view, the same viewpoint is also chosen for the purposes of this study (e.g. Tayi & Ballou 1998, Nelson et al. 2005, Wang & Strong 1996).

Relevancy of the data is also a contextual attribute as it links to how relevant the data is for the data consumer (e.g. Bovee, Srirastava & Mak 2003, Strong et al. 1997). The relevancy of the data is often described as how applicable the data is for the task the user is trying to perform (Wang & Strong 1996). On the other hand, the relevancy of data can

also be described as a higher level term compared to the other attributes: if the the data is irrelevant for the consumer, the other attributes do not have any impact on the perceived data quality (Miller 1996). With the Miller's (1996) definition, the data relevancy can be understood very similarly to the value-added attribute defined below.

Timeliness of the data is one of the basic data quality attributes identified already decades ago. Before the 1990's, timeliness was understood more as an intrinsic data quality: whether the data values are outdated, i.e. are different compared to the current value (Ballou & Pazer 1985). Timeliness itself can be divided into two sub attributes: currency, i.e. how new is the data, and volatility, i.e. whether the data has changed over time (Wang et al. 1995). On the other hand, Wang & Strong (1996) defined the timeliness not only through the intrinsic lens, but also as whether "the age of the data is appropriate for the task in hand". As a practical example, an analyst trying to understand how the oil prices correlated with the factory output has a different requirement for the timeliness of data compared to the doctor who assess whether there has been a change in patient's blood pressure during the past 12 hours.

Value-added as a data quality attribute refers to how advantageous the data is for the consumer (e.g. Wang & Strong 1996, Kahn et al. 2002). Providing added value is one of the fundamental data quality attributes, which is rather subjective: i.e. is measured by whether the user can perform the task in hand better with the data or not. It is, though, considered as one of the most important data quality features (Wang & Strong 1996). On the other hand, if the data consumers perceive the data to be of a little added value, the use of the data might be lower than expected: thus, the perception of the users about the value-added attribute is important in the accessibility of the data, i.e. whether the data is used by consumers or not (Strong et al. 1997).

2.2.1.3 Accessibility

Data accessibility has been well-established in the data quality literature for decades, as having too low data accessibility makes other data quality attributes irrelevant (Wang & Strong 1996). The data accessibility can be defined through two data quality attributes: access security and ease of access (Wang & Strong 1996). While the two topics have

been identified and researched a lot in the past (e.g. Gerstberger & Allen 1968), the new requirements of the big data era have posed new problems for the data accessibility: e.g. Nelson et al. (2005) identified correct integration of different datasets as one of the important data quality attributes. On the other hand, problems with other data quality attributes, such as too large data masses or poor representation, might cause problems with the data accessibility (Strong et al. 1997). Additionally, if the data consumer perceives some of the data quality attributes being poor, the data might be left unused even if it would be, in real life, of a good quality (Strong et al. 1997).

Access security is one of the key attributes of data accessibility, as security barriers might lead to long waiting times before being able to access the data (e.g. Wang & Strong 1996, Kahn et al. 2002). The problems with the access security are especially clear in the healthcare industry, where the patient records must be kept secret: thus, data consumers might not search for certain datasets as they are perceived to be hard to get (Strong et al. 1997). In addition to the access barriers, the access security can be understood positively as an attribute ensuring the quality of the data. Miller (1996) defines security through two lenses: logical security, i.e. protecting the data from people, and disaster recovery planning, i.e. securing the data in case of a natural disaster.

Ease of access is an attribute that takes a technical point-of-view to the data accessibility: whether the data is fast and easy to access (e.g. Wang & Strong 1996, Strong et al. 1997). The barriers for an easy access might arouse e.g. from having too large datasets which take a long time to open and edit (Strong et al. 1997): a problem Nelson et al. (2005) define as problems with the response times. On the other hand, data might be difficult to access due to the problems with the infrastructure, e.g. a slow or unreliable network connection (Strong et al. 1997). On the other hand, in the big data era a lot of datasets might not be integrated, and thus require users to access several different databases: accessing different databases might, in turn, become a barrier for accessibility as it raises concerns of combining data sets with different freshness (Peralta 2008).

2.2.1.4 Representational

Representational data quality attributes highlight the importance of the systems, as poor representation of correct data might lead to false conclusions or barriers for accessibility (e.g. Wang & Strong 1996). The representational attributes describe either the ease of finding the correct meaning in the data - interpretability, and ease of understanding - or having a correct formatting of the data - concise representation, and consistency (Wang & Strong 1996). The representational qualities of data have become increasingly important due to the increase in the amount of data: e.g. the ability for the concise representation (e.g. Sonntag 2004, Wang et al. 2003), and smooth integration between different datasets are crucial for guiding correct understanding of the data (Strong et al. 1997, Miller 1996).

Interpretability of the data can be defined through whether the data consumer is able to understand the data, and interpret the meaning of the data correctly (Bovee et al. 2003). The interpretability of the data can be decomposed to the indicators of quality, e.g. units and scale used in the data (Wang et al. 1995). The interpretations of these indicators can differ based on the context of the data usage: e.g. a cold weather has a different meaning for an inuit compared to an Israeli farmer (Agmon & Ahituv 1987). As the context of the data usage differs from person to person, conforming to the standard formatting and representation of the data is crucial for the guide the correct interpretation (Stvilia et al. 2007). Additionally, using subjective instead of numeric labels can likely cause problems with the data interpretability (Agmon & Ahituv 1987, Stvilia et al. 2007).

Ease of understanding refers to whether the data can be interpreted easily and unambiguously (Wang & Strong 1996). The unambiguous understanding often links tightly to the formatting of the data: whether the data consumer is familiar with the formatting and can understand it with ease (Nelson et al. 2005). On the other hand, the Wand & Wang (1996) suggest three steps for correct representation of the data to guide easy and unambiguous interpretation; Firstly, the data should be unambiguously represented, i.e. each state in the information system should refer to only one state in the real world. Secondly, all states in the information system should be meaningful, i.e. all

the states in the information system should be linkable back to the real world. Finally, the representation should be complete, i.e. all real world states should be linked to a state in the information system.

Concise representation means that the data should be compactly and briefly represented (e.g. Wang & Strong 1996, Pipino et al. 2002). While the compactness of representation is a virtue, the data should still be complete and derive the correct meaning (Wang & Strong 1996). Furthermore, the conciseness of the representation should be assessed through the lens of whether the data is useful for the data consumer's daily job (Kahn et al. 2002). One of the subattributes affecting the conciseness of the data is the resolution of the graphics (Wang et al. 2003): the data should be aesthetically pleasing and easy to interpret (Wang & Strong 1996). The concise representation has become increasingly important as the amount of the data has been on the rise: the data conciseness has to be taken into the account e.g. when assessing the natural language processing, NLP, data (Sonntag 2004).

Consistency of the data links to whether all representations of the data value are the same or not (e.g. Ballou & Pazer 1985, Tayi & Ballou 1998). While Ballou & Pazer (1985) focus on the intrinsic consistency of the data, i.e. whether the data value is in fact the same in different cases, Wang & Strong (1996) define the data consistency as a representational attribute. As a representational attribute, the consistency can be defined as whether the data is presented in the same format in different cases, and can be combined with the data from previous sources (Wang & Strong 1996, Tayi & Ballou 1998). Furthermore, having inconsistently represented data might lead to difficulties in utilizing the data, as integrating the different data sets becomes vastly more difficult (Strong et al. 1997). Miller (1996) suggests that the quality of data consists not only of of the quality of the data itself, but also how the datasets can be combined and easily delivered to the consumer.

2.3 Summary

The literature review focused on how the understanding of data quality changed over time - from the 1970's to the modern day. The findings of the literature review are summarized below in Table 1. The data quality attributes proposed in the seminal article by Wang & Strong (1996), and how the data quality literature links to those attributes, is described in the Appendix 1.

	Before the internet age - 1970's and 1980's	Rise of the internet - 1990's and early 2000's	Platform-driven businesses and big data - 2007 onwards
Description of the era	Computing power expensive Lack of proper information systems Reliance on hard-copy reports	Rapid expansion in computing power Access to Internet increases the amount of data	Soaring amount of data and cheap analytical power New types of data become more common, e.g. unstructured data
Focus areas of data quality literature	Intrinsic qualities of data Systemic understanding of data quality	Data quality as a "fitness for use" More holistic view for data quality, interactions between different quality attributes	Data quality in specific domain, e.g. unstructured data or data-as-a-service More focus on interactions between different data quality attributes
Key attributes of data quality	Accuracy, completeness, reliability Accessibility as a separate factor	Intrinsic attributes as critical to quality Accessibility as an enabler of use Contextual attributes about the context of use Representational attributes focus on correct interpretation of data	Majority of basic research already established - a lot of studies were based on basic intrinsic attributes Some new attributes rose, e.g. data quality as a cognitive attribute

Table 1: Summary of the findings from the academic framework

3 Methods

The findings from the literature review were supported by qualitative evidence collected via the case study method. This section will describe the academic foundations of the case study method, and then shed light on the chosen research questions as well as the data collection process itself.

3.1 Case study method

Case study method can be described as a research strategy which aims to build new or extend existing theories from the qualitative evidence from a real world case (Eisenhardt 1989, Eisenhardt & Graebner 2007). The case studies should follow a logic of replication: each conducted case study serves as one unit of evidence which can elaborate, replicate, or contrast an existing theory (Yin 2013, Eisenhardt & Graebner 2007). On the other hand, case studies should satisfy the duality criterion: the case study research should be situationally grounded, i.e. be disciplined with the empirical research from the beginning, and seek a sense of generality, i.e. aim to create broader theoretical understanding based on the findings from the case (Ketokivi & Choi 2014). This section will shed more light on the case study method, why it was chosen to be the basis for the study, and how the study itself was conducted.

Researchers often face a choice whether to test the research hypothesis via hypothesis testing or a case-study method. The case study method is especially appropriate when the phenomena under scrutiny is complex, and hard to assess separately from the wealth of factors present in the real world (Yin 2013, Eisenhardt & Graebner 2007). Whereas the laboratory experiments aim to hold as many noise factors constant as possible, the case study method emphasizes the richness and variety of the real-world phenomena without trying to separate it from the actual context (Yin 2013, Eisenhardt & Graebner 2007). As the research building on top of the human interactions is by nature complex and hard to predict, researchers often incline in those cases towards choosing the case study method (Eisenhardt & Graebner 2007). For the purpose of this study, the case

study method was chosen as the context of the data usage in improvement projects is highly volatile, and building laboratory experiments would likely fail to depict the real interactions and importance of different data quality elements.

Case studies can be used to address different types of research problems. According to Yin (2013), the case studies can be divided into three separate groups: exploratory, descriptive, and explanatory case studies. Exploratory case studies focus on finding patterns from the evidenced data without narrowing the question too much beforehand (Yin 2013); the goal of such an approach is often to generate new theories from the case studies (Ketokivi & Choi 2014). Descriptive case studies, on the other hand, often start with a clear direction for the inquiry from the beginning trying to expand an emerging theory (Yin 2013). The explanatory case studies take the approach a step further to understand the mechanisms why and how the phenomena happened (Yin 2013). The case studies can also be used to test the validity of an emerging theory (Ketokivi & Choi 2014). For the use of this study, a descriptive approach was chosen with a goal of elaborating an existing theory. The descriptive approach was chosen as the basic data quality research is well established, but not yet focused on the use of improvement projects inside organizations.

The case study research can comprise one or more case studies which serve as evidence for the research question. Whereas in most mathematical approaches choosing the samples randomly is often the best choice, that is rarely the case in the case study research (Eisenhardt 1989). In case study research, the number of cases which can be studied is often limited: thus, the focus should be more on cases where the phenomena under scrutiny is easily observable (Eisenhardt 1989). Bearing in mind the time limitations when building cases, having more than one case study often creates a more solid base for generalizing the conclusions (Eisenhardt & Graebner 2007). According to Eisenhardt (1989), the optimal number of cases often lies from 4 to 10. On the other hand, single-case studies are common in literature and they have their place in certain situations: using a single case can be the right choice if the case describes the phenomena unusually well, serves as an extreme example, or opens up a rare access for research (Yin 2013). Generalizing the conclusions from a single-case study can,

however, lack a solid empirical grounding, if the case itself is not consisted of several mini-cases.

3.2 Research questions

Based on the findings from the literature review the research question was formulated. To support the research question, three subquestions were chosen to guide the inquiry. The research question for this study goes as follows:

Research question: What is data quality in healthcare improvement projects?

Q1: Why data quality is important in healthcare improvement projects?
Q2: How is data quality organized in healthcare improvement projects?
Q3: What are the impacts of data quality in healthcare improvement projects?

3.3 Data collection process

In order to gather qualitative evidence for answering the research questions, a case study was conducted in a Finnish healthcare company. Based on the work of Yin (2013), a single-case study is appropriate method when the case serves as an extreme example or opens up a rare access for research. In this case, the decision was made to focus rather on one project which had five clear distinctive phases with decision points in between. In this case, understanding the actual decision-making processes is key: thus, deep understanding of the different phases is likely to add more insight rather than mere questionnaires or interviews. The chosen case study focuses on an operational improvement project in an invoicing process which is handled in collaboration with several different functions.

The case study was conducted using an action research method, where the author was actually part of the case study himself. The action research has been well established in the academic literature (e.g. Reason & Bradbury 2001, Coghlan & Brannick 2005): the roots of action research stem from as early as 1940's of the work of Kurt Lewin (Reason & Bradbury 2001). The modern theory in action research is not built on a single set of

papers; rather it combines several different streams of literature ranging from social anthropology to psychological experiments on education (Brydon-Miller et al. 2003). The term action research is often used to describe the research happening inside one organization: where the researcher actually is one of the actors inside the process (Coghlan & Brannick 2005). Rather than accepting that the theory creation should be done by observing the past, action research seeks to inform the theory creation process by tightly linking to the observations about the practice (Brydon-Miller et al. 2003). In the action research, the focus on creating actual, valuable knowledge in a highly participatory manner (Reason & Bradbury 2001). Thus, a term "participatory action research" is often used to emphasize the collaborative nature of the action research (e.g. Koch & Kralik 2009).

Action research is often especially topical when the phenomena under scrutiny is highly linked to human interaction, e.g. the human decision-making processes (Reason & Bradbury 2001). Due to its collaborative nature, the action research guides the theory creation process by seeking to deeply understand the actual occurrences in the real-world context: not only asking post hoc about the decision points, but rather being part of the situation where the decisions are made (Brydon-Miller et al. 2003). On the other hand, the participatory approach to research can be highly educative for the persons involved in the study: the goal of the study is often to reach higher understanding together with the persons involved in the process (Koch & Kralik 2009). As the purpose of this research was to understand better how the data quality affects healthcare improvement project, being part of the decisions was chosen as a superior method to post hoc reasoning. In hindsight, the decision made by the actors might seem to them very different compared to how they felt in the actual moment: thus, the reflection was chosen to be made rather on the work of a holistic project where the author was a part of rather than inquiring later on about several other projects.

The data collection process started on the 1.3.2016 and lasted until the 3.10.2016. During the data collection process, ~40 meetings and workshops were attended in which the data was gathered for further analysis. The full list of those meeting and workshops can be found from the Appendix 2. The data collected for the case study comprised two

main types of evidence: the qualitative evidence, e.g. meeting memos and presentations, and quantitative evidence, e.g. data analyses and graphs presented to the management. On the other hand, some of the evidence is used as a part of formal reporting, e.g. presentations or recommendations to units, and some as informal communications, e.g. meeting memos or the draft analyses. Based on the two identified axes, the Table 2 illustrated the gathered evidence.

Category	Evidence item	Description
Formal qualitative	Presentations	The presentations used to communicate outside the team, e.g. steering committee or units
	Workshop outcomes	The outcomes communicated outside the team to units and steering committee
Informal qualitative	Meeting memos	Internal meeting memos which were used inside the team
	Emails	Selected emails which were sent to the team or steering committee illustrating the current process.
	Draft presentations	The slides which were used for internal communications or material which was never shown.
Formal quantitative	Presentation graphs	Graphs and analyses shown outside the team, e.g. to units or steering committee
	Recommendations	Recommendations shown to persons outside the team based on the data analyses
Informal quantitative	Draft data analysis	The data analyses and graphs which were in progress or invalidated at the time
	Raw data	The actual raw data sets used in different phases of the process

Table 2: Evidence gathered to support the case

Due to the inherent complexity of the invoicing process, the project included several stakeholders from different functions. For the purpose of this academic report, three main groups of people emerged who were critical for the success of the project: steering

committee, project team, and the outside experts. For the sake of anonymity, the real names of the persons are left out, and they are called with the names expressed in the Table 3.

Category	Title	Role / Specialty
Project team	Project manager	Project manager
	Invoicer 1	Invoicing function
	Invoicer 2	Invoicing function
	Customer service 1	Customer service
	Customer service 2	Customer service
	Function manager 1	Manager in one of the functions inside a unit
	Invoicing support 1	Invoicing support
Steering committee	Manager	Representative(s) of the management team
	Project owner	Project owner
Experts	Expert	Wide variety of experts called when needed. The specialty specified when appropriate in the text.

Table 3: Project team members

Having described the data collection process and the methodology for the research, next section will describe the actual case: how it flowed over time, what decision gates existed in the process, and how the data quality was linked to the success of the project.

4 Case

The case study focuses on a project conducted in a Finnish healthcare company between 1.3.2016 and 3.10.2016. This section focuses on the five distinct phases of the project, and how the used data linked to the decisions and perceptions in each of the phases. Each of the five phases are described in detail in their own chapter. The values of the data are, however, all indexed and do not represent the actual amounts used for the conclusions. After this section, the results section will shed more light on the findings for the research questions, i.e. what kind findings the study suggests for the usage of data quality in healthcare improvement projects.

While the focus of the case study is on one project, the project itself comprised 5 different and distinct phases: the phases are shown in a Figure 1. The project started with combining previous efforts with qualitative data, and trying to push the found suggestions to the units. Second, an access to high level data was granted which was able to guide the project forward as a summary statistic - but did not help in finding the root causes. Third, a more detailed data was used to analyze the assumed correlation - if the invoices were slow, they were likely to be defective. Next, more detailed data about the claims was gathered: data which was very interesting but did not lead to actionable insights. Finally, the actual progress was made using the manually gathered data on the actual causes rather than relying on the good quality, high level data.

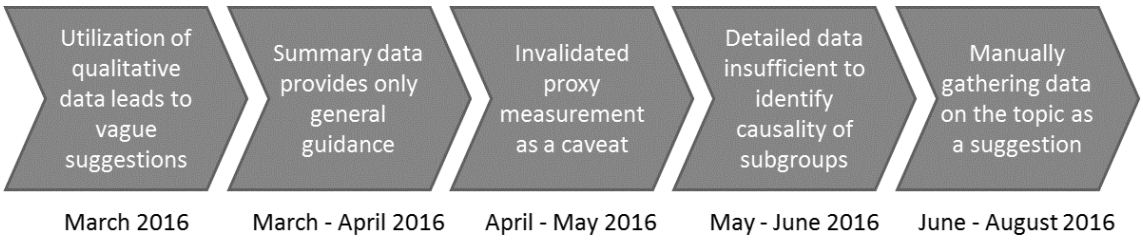

Figure 1: Timeline of the case

The project itself was focused on decreasing the amount of problems in the invoicing process. The problems were observed through three lenses: customer claims, extra work and hassle due to rework, and billing delay. Of those three goals, the main goal was initially chosen to focus on the claims and extra work, and the delay was left as a secondary goal to be tackled if easy wins emerged. Previous efforts were already conducted on the topic, and the data and findings from those efforts were used as the preliminary material for the project team. With this context, the project embarked on its seven-month-long journey aiming to decrease the invoicing problems.

4.1 Utilization of qualitative data leads to vague recommendations

The project started with assessing the work that had been previously done on the topic. Several workshops had been held, and preliminary data analysis was conducted. Based on those, several pages of recommendations were formulated, but hardly implemented. This chapter goes into detail on the material that was handed to support the decisions as well as describes how the project team started formulate, and build on those findings.

4.1.1 Creating the foundation for a successful project

The first step for the project manager in tackling the problem was to formulate a project team. The process itself was complex, and involved several different stakeholders situated geographically in different places. The high-level view of the process is illustrated below in the Figure 2. In order to have a good understanding on the process and its problems, the team needed to comprise people from very different background - both analytical and practice-oriented, both support functions and people doing the actual work. In order to build such a team, the first step was to create a high-level support for the project in order to secure resources.

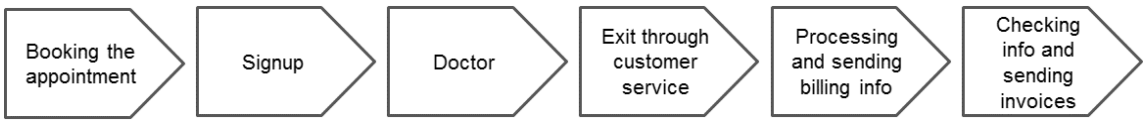

Figure 2: Process steps in the case project

The project was really kicked-off in the first steering committee meeting held 4.3.2016. The process owner, the person responsible for the daily management implementation, was chosen as well as the project sponsor, who was ensuring the high-level support and resources for the project. Present was also a person who had successfully conducted improvement projects in the company. In the first steering committee meeting, the scope of the project was clarified: the project focused on the claims rather than the invoicing delay. Though an assumption was raised that slow invoices could serve as a proxy measurement for the defective invoices. All participants in the meeting agreed on the assumption, and that was set aside for a while to be revisited in the very near future. In the end of the meeting, the discussion turned into the work which had already been conducted on the topic in the past, and who should actually be involved in the process.

After the steering committee, several calls and emails were exchanged on the composition of the team. The decision was made to divide the people involved in the project into three separate groups: steering committee, project team, and project support. While the steering committee had already been set, negotiations were going on who could be the best one to advance the work of the project team. After the period of several informal discussions face-to-face, via email, and via phone calls, the project team took shape. In the end, 8 persons were included in the team: the listing of the persons can be found above in the Table 3. The first team meeting was held 11.4.2016, but a lot of work happened already before that.

4.1.2 Using the material already gathered

While the process of formulating the project team was still in process, several discussions were held to start tackling the problem. The first step was to gather the material already done with the people who were involved in the past efforts. The project manager started by going through the data analysis, outcomes of workshops, and the

recommendations for improvement which were formulated at the end of the past effort. Based on those materials, two longer meetings were held on top of the calls and emails: one with the persons who had been involved in the analysis and gathering the data, and one with the person who had led the former improvement effort. The next paragraphs will explore the used material through the outcomes of those two meetings.

The first longer meeting focused on the data what was used to guide the former improvement effort. The difficulties of finding a lot data on the topic were depicted by a former improvement team member who said that “No useful data exist on the topic”. To support the quest for the reasons for invoicing problems, the former improvement team had basically had one useful dataset. The used dataset comprised of the numbers of invoices per month and the causes that were listed in the system. The output seemed very convincing: understanding the causes would greatly help in the focusing of the project to the biggest causes for defects. Furthermore, two reasons seemed to raise above others as denoted in the graph A: Type A with 100 defects in a month (indexed), and type E with 68 defects in a month (indexed). The illustration of the output graph can be seen below in the Figure 3.

When the discussions were taken forward, concerns started to be raised when no one in the room actually knew how the data was produced. The dataset was sent by a third person inside the company, which meant that the numbers were not easily checkable, let alone reproducible. After the meeting, a few phone calls revealed the source of the material, and how the causes were created. Seemingly legit, the resolution was in the end to aim to get more accurate data to validate the dataset. Though, there was no clear evidence at the time to suggest that the previous dataset did not reflect the best guess of the truth at the time.

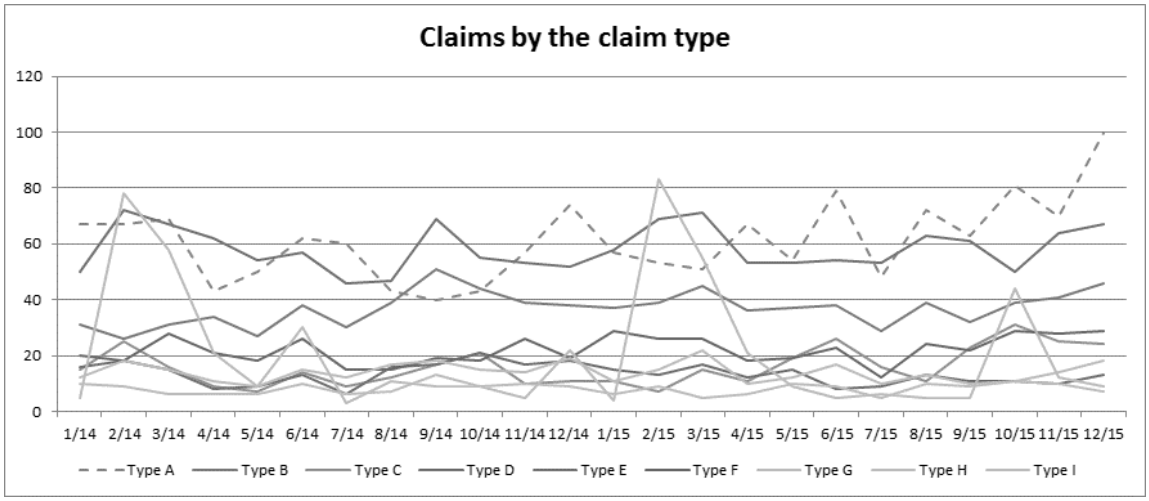

Figure 3: Initial data about the claims

After the meeting about the previous data, a handover meeting was held with the person who had actually pushed forward the last improvement effort. The material produced in the last effort was extensive: it included analysis of different factors that could cause variation in the invoicing process, workshopping about the reasons with unit-level and support function employees, as well as solutions suggested based on those findings. The identified possible reasons included over 50 different possibilities grouped under the 5 identified process steps. The reasons were then prioritized through a series of voting procedures held in the workshop sessions. The improvement suggestions were, in turn, based on the prioritized reasons, and looking at the material, they seemed to make a lot of sense. For some reason, there seemed to still be a lot of problems to push the long list of improvement suggestions forward to the units.

After going through the qualitative material produced in the past, the discussion turned into the data. In the discussions, it became clear that the data did not serve as the guiding factor of the improvement project, but rather the workshops and observations were the crucial bits of information. Digging deeper why so, the data on the causes seemed not to give enough support for prioritizing the improvement efforts. Among the most common data labels, there were “data not coded”, and a “process error in the unit”. In other words, several data labels were barely insightful to describe the real causes of problems in the units - a place where the actual work was done.

After the discussion with the former improvement effort leader, the project manager focused on understanding the source and reliability of the cause data in more detail. A call with an informal meeting with a person doing the invoicing revealed that actually some of the codes used in the invoicing were not part of the invoicing process the company focused on. Rather the data seemed to include also causes which should not be even possible in the invoicing process. After a call with the persons analyzing the data, the conclusion was that the data itself was not to be trusted. Before the conclusion, the perception of the project manager was that the data showed a detailed truth of the causes over the course of time. The finding motivated the project manager as well as the persons involved in the data analysis to search for better data on the topic which could actually be helpful in directing the project.

4.2 Summary data provides only general guidance

After the initial meetings and discussions about what had been done in the past, project team concluded that more quantitative approach to solving the problem was required. As the former improvement team member said, “the units are very different - some problems they have in common, but some are quite unique”. In order to have get a clear understanding on where to focus, the data could guide on grouping units with similar problems or units which had the most problems in a certain area. Thus, calls were made and emails exchanged to find if such a data existed which could help guiding the project.

4.2.1 Getting hands on the first detailed dataset

In few days, a dataset was found from one of the managers in the invoicing process. The dataset looked very promising: it comprised the information about each of the reimbursements with some additional information - including from which unit they came from. The mood among the project team members started to go up, and the project manager started to analyze the preliminary findings from the data. Everything looked very promising, and the initial qualitative findings might finally be backed up with actual data analysis. The initial graphs were made, and findings showed the total sum of reimbursed value per month as well as the contribution of the individual reimbursed

invoices towards the total sum - an illustration of the graphs can be seen in the Figure 4. The findings suggested that few invoices contributed half of the total sum of the invoices suggesting that focusing on vital few groups of invoices might solve most of the problem. On the other hand, two months with clear spikes of reimbursed value were observed - one in June and one in October (illustrative).

The initial data analysis findings were gone through with the process owner and the insurance specialist. Everyone seemed to be happy about the initial progress: understanding which units to focus made it easier to achieve real progress. In addition to that, the data helped in quantifying the actual monetary impact the project might be able achieve. The discussions focused on how the data should be used to guide the improvement project forward: one of the suggestions raised was to focus on few units which had the most problems, and then leverage the findings to all other units. Everyone seemed to agree, though, that the current data analysis itself was not helping the units on how they should do the change. Rather, the data analysis served as an illustration of the magnitude of the problem.

After the initial data analysis, an informal discussion was made with the controllers. A concern was raised that the dataset was not in the standard output format which should be coming out from the basic reporting systems. The output of the dataset looked, though, to be plausible when checking the unit and time data. A decision was made, however, that the facts presented should be double-checked by creating access to detailed data on the topic rather than relying solely on the unverified dataset. The work on creating the access to a detailed dataset was then kicked off, and the detailed specs were discussed further in a meeting 31.3.2016.

Uneasy about the discussion with the controllers, the data analysis continued but a decision was made not to publish the results yet to persons outside the project team. Inside the team, however, several discussions were held on why certain months were the worst, and why certain units seemed to have more claims compared to the other units. Initial percentages were calculated for the error rates, and they seemed to suggest significant differences between the units. On the other hand, all units seemed to have the

problems suggesting that at least some of the problems were in the system itself rather than how the units were doing their work. Efforts were also made on combining the qualitative findings from before to the data-based findings, but little advancement was made there. The problem was that the data was able to portray only a very high-level picture of the months and units, while the possibilities for the variation were highly complex and were observable only on actual work place. The decision was, thus, made to go to understand the process better and aim to build an access to better quality data.

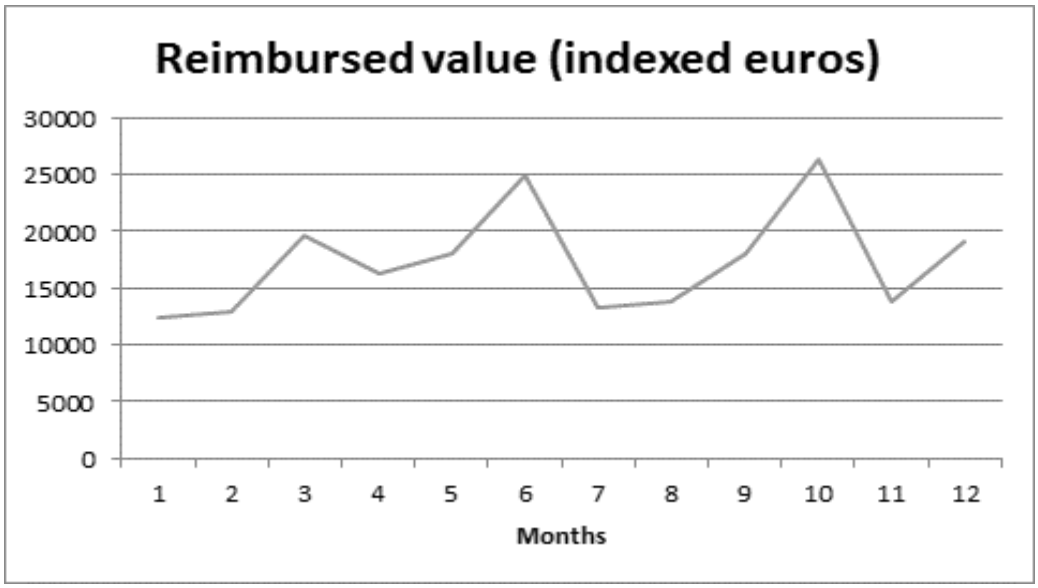

Figure 4: Reimbursed value per month - Initial data

4.3 Invalidated proxy measurement as a caveat

While the search for a better dataset was going on, and the results of that search were still uncertain, the project manager came back to the suggestion presented in the first steering committee meeting: there could be a correlation between the invoicing delay and the number of problems in the invoicing process. The idea was discussed with both the process owner, and the controllers after which everyone seemed to agree that the correlation could be a real thing. Thus, a dataset was gathered which had detailed level data about the billing delay, and its possible causes.

The first step of the detailed data analysis was to understand which kind of invoices caused the most problems. The data analysis showed that most of the problematic

invoices were, both percentage-wise and by actual values, in fact caused by simple operations rather than highly complex ones. The team went on to discuss the matter, and came up with several possible explanations for the finding. One of the possible reasonings was that the larger invoices receive special care in the units, as the time used to prepare those invoices is higher. On the other hand, smaller invoices might be left with less notice as they are often simple and straightforward operations - if something is missing, there might not be anyone who immediately spots the mistake. Motivated of the result, the team concluded that the difference between complex and simple operations was significant: thus, being able to focus on one size of operations was likely to help in guiding qualitative inquiry forward.

After the finding that the most simple operations seemed to cause the most problems and variation, next step was to understand what kind of operations were the most difficult ones. Consequently, the next step of the data analysis was to understand the average delays in different types of operations. On the other hand, just stating that one of the operation types causes most problems would be insufficient - it was also very important to understand the magnitude of the problem. Thus, the discussion turned into finding the most problematic types which were meaningful in terms of reducing the amount of extra work and claims. Thus, an analysis on different types of operations was conducted - an illustration of the analysis is presented in the Figure 5. The mood among the team was rising, as the outcome had so far suggested that focusing on certain types of operations, which would be of a simple nature, could result in a meaningful remedy for the problem.

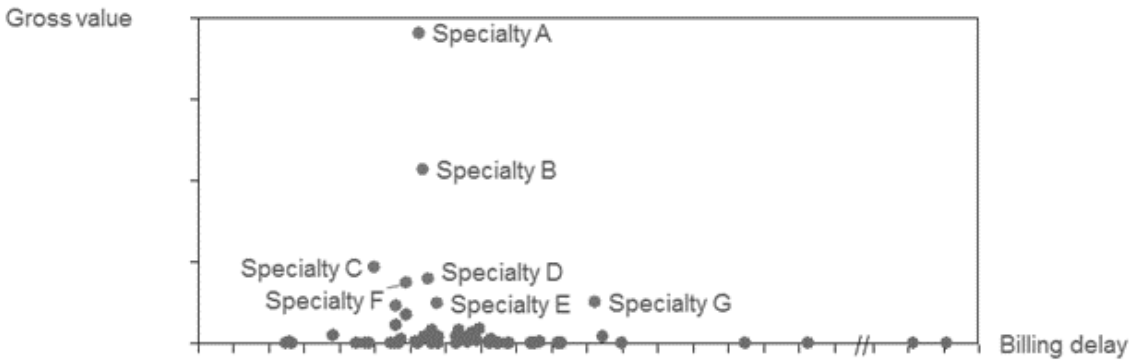

Figure 5: Reimbursed value per specialty

Intrigued by the results, more data was gathered to understand how big an impact could be gained by doing such changes. One part of the analysis was to understand the magnitude of the "slow invoices", and how much of the invoicing claims they could have caused. The illustration of the graph can be found below from the Figure 6. The findings from the data analysis were disencouraging: the total amount of simple invoices was not as high as assumed, and that number was hardly able to explain the variation in the reimbursements found in the previous phase. Even smaller was the proportion of slow and simple invoices suggesting that the cause of claims was hardly found from the small and slow invoices. While correlation between the invoicing delay and the proportion of invoicing problems was very likely to exist, it seemed that the majority of the problems were likely to be caused by normal invoices which were sent relatively fast. Thus, the team was sent back to drawing board to find out more information about the claims rather than focusing more on the invoicing delay problem.

Billing delay / Invoice value (€)	Fast	Slow	Very Slow	Instructions: Value in k€
Large	79 k€	13 k€	13 k€	**105 k€**
Mid-sized	32 k€	7 k€	14 k€	**53 k€**
Small	7 k€	2 k€	5 k€	**14 k€**
	118 k€	**22 k€**	**32 k€**	

Figure 6: Summary of the invoice values per billing delay category (indexed)

4.4 Detailed data insufficient to identify causality or subgroups

Soon after the setback of analyzing the invalidated proxy measurement, i.e. billing delay, an access to new dataset was available. The project manager had few meetings with the controllers after which the new dataset was validated to be correct, and ready to use. The new dataset provided information on the actual reimbursement numbers compared to the total invoicing. In addition, the new dataset included very detailed information of each claim – from which type of operations, units, and products the claims were coming from. Ready for the new challenge, the project team set on to analyze the new data to find more information about the actual problem.

4.4.1 Analysis of the detailed dataset

The new dataset revealed the real magnitude of the problem – rather than 1,5% (indexed) proposed earlier, the actual proportion of the defects was 3% (indexed). The data analysis began by assessing different units to see their performance both in relative and absolute terms. The differences existed, as predicted from the previous analyses, but the order of the units was not the same as in the delay ranking. The initial unit level findings are summarized below in the Figure 7**.** The project team gathered up 11.4. to discuss about the findings from the data analysis. A consensus emerged that taking 2-3 units as the pilot partners would make sense, as finding the solution that would

immediately fit all different units would be difficult. The data analysis guided the team to choose the most important unit as the focus of their data analysis.

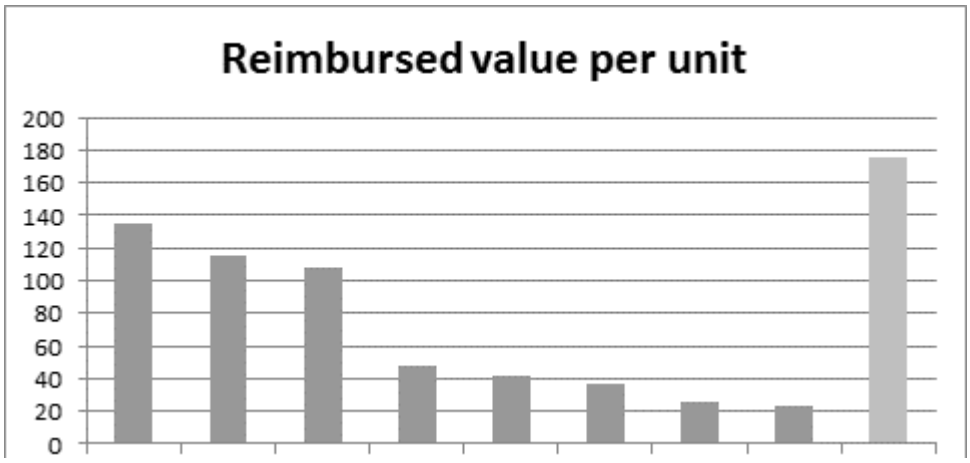

Figure 7: Reimbursed value per unit (indexed)

The more detailed data analysis went on quickly, and the first outcomes seemed very promising. The analysis showed, in fact, that differences existed between different units, especially in the biggest invoices - an illustration of the finding can be found from the Figure 8. The meetings were then held and reasons gathered on why this could be happening. The meeting outcomes were that the operations with the most defects are the most complex needing the most extra material and work. Finding that the differences actually were in these biggest invoices would help the project team to focus their improvement efforts on the highly complex invoices, which were different from all the others. Keeping in mind the previous findings which suggested the problem to be in fact in the small invoices, the project team prepared to find out very diverse set of reasons for the claims. Motivated from the current findings, the project team continued to dig deeper in the data analysis.

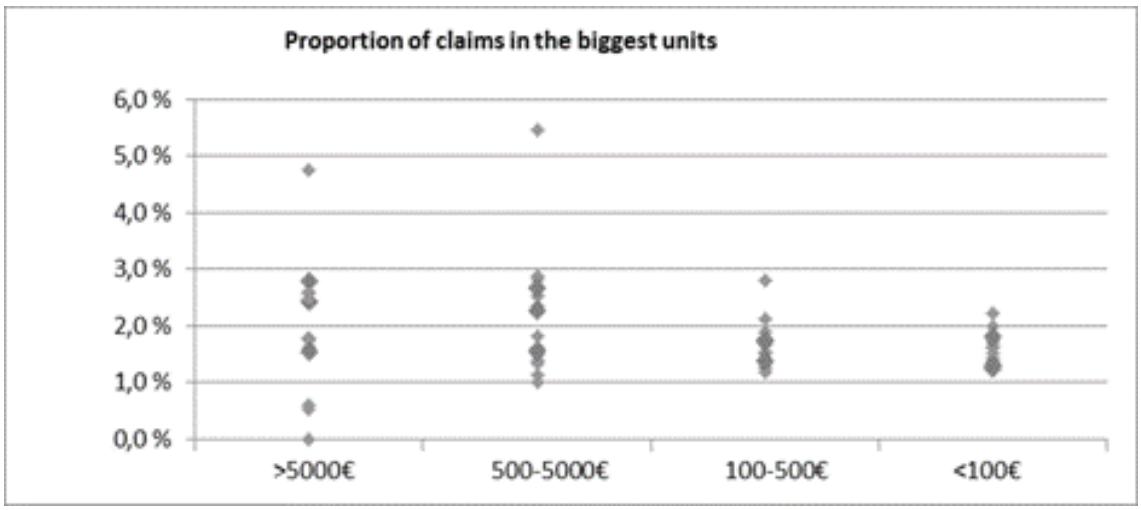

Figure 8: Proportion of claims in the biggest units (indexed)

Continuing the search for better focus and understanding, the project team found out now the same graph that they encountered before – but with very different numbers. The new data had codified also the reasons for the reimbursements – the illustration of the graph used can be found below in the Figure 9. The initially found spikes in certain months were still there in the same places as in the initial data - i.e. the 3rd, 6th, and the 10th month. On the other hand, the identified reasons told a totally different story this time compared to the initial findings. The new dataset revealed two new categories of claims, i.e. types J and K, both of which did not exist in the initial dataset. Furthermore, the magnitude of the two new types of claims was significantly bigger compared to the other types of claims. Thus, the conclusions made by the data would have been vastly different if they would have been based on the initial data.

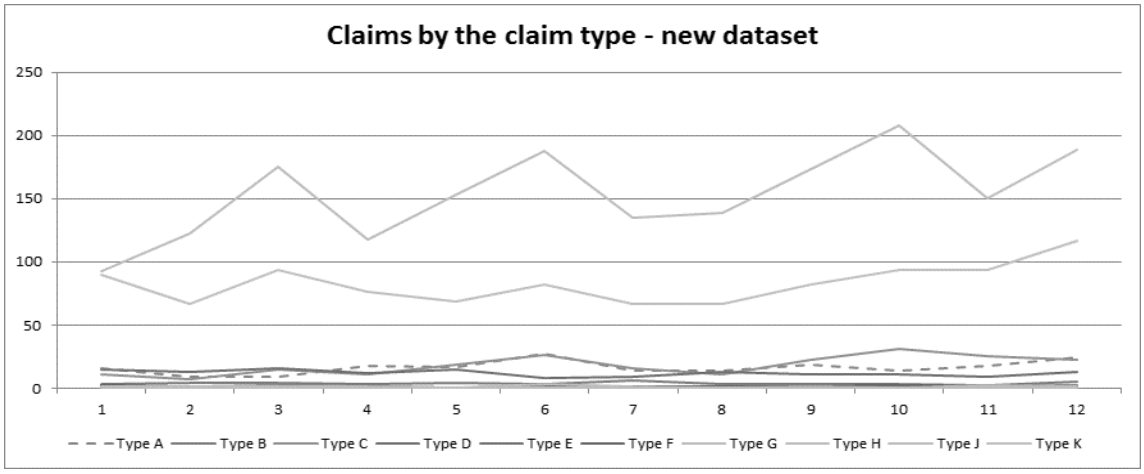

Figure 9: Claims by the claim type - new dataset (indexed)

4.4.2 Workshopping and qualitative tools to support understanding

In parallel with the data analysis, a rigorous effort was started to create a holistic understanding of the process, its failure modes, and what could be done to mitigate them. Dozens of small and large meetings were held which focused on understanding the actual process from the beginning to the end. At this moment, one pilot unit became an important part of the rest of the improvement efforts, as all the work was done in collaboration with them. The goal of the qualitative efforts was to use the data analysis as a guide, and let the data-backed qualitative insights to lead the project team to the root causes.

The efforts to understand the process started with ~10 one-on-one meetings which aimed to create a holistic understanding of the invoicing process with its handovers between different stakeholders. The process map helped the project team to ask questions about the data: if the most difficult invoices have the most problems in relative terms, what kind of loops those invoices share what the simpler invoices do not have. In the case of the difficult invoices, a separate loop was identified which could cause problems in the invoicing process. With the actual reimbursed invoices, the project team was able to have conversations with the workers about the possible problems the process had.

Having identified the process, and how it linked to the findings in the data, the process team started to list the failure modes. Using the process map, ~5 meetings were held on

building the failure analysis further. The goal of the step was to identify as many of the possible failure modes as possible, and then use the data, if possible, to prioritize the improvement targets among these possible failures. Having done the preliminary identification of the failure modes, a half-a-day workshop was organized to gather up the right people to actually decide on what to do with the problems. Armed with the data analysis done on the topic, the group of 6 persons representing all main function in the process started to prioritize the failures and list the possible solutions for them. In the end of the workshop, participants were happy that actual progress was done. In practice, the outcome included a list of 3 main problems, and illustrations of sequences on how they could happen. Then, a list of 10 solutions for those problems was chosen for further assessment. The mood among the project team was high; a consensus seemed to exist that the solutions were real and could have an effect.

The results were presented to both the steering committee and the unit level management team, and a good discussion started on how these solutions could be implemented. There seemed to be an agreement that the identified problems relied on good logic and were likely to be the most important ones. Several meetings were held which aimed to formulate the solutions further – some were IT-based, some needed a change in the daily work routines. Drafting the solutions proceeded initially in a good pace; detailed descriptions of the problems were created along with how the solutions could solve the problems. Doubts started to arise, though, whether investing in the few most important solutions would actually make the difference: in other words, the material was not conclusive that the reasons found out were actually the most important ones. The management supported pushing forward the improvement suggestions, but the project team decided to take a one more break before concluding the analytical work done.

4.5 Manually gathering the data on the topic as a solution

With a list of solutions in progress, the project manager decided to initiate a discussion about how could data be gathered from the actual reasons, not just about the financial view of the invoices. During a meeting with a unit-level specialist a new idea came up: a temporary manual gathering of the reasons for invoicing problems might in fact add

value on top of the current findings. The reasoning was supported by the feeling from the unit-level people that the real problem might not, in fact, lie in only the claims, but also the incorrect invoices which are never sent to the customer. A plan on gathering new manual data was created and an approval was gotten soon from the steering committee and unit management team.

4.5.1 Gathering manual data to find the root causes

The plan for manual gathering relied on capturing all problematic invoices, independent on whether they were sent to the customer or not, and codifying the problem to an easy VBA-based tool. The reasons were relatively well predefined based on the previous data analysis; the list was filled in meetings with unit-level specialists as well as the invoicing employees. The tool went through the testing period, and was very soon in use for the first measurement period of three weeks. The invoicing employees were generally happy about the tool, as they could now actually contribute on finding the amounts of defects: a topic which they had a strong hunch on but never had clear data about it.

The results of the manual gathering were surprising in the beginning – the results are illustrated below in the Figure 10. The data clearly showed that five reasons caused approximately half of the invoicing problems – a finding that is itself hardly shocking. But what was surprising was that the five reasons included only one of the previously prioritized reasons; from the other four, only one actually made even to the top ten reasons. Everyone knew that the manual gathering was unlikely to be absolutely precise, but what it managed to do was to codify the magnitude of the problem as well as the real reasons behind them. The previous hypothesis was true: most of the invoicing problems were corrected before reaching the customer. Thus, the biggest overall improvement was likely to be found in searching for the solutions with most leverage in decreasing the invoicing problems as a whole, rather than just the reimbursements.

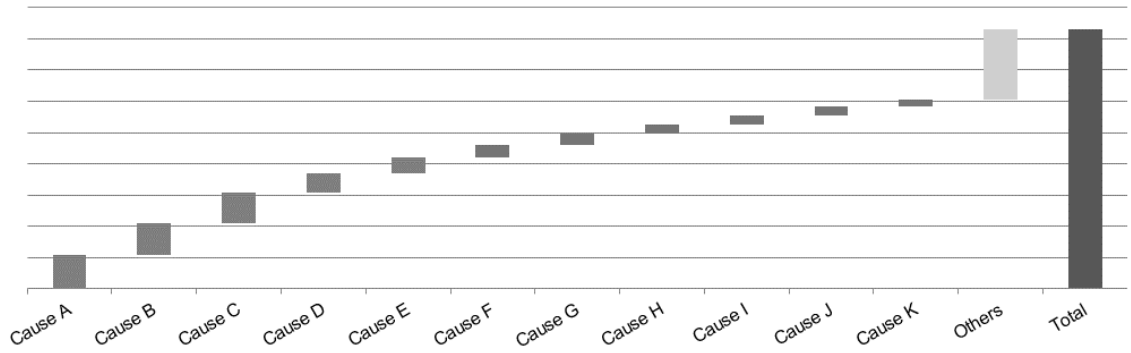

Figure 10: The most common causes for defects (indexed)

On top of understanding the causes for the problems, the manual gathering enabled discussing with units which were their own biggest problems. The project started with the hypothesis, "the units are very different - some problems they have in common, but some are quite unique". The new data was able to tackle such comments by pointing out the problems that actually were the most important for the unit under scrutiny. To test the effectiveness of the new data, meetings with nine different units were organized. Majority of the units described the data to be highly useful, and approximately half of the units drafted their own action plans already during the first one hour meeting. One of the unit managers said: "Finally we have clear data about the problems in our unit. Something we can immediately act upon".

Based on the findings from the manual gathering, a list of both unit-level and group-level solutions was formulated. The unit-level solutions were created for the most common problems, and shared with the units alongside the actual data. The goal of combining the data with the solutions was to draw a clear path for the units so that they could try to improve their biggest problems right away. The group-level improvements were, on the other hand, targeted to the five most common problems which were difficult for the units to solve themselves. Most of the group-level solutions were implemented in the end, and some were separated for their own projects which went on after the end of this project. In the end of the day, getting an approval from the different stakeholders seemed to be a lot easier with data which was actually showing the root causes for the problems.

The closing of the project happened on the 3.10. with the last steering committee meeting. The responsibility of ensuring that the gains were lasting had been transferred to the daily management, and the control of the number of the defects made part of another decision-making process. In the end, the project managed to decrease the proportion of billing defects by double digits compared to the baseline measurement done before the implementation of the improvements.

5 Results

The case study comprised several key points of time where the project team had to decide whether to invest in finding more support for the suggestions or not. In all these decision points, the data quality had affected significantly the quality of suggestions made by the project team. Whereas looking back to the project, finding out whether the data was of a good quality was easy in hindsight, but very difficult during the actual moment the decision was required. Thus, this section focuses on describing the data quality attributes present in those five phases as well as the implications for decision-making and how the problems with the data were identified. The findings from the case study are summarized in the Table 4 below, and described in more detail in the following chapter.

Phase	Description of the data quality	Implications for the decision-making	How the problems with the data were spotted?
Utilization of qualitative data leads to vague recommendations	Data of the wrong topic, perceived to be correct Low accessibility due to lack of trust to the data	Decisions made without real data A long list of suggestions which were not implemented	Items in data which should not have been there Problems spotted when speaking with people doing the work
Summary data provides only general guidance	High-level data, did not lead to action Incomplete data, though spotted only later	Helped to focus, but not to find solutions The incorrect data led to bad business case estimates	Hunch when the data source could not be validated Spotted only when new data was available
Invalidated proxy measurement as a caveat	Accurate data with a lot of details Was, in the end, of a wrong topic	The focusing decisions would have led the project to the wrong direction	Hunches arose when validated the data against the preliminary numbers Truth revealed with the new dataset
Detailed data insufficient to identify causality	Intrinsically good quality with several dimensions Helped to focus but not to find the ultimate problem	Data with workshops looked convincing, but the solutions were wrong	Concerns arose when the implementation faced problems Solutions were deemed wrong only after the new dataset
Manually gathering the data on the topic as a solution	Accuracy of the data modest Revealed the likely causes and actionable input for units	Findings resonated with the units Significant improvement achieved	*Not applicable*

Table 4: Summary of the case results

5.1 Data quality in the case and its implications for the project

5.1.1 Utilization of qualitative data leads to vague recommendations

The project began with comments that no real data existed on the topic. The previous improvement efforts had, on the other hand, in their hands data which could have helped to find the possible causes for the problem. From an intrinsic quality point-of-view, the quality of the data was likely to be relatively high: problems with accuracy, reliability, consistency, and objectivity were not spotted. After all, the data was produced by filtering data from the invoicing database. On the other hand, more concerns can be raised about the contextual qualities of the initial dataset: the data was not relevant for the purpose, but it was perceived to be so in the beginning. The trickier part comes with the accessibility problems: The dataset was not used in the end of the day, because the project team had doubts about whether the dataset was value-adding. One could, thus, argue that the perceived low contextual quality of the data could be a cause for the low accessibility, i.e. the lack of usage of the dataset. Finally, the representational quality of the dataset was adequate - the data had clear labels for the problem types, and was simple enough to guide the right interpretation.

The lack of data usage led the previous improvement efforts to use their best qualitative understanding of the process to formulate solutions. Being a very complex process with several stakeholders, the qualitative approach led to a long list of improvement suggestions which were in a large scale not implemented. The previous efforts had faced a problem with trying to pursue forward without data: wide-spread agreement on the prioritization of the improvements was very difficult to achieve. Most seemed to nod that the improvement suggestions made sense, but they hardly drove action in the units. Even with the data available, it was hardly used to back the improvement suggestions for the reasons mentioned in the previous paragraph. The bad quality data was, however, not spotted in the first meetings of the new project team, but only later meetings pointed out the likely problems. The data was of a different topic that it was perceived to be - a problem difficult to spot by the project team.

5.1.2 Summary data provides only general guidance

The second phase of the project started with gaining access to the summary data on the invoicing claims. The intrinsic quality of the dataset was perceived to be high in the beginning: thus, it was used. The reality was, though, that the dataset was incomplete and could not tell the real magnitude of the problem. From the contextual data quality point-of-view, the dataset could still guide the project forward, and thus was value-adding. On the other hand, the dataset did not reveal any hints of the causes for the invoicing problems, and thus should be labelled as summary data. The accessibility of the data set was high, and it was highly relied upon: interestingly, even if the data was less accurate than in the first phase, it was perceived to be good and thus it was also used. Only after the cautions were raised about the source of the data, the accessibility started to decrease with perception of possibly worse quality data. The representational attributes were, again, good as the data was easy to approach and clear.

The dataset helped the project team to discuss on which units they should focus their improvement efforts. The dataset was clear, and gave a perception of understanding the magnitude of the business problem. The reality was, however, that the dataset was incomplete, and the number of invoicing problems was a lot higher than proposed by the dataset. From the project team point-of-view, the trust towards the dataset was high as the amounts seemed plausible based on the initial assumptions. In reality, the problems with the dataset were actually seen only after the access to the new dataset was built. Hunch about the possible concerns arose, however, already when the source of the dataset was unclear. The unclear source pushed the project team to seek for new data to validate the current findings. The reality would have been, though, that without the new data, most likely the current situation would have been depicted as far better than it was.

5.1.3 Invalidated proxy measurement as a caveat

After reaching the limits of analyzing the summary dataset, the next step was to get hands on very detailed dataset about the invoicing delays. Intrinsically, the dataset was of a very good quality: being validated and used several times before, it was accurate, reliable, and believable. From the contextual point-of-view, the dataset was perceived to

be of a good quality in the beginning. The project team expected the dataset to be guiding the inquiry forward, as it provided clear hints on where to focus. Only after deeming the dataset inadequate, it became clear that the dataset was not value-adding for the purpose. The accessibility of the dataset was also relatively high: the updated information was easy to reach, and it was extensively used. On the other hand, the representational data quality was adequate: the dimensions were defined, but they were difficult to immediately comprehend without extra effort.

The detailed data on the invoicing delay led the project team to focus the more detailed inquiry on the simplest invoices. Furthermore, the data had clear evidence on which units should the project be focused. While the storyline was compelling and widely accepted, it became clear, in the end, that the focusing decisions made based on the data would have guided the project to a wrong direction. The underlying problem was the acknowledged but accepted hypothesis that the billing delay would be an adequate proxy measurement of the invoicing problems: a claim which was supported by stories from manufacturing companies where longer lead times are often a sign of production problems. The hypothesis was, though, in the end deemed false, as the problems did not seem to correlate well with the longer invoicing delays. The first hints about the data problems were gotten when the data showed the number of the problematic invoices with long delays was vastly smaller than the number of invoices. The assumption was then proved with data later in the next phases.

5.1.4 Detailed data insufficient to identify causality of subgroups

The fourth phase of the project started with getting access to very detailed level on the invoicing reimbursements. The intrinsic data quality was, again, very high: the reimbursements were identified straight from the invoicing data which was validated several times. The dataset was, thus, highly accurate, reliable, and consistent. The contextual data quality, on the other hand, was also high: the data helped the project team to have actual guidance on where the problems were coming from. The accessibility of the dataset was also very high, as the data could be easily reproduced for

any given time period. The representational problems were similar than with the invoicing delay data: the dimensions were defined, but not all intuitively understandable.

The detailed dataset got the project team to full speed quickly. The data was used to help guiding the project team forward to focus on the most pressing issues. Furthermore, the project team used the data analysis to support the qualitative tools which aimed to understand the process and its possible failure modes. The data analysis was often present in all the discussions: e.g. a comment from the workshop by a unit-level specialist “I know that missing ticks in the payer checkboxes are a problem. But that problem should not be causing the problems in the most pressing specialties.”. The data-backed solutions created in workshops and meetings were accepted to be worth pursuing further in several different groups: the project seemed to be well on its way towards an actual impact. The reasons were identified as not the most pressing only after gaining access to better data in the last phase of the project.

5.1.5 Manually gathering data on the topic as a suggestion

The final phase of the project relied on manually gathered data from the invoicing problems. The intrinsic quality of the data set was modest at the best: everyone knew that not all of the problems were listed, and that human errors could exist due to subjectivity. The contextual data quality was, on the other hand, very high as it helped the units to actually focus also their effort on solving the most important few problems. The accessibility of the data was relatively low: gaining updated data required invested time in getting all invoicing employees to record their findings for a set period. Lastly, the representational quality was high, as the unit-level problems as well as the big picture were easily separable; on the other hand, some of the data labels needed more explanation before being immediately obvious.

The manually gathered data got a positive reception from both the group and unit-level employees, as it gave a clear prioritization agenda on where to focus. The data helped to prioritize the group level improvement efforts as a monetary impact could be directly calculated for the business case. In the unit-level, the dataset provided guidance creating

a focused task list which could be used to mitigate the problem. The units also used the succeeding datasets to see how their improvement efforts had impacted the current situation, and what they should focus on next. The dataset also proved a significant decrease in the number of invoicing problems during the first six months - a decrease that had not been successful by the last improvement efforts.

6 Conclusions

Based on the findings from the case study and the academic framework, this section aims to create a linkage between the literature and the case study. First, the findings from the case are portrayed against the data quality attributes found from the case. Second, the dimensions which affected the decisions made in the case study are discussed separately, and compared against the data quality attributes found from the literature.

6.1 Comparing the data quality attributes with the case study

The academic framework described how from several data quality models, Wang & Strong (1996) was chosen as the model to group the data quality attributes. In the model, the data quality attributes should be categorized into four main groups: intrinsic, contextual, representational and accessibility data quality. On the other hand, the findings from the case study portrayed how the data quality was in the case study, and what were its effects on the decisions made. Furthermore, the data quality was often perceived differently in the beginning compared to what was the actual case: thus, the division is made between perceived and actual data quality. The data quality is summarized using subjective findings, which were collected from the comments and findings from the project team. To categorize the data quality attributes a threefold scale is used: high, modest or low data quality. The Table 5 summarizes the case study findings against the data quality dimensions found from the academic framework.

		Utilization of qualitative data leads to vague suggestions		Summary data provides only general guidance		Invalidated proxy measurement as a caveat		Detailed data insufficient to identify causality of subgroups		Manually gathering data on the topic as a suggestion	
		Perceived	Actual	Perceived	Actual	Perceived	Actual	Perceived	Actual	Perceived	Actual
Intrinsic	Accuracy	⇧	⇧	⇧	⇨	⇧	⇧	⇧	⇧	⇨	⇨
	Reliability	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇩	⇩
	Consistency	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇨	⇨
	Believability	⇨	⇩	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧
	Objectivity	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇨	⇨
	Reputation	⇩	⇩	⇨	⇨	⇧	⇧	⇧	⇧	⇧	⇧
Contextual	Relevancy	⇧	⇩	⇨	⇨	⇧	⇩	⇧	⇧	⇧	⇧
	Value-added	⇩	⇩	⇨	⇩	⇧	⇩	⇧	⇨	⇧	⇧
	Appropriate amount of data	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧
	Completeness	⇨	⇩	⇧	⇩	⇧	⇧	⇧	⇧	⇧	⇧
	Timeliness	⇧	⇧	⇨	⇨	⇧	⇧	⇧	⇧	⇩	⇩
Accessibility	Ease of access	⇨	⇨	⇨	⇨	⇧	⇧	⇧	⇧	⇧	⇧
	Access security	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧
Representational	Interpretability	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧
	Ease of understanding	⇧	⇧	⇧	⇧	⇨	⇨	⇨	⇨	⇧	⇧
	Consistency	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧
	Concise representation	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧	⇧

⇧ High ⇨ Modest ⇩ Low

Table 5: Summary of the case results compared to the findings from the academic framework

Based on the data quality attributes from the literature review, the data quality in the first phase of the project was relatively poor. The data quality was perceived to be of a low reputation, as the source was not identified. On the other hand, the data was not considered to be value-adding - and thus not used - as the data was unable to help the project team forward in the process. The data was though perceived to be relevant for the purpose; a claim that was later deemed to be false, as the data was of a wrong topic. All in all, the data quality attributes predicted well the lack of the data usage, and the problems the data had.

Getting an access to the summary data enabled the project team to get their hands-on data that was of a right topic. The data was perceived to be initially relatively value-adding in describing the magnitude of the problem, and possibly helping with the focus. In the end, the data was, though, proven to be incomplete undermining all relevance from the data, and thus turning the data to be practically non-value-adding. In other words, the team's perception of the data quality guided the team to use the data to better understand the magnitude and focus - a direction which was later proven to be inaccurate.

The third phase began by getting an access to the clear and validated data on the perceived proxy measurement. The data quality was perceived to be excellent: it was

already widely used around the company, and had been tested to be accurate. Even in the end of the day, there were practically no concerns about the intrinsic or representational data qualities. The project team also trusted the contextual data quality based on the common acceptance for the idea of analyzing the billing delay as a good proxy measurement for the defects. Later on, the proxy relations were shown to be improbable based on the data analysis, and thus, the decisions made by the data would have guided the project to the wrong direction. In other words, using the data quality attributes as a thinking model did not reveal the actual data problems which existed in the fundamental question: is this dataset helpful for the question the team tries to solve.

The dataset used in the fourth phase of the project was perceived to be of a right topic and of a good quality. The initial perception of the data quality held until the end of the project: even the validation checks of the data showed no evidence against the data quality. Furthermore, the project team perceived that the data supported with qualitative analysis would be the right solution to drive action in the units. The data would be guiding the team to the most important improvement opportunities, and the qualitative understanding of the problems would then help to formulate robust solutions for those problems. In the end, the story behind solutions was compelling, but later was revealed to be insufficient to correct the problem. The data did not help the project team to understand the actual reasons in a complex process with several stakeholders: the focus areas were interesting, but did not lead to actionable insights. Similarly for the units, the response was a lot of interest but very little action to correct those problems. Rather, the units saw the data as a motivational tool which did not help them to find the root causes per se.

The fifth dataset was based on the manually gathered data on the actual defects and their reasons. Assessing the data quality dimensions, the intrinsic data quality was very poor: the data was known to be inaccurate and the same results could not have been produced again reliably. Furthermore, the data could not be gathered in a timely manner: as it was manually gathered, the data could not be analyzed from other periods of time than the actual period of the data gathering. On the other hand, the data was seen of a very high value by the project team and the unit level employees. In the meetings, the unit level

professionals often managed to create task lists or themselves already during the meeting. Furthermore, the units were later asking for the data meaning that they actually were interested in following up on how the improvements had been working. From the project team point-of-view, the dataset helped to choose the most important improvement opportunities, and to create a clear business case for the suggested improvements. All in all, despite the problems in the intrinsic data quality, the data was able to drive action in different levels of organization and led to actual results.

6.2 The model for the data quality attributes in improvement projects

The findings from the academic framework and the case study suggest that the data quality attributes are not of an equal importance for the improvement project teams. The majority of the data quality literature deems the intrinsic data qualities as the foundation of the data quality. In the seminal work of Wang & Strong (1996), value-added was proposed to be the second most important feature after the accuracy of the data. The findings from the case study provide evidence, however, for expanding the notion of value-added, and increasing its importance for the improvement project teams. Summarizing the findings from the case study and the academic research, a thinking model for the data quality in improvement projects is depicted below in the Figure 11.

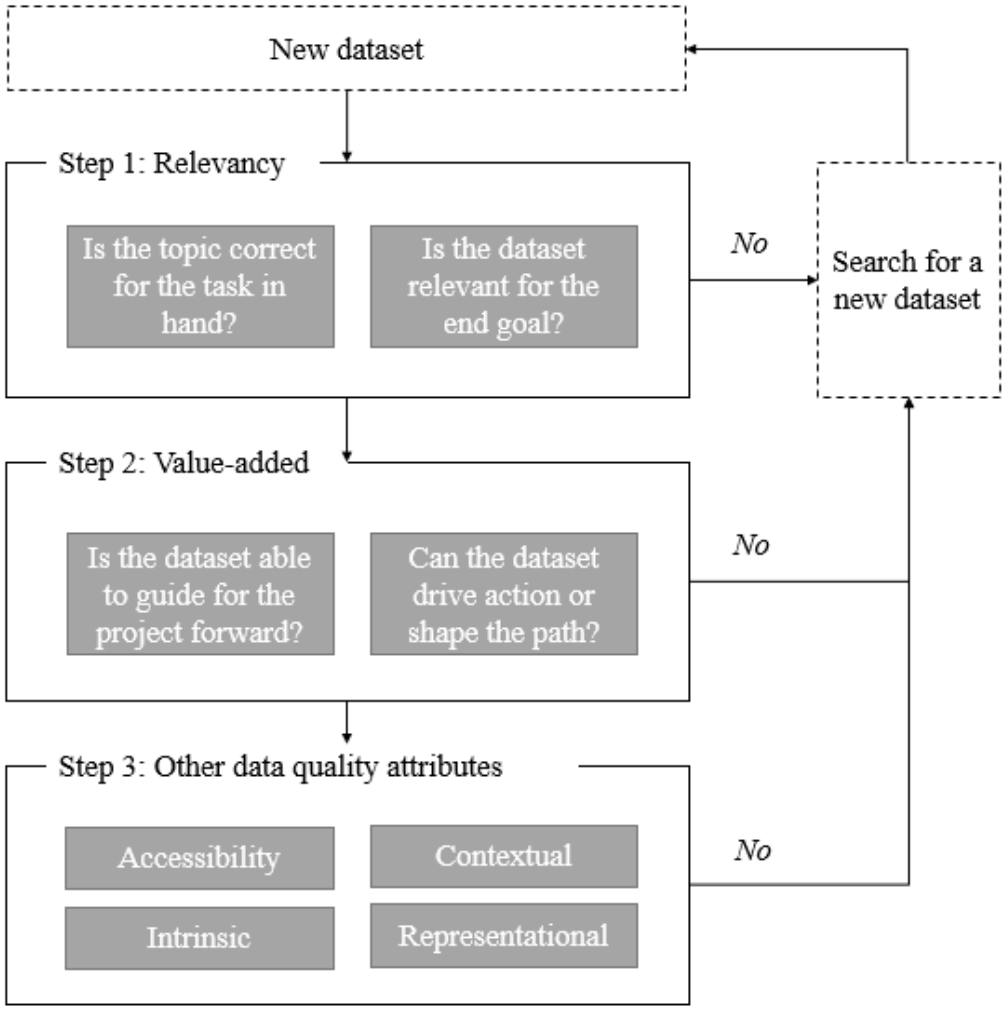

Figure 11: The model for data quality in the improvement project

The improvement projects very often have clear goal on what they are trying to achieve. The questions which need to be answered will change over time when insights become available from new data analysis. For assessing whether the data is of a good quality, it is necessary, though, to understand the end goal and use it as the starting point to assess the data quality. If the data is not relevant for the end goal, the other data qualities are irrelevant for the improvement project team. Another lens to understand the relevancy of the data is to understand whether the topic of the data matches is correct for the problem in hand. As a practical example, the project team used the data on invoicing delays as a tool to guide the project forward. The dataset turned out to be irrelevant for the question

in hand, thus being useless even if it would have driven action, and been otherwise of a good quality.

After confirming that the dataset is relevant for improving the end goal of the project, the project team should confirm that the dataset can provide value. The dataset is of a good value, when it can provide guidance for the project team on where to focus or where to continue the search for the root causes. On the other hand, the best datasets do not only provide summary statistics on the problem, but rather shed actual light on the most pressing problems worth solving. Thus, another lens to understand data quality should be to assess the dataset's ability to drive concrete action or shape the path to be able to focus on the critical few problems. If the dataset does not help the project team to push action or take the project to the next level, the other data quality attributes are of little importance. The importance of assessing value-added attributes first was present in the fifth phase, i.e. the gathering of the manual data: the dataset might have been of a below par quality, but it was the only dataset of the five that actually drove action among the persons who saw the data. Compared to the other datasets, the manually gathered data showed a relatively unambiguous path on how to improve the overall performance by focusing on the critical few solutions.

After the steps 1 and 2 are confirmed, only then the rest of the identified data quality attributes should be considered. Even if the dataset is relevant and value-adding, having too poor of an accuracy leads to unusable data. For the context of the improvement projects, the excellent quality of the data quality attributes in the category three is not sufficient to cover up for the lack of quality in either the category 1 or 2. As a practical example, the fourth dataset used in the case study, i.e. the detailed data about the invoicing claims, could have been deemed of a good quality based on the ordinary data quality dimensions. The reality was, however, that the dataset was unable to neither drive the units to action or shape the path for the improvement team to focus on the most pressing few improvement points.

7 Implications

Having described the conclusions from the case study, this section turns to the implications of the findings. The first chapter focuses on the implications for the management, i.e. what are the most important takeaways for the improvement teams, and organizations performing the improvement projects. The second chapter will, then, focus on the implications for the future research, and the limitations of making general conclusions from a single case study.

7.1 For management

The findings of the case study suggest that the improvement teams should have a different thinking model for the data quality attributes compared to the one used for assessing the information system quality in the academic literature. Rather than trying to find the data with the best intrinsic quality, the project teams should understand the whether the dataset is relevant for the end goal, and whether the dataset helps to gain actual progress in the project. The thinking model presented in the conclusions chapter could help the project teams in asking the correct questions to assess the actual value of the data for the project.

The case study also suggested that the perceived data quality does not always match the actual data quality. The literature has identified the phenomena by creating the linkages between the data quality attributes: e.g. low perceived accuracy might lead to low accessibility. From the improvement project point-of-view, the more important consequence is, however, that utilizing a bad quality dataset might lead to compelling solutions which do not have actual business value. The project teams should, thus, keep in mind that cross-validating the dataset can save a lot of lost time and resources if there is a risk of using irrelevant or non-value-adding data. Based on the case study, assumptions about the correlation between two factors, e.g. using proxy measurements, can easily lead to wrong decisions if they are not cross-validated with other datasets. Furthermore, the second phase of the case study showed how even good-looking dataset

can be leading to project team to falsely perceive the magnitude of the problem if it is incomplete.

The case study also supports the common and widely accepted claim that supporting decisions with data helps to get agreement around different people and thus, gain results. The improvement project teams should not, though, perceive the different types of datasets equally valuable in different phases of a project. A very high-level statistics might be very helpful in helping the project team to focus on the most important issues in the beginning of the project. The summary statistics, on the other hand, can be very helpful in guiding the structured inquiry about the possible problems and defects. Especially in complex processes, the data shedding light on the actual causes can, however, be vastly superior, as it is likely to be motivating and help to shape the path to achieve significant gains by focusing on the critical few data items.

As suggested by the thinking model for the improvement teams, the first and the important attribute of the data quality is suggested to be relevancy of the data for the topic in hand. If the dataset describes a topic which is only remotely linked to the end goal, it is very unlikely that the dataset could provide actual value for the improvement teams. In order to assess the relevancy of the dataset the project team must have specified the problem, and the end goal well enough: a feature often identified as an ingredient of a successful project. According to the case study, the caveat lies especially in using the proxy measurements: the project teams should seek to understand whether the correlation is real between the two factors.

To support achieving real change in processes, having accurate data on the right topic is sometimes far from enough. In the best case, the data should drive people towards action - it should provide a clear prioritization focus pointing to concrete problems. Furthermore, the data is likely to be at its best when it arouses strong emotions among the employees: e.g. in the case study, the unit-level employees had a strong will to improve their own processes, as they felt that the problem was real, and the solutions could reduce the hassle linked to their own work. The project teams should, thus, search

for a combination of data that motivates, and guides people to the most pressing problems, and if possible, a solid path on how to start to tackle the problems.

7.2 For research

A lot of research in the data quality stream have focused on the information system point-of-view: how the data quality should be improved in order to achieve better overall quality in the information systems. A different set of requirements arise, however, when the data users must assess the fit between the available data and needs of the improvement project. This paper suggests that the current data quality models need to be adjusted to be useful for the individual data users. The 3-step thinking model for organizing data quality proposed in the conclusions chapter serves as a preliminary step towards creating such a model. As this paper was based on a case study, more research needs to be, though, conducted on validating and building on the proposed model.

In order to give more guidance to the individual project managers, more literature linking the cognitive science to the data quality should be conducted. The first steps have been taken by Watts et al. (2009), who suggested that the contextual attributes should be assessed together with the objective attributes: e.g. if the data user perceives the data to be more ambiguous, they tend to take a more structured approach to analyzing data. While a lot of the required pieces of research have already been conducted in the fields of data quality, cognitive science, and change management, a shortage of research seems to lie arching over these three areas. In other words, too little research has been written on what kind of data qualities should the project teams seek for in order to drive actual change in units, and motivate employees to support the improvement initiative.

In addition to the cognitive linkage, the case study suggested that a mismatch seems to lie between the perceived and actual data quality. While the effects of the perceived data quality have been well studied for decades (e.g. Gerstberger & Allen 1968, Wang, Reddy & Kon 1995), the focus has been more on how the individual perceptions affect the usage of data. The case study pointed out, however, that another caveat lies in the

wrong perceptions of the data quality attributes, which could lead to decisions which could have either no or even negative effect on the output. More research on how the project teams can spot or validate the quality of the data a priori would, thus, help in the world of dirty data in messy processes.

All in all, benefits would be clear in focusing more data quality research on the pragmatic side of the operational improvement projects. In order to help the project teams, a data quality attribute "Ability to drive action" should be researched forward, as the attribute had a visible impact in the case study, but it is not included in the current data quality models. Based on the findings from this paper, the definition of the data quality attribute should include whether the data guides actionable insights or provides a clear path for continuing for the data consumers. As this paper was the first step towards linking data quality literature to the operational improvement projects, more research is needed to validate these results, including the definition of the proposed new data quality attribute. In the next section, more possible avenues for building on this work are considered.

8 Discussion

This last section focuses on discussing the results of the study complemented with the own experiences of the author. The chapter also presents thoughts on the applicability of these findings for a more general use.

This study found that the improvement teams should consider the relevancy of the dataset before other data quality attributes. The improvement teams often face a problem: digging deeper into the problems often surfaces several new avenues for potential improvement which are not included in the original scope. Same goes with the datasets; several datasets exist which can be very interesting for the company, but do not serve the goal of the improvement project. In a management consulting world, a similar phenomenon is called “boiling the ocean”: trying to analyze everything rather than focusing on the vital few analyses which actually make a difference.

A second step in the presented model was to consider that even if the dataset is of a right topic, the dataset might still not add immediate value for the improvement project. In healthcare, a lot of data is gathered, but the accessibility and the structure of the data is often subpar. Before delving into the data analysis, the improvement teams should do a moment of reflection: can this dataset add value to the goal we are trying to reach? Many times, the improvement project teams are open to all insights in beginning of the project, but tend to be more precise on their inquiry when the project goes forward. Thus, the requirements for the dataset change, in order to be value-adding for the improvement projects. A good guideline for the improvement teams would be the saying: stay as high level as you can as long as you can. Thus, try to find the datasets that match the current phase of your process of understanding the problems.

Sometimes, understanding whether a dataset is value-adding a priori can be notoriously difficult. A few guidelines exist, however, which can help the improvement teams to better assess the quality of the datasets. Firstly, the improvement teams would benefit from understanding the difference between the summary statistics, and the actual data showing the variance. Two key differences exist between the two based on my

experience working in healthcare improvement: The summary statistics talk often about the money or the process output, and hide the actual levers that can make a difference. Additionally, the summary statistics hide the variation between the best and worst performances, which are often the most interesting points of inquiry for the improvement teams.

Second way on how the improvement teams can assess the data quality beforehand is to think the usefulness of the dataset through two lenses. The first lens should be, which are my hypothesis on the problem, and what kind of data would I need to support or disprove my hypothesis. The hypothesis should reflect the questions of the project phase in hand: in the initial stage, they might focus on the differences in time or units, whereas in the later stages the focus should be on understanding the causal structures of individual defects. On the other hand, what kind of questions could I answer with my dataset, and how do they overlap with the questions posed by my initial hypothesis? If the two lenses do not match, the question arises whether the dataset is able to provide actionable insights for the current phase of the improvement project.

Finally, the improvement projects should focus on finding data that can drive change in the operational level. In the book Switch authored by Chip and Dan Heath, they describe that the change requires directing rider - the rational mind -, motivating the elephant - the emotional mind -, and shaping the path for the change. In order to achieve change, the book suggests that a simple path with few clear steps should be depicted in order to create action. For the improvement project managers, this implies that the dataset should aim to reveal the vital few problems that are both actionable and understandable for the operational level. If the improvement teams manage to find simple data showing a clear direction, achieving an actual change in the organization can be a lot more probable.

9 References

Agmon, N. and Ahituv, N., 1987. *Assessing data reliability in an information system*. Journal of Management Information Systems, 4(2), pp.34-44.

Alizamini, F.G., Pedram, M.M., Alishahi, M. and Badie, K., 2010, August. *Data quality improvement using fuzzy association rules*. In Electronics and Information Engineering (ICEIE), 2010 International Conference On (Vol. 1, pp. V1-468). IEEE.

Allen, T.J., 1966. *Performance of information channels in the transfer of technology*. IMR; Industrial Management Review (pre-1986), 8(1), p.87.

Bailey, J.E. and Pearson, S.W., 1983. *Development of a tool for measuring and analyzing computer user satisfaction*. Management science, 29(5), pp.530-545.

Ballou, D.P. and Pazer, H.L., 1985. *Modeling data and process quality in multi-input, multi-output information systems*. Management science, 31(2), pp.150-162.

Ballou, D.P. and Pazer, H.L., 1995. *Designing information systems to optimize the accuracy-timeliness tradeoff*. Information Systems Research, 6(1), pp.51-72.

Barone, D., Stella, F. and Batini, C., 2010, June. *Dependency discovery in data quality*. In International Conference on Advanced Information Systems Engineering (pp. 53-67). Springer Berlin Heidelberg.

Batini, C., Cappiello, C., Francalanci, C. and Maurino, A., 2009. *Methodologies for data quality assessment and improvement*. ACM computing surveys (CSUR), 41(3), p.16.

Batini, C. and Scannapieco, M., 2016. *Data Quality Dimensions*. In Data and Information Quality (pp. 21-51). Springer International Publishing.

Benbasat, I. and Dexter, A.S., 1985. *An experimental evaluation of graphical and color-enhanced information presentation*. Management science, 31(11), pp.1348-1364.

Bodnar, G., 1975. *Reliability modeling of internal control systems*. The Accounting Review, 50(4), pp.747-757.

Bovee, M., Srivastava, R.P. and Mak, B., 2003. *A conceptual framework and belief-function approach to assessing overall information quality*. International journal of intelligent systems, 18(1), pp.51-74.

Brydon-Miller, M., Greenwood, D. and Maguire, P., 2003. *Why action research?*. Action research, 1(1), pp.9-28.

Buhrmester, M., Kwang, T. and Gosling, S.D., 2011. *Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?*. Perspectives on psychological science, 6(1), pp.3-5.

Coghlan, D. and Brannick, T., 2005. *Doing action research in your own organization*. Sage.

Culnan, M.J., 1983. *Environmental scanning: The effects of task complexity and source accessibility on information gathering behavior*. Decision Sciences, 14(2), pp.194-206.

Culnan, M.J., 1984. *The dimensions of accessibility to online information: Implications for implementing office information systems*. ACM Transactions on Information Systems (TOIS), 2(2), pp.141-150.

Curry, E., O'Donnell, J., Corry, E., Hasan, S., Keane, M. and O'Riain, S., 2013. *Linking building data in the cloud: Integrating cross-domain building data using linked data*. Advanced Engineering Informatics, 27(2), pp.206-219.

Cushing, B.E., 1974. *A mathematical approach to the analysis and design of internal control systems*. The Accounting Review, 49(1), pp.24-41.

DeLone, W.H. and McLean, E.R., 1992. *Information systems success: The quest for the dependent variable*. Information systems research, 3(1), pp.60-95.

Doll, W.J. and Torkzadeh, G., 1988. *The measurement of end-user computing satisfaction*. MIS quarterly, pp.259-274.

Eisenhardt, K.M., 1989. *Building theories from case study research*. Academy of management review, 14(4), pp.532-550.

Eisenhardt, K.M. and Graebner, M.E., 2007. *Theory building from cases: Opportunities and challenges*. Academy of management journal, 50(1), pp.25-32.

Eppler, M. and Helfert, M., 2004, November. *A classification and analysis of data quality costs*. In International Conference on Information Quality (pp. 311-325).

Ge, M. and Helfert, M., 2013. *Impact of information quality on supply chain decisions*. Journal of Computer Information Systems, 53(4), pp.59-67.

George, M.L. and George, M., 2003. *Lean six sigma for service* (p. 273). New York, NY: McGraw-Hill.

Gerstberger, P.G. and Allen, T.J., 1968. *Criteria used by research and development engineers in the selection of an information source*. Journal of applied psychology, 52(4), p.272.

Gorla, N., Somers, T.M. and Wong, B., 2010. *Organizational impact of system quality, information quality, and service quality*. The Journal of Strategic Information Systems, 19(3), pp.207-228.

Haug, A., Zachariassen, F. and Van Liempd, D., 2011. *The costs of poor data quality*. Journal of Industrial Engineering and Management, 4(2), pp.168-193.

Hsueh, P.Y., Melville, P. and Sindhwani, V., 2009, June. *Data quality from crowdsourcing: a study of annotation selection criteria*. In Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing (pp. 27-35). Association for Computational Linguistics.

Ives, B., Olson, M.H. and Baroudi, J.J., 1983. *The measurement of user information satisfaction*. Communications of the ACM, 26(10), pp.785-793.

Johnson, J.R., Leitch, R.A. and Neter, J., 1981. *Characteristics of errors in accounts receivable and inventory audits*. Accounting Review, pp.270-293.

Järvenpää, S.L., 1989. *The effect of task demands and graphical format on information processing strategies*. Management Science, 35(3), pp.285-303.

Kahn, B.K., Strong, D.M. and Wang, R.Y., 2002. *Information quality benchmarks: product and service performance*. Communications of the ACM, 45(4), pp.184-192.

Ketokivi, M. and Choi, T., 2014. *Renaissance of case research as a scientific method*. Journal of Operations Management, 32(5), pp.232-240.

Kim, W. and Choi, B., 2003. *Towards Quantifying Data Quality Costs*. Journal of Object Technology, 2(4), pp.69-76.

Klein, B.D., 2001. *User perceptions of data quality: Internet and traditional text sources*. Journal of Computer Information Systems, 41(4), pp.9-15.

Knechel, W.R., 1985. *A simulation-model for evaluating accounting system reliability*. Auditing-A journal of Practice & Theory, 4(2), pp.38-62.

Koch, T. and Kralik, D., 2009. *Participatory action research in health care*. John Wiley & Sons.

Laudon, K.C., 1986. *Data quality and due process in large interorganizational record systems*. Communications of the ACM, 29(1), pp.4-11.

Lee, Y.W., Strong, D.M., Kahn, B.K. and Wang, R.Y., 2002. *AIMQ: a methodology for information quality assessment*. Information & management, 40(2), pp.133-146.

Li, G., Ooi, B.C., Feng, J., Wang, J. and Zhou, L., 2008, June. *EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data*. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 903-914). ACM.

Madnick, S.E., 1995. *Integrating information from global systems: Dealing with the "on-and off-ramps"; of the information superhighway*. Journal of Organizational Computing and Electronic Commerce, 5(2), pp.69-82.

Madnick, S.E., Wang, R.Y., Lee, Y.W. and Zhu, H., 2009. *Overview and framework for data and information quality research*. Journal of Data and Information Quality (JDIQ), 1(1), p.2.

Marsh, R., 2005. *Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management*. Journal of Database Marketing & Customer Strategy Management, 12(2), pp.105-112.

Miller, H., 1996. *The multiple dimensions of information quality*. Information Systems Management, 13(2), pp.79-82.

Miller, J. and Doyle, B.A., 1987. *Measuring the effectiveness of computer-based information systems in the financial services sector*. MIS quarterly, pp.107-124.

Morey, R.C., 1982. *Estimating and improving the quality of information in a MIS*. Communications of the ACM, 25(5), pp.337-342.

Nelson, R.R., Todd, P.A. and Wixom, B.H., 2005. *Antecedents of information and system quality: an empirical examination within the context of data warehousing*. Journal of management information systems, 21(4), pp.199-235.

Panahy, P.H.S., Sidi, F., Affendey, L.S., Jabar, M.A., Ibrahim, H. and Mustapha, A., 2013. *A framework to construct data quality dimensions relationships*. Indian Journal of Science and Technology, 6(5), pp.4422-4431.

Peralta, V., 2008. *Data quality evaluation in data integration systems*. Doctoral dissertation, Universidad de la República, Uruguay.

Pipino, L.L., Lee, Y.W. and Wang, R.Y., 2002. *Data quality assessment*. Communications of the ACM, 45(4), pp.211-218.

Reason, P. and Bradbury, H. eds., 2001. *Handbook of action research: Participative inquiry and practice*. Sage.

Redman, T.C., 1998. *The impact of poor data quality on the typical enterprise.* Communications of the ACM, 41(2), pp.79-82.

Rosenberg, V., 1966. *The application of psychometric techniques to determine the attitudes of individuals toward information seeking and the effect of the individual's organizational status on these attitudes*. Lehigh University of Bethlehem PA Center for Information Science.

Santos, B.L.D. and Bariff, M.L., 1988. *A study of user interface aids for model-oriented decision support systems*. Management Science, 34(4), pp.461-468.

Sonntag, D., 2004. *Assessing the Quality of Natural Language Text Data*. In GI Jahrestagung (1) (pp. 259-263).

Strong, D.M., Lee, Y.W. and Wang, R.Y., 1997. *Data quality in context*. Communications of the ACM, 40(5), pp.103-110.

Stvilia, B., Gasser, L., Twidale, M.B. and Smith, L.C., 2007. *A framework for information quality assessment.* Journal of the Association for Information Science and Technology, 58(12), pp.1720-1733.

Tan, J.K. and Benbasat, I., 1990. *Processing of graphical information: A decomposition taxonomy to match data extraction tasks and graphical representations*. Information Systems Research, pp.416-439.

Tayi, G.K. and Ballou, D.P., 1998. *Examining data quality*. Communications of the ACM, 41(2), pp.54-57.

Tee, S.W., Bowen, P.L., Doyle, P. and Rohde, F.H., 2007. *Factors influencing organizations to improve data quality in their information systems*. Accounting & Finance, 47(2), pp.335-355.

Truong, H.L. and Dustdar, S., 2009, December. *On analyzing and specifying concerns for data as a service*. In Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific (pp. 87-94). IEEE.

Wand, Y. and Wang, R.Y., 1996. *Anchoring data quality dimensions in ontological foundations*. Communications of the ACM, 39(11), pp.86-95.

Wang, K.Q., Tong, S.R., Roucoules, L. and Eynard, B., 2008, September. *Analysis of data quality and information quality problems in digital manufacturing*. In Management of Innovation and Technology, 2008. ICMIT 2008. 4th IEEE International Conference on (pp. 439-443). IEEE.

Wang, R.Y., Kon, H.B. and Madnick, S.E., 1993, April. *Data quality requirements analysis and modeling*. In Data Engineering, 1993. Proceedings. Ninth International Conference on (pp. 670-677). IEEE.

Wang, R.Y., Reddy, M.P. and Kon, H.B., 1995. *Toward quality data: An attribute-based approach*. Decision Support Systems, 13(3), pp.349-372.

Wang, R.Y. and Strong, D.M., 1996. *Beyond accuracy: What data quality means to data consumers*. Journal of management information systems, 12(4), pp.5-33.

Watts, S., Shankaranarayanan, G. and Even, A., 2009. *Data quality assessment in context: A cognitive perspective*. Decision Support Systems, 48(1), pp.202-211.

Yin, R.K., 2013. *Case study research: Design and methods*. Sage publications.

Yu, S. and Neter, J., 1973. *A stochastic model of the internal control system*. Journal of Accounting Research, pp.273-295.

Appendices

Appendix A: List of the data quality attributes

Category	Data quality dimension	Description	Articles
Intrinsic	Accuracy	Are the data items error-free, correct and flawless?	Wang & Strong 1996, Knechel 1985, Peralta 2008
	Believability	Do the data users see data as credible and true?	Wang & Strong 1996, Kahn et al. 2002, Wang & al. 1995
	Objectivity	Is the data unbiased and without prejudice?	Wang & Strong 1996, Strong et al. 1997, Kahn et al. 2002
	Reliability	Is the data accuracy sustained over time?	Ives et al. 1983, Wang & Strong 1996, Nelson et al. 2005
	Reputation	Is the data trusted and kept in high regard by the data users?	Wang & Strong 1996, Strong et al. 1997, Stvilia et al. 2007
Contextual	Appropriate amount of data	Is the amount of data appropriate to conduct proper analysis?	Wang & Strong 1996, Strong et al. 1997, Srirastava & Mak 2003
	Completeness	Are all relevant values stored in the data?	Ballou & Pazer 1985, Tayi & Ballou 1998, Nelson et al. 2005
	Relevancy	Is the data relevant for the data consumer?	Miller 1996, Wang & Strong 1996, Bovee et al. 2003
	Timeliness	Is the age of the data appropriate for the task in hand?	Ballou & Pazer 1985, Wang et al. 1995, Wang & Strong 1996,
	Value-added	How advantageous the data is for the data consumers task in hand?	Wang & Strong 1996, Strong et al. 1997, Kahn et al. 2002
Accessibility	Access security	Does the data security cause a barrier for the data accessibility?	Miller 1996, Wang & Strong 1996, Kahn et al. 2002
	Ease of access	Is the data fast and easy to access?	Wang & Strong 1996, Strong et al. 1997, Nelson et al. 2005

Category	Data quality dimension	Description	Articles
Representational	Concise representation	Is the data represented compactly and briefly?	Wang & Strong 1996, Pipino et al. 2002, Kahn et al. 2002
	Consistency	Are all representations of the data values the same?	Ballou & Pazer 1985, Miller 1996, Tayi & Ballou 1998,
	Ease of understanding	Can the data be understood easily and unambiguously?	Wang & Strong 1996, Wand & Wang 1996, Nelson et al. 2005
	Interpretability	Do the data consumers interpret the data correctly without errors?	Agmon & Ahituv 1987, Bovee et al. 2003, Stvilia et al. 2007

Table 6: Summary of the data quality attributes

Appendix B: List of meetings used as a reference material

Date	Topic	Persons involved
4.3.2016	Steering Committee – kickoff	4
10.3.2016	Previously used data - with controllers	3
21.3.2016	Previously used data - former project manager	2
21.3.2016	Kickoff - invoicing function	2
23.3.2016	Kickoff – controllers	3
28.3.2016	Meeting - initial data analysis	3
31.3.2016	Building access to the new data set	3
31.3.2016	Kickoff - insurance specialist	2
11.4.2016	Kickoff - project team	8
20.4.2016	Steering committee	3
25.4.2016	Meeting – insurance	2
27.4.2016	Meeting – invoicing	2
28.4.2016	Meeting - unit-level specialists	4
9.5.2016	Workshop - unit-level specialists	4
10.5.2016	Decision gate	10
11.5.2016	Meeting – insurance	2
12.5.2016	Meeting - unit-level specialists	3
16.5.2016	Meeting - customer service	3
17.5.2016	Meeting – invoicing	3
23.5.2016	Meeting - data access with controllers	4
24.5.2016	Meeting - unit-level specialist	2
24.5.2016	Workshop - project team	6
28.5.2016	Meeting – invoicing	2
31.5.2016	Team meeting	7
31.5.2016	Unit mgmt meeting	9
1.6.2016	Meeting – invoicing	3
1.6.2016	Unit meeting - solution/implementation	4
3.6.2016	Meeting – invoicing	2
13.6.2016	Meeting – solutions	3
15.6.2016	Meeting – solutions	3
16.6.2016	Meeting – invoicing	2
21.6.2016	Solution workshop - unit-level specialists	5
30.6.2016	Meeting – solutions	2
4.8.2016	Meeting - direction of the project	3
10.8.2016	Meeting – solutions	3
16.8.2016	Steering committee	3
16.8.2016	Meeting – solutions	2
19.8.2016	Meeting – solutions	3
6.9.2016	Decision gate	9
14.9.2016	Meeting – implementation	2
23.9.2016	Meeting – implementation	2
3.10.2016	Steering committee	3

Table 7: The meetings used as material for the thesis