UMEÅ UNIVERSITY

DATA MINING IN PRACTICE

An application of the CRISP-DM framework in healthcare

Emma Lind and Sofi Glas

Department of informatics
Magister thesis, 15 hp
Master’s Programme in IT Management
SPM 2022.13

Abstract

With extensive data available in today's organizations, it has become increasingly important to secure valuable insights through data. As a result, the management of data to support decision-making processes is receiving increasing attention in organizations' IT strategies. The healthcare sector is no exception. However, there is an urgent need for tools that help organizations extract valuable insights from the rapidly growing volumes of data, one of the most important steps of which is data mining. So far, the healthcare sector has not found a way to harness its full potential, due to limited methods to extract useful knowledge hidden in large data sets. Knowledge gained from data mining can help healthcare to better serve patients, but there is a limited comprehensive picture of applications regarding data mining processes in healthcare. Against this background, the purpose of this study is to investigate practical dimensions of the data mining process in healthcare and further identify barriers that can inhibit this process. To answer our research question, we used a qualitative case study with semi structured interviews based on the CRISP-DM framework. Our findings indicate barriers that can inhibit the data mining process, which are related to the objectives, data availability and final reports.

Keywords: Data governance, Data mining methodology, The CRISP-DM Framework, Data mining in healthcare

1. Introduction

It has become increasingly important to create value with data in organizations, which is a result of the vast amount of data available today (Provost and Fawcett, 2013). Capturing value from data requires mastering the fundamentals of data management and is not at job for a few IT functions alone, it must be an overarching direction for the entire organization (Lee, Madnick, Wang, Wang & Zhang, 2014). Data science is an area that has contributed to improved results when it comes to data management (Provost and Fawcett, 2013) where the concept of data governance is central. Data governance is a data science initiative that creates conditions for organizations to succeed in maintaining opportunities from data but also to structure organizations to solve data-related problems (Bendfeldt, Persson & Madsen, 2018). There is an assumed direct relationship between data and data governance as an organizational capability. Therefore, there is an effort on the research front to better understand the practical dimension of how organizations organize their data management (Alhassan, Sammon & Daly, 2016). By comparing different results and frameworks for data management, research in this area can be enhanced (Alofaysan, Alhaqbani, Alseghayyir & Omar, 2016).

The public sector faces several challenges when it comes to utilizing its data resources with data management, where three main factors have been identified. Difficulties to communicate the value created by data and data management, local methods that make it difficult to design and implement standardized methods for data and as well as different levels of data capability, which makes it difficult to see a strategic direction for the use of data (Bendfeldt et al., 2018). Challenges in creating value with data within the public sector are apparent, where healthcare is no exception (Bendfeldt et al., 2018; Witjas-Paalberends, van Laarhoven, van de Burgwal, Feilzer, Swart, Claassen & Jansen, 2018)

Healthcare systems hold enormous amounts of data (Bhavnani and Sitapati, 2019) and due to the digital age, healthcare and technology have become intertwined which has given rise to data-related applications (Subrahmanya, 2021). The vast amount of data creates opportunities to improve the quality and efficiency of healthcare, if its value can be exploited (Teede et al., 2019). Hence, increased efforts to enhance the quality of care have contributed to a growing interest in quality improvement for data mining (Chae, Kim, Tark, Park, & Ho, 2003). By applying data governance disciplines in healthcare, valuable starting point for data-driven projects such as improving data quality, data mining, healthcare analytics and strategic decision-making efficiency can be improved (Alofaysan et al., 2016). Still, there is limited research on how different data mining approaches are used in different application areas (Plotnikova et al., 2020). Knowledge gained through data mining can help healthcare to better serve patients (Yoo, Alafaireet, Marinov, Pena-Hernandez, Gopidi, Chang & Hua, 2011) but there is a limitation in terms of a specific and detailed framework for conducting data mining analysis in healthcare (Niaksu, 2015). In addition, the healthcare data are underutilized due to limited methods to extract useful knowledge hidden in large data sets, (Jothi, Rashid & Husain, 2015; Teede, Johnson, Buttery, Jones, Boyle, Jennings & Shaw, T., 2019).

Against this background, the purpose of this study is to create an understanding of the data mining process in healthcare, through the CRISP-DM framework. Furthermore, we hope that this study will contribute with an understanding of how practical insights may inhibit the data mining process in healthcare. To analyze the data mining process in healthcare, we have developed the following question: ***What barriers can be identified in the data mining process in healthcare, by using the CRISP-DM framework?*** To answer our research question, we have used a qualitative case study with semi-structured interviews. The interview questions were answered by a number of respondents with different roles at an IT department in healthcare. In the analysis of the empirical data, we have used a theoretical framework based on the CRISP-DM Framework. The CRISP-DM framework is an analysis model which describes a step-by-step process that sheds light on relevant phases to structure the data mining process (Schröer, Kruse & Gómez, 2021).

2. Related research

The section of related research will contain research mainly on data governance and data mining as a process to create value with data. This body of research will later help us investigate and analyze our findings from the study conducted at an IT department in healthcare.

2.1 Data Governance

The use of data science in business infrastructure has improved the ability to manage data throughout the enterprise. Data science can be described as a set of core principles that support and guide the extraction of information and knowledge from data (Provost and Fawcett, 2013).

The use of open data, big data and predictive data analytics has long offered the promise of transforming entire industries and societies (Bendfeldt et al., 2018).
With the vast amount of data now available in organizations, it has become more important than ever to ensure that data can create opportunities, such as increasing organizational profit, customized marketing, and customer service (Provost and Fawcett, 2013). Hence, organizations are introducing data science initiatives to support predictive maintenance for increasing better decision-making (Bendfeldt et al., 2018).

One data science initiative is Data Governance and has emerged as a promising approach for maintaining and gaining opportunities from data and transforming organizations by solving organizational data issues (Bendfeldt et al., 2018). Extant research does not provide a common definition of what data governance main objectives are but agrees that data will create opportunities for organization (Parmiggiani and Grisot, 2020). Studies have addressed the importance of data quality and emphasize the improvement of quality as a key objective (Begg and Caira, 2011; Alofaysan et al., 2016). Although quality is important, it is only one objective of effective data governance, which other studies show must be driven by and aligned with business priorities and objectives (Lee et al., 2014; Bendfeldt, 2017; Brous, Janssen & Vilminko-Heikkinen, 2016). Data governance has to involve clearly defined authority to create and maintain data policies and procedures (Brous et al., 2016). Management plays a crucial part in the control of planning, supervision, enforcement, and availability over the data assets and sets the direction for the organization's data practices overall (Kahtri and Brown, 2010; Brous, Janssen & Krans, 2020). Other studies stress the importance of a common understanding and argue that to be able to govern data appropriately it has to be possible to understand what the data to be managed means, and why it is important to the organization (Brous et al., 2016, Bendfeldt et al., 2018).

A study at a Danish Local Government, explores why governing data can be difficult (Bendfeldt et al., 2018) and finds that it is a challenge for public sector organizations to explore and leverage their data assets and concludes three main reasons. Struggling to communicate the value that data and data governance might be able to create, local practices tend to complicate the design and implementation of standardized approaches to data. Third, different levels of data capability among departments and managers makes it difficult to see a strategic direction for the use of data in the organization (Bendfeldt et al., 2018).

Data governance has also been studied in the healthcare sector (Alofaysan et al., 2016) and shows that data governance disciplines in healthcare provides a valuable starting point for data-driven projects such as data warehousing, data quality improvement, data mining, healthcare analytics, business governance, strategic decisions effectiveness and business intelligence (Alofaysan et al., 2016). However, even though these disciplines are valuable, it has only been investigated as a starting point for achieving opportunities with data. Studying the outcomes of data governance is an essential piece of future work (Alofaysan et al., 2016). While extant research on data governance is rich and extensive, they assumed a direct relationship between data and data governance as an organizational capability and scant attention has been paid towards the practical dimension regarding the everyday practices such as the data mining process (Parmiggiani and Grisot 2020; Mikalef, Pappas, Krogstie & Pavlou, 2020).

As this is an aspect of data governance that has yet to be researched it stands as a further motivation for this thesis. By engaging with the users who actually work with the data mining process, a strategic value can be made for the organizations (Alhassan et al., 2016; Mikalef et al., 2020). Data mining in a specific context remains problematic for organizations and there is limited research helping organizations to establish a well working data mining process (Mikalef et al., 2020; Parmiggiani and Grisot, 2020).

2.2 Data mining in practice: Evidence from healthcare

Management of large amounts of data to support decision-making processes is receiving increasing attention in organizations' IT strategies (Schröer et., 2021). Consequently, there is an urgent need for new tools that help organizations extract valuable information from the rapidly growing volumes of data (Jothi et al., 2015), where one of the key aspects is data mining (Jothi et al., 2015). Data mining is the extraction of knowledge from data using different techniques (Provost and Fawcett, 2013).

There is a recurring focus on organizations embed data mining solutions into knowledge-based decision-making processes in order to support rapid and efficient knowledge discovery (Bohanec, Robnik-Sikonja & Borstnar, 2017). Data mining methods involve a set of guidelines to perform several tasks in order to achieve the goals of the data mining project itself (Mariscal and Marbán et al., 2010). These guidelines thus need to be adapted to different organizational contexts and business objectives (Plotnikova, Dumas & Milani, 2020). For that, it requires effective strategies to yield valuable insights for the organization (Subrahmanya, 2021). Hence, there is limited research on how different approaches are used in different application areas (Plotnikova et al., 2020).

The healthcare industry is one of the world's largest, most critical and fastest growing industries, which has recently undergone major challenges (Nambiar, Sethi, Bhardwaj & Vargheese, 2013). Health systems contain enormous amounts of data (Bhavnani and Sitapati, 2019), which creates conditions to improve the quality and efficiency of health care (Teede et al., 2019). However, the challenges lie in the fact that data in healthcare are underutilized due to limited methods to extract useful knowledge hidden in large data sets (Jothi et al., 2015; Teede et al., 2019).

Further common challenges are associated with poor data quality and missing values (Yoo, et., 2011; Niaksu, 2015; Witjas-Paalberends et al., 2018). To improve the quality of data it is important to understand the key aspects of data-driven value creation (Galetsi et al., 2020), of which Healthcare is behind (Witjas-Paalberends et al., 2018). This can be further explained by the fact that much of the medical data is historical paper-based patient data, which in turn results in medical data that is often incomplete in terms of electronic accessibility (Niaksu, 2015). Most data-driven decisions in healthcare ultimately affect patients, which probably explains why data quality is perceived as an important challenge (Witjas-Paalberends et al., 2018). Data-driven decision-making opens new possibilities to enhance healthcare quality (Subrahmanya, 2021).

Hence, the healthcare sector has not yet found a way to harness its full potential, which is a result of the fact that healthcare must consider challenges related to accountability and complex healthcare systems (Witjas-Paalberends et al., 2018). Knowledge gained through data mining can help healthcare to better serve patients (Yoo et al., 2011), but application of data mining in healthcare has limited standards (Niaksu, 2015) and are missing a comprehensive picture of the process of knowledge discovery (Islam, Hasan, Wang, Germack & Noor-E-alam, 2018).

Analytics thinking provides techniques to extract information from healthcare's complex and voluminous data and furthermore transform it into valuable information and facilitate evidence-based decision-making for planning, management and learning (Islam et al., 2018). This philosophy is moreover used by practitioners in many industries (Jaggia, Kelly, Lertwachara & Chen, 2020) but current analytics teaching tends to focus on the modeling phase to a greater extent and limited comprehensive understanding of the whole analysis process (Jaggia et al., 2020). However, there is no specific framework for conducting data mining analysis in health care (Niaksu, 2015). End-to-end data mining strategies, such as the CRISP-DM framework, are a data mining analysis model that represent the whole analysis process (Jaggia et al., 2020).

3. Theoretical framework

For this study, we have applied the CRISP-DM framework to best explore the area of concern presented earlier. By using this analysis model for data mining, we have investigated the data mining process and moreover barriers that can inhibit this process at an IT department in healthcare. We begin with defining the CRISP-DM framework and later explain how this framework is applied in this study.

3.1 Cross-industry standard process for data mining

The framework of the Cross Industry Standard Process for Data Mining or CRISP-DM is the most relevant and used data mining method (Niaksu, 2015; Abbasi, Sarker & Chiang, 2016). This analysis model offers organizations the structure needed to achieve better and faster results from data mining (Shearer, 2000).

The CRISP-DM framework emerged at the end of 1996 and is based on previous attempts to define methods for knowledge discovery (Shearer, 2000; Yoo et al., 2011). Initially, the CRISP-DM framework was designed to provide guidance to beginners in data mining that can be adapted to the needs of a particular industry. Therefore, the framework is based on practical experience of data mining, where professionals have provided input and ideas for developing the process. The CRISP-DM framework is an organizational process model and is not limited to one particular technology and therefore several technologies can support the process (Schröer et al., 2021). Hence, the iterative process is also a distinguishing feature which makes the CRISP-DM framework unique (Plotnikova et al., 2020). For this reason, the analysis model has become a success and it is said to be the "de facto" standard for data mining (Shearer, 2000; Azevedo and Santos, 2008).

The CRISP-DM framework consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. These phases help to understand the data mining process by guiding organizations that plan a data mining project (Shearer, 2000; Rivo, de la Fuente, Rivo, García-Fontán, Cañizares, & Gil, 2012; Schröer et al., 2021).

3.2 Application of the CRISP-DM framework

CRISP-DM is a dynamic framework based on a hierarchical and iterative process model that can be used from a specific to an overall approach (Niaksu, 2015). In our study, we used the CRISP-DM framework from an overall approach. This is motivated by the fact that we intend to investigate the organization's data mining process at a comprehensive level and not according to a specific objective in relation to every specific phase. Based on this approach, we use the following three phases from the CRISP-DM framework; *business understanding*, which emphasizes the importance of understanding the problem to be investigated, *data understanding* which emphasizes the importance of understanding the data and owning the right data and *deployment,* which emphasizes the importance of reviewing the final reports (Provost and Fawcett, 2013) *(see figure 1).* These phases constitute our theoretical framework in this study which are described in more detail below. By using this theoretical framework, we have investigated the data mining process and further identify barriers related to these phases.

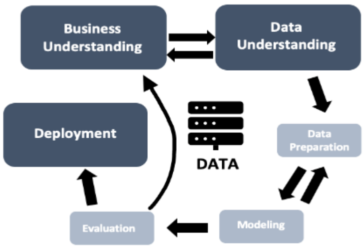


Figure 1. The CRISP-DM framework with emphasis on the theoretical framework.

The *Business Understanding* phase emphasizes the understanding of the objective to be investigated (Provost and Fawcett, 2013; Niaksu, 2015). To understand the objective, the organization needs to transform this knowledge into a definition of data mining problems and then develop a preliminary plan designed to achieve the goals. To understand the data to be analyzed it is important for data mining practitioners to fully understand the context for which they are finding a solution (Shearer, 2000). Setting the goal of data recovery is one of the most important aspects of this phase (Shearer, 2000; Schröer et al., 2021). In addition, it is also important to have an overview of available and necessary resources. Furthermore, the business understanding phase can be described in different ways, as the CRISP-DM framework can be applied in different areas. Each objective is unique and consists of its own combination of goals and limitations. Data scientists divide a business problem into sub-tasks, the solutions to the sub-tasks can then be compiled to solve the overall problem (Provost and Fawcett, 2013). In the business understanding phase, analysts' creativity plays a major role, where the key to great success is iterative process rewording the objectives. A challenge related to this phase is to link the results of the data mining to the objectives (Provost and Fawcett, 2013).

It is important to spend as much time as possible in the phase of understanding the business and the data, to create a concrete definition of the objective and achieve it (Provost and Fawcett, 2013).

The second phase is *Data understanding*. This involves familiarizing with existing data, identifying problems with data quality, creating an initial insight into the data, and forming hypotheses from the data (Niaksu, 2015; Plotnikova et al., 2020). In this phase, it is important to form an idea of strengths and limitations related to the data, as there is rarely an exact fit to the problem (Provost and Fawcett, 2013). Therefore, is it important to discover the data that is available to meet the primary business objectives (Shearer, 2000). The data available has to contain the necessary information to create the right prerequisites for the data mining process forward (Provost and Fawcett, 2013). If the data quality is poor, the users must improve it. This covers an important aspect that often proceeds along with data understanding, the preparation of data where the data is manipulated and transformed into formats that produce better quality (Provost and Fawcett, 2013). This conversion of data is necessary because data usually has different formats and structures depending on where it is been extracted.

The third phase is *Deployment* and includes the results presented in the form of a report and hence the main findings of the data mining process (Schröer et al., 2021). The report should be shared with decision makers so that they can use the report as a basis for decision making (Provost and Fawcett, 2013; Plotnikova et al., 2020). Although the purpose of the CRISP-DM framework is to increase knowledge of the data, the knowledge obtained must be organized, presented, and distributed in such a way that the end user can use it (Plotnikova et al., 2020). This phase also stresses the importance of communicating the results of the analysis to the intended audience (Jaggia et al., 2020). Depending on the requirements, the installation phase may involve both creating a report but also repeating the data mining process (Niaksu, 2015; Plotnikova et al., 2020). This phase highlights the importance of the end-user being able to set the conditions for the actions that need to be taken to create value with the CRISP-DM framework (Niaksu, 2015).

4. Method

The purpose of this study is to create an understanding of the data mining process in healthcare. For that, we consider it appropriate to use a qualitative method with an interpretive approach to carry out a case study with semi-structured interviews. The motivation for using an interpretative approach is based on our research investigating the perceptions of real people in their real work organization. While conducting research it is of the greatest importance to strive for transparency. The following section aims to do just that, describing our decisions and arguing for the choices. We begin with how we chose to design the study, followed by how we gathered and analyzed the data, how we approached the ethical parts, and lastly methodological discussion.

4.1 Research strategy

To evaluate the reliability and validity of our research, we have conducted a methodological review of previous literature, as this is an essential feature for any academic research (Webster and Watson, 2002).

For the first selection of past research, we used the Google Scholar, Umu UB and AIS database and the selected literature had phrases similar to "data governance", "CRISP-DM" and "data mining" in the title, abstract or keyword. From the large number of publications that met this criterion, we screened the articles briefly to identify articles with digital governance, data mining or CRISP-DM as a major topic of study. Later, we analyzed the selected literature to identify key concepts that were characteristic of their content. Furthermore, these concepts were categorized to identify research streams, inspired by Henfridsson and Bygstad (2013), in the selected literature used. The identified research streams were labeled Data Governance, Data Mining in Healthcare and the CRISP-DM Framework which represent our primary theoretical orientation for this study. The chosen literature shown are those considered most relevant and useful for fulfilling the study's purpose and answering the research question (see appendix 1).

4.2 Case study

The choice to conduct a case study in the health care sector is motivated by the fact that we want to investigate the organization's data mining process based on several current respondents at an IT department. The starting point of a case study is the respondents and their description of a certain phenomenon, which is the interesting part of the study (Sohlberg and Sohlberg, 2019). Furthermore, a case study is a strategy for qualitative empirical research that allows for an in-depth investigation of a contemporary phenomenon (De Massis and Kotlar, 2014) from real people in contemporary real organizations (Myers, 2020). Hence, these respondents represent our empirical base as well as the context of the study's findings.

4.3 Good research practice

This study respects the four requirements of research ethical principles stated by Vetenskapsrådet (2002; 2017): the information requirement, the confidentiality requirement, the utilization requirement, and the consent requirement. Before each interview, all participants received an email about the study's ethical requirements and information about the purpose of the study. Participation was optional and the participants were informed that they could withdraw at any time, for any reason. If a respondent chooses to discontinue participation, the material collected will be removed and further excluded from the results. We asked permission to record the interview and informed all respondents that the recorded material nor the transcription will be shared or used for purposes other than this study (Vetenskapsrådet, 2017). To guarantee confidentiality, all names and places will be anonymized. The anonymity will also be protected through the transcribed interview material. The field of research ethics is broad due to the existing variety of laws, guidelines and professional ethics regulations (Vetenskapsrådet, 2017). Therefore, we find it important to strive for good research practice. Informants has also been given a form with information about the purpose of the study, management of data and how long the information is stored. This is a standard procedure at Umeå University to ensure the integrity of personal data and thus compliant with the GDPR.

4.4 Data collection process

When conducting research within the field of information systems there are different approaches to produce knowledge. Interpretative research is a method of idiographic orientation in which the researcher wants to access meanings that people attribute to various events and which furthermore are a result of the interaction between actors in the social world (Fejes and Thornberg, 2019).

As this study aims to investigate the perceptions of real people in their real work organization through a case study, we engaged with the organization via employees. First, we had a meeting with our contact person where an explanation of the context of the organization was provided along with establishing the topic of the study. The selection of respondents was of targeted selection character and made by our contact person to best suit our research agenda. We were referred to several respondents with different roles and competences within the organization, something we had asked for in order to get as varied a picture of the organization as possible. Another requirement we set was that all respondents should have experience of working with data in the organization. The respondents undertake work associated with data in different ways (see table 2.) We also had access and read up on documents and policies connected to digital strategy that we chose to include to our empirical data. Documents are a good source of additional data that are not obtained from interviews (Myers, 2020). The advantages of using documents are that there may be invaluable things in the documents that do not emerge in interviews. However, we were aware that it may also be problematic to access these results and to confirm their authenticity (Myers, 2020). We believe that access to the documentation provided us with a better contextual understanding of the case by strengthening our empirical data. Hence, the document *Revised Digital Strategy document 2025* demonstrates the conditions and resources available to respondents when working with data (The organization, 2021).

The interview guide is designed in accordance with our theoretical framework and its three categories: Business Understanding, Data Understanding and Deployment (see appendix 2). From the Business Understanding part, we intend to capture respondents' perception of the organization's goals, from the Data Understanding part, we intend to capture respondents' perception of the organization's data and in the deployment part, we intend to capture respondents' perception of final reports. In order to create good conditions for the interviews, we have carefully analyzed the interview guide. As the respondents have different roles, we have adapted the interview material accordingly to capture different interpretations. In a qualitative research interview, knowledge is produced socially in an interaction between interviewer and respondent (Kvale & Brinkmann, 2014). We choose semi-structured interviews for data collection which are motivated by the need to capture in-depth knowledge about the individual's experience in each context (Fejes and Thornberg, 2019). The semi-structured type of qualitative interview is the one that is most commonly used in business and management research (Myers, 2020). It consists of a structure with some predefined questions, while as well allowing some improvisation. It opens the opportunity to add important insights as they arise during the conversation. We conducted 7 interviews individually, via teams with both authors participating. All interviews lasted approximately 40 minutes.

To receive a broader understanding of the organization's strategy and objectives we also coded the document (The organization, 2021) through the same approach we used for the interviews (see appendix 3).

Respondent:	**Role:**	**Length:**
Respondent 1	System developer	30:59
Respondent 2	System developer	40:43
Respondent 3	Business developer	39:12
Respondent 4	BI team	38:14
Respondent 5	BI team	25:57
Respondent 6	System developer	48:27
Respondent 7	Head of Section, IT	22:15

Documents:	**Pages:**	**Author:**
Revised Digital Strategy 2025	20	The organization, 2021

Table 1. Respondents' role and interview lengths

4.5 Data analysis

In order to analyze the data collected, we have used thematic analysis as a methodological approach. By following Braun and Clarke (2006) and Braun, Clarke, Hayfield & Terryet (2019) guidelines, we have identified the existence of themes and patterns within our empirical data (see appendix 3). The following guidelines were carried out: familiarization, generating initial codes, searching for themes, reviewing themes, defining and naming themes and producing the report.

1. When *familiarizing* with the data and in order to prepare the collected data for further analysis, we transcribed each recorded interview, in our case from audio to text. During this process, we immersed ourselves in our material by carefully reading through it, taking notes and commenting on interesting elements that make us more familiar with the data.
2. When engaging in *generating initial code,* we moved to a more systematic engagement of our collected data. During this process, we utilize codes that we identified as descriptive patterns from all respondents that corresponded to the questions they answered. These were later reflected in relation to our theoretical framework and research question before moving on to the next phase of constructing themes (Braun and Clarke, 2006; Braun et al., 2019).
3. After the previous processes of mapping the generated data into codes, we needed to test how our selection of data relates to the purpose of the study by *searching for themes*. We had in mind to identify elements in the coded material that were considered significant in relation to our theoretical framework and research question.

4. In the step of *reviewing themes*, we categorized our findings and candidate themes against our initial codes to exclude, revise or further define sub-themes to fit the categories (see appendix 3). As we intended to categorize our findings to the three categories from our theoretical framework we needed to start excluding or merge themes that overlapped each other.
5. During the *defining and naming themes* we worked iteratively as we needed to go back and forth to discuss what our codes and themes meant in relation to our theoretical framework.
6. In this final phase of thematic analysis, *producing the report*, (Braun and Clarke, 2006; Braun et al., 2019) the iterative process continues, revisiting the research question, making sure that our analysis has a continuity to it and relates to our theoretical framework.

4.6 Methodological discussion

Through a case study, we have been able to investigate the data mining process in a specific environment related to healthcare. The limitation of a case study is that the method does not capture a broader perspective from diverse organizations.

The motivation for conducting a single case study, as to multiple case studies, was due to concerns about obtaining superficial results from a variety of research subjects. We rather thoroughly examine one department and its contribution to the field in question in depth. We believe that focusing on it would be more appropriate for our research strategy. A multiple case study could have provided us with a more generalizable result but given the timeframe of this thesis, we decided to focus on a single case study.

The choice of using semi-structured interviews as a data collection method proved to be an appropriate approach allowing us to adapt the interview material to each individual respondent. This is because respondents occupy different roles in the organization and therefore work with different tasks in relation to data mining. During the interview process, we chose to delete or reformulate certain questions in order to create the best possible support for the purpose of this study. Furthermore, we did not get the opportunity to interview all the employees in the organization that were connected to our research case. If we had, we presumably would have had more nuanced data, which additionally would have strengthened the validity and made the findings more generalizable. To create the best possible conditions for our interview guide, we also chose to discuss and receive feedback during our first interview. Our first interview was with our contact person who had insight into our work. For this reason, we felt it was most appropriate to conduct the first interview with the person in question.

The CRISP-DM framework as a theoretical framework, has provided a structured approach to investigate the organization's data mining process and further barriers according to this process. The CRISP-DM framework can be used at a more specific level by following each phase of the data mining process, however, in this study we chose to start from a more comprehensive level. A criticism of this approach is that we do not get a deeper understanding of all the phases. This is further defended by the fact that this approach seemed more appropriate as we got the impression that the organization was not even familiar with CRISP - DM. Another criticism of

CRISP-DM as a framework is that although it is dynamic and adaptable, it may need to be more compatible to better fit different contexts.

5. Findings

In the following section, we present the investigative context of this study. Subsequently, we first present the results from the document regarding the organization's digital strategy (The organization, 2021). Secondly, we present the empirical data of our interviews following our theoretical framework. Since our interview guide was adapted to meet the different work roles of our respondents there is a variation in the questions asked. As a result, all respondents are not asked the same questions.

5.1 Investigative context

Healthcare has long been an arena for digitization issues, where data science challenges have long been a topic of discussion (Silver, Sakata, Su, Herman, Dolins & Shea, 2001; Teede et al., 2019).

The investigating purpose of the study is to further explore the research field of data mining in healthcare. More specifically, we find it interesting to explore the public sector's everyday practices of data mining to identify barriers.

The investigated organization is a public authority, and one of their primary functions is the governance of the public health care system. The context of our case study lies within one unit that supports IT. The organization has produced a document, *Revised Digital Strategy document 2025* (The organization, 2021), that outlines key strategic objectives as a desirable state to be achieved by 2030. In this document, four strategic focus areas have been identified that are considered particularly important to provide clearer direction for the transformation. These focuses are: (1) *Improving access to care,* (2) *Strengthening the conditions for individual participation in care and health,* (3) *Increased digitization of infrastructure and support processes for more time for care* and (4) *more efficient management and advanced analysis of data for better decision support.* For the context of our study, we find strategic focus 4 as the most relevant area.

5.2 Business understanding

There is an overall common understanding of what the respondents know about the organization's objectives. Respondent 2 believes that there is a lack of good understanding about the objectives, which is also suggested by respondent 4. Respondent 4 further means that it is difficult to define the objectives, which is also confirmed by respondent 6 who means that the objectives are very vague. Respondent 3 believes that the objectives are not entirely clear and furthermore would not be able to say what they are. Respondent 6 also experiences difficulties in defining the objectives.

> *"I don't have a good idea of all the objectives...it usually comes when you have to sit with it don't know the umbrella objectives...don't have a very good grasp"* - Respondent 2

> *“I'm not sure how to call in the targets”* - Respondent 4
>
> *“Well...what are the objectives...can't rake them up off the top of my head”* - Respondent 6

Furthermore, there are differences of opinion regarding the current objectives. Respondent 1 believes that the objectives are quite clearly described and that they are good to work towards. However, respondent 2 believes that there may be objectives that are not reasonable. Respondent 3 further argues that the objectives may be perceived as "ambitious and forward-looking" by some, but that they are not crystal clear to the person in question, which is also confirmed by respondent 4.

> *“But now it's about freeing up more time for care and I think that's perhaps clearly described ... they're nice to work with...good as guides in some way”* - Respondent 1
>
> *“There are so many good ideas and needs, but we see a huge power if we can start to hang needs on impact goals... “*- Respondent 3

Respondents 4 and 7 also believe that the objectives should be broken down one more level to create a clearer understanding of them. Respondent 6 further believes that the objectives should be evaluated from a holistic perspective to clarify what they ultimately mean for the organization.

> *“Description on an abstract level...in order for them to give a value to each employee you need to break them down...clarify how you can derive upwards...so you need to work more on that.”* - Respondent 7

There is a split opinion among the respondents as to why the organization has chosen to focus on these impact goals. Respondent 1 believes that there is a need to streamline and personalize the organization, which is also confirmed by respondent 6. Respondent 6 further believes that there is an ambition on the part of the organization to understand what the goal means for the people out in the field. Respondent 2 partly agrees, stating that it is based on economics and a desire to create better care. Respondent 3 claims not to know why the organization has chosen to focus on these objectives.

> *“What I know is that we are trying to get a funnel down...to get down to what it means for us out in the field...”* - Respondent 6

There is an overall common perception among the respondents regarding the way in which the objectives were communicated to them. Respondent 1 and 4 believes that the objectives were disseminated in several ways and describes that they were communicated at departmental meetings and through senior managers. This is also confirmed by respondent 2 who believes that the objectives were communicated when the person was in contact with the organization or through meetings. Respondent 3 and 6 believes that the objectives were communicated through information at the workplace or monthly IT meetings. However, respondent 3 believes that it would be easier to access the targets if they were available in a system. Respondent 5 believes that they come as an order to the team which they further manage together.

> *"Communicated to my type of role... an easier way to access them would be to put them in a system if you want to see them..."*- Respondent 3

> *"Via presentations and mediations from 'top' people...a bit fuzzy on how it is communicated..."* - Respondent 4

On the question of whether respondents believe that they have the necessary conditions and resources to achieve the objectives, there are both common and divided opinions. Respondent 1 believes that there are good conditions for developing the architecture for a technical platform but poor conditions for interpreting information. The respondent further argues that this is partly since the people sitting on the information are busy dealing with people.

> *"I'm in a bad position to interpret the information...those who are sitting on that information are busy caring for people...basically that's how it is" - Respondent* 1

Respondent 2 further believes that the right resources are available and that the support within the team is an important resource. Respondent 3 believes that the Power-BI and the data available create good conditions and that collegial work is highly valued. Respondent 4 believes that there is a lack of analysts but also to think more broadly and hence cross-functionally. The respondent further explains that there is a need for resources to understand data based on business needs with an analytical mindset, which is also confirmed by respondent 7.

> *"We have the gap in between which is about understanding the business needs and understanding data and analyzing business development with the analytics hat on...there is a lack of people...know how to understand data from the business point of view "*- Respondent 4

> *"We really need more resources if we're going to do it...very small percentage working with IT and technology and analysis follow-up compared to other organizations...so we're very thin on this side "*- Respondent 7

There are both common and divided views on whether there is a strategy for data mining processes. Respondent 1 believes that there are documents on data as a strategic resource but that this is under development. This is confirmed by respondent 4 who believes that the strategy is described at a high level of abstraction. Respondent 6 partially agrees and believes that there are various projects for this but that the organization is not there yet. Respondent 2 further claims that there is a strategy for decision making and how to extract data from other systems while respondent 3 claims that there is no strategy.

> *"Well...I don't think so...and anyway it's hard to see a common thread in the way we work with it anyway"* - Respondent 3

> *"It is written down but it's at a high level...and you work on it...clearly described at a high level but has to be translated to 'make sense' when you work on it too"* - Respondent 4

To conclude the first part of the interview guide, we asked all respondents if they were aware of the CRISP-DM framework. Thus, the answer of all except respondent 4, was no.

> *"Yes... what we really lack is analysts...who work with analyses...there is a gap...something that I think we should get to"* - Respondent 4

5.3 Data understanding

Based on the findings of the document, the organization demonstrates an understanding that data should be a strategic resource (*Revised Digitalization Strategy (2021, Strategic focus '4').* Although this strategic focus is something that the organization seems to be achieved by 2030, there is a description for the guiding principles which has been in use since 2021. They describe the importance of:

> *"Having a holistic approach to information management...the quality of the data must be appropriately high...opening up the data to create a common strategic resource that should be used for a primary purpose..." - Revised Digital Strategy document 2025* (The organization, 2021, p.16)

These guiding principles indicate that there should be clarity on how to work with data in the organization. In the context of data understanding from our empirical data this is not agreed upon, as a somewhat diverse picture emerged of how data is currently managed. There is an expressed need by respondent 1 to create a holistic approach to understand how data should be managed within the organization. A concern that respondent 2 further described as a lack of a central place where all the data can easily be accessed. As respondent 1 and respondent 2 notes, it creates some difficulties when these needs are not satisfied.

> *"What we have some struggle with is getting a holistic perspective and linking the definition: what is the data supposed to answer?"* - Respondent 1

> *"We don't have a central place where we keep all the data yet which means that it can be a problem sometimes if you want to retrieve certain data"* - Respondent 2

Overall, respondents reflected that there is a lot to improve for the organization to succeed in meeting the needs of such a large operation. Respondent 4 argues that the organization has to figure out how data should be used most efficiently. Respondent 3 agrees upon having things to be improved and notes that more modern solutions are in demand. Respondent 6 further expresses the concern that decisions based on the information do not have sufficient support.

> *"In many parts we are very immature in the perspective of data mining...we extract information that is not really supportive...you can't make decisions based on the information"*- Respondent 6

The general understanding of how data is managed is shaped by how well respondents receive data has been established in the organization. Respondent 5 emphasized that management of data in the organization feels generally forward-looking and progressive, hence taking another stance of how data is managed in the organization.

> *"I think they are pretty good at data...I think the organization is quite forward-thinking in general...I think they are at the forefront of IT development, taking big steps forward and testing lots of new solutions"*- Respondent 5

When asked how respondents view the data available to them, there are different reflections. Respondents 1, 3, 5, 6 and 7 all called attention to the fact that there is a width to the data meaning that it contains both high and low quality, which affects whether data is reliable or not. Both respondent 4 and respondent 7 notes that managing the extensive amount of data is difficult which in their experience leads to a loss of some of the data quality. Respondent 5 expressed a concern that even though the data is extensive there are still only a few different ways to find it.

> *"The data we have available is extensive... there are quite a lot of streams that are not always connected...it's a sea you can dive into but only find information in a few different ways"* - Respondent 5

> *"Which makes it difficult to get data available so that you can then use it, so the data quality can sometimes be a little shaky and makes it difficult to analyze and use the data"* - Respondent 7

The respondents note that there are many different systems they work within, some of which are inadequate to support the data they have available. The limited control in some systems leads to not knowing if the data is sufficiently structured. A concern expressed by respondent 1, 2 and 6. For example, respondent 2 reports that there is limited control over what data they have access to, which creates problems with their own systems, especially when it comes to external data.

> *"When it comes from third party providers we don't have all the control over the data we get...it feels unstructured...there's a lot of stuff left behind that's redundant that clogs up our system"* - Respondent 2

Further, when respondents are asked about the data availability in relation to performing their tasks, respondent 3 is the only one answering by agreeing that all the data necessary is provided. The other respondents reported a different view of data availability:

> *"The data is there but that doesn't necessarily mean that we have access to it...it is there but definitely it is locked"* - Respondent 1

> *"We have a lot of data in our systems, but we may not be using all that data in the way that we could, we have a lot of hidden assets in our data today...in some areas we have great access to data but in some areas we have less access"* - Respondent 7

Respondent 2 expresses a concern of not having the data required, something that respondent 4 additionally expresses has led to making guesses of the data.

> *"We don't really have that...then you see that there are gaps and we have to guess things together...try to match instead of trusting that it is clearly handled in a structured way"* - Respondent 4

Both respondent 6 and respondent 5 confirmed the expressed concern that even though data is available a lot is still missing to accommodate some requests. Respondent 6 further reports that a lot of information is missing to make decisions based on them.

> *"A lot of information is missing in order for the management to make decisions and be able to steer on the information that exists...we want to create data to get better information...we don't have that today it's missing...data is there but you want to improve things"* - Respondent 6

To capture a further reflection on the data available in the organization, respondents were asked whether they trust the data available to be reliable and cover the needs to meet the objectives of the organization. Respondents 1, 2 and 4 answer sharply no at first and then reflect why they feel so strongly that they don't trust the data. Respondent 1 notes that to make safe decisions with the data it should have been collected years ago.

> *"We have a lot of years of work ahead of us to get to a level where you can make safe and good decisions on data"* - Respondent 1

An additional reflection on the data collection was made by respondent 6 who argues that the registration of certain data each month is delayed. As described, there are a lot of different units within the organization, and each manages data in their own way which creates difficulties with the overall flow. This further leads to a lot of information missing, making it difficult to understand and rely on the data, also confirmed by respondent 5 and respondent 7 who additionally agrees that the data must be applicable.

> *"The trust is very varied...what did it look like last month...then a very large percentage is missing because they have not been able to keep up with and register the activities, they have done...there are a lot of delays...this means that it is very difficult to understand"- Respondent 6*

The only respondent who expresses trust towards the data available, later points out that within the organization they would need to look at other types of data to achieve the goals more effectively. Respondent 3 describes that more could be added:

> *"I trust the data... but from an organizational point of view we would probably need to look at other things to achieve our goals more effectively" - Respondent 3*

5.4 Deployment

There is uncertainty among respondents as to whether there is a plan on how to use the reports. Respondent 2 does not answer whether there is a plan but that the reports should be a basis for decisions and external reporting. Respondent 3 believes that there has been and is a plan forward but it has been unclear going forward. Respondent 4 believes that there are challenges with this element, which further describes that this is since reports have been built based on the needs of different clients and thus lack a holistic approach. This is also confirmed by Respondent 6 who says that there have been discussions back and forth about this plan but says that he has no idea how structured the plan is. Respondent 7 means that there is a plan on an overall level.

> *"No, that's what we have challenges with...something we need to work on as well...think we should move more towards standardized reports...in the past we've had a bit of a spread...need a specific report instead of a common report for a unit..."* - Respondent 4

> *"At an overall level, there are reports divided into strategic, tactical and operational reports...there are different levels in the organization, so they have different needs for different types of reports."* - Respondent 7

On the question of how the reports are used today, there is also uncertainty. Respondent 2 further means that the reports are used to make certain decisions. Respondent 4 further means that there is not a very clear way to work our way forward but it's something that the organization is trying to develop. This is also confirmed by Respondent 6, who states that most reports are a description of the production in different ways, but that it is not so upsetting and that there is no report that looks into the future. Respondent 7 means that the reports are used in a variety of fields.

> *"Works towards different clients where it is about getting the client to think more long-term, a little broader than just their specific needs... there is no very clear way to work our way forward there... it is something we try to develop..."*
> - Respondent 4

> *'They are used in a variety of fields... we have reports on a more tactical level that show how the production works and above all the reports answer how it has gone, we have perhaps fewer reports that are produced forward how it goes or how it should go"* - Respondent 7

There are different views on the extent to which the reports are used. Respondent 1 means that there is uncertainty about what counts as used but that you can see how many people have accessed a report and how many have used the report. Respondent 2 also states that they are working on getting everything over to Power-BI to see how the statistics are used. Respondent 3 argues instead that very few reports are used for the purpose for which they were produced. Respondent 4 further states that the business uses the reports very much and that it belongs to an important area. Respondents 6 and 7 argue that the reports are included in many forums in different ways and that the information is used but further describes that he does not know more than that.

> *"My picture is that very few reports are used for the purpose for which they are designed... we have had a very low threshold for building reports so we have very many reports that are probably used very little"* - Respondent 3

When we ask respondents whether old reports are reused, we get quite different answers. Respondent 3 argues that some reports are built upon and further argues that it is very unclear when it comes to ownership. Respondent 4 thus agrees, but also describes that there is an ambition to have a report manager. However, respondent 6 believes that the reports are widely used, but that they should be adapted to each activity. Respondent 7 partly agrees and further means that there are a lot of reports in the organization.

> *"It's probably a mixture...I think the threshold for building reports has been very low and there hasn't been real ownership, and I think very much with reports it's like any product that there's a life cycle."* - Respondent 3

Respondents 3 and 6 further argue that there is a lack of trust in reports that are produced while respondents 4 and 7 don't agree, who further trust the reports. Furthermore, respondent 6 means that concepts like data governance are important and something that the organization tries to highlight. However, respondent 4 argues that the client needs to validate and know what the truth is, which can sometimes be underestimated.

> *"It's not that reliable...you have to have a certain feel for it...to make decisions...of course there are flaws...there are quality gaps in our reports...they need to be improved"*- Respondent 6

In addition, we have two questions regarding the respondents' perceptions of an optimal data mining process and what challenges they believe may exist related to this optimal process in the organization. The purpose of these two questions is to get the respondents to reflect further on the data mining process, to capture potential barriers that may inhibit the data mining process. An optimal data mining process according to the respondent is hence related to preventively work, automation, standards, clear structure, resources, functional teams, and objectives.

> *“Standard reports that are simple... be able to have analysis groups that can really twist and turn questions around objectives that are relevant. “*- Respondent 3

> *“Look at the need... needs to be described of course...dare to collect the data......dare to look a bit broader... what we have for data and use data and analyze data too I believe in.”* - Respondent 7

Challenges mentioned regarding the optimal process are related to lack of data, locked in systems, maturity, historical reasons, hierarchical levels, fear that it will take too long and lack of resources.

> *“Many chefs and hierarchical levels, so that always makes things difficult in any context...”* - Respondent 5
> *“Biggest issue is the resource issue...there is an old architecture...about scaling things off...a barrier to not being able to remove such things...missing in the core system today...”* - Respondent 6

6. Analysis

In this section we have summarized and compared the empirical data in relation to our theoretical framework and its associated categories, Business understanding, Data understanding and Deployment. Through this process we have identified significant barriers for the organization's data mining process.

6.1 Business Understanding

Based on the findings from the business understanding phase we have identified the following. There is an overall common perception that there is a restricted understanding of the objectives among the respondents. Hence, the respondents believe that the objectives are not clearly stated and therefore have difficulties defining them. Most of the respondents also note that they have been able to take part of the objectives at departmental meetings and from senior managers. There is furthermore a perception that it would be easier to access the objectives if they were available in a system. Based on whether there is a strategy for the data mining process, some stated that there are various projects for this but that the organization is not there yet. Hence, there is a perception that there is no strategy for this process and furthermore everyone, except one, knows about the CRISP-DM framework. According to the CRISP-DM framework, it's important to create a preliminary plan for how the objectives are to be achieved, to understand the type of data to be analyzed in relation to objectives (Shearer, 2000). Additionally, it is important to spend as much time as possible in the business and data understanding phase (Provost and Fawcett, 2013).

Furthermore, some respondents argue that the objectives are unreasonable and that they should be broken down further to create a clearer understanding of them. However, there is an overall common understanding among respondents as to why the organization has chosen to focus on these objectives. Furthermore, it is felt that the objectives should be evaluated from a holistic perspective in order to create an understanding of what they ultimately mean for the organization. According to the CRISP-DM framework, objectives should be transformed into data science tasks, which is an iterative discovery process (Provost and Fawcett, 2013). Furthermore, the CRISP-DM framework argues that the subtasks are unique to the specific business problem, but these should then be compiled to solve the overall goal (Provost and Fawcett, 2013).

There is a perception that there is a lack of conditions and resources to achieve the objectives. This is further described by the limited number of analysts and the need to understand data based on business needs with an analytical mindset. According to the CRISP-DM framework, it is important that there are available and necessary resources (Shearer, 2000; Schröer et al., 2021) to achieve the objectives. The CRISP-DM framework highlights that the key to success is grounded in creative objectives formulated by an analyst, which plays a major role in this phase (Provost and Fawcett, 2013).

6.2 Data Understanding

Based on the findings from the data understanding phase we have identified the following. From the findings in the document (The organization, 2021), the organization demonstrates an understanding that data should be a strategic resource forward. However, what our empirical data shows is a varied picture of how the strategic resources are expressed. What can be summarized by most of the respondents is that they do not perceive that there is a standardized procedure for the data mining process. The data management in the organization is described as immature and does not extract the necessary data to make confident decisions. This implies that there is no clear approach on how to use data to achieve the objectives, something that the CRISP-DM framework advocates as necessary for successful data mining (Provost and Fawcett, 2013). A further finding is that there is a common agreement among respondents that the data available is not enough for meeting the objectives. Data availability is a key aspect in the data mining process of the CRISP-DM framework, since data is the main component to ensure good insights (Shearer, 2000; Provost and Fawcett, 2013).

Another essential aspect that the CRISP-DM framework covers is having the competence to structure the data, creating initial insights and an ability to identify problems with data quality (Niaksu, 2015; Plotnikova et al., 2020). All respondents note that there is a lot of data within the organization but that doesn't necessarily mean that all data is of high quality and a reliable source for decision-making. The CRISP-DM framework highlights the importance of having high quality data as it represents the available raw material from which the solution will be built (Provost and Fawcett, 2013). The better the quality of the data the better the quality of the insights.

If the available data is of low quality, it becomes of utmost importance to have the appropriate competencies and prerequisites to transform the data to enhance its usability (Provost and Fawcett, 2013; Plotnikova et al., 2020). Some respondents argue that there are insufficient prerequisites to understand the data and need a lot of support in interpreting and doing translations to structure the data. This is a perceived need that is not fully met by the organization due to other priorities. As a result, the data cannot be ensured nor is it reliable enough to meet the objectives.

6.3 Deployment

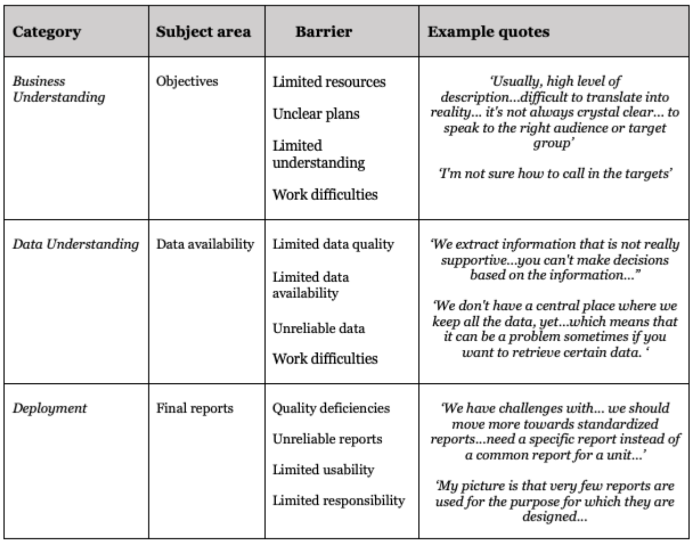
Based on the result from the deployment part we have identified the following. Deployment includes the results of the data mining process (Provost and Fawcett, 2013; Plotnikova et al., 2020). There is uncertainty among respondents as to whether there is a plan for how the reports will be used. It is further claimed that there is a plan, but that it is unstructured. There are also perceptions that there are quality deficiencies in the reports produced, making them unreliable. This is explained by the fact that it may be based on the ownership of the reports and that the work of validating the reports is underestimated. According to the CRISP-DM framework, the findings should be shared with the decision makers so that they can use the report for decision support (Provost and Fawcett, 2013; Plotnikova et al., 2020). The final phase of the project therefore includes a written report and a presentation of the main findings (Schröer et al., 2021). Although the aim of the CRISP- DM framework is to increase knowledge of the data, the knowledge obtained must be organized, presented, and distributed in such a way that the end user can use it (Plotnikova et al., 2020).

There is uncertainty among respondents regarding how the reports are currently used. Thus, some respondents believe that there is no clear way forward but that it is something that the organization is working on. According to the CRISP-DM framework, it is important that the end user anticipates the actions that need to be taken in relation to the practical benefits of the framework (Niaksu, 2015). Depending on the requirements, the deployment phase may involve either generating a report or even conducting a repeatable data mining process (Niaksu, 2015; Plotnikova et al., 2020). However, there is a perception that very few reports are used for the purpose for which they were produced.

There are divided opinions on whether old reports are reused. Respondents believe that some reports are reused but that there is uncertainty regarding ownership and that reports should be adapted to each business. This phase also stresses the importance of communicating analytical findings to the intended audience (Jaggia et al., 2020). There are different views on the extent to which the reports are used. Some say that there is uncertainty about what counts as used but.

6.4 Identified barriers

The significant subject areas that present barriers to the data mining process are identified as the following (see table 2). Through the category business understanding, we have identified that the respondents have a limited understanding regarding the objectives of the organization which makes it difficult to understand and how to work in practice. Furthermore, limited resources and unclear plans for the data mining process have been identified as potential barriers. Through the category data understanding, we have identified that the data is not supportive enough to meet the objectives, which reflects upon bad data availability in the organization. The data available is of low quality and is not a reliable source for confident decision-making which creates barriers. Through the category deployment, we have identified that the final reports have quality deficiencies making them unreliable and a barrier. Furthermore, few reports are used for the purpose for which they were produced.



Category	Subject area	Barrier	Example quotes
Business Understanding	Objectives	Limited resources Unclear plans Limited understanding Work difficulties	*'Usually, high level of description...difficult to translate into reality... it's not always crystal clear... to speak to the right audience or target group'* *'I'm not sure how to call in the targets'*
Data Understanding	Data availability	Limited data quality Limited data availability Unreliable data Work difficulties	*'We extract information that is not really supportive...you can't make decisions based on the information..."* *'We don't have a central place where we keep all the data, yet...which means that it can be a problem sometimes if you want to retrieve certain data. '*
Deployment	Final reports	Quality deficiencies Unreliable reports Limited usability Limited responsibility	*'We have challenges with... we should move more towards standardized reports...need a specific report instead of a common report for a unit...'* *'My picture is that very few reports are used for the purpose for which they are designed...*

Table 2. Identified significant subject areas and barriers

7. Discussion

There is a need to create a greater understanding of how organizations organize their data management in practice (Alhassan et al., 2016). As such, we have witnessed many efforts directed towards data governance in general and creating better value from data (Bendfeldt et al., 2018). As a result of the analysis between the collected empirical data and the CRISP-DM framework, this section will discuss the barriers identified.

7.1 Objectives

Barriers identified are based on a limited understanding of the objectives and therefore hard to work towards them in practice. This is also reflected in research showing that the public sector has difficulty communicating the value that data and data governance can create (Bendfeldt et al., 2018). Furthermore, to manage data appropriately, it is necessary to understand the data to be managed and why it is important to the organization (Brous et al., 2016). Data management must be further driven by and aligned with the organizational objectives (Bendfeldt 2017; Brous et al., 2016). Highlighting concepts such as data governance within the organization is confirmed by the respondents, who believe that data governance is an important aspect and something that the organization is striving for. The CRISP-DM framework highlights the importance of understanding the objective to be investigated (Provost and Fawcett, 2013; Niaksu, 2015) since setting the goal of data mining is one of the most important aspects of the business understanding phase (Shearer, 2000; Schröer et al., 2021). However, what this should look like can vary since the business understanding phase can be described in different ways (Provost and Fawcett, 2013).

Respondents thus believe that the objectives are described at a high abstract level, resulting in challenges in understanding how to work towards them in practice. According to the CRISP-DM framework, each decision problem should be unique and consists of a combination of different objectives and constraints. Together with data scientists, an iterative process can thus break down the objective formulation into sub-objectives and later synthesize them to solve the overall problem (Provost and Fawcett, 2013). Transforming an overall objective formulation into several data science problems requires high-level knowledge, which can be obtained from creative analysts (Provost and Fawcett, 2013).

Based on creative analysts' thinking, limited resources are identified as a barrier and more specifically missing analytical mindset. Hence, research shows that analytical thinking can help to extract valuable information from healthcare data and further transform it into valuable information that can facilitate decision-making in healthcare. Analytical thinking can help to provide insights into the effective use of data and application of analysis (Islam et al., 2018). Further, barriers are related to an unclear plan or strategy for the data mining process, which according to the CRISP-DM framework is an important aspect, to understand the type of data to be analyzed in relation to the objective (Shearer, 2000).

Different levels of data literacy among departments and managers in the public sector make it difficult to see a strategic direction for the use of data in the organization (Bendfeldt et al., 2018). This is thus reflected in our empirical data, which shows that there is a need to work in cross-functional teams to get a more holistic view of existing needs that the organization should strive for. By creating a more unified picture of the organization’s needs, a clearer strategic direction can emerge.

7.2 Data availability

Other significant barriers identified are based on the organization's current limited data availability and data quality. It is repeatedly pointed out that there is an immense amount of data within the organization, but that this does not necessarily mean that this data is suitable for meeting the objectives. Hence, their availability of appropriate data is currently low. What is clearly indicated in the CRISP-DM framework is the importance of having the appropriate data available for creating valuable insights (Shearer, 2000; Provost and Fawcett, 2013). Extant research argues for the importance that data should be managed in a way that makes it accessible to gain opportunities (Brous et.al, 2016; Bendfeldt et.al, 2018). Having good data availability is desirable as it permits your business to run uninterrupted (Shearer, 2000; Kahtri and Brown, 2010). The limitation of appropriate data is something that complicates the work and eventually slows down the process of data mining. Medical data that exist within the organization is usually historical paper-based patient data, which in turn can result in the data often being incomplete in terms of electronic accessibility (Niaksu, 2015). Despite this, poor data availability must be reinforced if the organization wants to gain and retain opportunities with data (Begg and Caira, 2011; Bendfeldt, 2017; Brous et al, 2020).

An alternative to enforcing low data availability, is the possibility of bringing in third-party data. The use of third-party data, such as open data and big data, can positively empower the data sets by creating greater opportunities for potential insights and transformation of an organization (Bendfeldt et.al, 2018). However, the third-party data used in the organization has been described as problematic of its unstructured nature. Unstructured data requires more work for system developers to clean and transform the data to make it useful (Provost and Fawcett, 2013). Further, this complicates and creates a barrier to work within the data mining process.

The findings show that the data available is not a reliable source for confident decisions. The quality of the data is of great importance and stands as one of the key objectives within data governance, data mining and the CRISP-DM framework (Provost and Fawcett, 2013; Brous et al., 2016; Bendfeldt et al., 2018). The data quality of the organization constitutes a barrier to the data mining process. Data mining challenges are often associated with low data quality and missing values (Yoo et al., 2011; Niaksu, 2015; Witjas-Paalberends et al., 2018), which is further confirmed from the findings as a continued challenge. Most data-driven decisions in healthcare ultimately affect their patients, (Witjas-Paalberends et al., 2018) which reinforces the importance of the need for high data quality in healthcare to make confident decisions. Also, there is a great initiative from the organization to use advanced analysis of data for better decision support (The organization, 2021). However, the findings show that there is poor data availability and low data quality. Based on the document *"the quality of the data must be appropriately high* "(The organization, 2021, p.16), which is not reflected in the daily activities at present. As a result, this can be identified as a barrier for the data mining process.

7.3 Final reports

Other significant barriers identified are based on limited applicability of the final reports (Shearer, 2000; Provost and Fawcett 2013). Due to quality deficiencies and that few reports are used for its purpose the final reports do not fulfill their intended purpose, as described in the CRISP-DM framework (Provost and Fawcett 2013). The final report does not have a clear strategy or plan on how to use them, something that creates work difficulties as a barrier. Research indicates that having a standardized use of data practices might help improve the chance of maintaining data opportunities (Alhassan et al., 2016; Mikalef et al., 2020). For the final reports to generate decisions for planning, management, and learning (Islam et al., 2018), they need to be reliable and comprehensible. The CRISP-DM framework highlights the importance of having the results of a data mining process shared in an organized, presented, and distributed way so that the decision-makers can use the data for decision support (Provost and Fawcett, 2013; Plotnikova et al., 2020). This also emphasizes the importance of communicating the analytical findings to the intended audience (Jaggia et al., 2020). The final report should have a high application possibility (Islam et al., 2018), which does not correspond to the findings. As a result, this may be due to lack of a strategy.

The final reports have also been described as unreliable due to quality gaps. The reports should generate evidence-based decisions (Islam et al., 2018), meaning that the quality of the reports has to be high. The fact that the reports are perceived as not high quality may be due to low quality of the data itself (Begg and Caira, 2011; Alofaysan et al., 2016). Nevertheless, it is still important to have someone responsible for ensuring that published reports are of high quality (Islam et al., 2018). This is reflected in the findings, showing that there is an ambition to have a report manager in place. As there is no clear strategy on how to deal with the final reports, no one can be held responsible for the reports. This may create barriers for the data mining process.

The final reports contain knowledge to be gained from the data mining process, which further could help healthcare to better serve patients (Yoo et al., 2011). Due to a limited strategy in the organization these valuable opportunities are somewhat lost. Data mining in healthcare lacks standardized procedures of knowledge discovery (Yoo et al., 2011), which is further confirmed from the findings. Therefore, it is of utmost importance for the organization to strengthen the deployment phase to ensure reliable and compressible decisions (Provost and Fawcett 2013).

7.4 Further research

Based on this study, the findings confirm that data mining methods are a highly important aspect to extract valuable information from data. Furthermore, the scope of this study has concerned only one IT department in the health sector. For this reason, we believe that further studies according to this topic should be conducted to highlight additional potential barriers. By illustrating barriers related to the data mining processes, stakeholders in the healthcare sector can develop strategies for the data mining process. In addition, we present the CRISP-DM framework from a comprehensive approach. This opens for further research based on a more specific approach that could usefully be presented. Further research on the CRISP-DM framework as an analysis model in practice, may provide guidance in the data mining process.

7.5 Acknowledgements

In acknowledgement, we want to give thanks to our supervisor Jonny Holmström, who guided us through this study. We would also like to give thanks to Ida Bodén, who gave us valuable insights on the data mining process and the CRISP-DM framework and kept our motivation high. To our contact persons at the collaborative organization, we are very grateful, they ensured the resources needed to make the study a success. Finally, we would like to give a special thanks to all the respondents who dedicated their time and insights to our study, without them this study would not have been possible.

8. Conclusion

After conducting this study, we conclude that organizations with extensive data available need to integrate data governance and moreover data mining methods in their IT strategy. This is to sustain and create opportunities from data but also to address organizational data problems. The healthcare sector is no exception, as this organization has high ambitions to become more data-driven to create more time for care. Knowledge gained from data mining can help healthcare providers better serve patients. However, there is a lack of frameworks for analyzing data mining in healthcare. By using frameworks for analyzing the data mining processes, knowledge of barriers that can inhibit the data mining process can be addressed.

The purpose of this study has been to investigate the data mining process in healthcare where we used the following research question: *What barriers in the data mining process in healthcare can be identified, by using the CRISP-DM framework?* Hence, we can conclude that there are a number of barriers that may inhibit this process according to the CRISP-DM framework. These are linked to the organization's objectives, data availability and final reports. We have also evaluated CRISP-DM as a framework and concluded that the analysis model provides a distinct guide to the data mining process and can be used to advantage at different levels of abstraction. Finally, we believe that the results of this study would have contributed to a more nuanced picture of the topic if more respondents had participated. As a result, the study would have contributed to an even greater understanding of how healthcare can create value through data mining methods.

9. References

Abbasi, A., Sarker, S., Chiang, R. H. (2016) Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, 17 (3).

Alhassan, I., Sammon, D., and Daly, M. (2016) Data Governance Activities: An Analysis of the Literature. *Journal of Decision Systems* (25:sup1): 64-75.

Alofaysan. S., Alhaqbani, B., Alseghayyir. R, and Omar, M. (2016) The Significance of Data Governance in Healthcare A Case Study in a Tertiary Care Hospital. College of Health Informatics, King Saud Bin AbdulAziz University for Health Sciences, Riyadh, Saudi Arabia.

Azevedo, A., Santos, M. (2008) KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European conference data mining,* 182–185.

Begg, C., Caira, T. (2011) Data Governance in Practice: The SME Quandary Reflections on the Reality of Data Governance in the Small to Medium Enterprise (SME) Sector, *In: 5th European Conference on Information Management and Evaluation (ECIME)*, pp. 75–83

Benfeldt Nielsen, Olivia (2017) A Comprehensive Review of Data Governance Literature. *Selected Papers of the IRIS*, Issue Nr 8 (2017) 3.

Benfeldt, O., Persson, J. S., Madsen, S. (2018) Why governing data is difficult: Findings from Danish Local Government. *In Smart Working, Living and Organizing: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2018, Portsmouth, UK, June 25, 2018, Proceedings* (pp. 15-29). Springer. I F I P Vol. 533.

Bhavnani, S.P. and Sitapati, A.M. (2019) Virtual Care 2.0—a Vision for the Future of Data-Driven Technology-Enabled Healthcare. *Current treatment options in cardiovascular medicine*. 21(5), pp.1–13.

Bohanec, M., Robnik-Sikonja, M., Borstnar, MK. (2017) Decision-making framework with double-loop learning through interpretable black-box machine learning models. *Industrial Management and Data Systems*. 2017;117(7):1389–1406

Braun, V., and Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology,* 3, 77–101. http://dx.doi.org/10.1191/1478088706qp063oa.

Braun, V., Clarke, V., Hayfield, N., Terry, G. (2019) Thematic Analysis, in: Liamputtong, P. (Ed.), *Handbook of Research Methods in Health Social Sciences*. Springer Singapore, Singapore, pp. 843–860. https://doi.org/10.1007/978-981-10-5251-4_103.

Brous, P., Janssen, M., Vilminko-Heikkinen, R. (2016) Coordinating Decision-Making in Data Management Activities: A Systematic Review of Data Governance Principles. *In: Electronic Government. EGOV. Lecture Notes in Computer Science,* vol 9820. Springer, Cham. https://doi.org/10.1007/978-3-319-44421-5_9.

Brous, P., Janssen, M., Krans, R. (2020) Data Governance as Success Factor for Data Science. In: Hattingh, M., Matthee, M., Smuts, H., Pappas, I., Dwivedi, Y., Mäntymäki, M. (eds) *Responsible Design, Implementation and Use of Information and Communication Technology*. I3E 2020. vol 12066. Springer, Cham. https://doi.org/10.1007/978-3-030-44999-5_36

Chae, Y.M., Kim, H.S., Tark, K.C., Park, H.J. and Ho, S.H. (2003) Analysis of healthcare quality indicator using data mining and decision support system. *Expert systems with applications*. 24(2), pp.167–172.

De Massis, A. and Kotlar, J. (2014) The case study method in family business research: Guidelines for qualitative scholarship. *Journal of family business strategy*. 5(1), pp.15–29.

Fejes, A. and Thornberg, R. (red.), (2019) Handbok i kvalitativ analys. (Upplaga 3). Stockholm: Liber.

Galetsi, P., Katsaliaki, K. and Kumar, S. (2020) Big data analytics in the health sector: Theoretical framework, techniques and prospects. International journal of information management. 50, öpp.206–216.

Henfridsson, O., and Bygstad, B. (2013) The generative mechanisms of digital infrastructure evolution. *MIS Quarterly* Vol. 37 No. 3, pp. 907-931/September 2013.

Islam, M.S., Hasan, M.M., Wang, X., Germack, H.D. and Noor-E-alam, M. (2018) A systematic review on healthcare analytics: Application and theoretical perspective of data mining. *Healthcare (Basel)*. 6(2), p.54–.

Jaggia, S., Kelly, A., Lertwachara, K. and Chen, L. (2020) Applying the CRISP-DM Framework for Teaching Business Analytics. *Decision sciences journal of innovative education*. 18(4), pp.612–634.

Jothi, N.,Rashid, N.A. and Husain, W. (2015) Data Mining in Healthcare – A Review. *Procedia computer science*. 72, pp.306–313.

Khatri, V., and Brown, C. V. (2010) Designing data governance. *Commun. ACM*. V. 53, 148.

Kvale, S. and Brinkmann, S. (2014) Den kvalitativa forskningsintervjun. 3 red. Lund: Studentlitteratur.

Lee, Y.W., Madnick, S.E., Wang, R.Y., Wang, F.L., Zhang, H. (2014): A cubic framework for the chief data officer: Succeeding in a world of big data. *MIS Q*. Exec. 13, 1–13.

Mariscal, G., Marbán, Ó., Fernández, C. A. (2010) survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review*. 2010;25(2):137–166.

Mikalef, P., Pappas, I. O., Krogstie, J., and Pavlou, P. A. (2020) Big Data and Business Analytics: A Research Agenda for Realizing Business Value. *Information & Management* (57:1): Article 103237.

Myers, Michael D. (2020) Qualitative research in business & management. Third edition London: SAGE Publications Ltd.

Nambiar, R., Sethi, A., Bhardwaj, R. and Vargheese, R. (2013) A look at challenges and opportunities of big data analytics in healthcare. *Institute of Electrical and Electronics Engineers 2013 International Conference on Big Data*.

Niaksu O. (2015) Crisp data mining methodology extension for medical domain. *Baltic Journal of Modern Computing*. 2015;3(2):92.

Parmiggiani, E., and Grisot, M. (2020) Data Curation as Governance Practice. *Scandinavian Journal of Information Systems* (32:1), p. 1.

Plotnikova, V., Dumas, M. and Milani, F. (2020) Adaptations of data mining methodologies: a systematic literature review. *PeerJ. Computer science*. 6, pp.e267–e267.

Provost, F and Fawcett, T. (2013) Data science for business: what you need to know about data mining and data-analytic thinking. 1. uppl. Sebastopol, Calif.: O'Reilly.

Rivo, E., de la Fuente, J., Rivo, Á., García-Fontán, E., Cañizares, M.-Á. and Gil, P. (2012) Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. *Clinical & translational oncology*. 14(1), pp.73–79.

Schröer, C., Kruse, F. and Gómez, J.M. (2021) A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia computer science*. 181, pp.526–534.

Shearer, C. (2000) The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing,* 5(4), 13–22.

Sohlberg, P. and Sohlberg, B. (2019) Kunskapens former: vetenskapsteori, forskningsmetod och forskningsetik. (Fjärde upplagan). Stockholm: Liber.

Silver, M., Sakata, T., Su, H. C., Herman, C., Dolins, S. B., and O Shea, M. J. (2001) Case study: how to apply data mining techniques in a healthcare data warehouse. *Journal of healthcare information management*, *15*(2), 155-164.

Subrahmanya, S.V.G., Shetty, D.K., Patil, V., Hameed, B.M.Z., Paul, R., Smriti, K., Naik, N. and Somani, B.K. (2021) The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish journal of medical science*.

Teede, H., Johnson, A., Buttery, J., Jones, C.A., Boyle, D., IR., Jennings, G., and Shaw, T., (2019) Australian Health Research Alliance: national priorities in data driven healthcare improvement. Med J Aust. ;211(11):494–7.

Vetenskapsrådet (2002) Forskningsetiska principer inom humanistisk- samhällsvetenskaplig forskning. Stockholm: Vetenskapsrådet.

Vetenskapsrådet (2017). God forskningssed. [Downloaded: 2022-05-20] https://www.vr.se/analys/rapporter/vara-rapporter/2017-08-29-god-forskningssed.html

Webster, J. and Watson, R. T (2002) Analyzing the past to prepare for the future: Writing a Literature Review. *MIS Quarter,* 26(2), pp. xiii-xxiii.

Witjas-Paalberends, E.R., van Laarhoven, L.P.M., van de Burgwal, L.H.M., Feilzer, J., de Swart, J., Claassen, E. and Jansen, W.T.M (2018) Challenges and best practices for big data-driven healthcare innovations conducted by profit-non-profit partnerships - a quantitative prioritization. *International journal of healthcare management*. 11(3), pp.171–181.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F. and Hua, L (2011) Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *Journal of medical systems*. 36(4), pp.2431–2448.

10. Appendix

Research streams

Research stream	Key concepts	Example literature	Example References
Data Governance	- Data governance objectives - Evidence: Healthcare - Challenges - Opportunities - Strategy	"Why governing data is difficult: *Findings from Danish Local Government"* (Benfeldt et al., 2018)	-Benfeldt et al (2018) -Brous et al (2016) -Mikalef et al (2020) -Parmiggiani and Grisot (2020)
Data mining in healthcare	- Data mining method - Challenges - Opportunists - Evidence: Healthcare	" Data Mining in Healthcare – A Review" (Jothi et al, 2015)	- Niaksu (2015) - Witjas-Paalberends et al (2018) - Yoo et al (2011) - Jothi et al (2015)
The CRISP-DM framework	- Data analysis - Data extraction - Data analysis method - Process	" The CRISP-DM model: *The new blueprint for data mining" (*Shearer, 2000)	- Schröer et al (2021) - Shearer (2000) - Provost and Fawcett (2013)

BACKGROUND ON RESPONDENT:

1. What is your job role?
2. In what way is data included in your job description?
3. What is your view on the organization's current data mining process?
4. Have you heard of CRISP - DM?

BUSINESS UNDERSTANDING

1. How do you work with/against the objectives of the organization?
2. Do you think you have the right resources or conditions to meet/work with these objectives of the organization?
3. What is your perception of the current objectives of the organization at the organization?
4. Do you know why the organization has chosen to focus on these objectives of the organization?
5. Based on the objectives of the organization you are working on, what do you think of these?
6. How did you learn about the objectives of the organization?
7. How are the objectives of the organization communicated to employees?

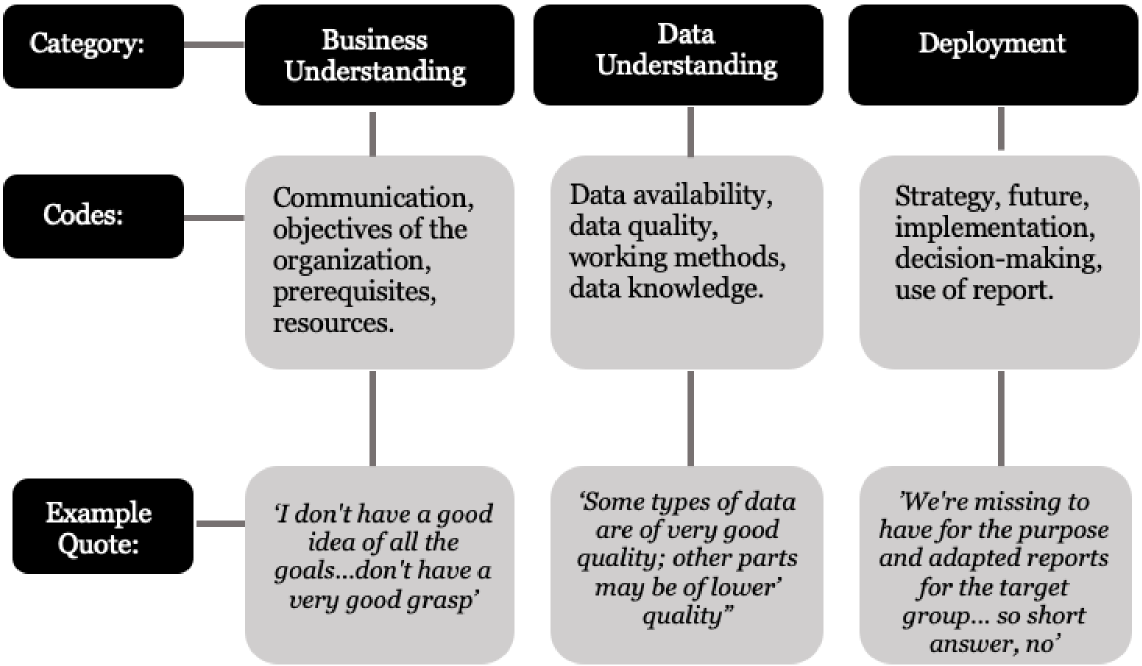
DATA UNDERSTANDING

1. What do you think about the data you have access to?
2. Do you think you have access to the data needed to meet the objectives of the organization/perform the tasks you have?
3. Do you trust that the data you have access to covers the needs to meet the objectives of the organizations?
4. Do you think you have the necessary skills to cover the needs?

DEPLOYMENT

1. Is there a plan for how the reports will be used?
2. In what way do you think the reports are currently used?
3. To what extent do you think that decisions are made from the reports and then implemented?
4. Do you capture new needs and create new reports, based on existing reports?
5. Do you think you have the necessary knowledge to read the content of the reports?
6. Do you have someone to turn to for questions about the report?
7. Do you trust the reports that are produced?
8. Do you know if there is a strategy for the data mining process?
9. Can you describe how you would define the optimal data mining process?
10. What challenges do you think there might be in working from this optimal approach in the organization?
11. What do you think the organization could look like if it started from this optimal process?

Appendix 3: Coding process model



Category:	Business Understanding	Data Understanding	Deployment
Codes:	Communication, objectives of the organization, prerequisites, resources.	Data availability, data quality, working methods, data knowledge.	Strategy, future, implementation, decision-making, use of report.
Example Quote:	*‘I don't have a good idea of all the goals...don't have a very good grasp’*	*‘Some types of data are of very good quality; other parts may be of lower’ quality”*	*’We're missing to have for the purpose and adapted reports for the target group... so short answer, no’*