Utrecht University

Master's Thesis

Submitted in fulfillment of the Master's Program

Business Informatics

Knowledge Discovery for Domain Experts: A Data Preparation Approach

Thomas Jan Dedding
t.j.dedding@students.uu.nl

supervised by

Dr. Marco Spruit (Utrecht University)
Dr. Matthieu Brinkhuis (Utrecht University
Dr. Daniel Vijlbrief (Wilhelmina Kinderziekenhuis)

10th July 2018

ABSTRACT

Knowledge Discovery (KD) and Data Mining are two well-known and still growing fields that, with the advancements of data collection and storage technologies, emerged and expanded with great strength by the many possibilities and benefits that exploring and analyzing data can bring. However, it is a task that requires great domain expertise to really achieve its full potential. Additionally, it is also an activity which is done nowadays mainly by data analysts and scientists, which most of the times knows little about specific domain subjects, like in the healthcare segment, for example.

The term Applied Data Science (ADS), recently introduced, focus on creating means, by using analytical methods and applications, for facilitating the daily life of domain experts. Thus, in this research, following an ADS orientation, we propose means for allowing domain experts from the healthcare segment (e.g. doctors and nurses) working in the Wilhelmina Kinderziekenhuis (WKZ), to also be actively part of the Knowledge Discovery process, focusing in the Data Preparation phase, and use the specific domain knowledge that they have in order to start unveiling useful information out of the data. Hence, a guideline based on the CRISP-DM framework, in the format of methods fragments is introduced to guide these professionals through the KD process, focusing in the data pre-processing stage. In order to build the model, an extensive literature review was performed, followed by interviews which aimed to understand what domain experts actually knew about KD, and what should be feasible for them to do when addressing an analytical problem. In addition to that, also to understand what types of problems domain experts would be dealing with in their daily routine, a data quality assessment from the available information within the databases from the WKZ was performed.

Regarding the evaluation of the proposed solution, five meetings with domain experts were held, where the model has introduced and extensively explained, and two case studies representing a real analytical project (using real data) were performed. The findings of this study were acquired by means of a survey, which extracted their opinions about the interpretability (understandability and accuracy), ease of use, perceived usefulness, and intention to use the MAM. The results (described in the previous section) showed that domain experts were very much satisfied about both understandability and accuracy of the model, as well as with its perceived usefulness. Additionally, regarding the model's ease of use, that is the effort it took to both understand it and to follow it, although not optimal the ratings were above average, which is considered to be normal since they were seeing and experiencing it for the first time. Finally, most participants said that they have the intention to use the model in future activities.

CONTENT

ACKNOWLEDGEMENTS

Before you lies my master’s thesis, representing the end point of a long period of learning and intensive work. This project was only possible by the great help that I had from different parties throughout the past eight months, that supported the execution and achievement of this research project, and therefore, deserve to be acknowledged and thanked.

First and most important, special thanks to Marco Spruit for the constant feedback and supervision. You gave me valuable insights which help me on pursuing the right paths and achieving better results. At the same time that you kept me motivated, you did that with a light and humorous attitude which was very much appreciated. Thus, thank you so much for all guidance, patience, and good talks these last eight months.

Second, thanks for Matthieu Brinkhuis for your supervision and feedback during this period. Although in a smaller proportion, all our meetings had a great impact on my research and helped me on achieving a better final product, and for that my special thanks.

Third, special thanks to Daniel Vijlbrief, which was my focal point within the Wilhelmina Kinderziekenhuis, and provided me with all the help that I needed. You always shared your time with me in moments of need, and you most certainly provided the means for making this project happened. For that, I am very thankful to you.

Finally, special thanks to all domain experts who agreed on helping me during the project and provided me with valuable insights and knowledge which allowed me to achieve the objective of this research. Moreover, to my family and girlfriend who supported me greatly during this period. A special thanks to you all.

1 INTRODUCTION

1.1 CONTEXT

Premature birth, defined as babies who are born under the gestational period of 37 weeks, is one of the major perinatal health issues in the world. According to Blencowe et al. (2013), the estimative is that more than 15 million babies are born prematurely every year worldwide, where more than a million die as a direct result of being born too early. Still according to Blencowe et al. (2013), that places premature birth as the second leading cause of deaths in children under the age of five, and the major cause of death during children's first month of life. For those who manage to survive, there are still risks of developing health problems and long-term disabilities, such as intellectual impairment, cerebral palsy, vision and hearing loss, and chronic lung disease (Howson, Kinney, McDougall, & Lawn, 2013). Additionally, many preterm babies also have behavioral sequelae, which affects other cognitive areas, such as: attention, academic progress, visual processing, cognitive functions, emotional control, and social interaction (Saigal & Doyle, 2008).

Preterm newborns are classified into three levels of risk, which are: extremely preterm (when the birth is in less than 28 weeks of gestation), very preterm (between 28 and 32 weeks of gestation), and moderate preterm (between 32 and 37 weeks of gestation), where a shorter gestation period is associated with a higher chance of major health complications (Blencowe et al., 2013). Those babies are normally admitted to a neonatal intensive care unit (NICU), where they receive all the care needed. During their hospitalization period, the newborns are monitored in almost every aspect, as they require real-time and clinical decision support medical data, to help on their treatment. Moreover, it also provides insights for physicians helping them on making decisions that could avoid major health issues (McGregor, 2013). Therefore, these babies are normally connected to a range of devices and monitors that constantly displays all health information data, assisting nurses and doctors in taking specific actions when needed. Those devices also give audible and visual alerts when the measurements surpass given thresholds, which indicates a potential risk to the patient. The patient's vital signs and other health information are recorded for future reference generating an enormous quantity of data every day. For example, babies usually have their heart activity monitored by electrocardiography (ECG), which can output up to one thousand readings per second, summing more than 80 million readings per day; patient's neurological activity can result in tens of million readings per day; smart infusion pumps can generate more than 3 gigabytes of data per month, and patients can be connected to more than one pump at the same time (McGregor, 2013). Those are just a few, among many other cases, of the data generated from these patients that could potentially hold undiscovered information, which can be used to: increase the knowledge and effects of the applied treatments, better understand cause and effect of diseases in newborns, and to apply predictive analyses, etc.

However, analyzing and extracting information of these huge amounts of data (most known as Big Data) is not a trivial activity. The term Big Data refers to really large, varied and complex sets of data that require most of the times sophisticated hardware and software solutions in order to store, analyze and visualize its information, and extract knowledge out of it (Belle et al., 2015). Big Data can be further explained by the three following aspects: Variety, Velocity, and Volume (McAfee & Brynjolfsson, 2012). Variety, makes reference to the huge amount of information that is part of Big Data, which may come from various sources (e.g. log files, web pages, written notes, images, e-mails, documents, sensor devices, etc.), in different data types (i.e. structured, semi-structured and unstructured), which makes it difficult to be handled by common analytic systems. Velocity, states that processes and applications working with big data should be built considering how fast new data is being generated. That creates new possibilities such as real-time or nearly real-time analysis, allowing the business to be more agile. Finally, Volume means that Big Data is not only large in variety but also in size. An IBM study ("10 Key Marketing Trends for 2017," n.d.) says that 25 Exabyte of new data is generated every day, and that 90% of the total amount of all human data was generated in the last two years alone.

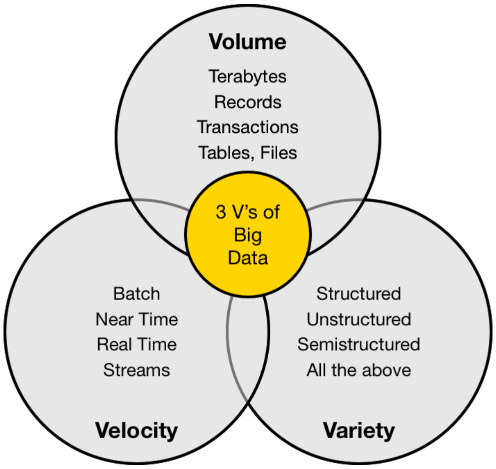


Figure 1 - The Three V's of Big Data (Russom, 2011)

Nevertheless, Big Data has many benefits and brings many possibilities to businesses within different industries such as retail, public sector, manufacturing and of course, healthcare (Sagiroglu & Sinanc, 2013). The importance of Big Data analytics has finally reached the healthcare years after its first appearance. The potential advantages for using data smartly and effectively includes: detecting and preventing potential diseases, predicting outcomes, estimating length of hospital stay, checking how likely a person will benefit from surgery, etc. (Raghupathi & Raghupathi, 2014). Additionally, the power of data has proven to be effective, not only to the patients (regarding their treatments), but also economically efficient. Belle et al. (2015) suggests that, if Big Data were used in a clever and innovative way, the USA healthcare sector would create more than $300 billion in value every year, where more than $200 billion would be made by reducing currently expenditure (e.g., detecting healthcare insurance fraud, creating more cost-efficient ways of diagnosing and treating patients, etc.). Additionally, a study done by Hackbarth (2012) illustrates the monetary waste in the healthcare segment from the United States, dived in six main categories: *failures of care delivery*, meaning the waste that is caused by bad treatment or lack of general adoption of better care procedures; *failures of care coordination*, which refers to the waste caused by fragmented care, resulting in patients readmission, complications after treatment, etc.; *overtreatment*, meaning the waste caused by excessive treatment over an patient, where it will not make it get any better; *administrative complexity*, which makes reference to the waste caused by excessive bureaucracy and misguided rules; *pricing failures*, which refers to the errors made in forecasting prices, or due to some political decision, that affects actual costs; and *fraud and abuse*, meaning the waste coming from frauds and scams in the billing, procedures, inspections, regulations, etc. Therefore, researches involving Big Data can, in the future, help in patients' treatment, and also on creating monetary value to hospitals, helping society in general.

The list of benefits of data science practices just grow in number, and every day it is being more used and recognized as a game changer in the way business is being done. However, most data are still being dismissed and underutilized, which hinders the possibilities for nurses and doctors to fully understand and better treat their patients. As per Wang & Hajli (2017), one of the reasons why the healthcare industry still struggling in implementing data analytics within their processes and departments, is that most businesses have difficulties to understand its economic potential. Nonetheless, applying data-driven approaches should not be only seen as something that will bring benefits to the business by creating monetary value, but as an essential and indispensable activity to actually save lives. As per Belle et al. (2015), due to the lack of efficiency in the gathering, processing, and using the information within the current healthcare systems, annually (only in the US), around 1 in 1000 people die as a result of some health condition that could have been treated. Currently, hospitals are still adapting and starting with data driven activities, where data analytics has the opportunity of being a transformative tool, that will improve the data exploration and knowledge discovery process, thus, helping in the delivery of care. Hence, that makes the field very promising, challenging, and full of potential for improvements and development.

1.2 Problem Statement

Wilhelmina Kinderziekenhuis (WKZ) is a children's hospital in the city of Utrecht – the Netherlands – that, for the past decade, has invested heavily in building an environment to collect, store, and manage huge amounts of patient's data gathered throughout the years. However, so far, not enough has being done to extract knowledge out of that data, what currently, is the hospital's main goal. The Neonatal department within the WKZ, experiences the issues and complications mentioned earlier, and for that, they seek experts to assist them in identifying potentially useful patterns in data that could improve patient's treatment. However, these professionals often do not have the specific background information to make the most of their analysis, like doctors would if the right tools were provided to them. Therefore, the data exploration process could take longer than needed only by the fact that the person would not have the proficiency, for example, to fully understand all the variables within the data. Thus, if field experts could also have a way and be supported to analyze the data, the knowledge discovery process could be greatly improved, as they are supposed to have a deeper understanding of the business, problems, and the variables and measurements within the data.

Extracting knowledge from data is not a trivial task and the process is composed of many phases and activities. These vary in complexity and importance for each specific and distinct scenario, dataset, and problem. For larger and massive datasets, like the ones within the WKZ, one of the most problematic and exhaustive tasks is to prepare the data, by removing inconsistences, integrating tables, transforming its variables and values, before applying the different statistical methods and techniques to obtain useful information from it. It requires a good understanding of the business goals and project's objectives, and also a good understanding of the data available. Moreover, since data preparation is an activity as important as data mining itself (as it will be further explained during this document), lacking in doing it properly can hinder or even compromise the entire data analysis activity. Additionally, this is a task that requires, most of the times, technical knowledge for transforming and integrating data, which is done usually by running pieces of code, such as in SQL, Python or R. Hence, even for simple questions and hypotheses that would not need deeper statistical knowledge nor the application of any complex DM method in order to find its answers, preparing the data is essential for a good analysis, and to avoid getting bias by not considering 'dirty' data into it. Therefore, if domain experts could do some data preparation tasks more easily and intuitively, the data analysis and exploratory knowledge discovery process could be facilitated.

1.3 Reading Guide

This document contains the entirety of a thesis project which had the objective of facilitating the knowledge discovery process for domain experts. Since some chapters may be more interesting than others for some readers, the report structure is described below with the purpose of helping on navigating throughout this thesis.

First, an extensive literature review (described in Chapter 2) was made, with the purpose of addressing state of the art, and important background information for further development of the project. Moreover, Chapter 3 contains the research process that was followed, as well as the objectives and research questions that were answered throughout the report. Next, in Chapter 4, a qualitative study is described which had the purpose of understanding what, in fact, domain experts know about the Knowledge Discovery process. In Chapter 5 a data quality assessment over the datasets available within the WKZ is described. Furthermore, Chapters 6, 7 and 8 were assigned to the actual development of a Meta-Algorithmic Model to be used as a guideline by domain experts to perform KD (main artifact from this project). Finally, the main findings and conclusions of this thesis can be seen in Chapter 10.

2 LITERATURE REVIEW

In this chapter, existing scientific studies are reviewed and analyzed to provide an overview of the relevant literature related to the thesis' subject of knowledge discovery for domain experts, and also to understand the relevant concepts and ideas that could support the development and the success of this thesis project. The main points discussed below are the knowledge discovery process and its importance for today's businesses, the exhaustive step of preparing data prior to the analysis, an overview of the existing analytical tools, and why domain experts should be allowed and empowered to analyze data themselves.

2.1 KNOWLEDGE DISCOVERY AND DATA MINING

Knowledge Discovery and Data Mining are two well-known and still growing fields that, with the advancements of data collection and storage technologies, emerged and expanded with great strength by the many possibilities and benefits that exploring and analyzing data can bring. Knowledge Discovery (KD) is defined by Fayyad, Piatetsky-shapiro, & Smyth (1996) as "the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". Thus, based on that definition, it is implicit that KD consists of more than one step. Additionally, Data Mining (DM) is defined by Luo (2008) as "the process of analyzing data from different perspectives and summarizing it into useful information". Hence, although sharing the same goal (turning data into knowledge), assuming that KD and DM are the same is a misconception. KD is an overall process of extracting information from data which can be turned into valuable insights, having the application of DM techniques within it. DM on the other hand, can be addressed as the application of methods, techniques and specific algorithms to extract those useful patterns out of the data. Nowadays, there are several knowledge discovery frameworks already well stablished in the market, such as KDD (Knowledge Discovery in Databases), Three-Phases Model (3PM), and CRISP-DM (Cross-Industry Process for Data Mining). The latter, for example, is composed of six phases, where only in the fourth step there are applications of Data Mining routines. Although DM is in the core of the KD process, it is estimated that it only takes approximately 20% of the total project time (Goebel & Gruenwald, 1999). All phases before DM activities (in short) aim to understand what has to be done, what data is available and how it is composed, and to prepare it and transform it to be finally analyzed.

The human race proved, throughout history, to never be satisfied by their past or current accomplishments for a long period of time. Hence, we are always seeking better ways of doing something, achieving better results, surpassing a given boundary, and accomplishing something unthinkable. With the technology progress, many things that fifty years ago seemed impossible are now being realized within a blink of an eye. In medicine, the advancements of technology have helped humanity to grow strong and healthy, however, it still far from optimal and far from human ambitions. Medicine and technology were once two distinct areas of research, but they are now closer than ever. Innovative solutions such as nanotechnology, robots, artificial intelligence, etc. are appearing and showing that medicine can be reinvented and become smarter and more efficient every day. As mentioned earlier, Data Analytics and Big Data are a multidisciplinary hot topic nowadays, which, by means of KD activities, help businesses and organizations to achieve better results, perform daily tasks more efficiently, understand their market and clients more clearly, etc. Although in medicine KD's application is currently not as advanced as in other areas, it is already being used to aid nurses and doctors on their daily activities and patient's treatment.

These professionals are requested and tested every day by their many patients with many health problems and stories. To come up with a diagnosis or treatment, many variables need to be taken into consideration. As mentioned in the beginning of this essay, data is growing in size, complexity, variability, and volume. All the equipment connected to a patient is outputting valuable data that could contain hints on how to treat the patient, or what could be the cause of that illness and why he is felling that way. Moreover, it is impossible for humans' brains to compile as much information as a computer, and thus, mistakes can happen (due to ignoring some important factor in the illness' analysis, for example). A recent study reported that medical errors is the third leading cause of deceases in the United States (Makary & Daniel, 2016). It states that in the year of 2008, 180.000 deaths were due to some medical error, and therefore, could have been prevented. Hence, the use of KD and DM activities within the healthcare segment have become a trend and a very promising line of research. It is being used for innumerous scenarios, such as prediction and/or detection of diseases, better diagnosis and decision support, adverse drug events, etc. Examples of these scenarios and applications of KD and DM can be seen in several studies, as follows: Soni, Ansari, Sharma, & Soni (2011) and Rani, Govrdhan, Srinivas, Rani, & Govrdhan (2010) illustrate how DM techniques, such as decision trees, clustering, neural networks, can be used to predict whether a patient is likely to have a heart attack; Kharya (2012) describes how extracting knowledge through data mining techniques can be used on breast cancer detection and diagnosis in a non-invasive manner. Additionally, Tsipouras et al. (2008) used KD and DM techniques to develop a Decision Support System (DSS) for diagnosis of coronary artery disease also in a non-invasive way, that is, only by looking at easily obtained data (e.g. patient's medical history, blood samples, physical evaluation, etc.); Harpaz et al. (2012) describe the importance of DM to Pharmacovigilance for the recognition of post-approval adverse drug effects (which is responsible for more than 2 million hospitalizations, injuries and deaths per year only in the United States), and present DM approaches for the detection and analysis of it. These are just a few of the many possible applications of KD within medicine and how it can positively affect physicians, nurses, patients, and society.

2.2 Data Mining Methods and Tools

The core of Knowledge Discovery is composed by Data Mining methods and techniques. Data Mining is an intersection of many disciplines, such as statistics, computer science, mathematics, and econometrics. (Raghupathi & Raghupathi, 2014). Its taxonomy, based on Tsoumakas, Katakis, & Vlahavas (2010), shown in Figure 2, depicts the variety of methods that compose DM. Additionally, it illustrates the two main DM orientations: verification-based and discovery-based. Verification methods are used when a hypothesis is proposed, and it needs to be evaluated and proven. On the other hand, Discovery-oriented methods are not based on hypothesis or previous assumptions. They are part of an exploratory approach to be later interpreted by users.

DM has many orientations and goals, which leads to many methods and techniques to perform a set of specific tasks. Under the DM Discovery orientation there are two main approaches: Prediction and Description. Predictive data mining refers to using variables within the dataset to predict unknown outcomes for variables of interest. Descriptive Data Mining (DDM), on the other hand, has the objective of finding patterns, associations, modifications, peculiarities and noteworthy structures, in order to describe the data for further human interpretation (Kantardzic, 2011). For the latter, the goal is to understand large sets of data by finding hidden patterns and relationships between the many variables, which is possible by applying different DM methods to the analysis. Moreover, DM has two main learning approaches: *Supervised* and *Unsupervised* learning. Supervised learning methods aim to discover the relationship between input variables (x) and output variables (Y), so data can be used to predict a target outcome from a set of input data. In this type of learning, both input variable (x) and its correspondent output (Y) are known and the model is trained to fit the mapping function Y = f(x), identifying suitable predictors of Y. Therefore, predictive DM uses a supervised learning approach, and some of the methods for doing so are Classification and Regression. Unsupervised learning on the other hand, is when no outcome variable (Y) is known. Hence, the goal is fuzzier, once that the model cannot be trained on an already known true answer. The objective then, is to learn more about the data trying to discover interesting things about the measurements in an exploratory manner (Tsoumakas et al., 2010). Therefore, DDM is oriented as an unsupervised learning approach where, rather than predicting outcomes for specific scenarios, it tries to explain what happened by providing data insights without much user direction. Some of these unsupervised exploratory methods are: Principal Components Analysis, Clustering, Association Rules, Summarization, Dependency Modeling, and Deviation Detection.

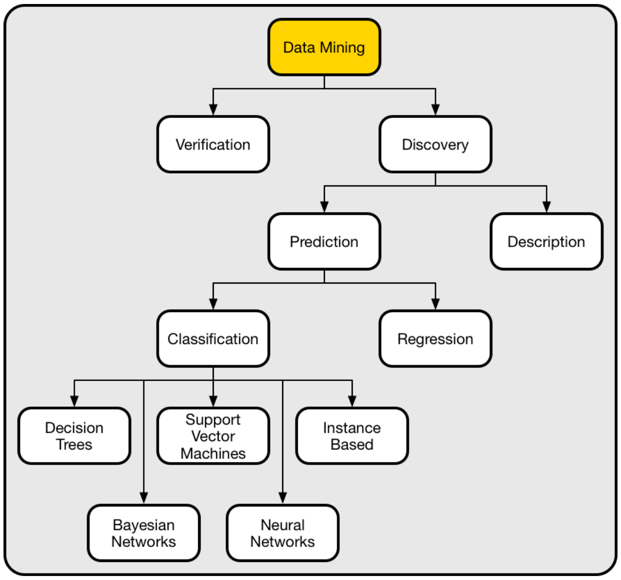


Figure 2 - Data Mining Taxonomy (Tsoumakas, Katakis, & Vlahavas, 2010)

Nowadays, there are many DM tools available with different characteristics, focusing in different user groups, with different analytical methods, etc. An recent study made by analysts from Gartner Inc. (Linden et al., 2017) evaluated top-rated commercial data science platform (software application that can produce all types of data science solutions, and integrate them into business processes, departments and products) vendors, based on the following fifteen capabilities: *data access*, *data preparation*, *data exploration and visualization*, *automation*, *user interface*, *machine learning*, *other advanced analytics*, *flexibility, extensibility and openness*, *performance and scalability*, *delivery*, *platform and project management*, *model management*, *pre-canned solutions*, *collaboration and coherence*. By the end of the research, sixteen vendors were highlighted, and their data science solutions' strengths and weaknesses were explained. The current vendors, which were defined by Gartner as leaders in this study are KNIME, RapidMiner, SAS and IBM. Moreover, it should be noted that pure open-source platforms were not included in the study (this will be discussed further in this chapter). An overview of the major technical strengths and weaknesses from the reviewed tools is shown in the following tables:

	Strengths
Many data science platforms	• Have a great variety of machine learning techniques
Some data science platforms	• Provide prescriptive analytics • Have good collaboration capabilities • Support great variety of data types • Have good flexibility and scalability • Have support to citizen data scientists • Have good project management and model management capabilities
Few data science platforms	• Support Big Data analytics

Table 1 - Gartner Data Science Platforms Strengths

	Weaknesses
Many data science platforms	• Lack of visualization and exploration features • Problems with data access and data preparation
Some data science platforms	• Have limited native analytic operators • Do not support (or not entirely support) some common programming languages such as Python and Scala • Have scalability and performance problems • Problems with large datasets • Require high technical knowledge • Hard to use • Undetailed documentation

Table 2 - Gartner Data Science Platforms Weaknesses

As seen in the tables above, most tools are equipped with several machine learning techniques to enhance DM capabilities, however, several platforms also have problems with data visualization and exploration, data access, and data preparation features, which automatically hinders the understanding of the given data or analysis, and therefore, the results for the DM activities. Moreover, only some of the tools support citizen data scientists (not totally experts in data science but still can use data science technologies to create data models and analyze data) while the rest still require a deeper technical knowledge from the user. Additionally, only few tools support Big Data analytics, which is a problem when dealing with large and continually growing datasets.

A similar study was made by Forrester (an American market research company focused in providing insights about the eventual impact of technology for the public) over predictive and machine learning solutions (Gualtieri, 2017). Those solutions are defined by the author as "*Software that provides data scientists with 1) tools to build predictive models using statistical and machine learning algorithms and 2) a platform to deploy and manage predictive production models*". Again, the strengths and weaknesses of data science platforms were evaluated based on some evaluation criteria, such as: *model training scalability, deployment options, workload scalability, data preparation, algorithms, model management, set of tools for data scientists, strategy, and market presence*. Fourteen vendors by the end of the study were highlighted. Moreover, differently than Gartner, Forrester was not really concerned about tools supporting citizen data scientists, as it states: "*... an enterprise should not think that this* [tools that support citizen data scientists] *will replace real data scientists, because there are too many complexities of model building, such as feature creation, model evaluation, overfitting, and creating ensembles*" (Gualtieri, 2017). Therefore, some of the evaluation criteria, such as user interface and data exploration were based on users with a more data-science expertise, which explains the reason that no word was said about a tool being hard to use or about the possibility of it being used by less technical users.

Both of the studies described earlier highlighted some of the data science platforms that are very well established in the market, however, no evaluation of pure open-source platforms such as R and Python was made. Given that, a survey from O'Reilly (King & Magoulas, 2016) with almost a thousand responders (working in data-driven activities), from different industries, from forty-five different countries explored, among other things, which tools data scientists and analysts are really using to perform their daily activities. From the sample, 69% of the responders said that they use Excel or SQL to analyze data; 57% use R and 54%, Python. Tools like Tableau and Teradata also appeared with a small share of the responders. Additionally, Spark was the most used tool for Big Data Analytics, and SAS, which was pointed as a leader from Gartner and Forrester, appeared with only 5% of the sample share. Nevertheless, 90% of the sample spend some time coding and 80% use at least Python or R. Moreover, nothing was mentioned about less technical users. Thus, although many data science platforms are being developed and enhanced, people still prefer traditional methods and open source tools. The reason behind that is not entirely proven, however, the fact of data-driven activities being performed mainly by data scientists (with higher technical skills), together with the tools being sometimes hard to use, overloaded with methods and buttons, not supporting open source platforms, and not being entirely free, may explain in part why that is the case.

2.3 CRISP-DM Framework for Knowledge Discovery

Knowledge Discovery is an overall process with the objective of extracting interesting and usable information from data. Hence, applying algorithms without a prior understanding of the business, data structure, the variables within it, etc., can hinder (or even totally inhibit) the benefits of KD. Thus, frameworks such as KDD, SEMMA (Sample, Explore, Modify, Model, and Assess) and CRISP-DM appeared to be used as a guide on KD activities. A comparative study between these three frameworks concluded that both CRISP-DM and SEMMA are extensions of the KDD (developed back in the 1990s), and therefore more complete (Azevedo & Santos, 2008). Also, as per the author, SEMMA and CRISP-DM are equivalent in completion to guide a user in a KD process. Thus, as CRISP-DM is used increasingly throughout the industry, and as a personal preference of the author of this thesis, it has been chosen as the supportive KD framework for this study. The CRISP-DM was built to make the knowledge discovery process less costly, more generalizable, reliable, manageable, and fast (Wirth, 2000). It consists of six phases, as shown in Figure 3. There are two aspects worth mentioning before diving into the purpose of each phase: (1) Following the order of the steps is not compulsory. It only shows the natural sequence and dependencies between phases. However, the choice to follow it or not always depends on the project at hand; (2) the outer cycle-circle represents the continuous work and life of a data mining project. It means that, even after the deployment of an artifact, new activities can emerge and the cycle can start over again (Wirth, 2000). Below, each phase from the CRISP-DM is shortly described.

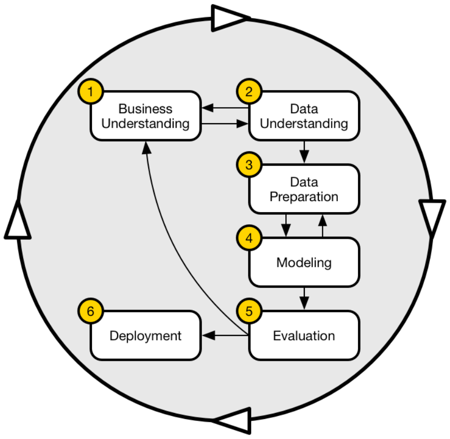


Figure 3 – The CRISP-DM Process

1. *Business Understanding:* Before the data mining activities can begin, an understanding of the business is needed. This first phase aims to collect, from a business perspective, the objective, all requirements, the success criteria of the project, and create a project's plan for achieving the desired results. That would be later extended as a knowledge discovery and data mining problem.
2. *Data Understanding:* After the project is defined, the next step is to collect and get familiarized with the available data. In this phase, the objective is to dive deep into the tables and databases to understand what data is accessible, its structure, inconsistences, quality problems, etc. Also, this phase allows to start playing with the data, by creating small subsets that may lead to some hidden information. Both Data Understanding and Business Understanding phases are closely connected, since that no data mining goal can be defined without knowing in advance what data is available to work with and extract information from.
3. *Data Preparation:* This is the phase where all activities focus on creating the final dataset, from the initial data, to be further analyzed. As per the framework, its main tasks are: data selection, data cleaning, data construction, data integration and data formatting.
4. *Modeling:* Once the data is ready to be analyzed, several modeling methods and techniques are applied for one specific problem at a time (since no perfect solution or guideline exists), and their optimal parameters are found. These parameters can then be used to, for example, predictive analysis on unseen data. As per the model in Figure 3, there is a close link between Data Preparation and Modeling. The reason for that is because sometimes, specific modeling methods require specific data formats, so the data has to be prepared again, and so on.

5. *Evaluation:* In this phase, all models and achieved results are evaluated to ensure its quality, that all requirements are met, and the business objectives achieved. Based on the results, a decision about whether the artifact should be deployed or not is given.
6. *Deployment:* Finally, the Deployment phase is where the knowledge obtained from all previous steps are organized and presented to the customer. There are several ways of doing this. It could be either in a report format or implementing the created model, so it would be possible to execute it over and over again.

2.4 DATA PREPARATION

The volume of data that is being generated and stored every day throughout the world is huge. In the healthcare segment for example, given the importance of data-driven activities for the business, enormous amounts of patient's data are being (at a very fast pace) recorded into their databases. However, a big part of that data can be considered to be dirty, since its gathering process still rely very much in human factors, for example, connecting a machine properly to a patient, and make sure it still connected whenever the patient moves around. A study made by Kim, Choi, Hong, Kim, & Lee (2003), in which a taxonomy for dirty data is proposed, says that some of sources for problematic data are input errors and update errors (either by a computer or human), data transmission errors, and either some bugs that may occur in data processing tasks. Hence, as per the taxonomy, three main types of dirty data are illustrated: missing data, wrong data, and unusable data. As the main focus of this research, the importance of Data Preparation is huge since it is an activity that can be more time consuming than data mining itself, and sometimes even more challenging, as stated by Zhang, Zhang, & Yang (2003). Still according to these authors, there are three main aspects to enforce the importance of Data Preparation: (1) the available data can be impure, that is, incomplete (e.g. containing missing values), noisy (including errors and outliers), and inconsistent (e.g. redundancy, and discrepancy in its values). That may reduce the chances of finding hidden and useful information; (2) by means of data preparation, the quality of data is enhanced, and the volume of the data that will be further analyzed will get smaller. Thus, the efficiency of data mining methods and techniques increases; (3) by increasing the quality of data, the quality of patterns also increases, which allows, for example, to get incomplete data recovered, to have data conflicts resolved, etc.

Trying to overcome the problems mentioned above and enhance the quality of the analysis, the CRISP-DM framework provides five main data preparation tasks, as shown in Figure 4, that could help on preparing the data for the modeling phase (DM activities). Below, each of these activities are going to better explained.

Data Selection: As its own name says, this activity includes addressing what data should be selected to be utilized in the analysis. Although it seems a rather simple task, many constraints should be considered before deciding which data to bring, and which data to leave out of the analysis. Some examples of selection criteria are: data mining objectives, quality of data, and technical constraints (i.e. data processing power). Many activities can be done within this step and there is no optimal one for all scenarios. As per the reference manual of CRISP-DM (Pete et al., 2000), some of these activities are: collect additional data when appropriate, perform tests to check whether a field should be included or excluded, make use of sampling techniques, and analyze the rationale behind the selection criteria.

Data Cleaning: As an extension of the previous activity, the data cleaning activity makes sure that all rationale and selection criteria that were defined are used, and it has the objective of reduce the noise and remove inconsistences from the data, ensuring its quality. Some of the activities listed by Pete et al. (2000) to be performed during this phase are: decide how to deal with a specific type of data noise and outliers (fixing, removing or ignoring them), decide how to deal with different types of variables and values (e.g. binary values, temporal data, etc.), and if needed, redefining selection criteria.

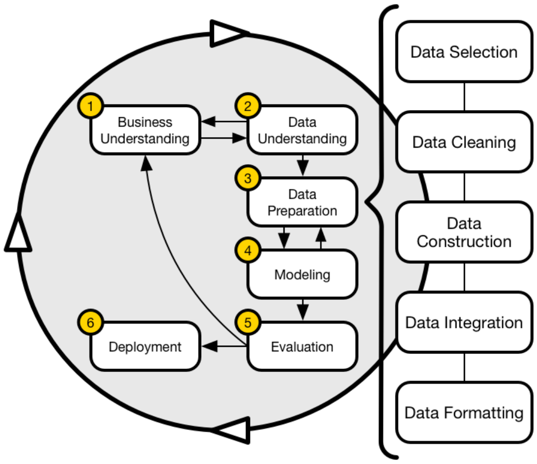


Figure 4 - CRISP-DM's Data Preparation Step

Data Construction: Data construction involves transformation activities such as: normalize the data, construct new attributes or derived attributes (such as calculating the body mass index, based on weight and height). Based on DM goals, data transformation may be needed to help on identifying patterns, either by applying DM methods or by using some visualization technique. Nevertheless, new attributes can be computed during the analysis, however, depending of the data size and/or computing processing capacity, letting "unnecessary things" to the modeling phase may cause performance issues.

Data Integration: Integrating data means combining data that may be spread across many tables within a database (or even different databases) so it can be more easily retrieved and analyzed. Again, when dealing with a small amount of data and only one table, data preparation activities may not be something over exhaustive and complicated. However, in the case of many hospitals, such as the WKZ (which has several highly intensive data generating environments), where there are innumerous tables and different databases, the necessity of integrating the data, so tables may be combined in order to extract knowledge from it, is huge. Moreover, this is also not a trivial task (especially for a domain expert with no technical knowledge) once that its databases schemas may not be optimal for it, and variables and values may differ from one table to the other. Therefore, data quality is also very important in this case.

Data Formatting: Formatting the data means syntactically changing values, without modifying their meaning. Activities for doing so can be rearranging attributes, reordering records, changing from upper to lower case, etc.

As seen in the image above, the data preparation step is the sequential phase after the understanding of the business and the understanding of the available data, in which, without knowing what the main business' objectives and the data mining goals are, and without a great understanding of the data at hand, preparing the data properly could be a shoot in the dark. Moreover, although there is no order to perform these tasks, they are pretty much linked, as many actions and activities within each task may require some of the others to be accomplished as well. Furthermore, a lot of time is being expended in these activities, which can be done repeatedly over and over again until the quality of dataset that will be used in the analysis is acceptable. Therefore, creating generic approaches that overcome some of the common problems and frequent situations within this phase can be very helpful to facilitate the whole KD process, especially for domain experts that usually do not have the knowledge nor the time for doing data preparation themselves. Some studies were made, in which methods, techniques, and algorithms for improving or automating the data preparation step were developed. Some examples are a study made by Hmamouche, Ernst, & Casali (2015) which aims to automatically detect and remove outliers (which are observation points that vary a lot from the others, which usually represents wrong data) from the data, and then apply discretization methods in the variables and values to facilitate further DM analysis; integrating data is also a big problem in data preparation, especially when dealing with large sets of data spread across many tables, in one or more databases. If the relationship between primary-keys (PK) and foreign-keys (FK) is not well defined, finding them may consume a very long time even for specialists. Based on that problem, some studies were made where they focus on discovering this relationship between PK-FK from the many tables of a database automatically, and therefore, facilitating the data integration process (M. Zhang, Hadjieleftheriou, Ooi, Procopiuc, & Srivastava, 2010; Rostin, Albrecht, Bauckmann, Naumann, & Leser, 2009); additionally, by integrating tables, there is a chance of having duplicated variables and columns, and even if the variables' names are different, its content can be the same, and therefore, they need to be removed. Thus, a study made by Shahri & Barforush (2004) try to overcome this issue by automatically detecting duplicated values based on a set of rules.

Nevertheless, knowing how to prepare, and when the data quality achieved is good enough is not a trivial task, as the term 'data quality' is very subjective to the project and data at hand, once that the data quality is only good when it can be used for the needs of a given purpose. More will be said about that in the following sections. Hence, most data preparation actions are taken after a project is defined, or a hypothesis or question that requires an answer. However, most of the data issues can be unrecognizable or hidden to the naked eye in daily practices. In conclusion, preparing the data still a very exhaustive and complicated task with a lot of room for improvements. It is also a determinant factor for how good, reliable, and valid the results of data analytics activities will be, and therefore, is a task that requires a very deep and careful analysis for ensuring that a good data quality comes out from the huge dirty data that is found within the databases.

2.5 Knowledge Discovery and Domain Experts

As mentioned above, KD is a complex and extensive process where DM is only a step within it. Even so, data driven activities keep on focusing specially in DM, while the other phases are underestimated and their importance is not really taken seriously (Tsai, Lai, Chao, & Vasilakos, 2015). That creates a deficiency in what is expected from the business and what is actually delivered. For example: data scientists and researches normally focus on using innovative solutions, while business analysts want optimal ones for their organization; data scientists and researches identify achievements and findings from a technical perspective, while business analysts need useful information that actually add some value to the business. A few ways of reducing some of these problems (which relates with the objective of this research) includes involving domain experts to the KD process; balance findings between technical and business perspectives; taking in consideration environment aspects in the KD process, such as domain, organizational, social factors, etc. (Cao, 2012). Moreover, organizations then seek to perform Actionable Knowledge Discovery (AKD) instead of simple KD, or in other words, extract knowledge from data that actually supports decision-making and action-taking activities. To do so, some researches came up with frameworks and solutions for applying a so-called *domain-driven data mining* (D^3M), which aims to overcome the gaps mentioned earlier by incorporating into the KD process environment factors, domain knowledge, human interactions, measurements to check whether the result is actionable or not, etc. (Cao & Zhang, 2007). Thus, data and domain knowledge are used together to identify and extract patterns that can be actually used as a decision-making information by final users.

As per Cao (2012), the concepts which create the basis for applying AKD are: domain problems that are too complex to be handled alone by existing data mining techniques; ubiquitous intelligence which makes reference to all knowledge and information surrounding the AKD process; theoretical foundation that allows the application of the AKD process; techniques that support and consolidate the environment's omnipresent intelligence; and finally, actionable computing, which makes reference to the actual power of discovering actionable knowledge. Furthermore, still according to Cao (2012), the real power of AKD is attributed to the **ubiquitous intelligence** concept and its inclusion to the process. It can be categorized by the following features: *In-depth data intelligence*, which refers not only to the task of extracting patterns from transactional or demographic data, but the power of adding into the analysis real-time data, multidimensional data, business performance data, environmental data, etc.; *domain intelligence*, which involves extracting all relevant knowledge from the project's domain, such as expert knowledge, background information, possible constraints, etc.; *organizational and social intelligence*, which refers to all organizational and social information that can be extracted to and added into the analysis, such as business processes and rules, organograms, etc.; *network and web intelligence*, which refers to hidden information throughout, for instance, distributed systems, network structures, online communities, emails, etc.; and *human intelligence*, which refers to the participation of domain experts into the knowledge discovery process, by means of supervising, evaluating, sharing knowledge, sharing expectations and priorities, etc.

Domain-driven data mining, and the intensive involvement of domain experts into the KD process, is proven to be very important in order to extract actionable knowledge from data. When performing supervised learning problems, models can be trained already knowing a true answer, as explained before. Even so, without prior domain knowledge, data scientists could miss or misjudge some of the variables, and even so, have a "good" outcome in a technical perspective. Therefore, the model evaluations are done by (or together with) a domain expert, mainly to see whether an outcome makes sense or not. If the evaluation finds something wrong with the model, the process has to start over, increasing the cost to the business. Furthermore, especially when dealing with an unsupervised learning problem, domain knowledge is indispensable in order to perform the exploratory analysis since no true answer is known, and many times the data scientist does not know what he is looking for.

3 RESEARCH DESIGN

As illustrated in the previous chapter, it is clear that KD is not a trivial activity and demands a lot of time and knowledge in order to be of some value. As it is still a very growing field of study, most of the tasks, tools, techniques and methods that were created aiming to facilitate the extraction of useful information from data, requires somehow a minimum technical level (e.g. some programming skills). Another important aspect shown in the previous chapter was regarding the Data Preparation phase from the CRISP-DM, which still very time consuming and complicated, especially for non-technical people. Finally, the difference between KD and AKD was also highlighted, and the importance of using domain knowledge within an analytical task for being able to extract useful and actionable knowledge out of the data was emphasized.

As defined by Spruit & Lytras (2018), Applied Data Science is "*the knowledge discovery process in which analytical applications are designed and evaluated to improve the daily practices of domain experts*". Given the power that data analytics has, and although data scientists and analysts are very much required in today's market, KD should not be an exclusive activity for those professionals, especially because (as stated above) they usually lack on ubiquitous knowledge when compared to domain experts. Hence, given the current situation mentioned in the previous chapter, the problems and drawbacks of the KD process regarding pre-processing the data, and the necessity of having more domain knowledge within the KD process, the overarching research question (MRQ) for this research is:

How can the data preparation phase, embedded within the knowledge discovery process, in an applied data science context, be facilitated so domain experts can explore an analytical problem more easily and intuitively?

3.1 OBJECTIVES AND RESEARCH QUESTIONS

To help domain experts to analyze, understand and extract knowledge from data, and by this, improve their daily practices, this thesis has the following objectives:

1. Develop a Meta-Algorithmic Model, based on CRISP-DM (Cross Industry Process for Data-Mining), that can help domain experts to follow a Knowledge Discovery process and prepare data for an exploratory data analytical task.
2. Analyze the data preparation phase of CRISP-DM, aiming to understand and verify the necessity, importance, and applicability of the all sub-steps provided by the framework, when focusing on domain experts.
3. Perform a qualitative study in order to understand what domain experts really know about KD.

Additionally, to succeed in delivering valuable results throughout this study, six sub-research questions were investigated and overcame. They are:

SQ1: *What is the current understanding that domain experts have on Knowledge Discovery and Data Mining activities?*

SQ2: *What are the risks and benefits of allowing domain experts to analyze data themselves?*

SQ3: *What is the current quality of the available data within Dutch (academic) Hospitals?*

SQ4: *What aspects have to be considered, regarding the Data Understanding phase of the CRISP-DM, so it can be adapted for the domain experts' needs?*

SQ5: *How and which data preparation step's activities should be included in the model, and how they can be adapted for the domain expert's needs?*

SQ6: *What is the best way to guide domain experts throughout the Knowledge Discovery process, so they can most likely successfully accomplish the analysis?*

3.2 Design Science Research

Throughout this study, a Design Science Research (DSR) approach is being followed. DRS is defined as a "research activity that invents or builds new, innovative artifacts for solving problems or achieving improvements" (Livari & Venable, 2009). It is based on three cycles that bridges environment, research design activities, and knowledge base foundations, as shown in Figure 5.

The DSR starts with the Relevance Cycle, supplying the project's context, the application's requirements and the acceptance criteria for the achieved results (Hevner, 2007). It bridges the application domain, consisting of people, organizational systems and technical systems, with the research development towards the realization of a common goal. Additionally, continuous interactions with the application domain are possible in order to gather domain knowledge to successfully attend to business expectations. This cycle starts and ends the whole process, since the produced artifact will also be field-tested in the application domain. Moreover, based on evaluation's outcomes, new interactions of the Relevance Cycle are possible.

The Rigor Cycle provides the knowledge base to the DSR. Scientific theories and methods are selected and applied in the development of the research project (Hevner, 2007). This has the objective of identifying opportunities, research gaps, methods to approach a given problem, and to create a baseline of what has been developed already, so innovative solutions can indeed be produced.

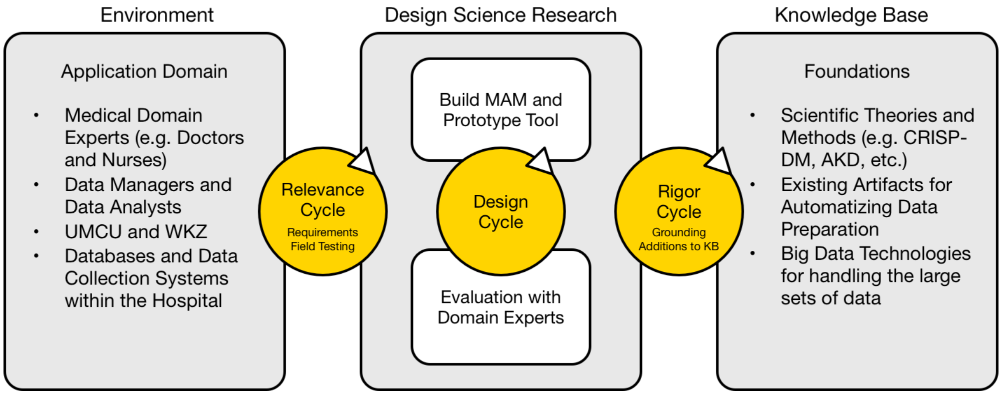

Environment
Application Domain
• Medical Domain Experts (e.g. Doctors and Nurses)
• Data Managers and Data Analysts
• UMCU and WKZ
• Databases and Data Collection Systems within the Hospital
Design Science Research
Build MAM and Prototype Tool
Design Cycle
Evaluation with Domain Experts
Relevance Cycle
Requirements
Field Testing
Rigor Cycle
Grounding
Additions to KB
Knowledge Base
Foundations
• Scientific Theories and Methods (e.g. CRISP-DM, AKD, etc.)
• Existing Artifacts for Automatizing Data Preparation
• Big Data Technologies for handling the large sets of data

Figure 5 - Design Science Research Framework

Finally, in the Design Cycle, all activities for constructing the project artifact are made, using the requirements and all knowledge collected from the Relevance cycle and the methods and scientific contributions selected and learned from the Rigor cycle (Hevner, 2007). In this phase, developing and evaluating the evolving artifact is mandatory, which means that several interactions of the Design Cycle may be needed in order to deliver, in the end, a useful and reliable solution.

3.2.1 Research Design and Research Questions

So far in this chapter the research method approach, proposed artifacts, and the evaluation method were introduced and explained. Furthermore, in order to answer the main research question of this thesis, first the sub-questions that were given in the beginning of this essay had to be answered, as they support and help on the achievement the goals from this research, including on answering the MRQ and in the development of the proposed deliverables. The following table illustrates, for each sub-question, which activity(s) is responsible for gathering information that was used to answer it.

Sub-Questions	*How to answer it*
SQ1: What is the current understanding that domain experts have on Knowledge Discovery and Data Mining activities?	Literature Review and Interviews
SQ2: What are the risks and benefits of allowing domain experts to analyze data themselves?	Literature Review and Interviews
SQ3: What is the current quality of the available data within Dutch (academic) Hospitals?	Interviews and Data Analysis Activities
SQ4: What aspects have to be considered, regarding the Data Understanding phase of the CRISP-DM, so it can be adapted for the domain experts needs?	Analysis of Scientific Theories and Methods, and Interviews
SQ5: How and which data preparation step's activities should be included in the model, and how they can be adapted for the domain expert's needs?	Analysis of Scientific Theories and Methods
SQ6: What is the best way to guide domain experts throughout the Knowledge Discovery process within a tool, so they can most likely successfully accomplish the analysis?	Analysis of Scientific Theories and Methods, Interviews, and Prototyping

Table 3 - Research questions, data collection, and development activities

The project was divided in two phases, the first consuming around 30% of the total project time has dedicated to the extensive literature study shown in the previous chapter. The second phase was dedicated to the artifact development and answering all research questions. Additionally, to better visualize how and when these activities were performed, and also the dependencies and parallelism between them, Figure 6 depicts an overview of the main activities within the two phases of the project, and their connection with each sub-question. As per the image, answering the first two sub-questions is crucial before diving into the model's construction (and also to answer the rest of the sub-questions), as they provide important insight from literature as well as information retrieved by domain experts by means of interview's content analysis. Furthermore, to answer sub-questions 3 to 6, constant analysis of scientific theories and methods were realized in parallel with the model construction. Furthermore, in order to help the evaluation procedure, a prototype tool was developed replicating the activities depicted within the MAM. This will be better explained throughout this document.

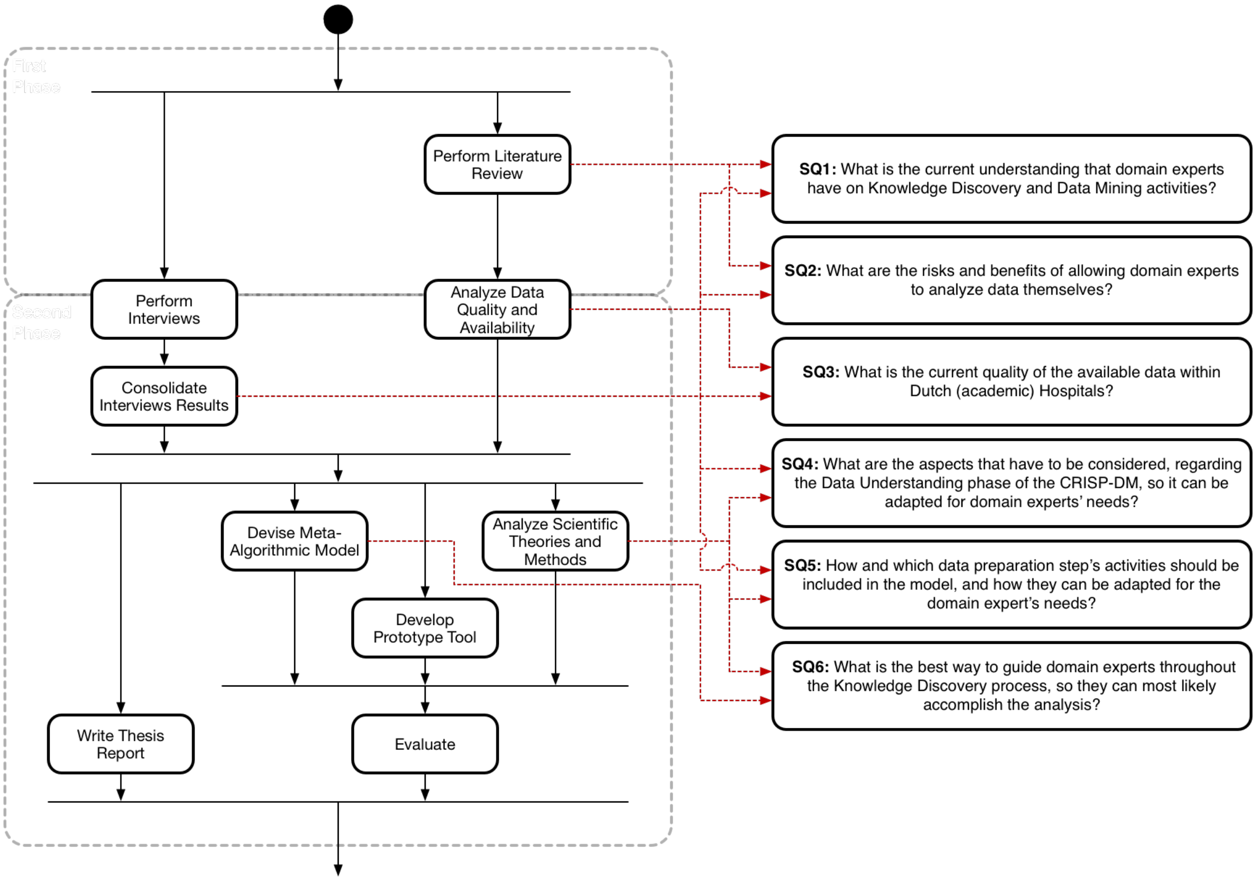

First Phase
Second Phase
Perform Literature Review
Perform Interviews
Analyze Data Quality and Availability
Consolidate Interviews Results
Devise Meta-Algorithmic Model
Analyze Scientific Theories and Methods
Develop Prototype Tool
Write Thesis Report
Evaluate
SQ1: What is the current understanding that domain experts have on Knowledge Discovery and Data Mining activities?
SQ2: What are the risks and benefits of allowing domain experts to analyze data themselves?
SQ3: What is the current quality of the available data within Dutch (academic) Hospitals?
SQ4: What are the aspects that have to be considered, regarding the Data Understanding phase of the CRISP-DM, so it can be adapted for domain experts' needs?
SQ5: How and which data preparation step's activities should be included in the model, and how they can be adapted for the domain expert's needs?
SQ6: What is the best way to guide domain experts throughout the Knowledge Discovery process, so they can most likely accomplish the analysis?

Figure 6 - Research Questions and the main activities from the project

3.3 RELEVANCE

Two main topics regarding this thesis are believed to provide a significant relevance for this research and support its development. They are explained below:

Facilitating knowledge discovery for domain experts: As seen in the prior sections, terms like Knowledge Discovery and Data Mining are already well-known and their importance in today's society and organizations has been proven. However, extracting information from large sets of data is not a trivial task and most people do not know how to do it. Knowledge Discovery emerged with the advancements of data management and storage technologies, and although its main purpose still is extracting knowledge out of data, the methods to do so evolved over time. Nowadays, KD and DM alone are not enough to fill the needs from the organizations, and AKD and the concept of domain-driven data mining, are filling their gaps. Hence, domain knowledge (i.e. domain experts) are being including and becoming essential to the KD process. Furthermore, although domain experts have an essential role to KD, to actually perform it, people still need to have (besides knowledge of the business and from the data) some technical skills. Thus, most of the times, those activities are attributed to specialists from the field of data science since they have expertise on it and so far, most analytical tools available do not provide an interface suited to domain experts (with no expertise in data science) to use. Nevertheless, as said in previous sections, relying on those professionals all data driven activities can be sometimes not entirely efficient. Hence, although domain experts could benefit from being able to do their own analysis, not much has been done so far to allow these users to analyze and try to discover some patterns by their own. Furthermore, good analysis, and therefore, a good outcome, depends very much on the quality of the data being analyzed. Good quality data provides a secure and reliable source of potential useful and valuable information ready to be discovered. However, a bad quality data can hinder the usefulness of the analysis, by adding many obstacles in the way, such as: outliers, missing data, wrong data, and useless data. In an intensive data-generation environment (such a hospital), the probability of having lots of dirty data within the databases is very high. Therefore, preparing the data by carefully analyzing, cleaning, transforming, integrating and formatting it is essential in order to trust in the analysis' outcomes, avoiding getting bias from it. Additionally, as mentioned earlier, this task can be sometimes even more time consuming and trickier than DM activities, and therefore, they represent a big obstacle for domain experts when trying to analyze data. Hence, by devising a MAM for letting medical domain experts to follow and perform data preparation, the KD process can be facilitated and the problems mentioned above, addressed. Additionally, the second thesis' objective of study the importance and usefulness of the current CRISP-DM framework's data preparation steps, should be of value for the scientific community.

Understanding domain expert's comprehension and knowledge of KD and DM: This worth mentioning subject follows the same line as the one just covered above. In order address the problem and dive into the objective of this research, a prior investigation needs to be done to understand and discover what domain experts really know about Knowledge Discovery and Data Mining activities. This will be made in a format of a qualitative study where semi-structured interviews guided by a interview protocol will be conducted with medical domain experts (doctors and nurses) from the UMCU and WKZ, aiming to make sense of question such as: how domain experts see DM and KD importance to their business; if they would like to analyze the data of their own; what benefits they think DM and KD can bring to the business, daily activities, and in the patients' treatment; how much technical knowledge (necessary to analyze the data) do they have; etc. Thus, this understanding has the objective of creating a baseline that will guide the rest of the thesis. Additionally, a study like this was not found and it is believed that it has not been done before; therefore, it would be of value for this and future research.

4 WHAT DOMAIN EXPERTS KNOW ABOUT KD

To successfully achieve the objective of this research, and also answer the research questions, first a study had to be made to understand what domain experts really know about Knowledge Discovery and Data Mining. To do that, semi-structured interviews were conducted with members of the UMCU and WKZ to gather information on how they see the importance of extracting knowledge from data to their daily work, patients and to the whole organization; if they can explain what KD is; their technical skills; their ambitions and expectations of extracting knowledge from data; and their willingness and excitement about doing data analysis themselves. In total, seven interviews were conducted. The number of participants was chosen following the 'data saturation' theory (Francis et al., 2010), in which data saturation represents the data collection point where there are no new ideas or relevant concepts being introduced anymore. Therefore, after the seventh interview data saturation was considered reached and no further data collection was required.

4.1 A QUALITATIVE STUDY

The overall process, from conducting the interviews to analyzing all the data collected, has been identified with four main steps, as shown in Figure 7.

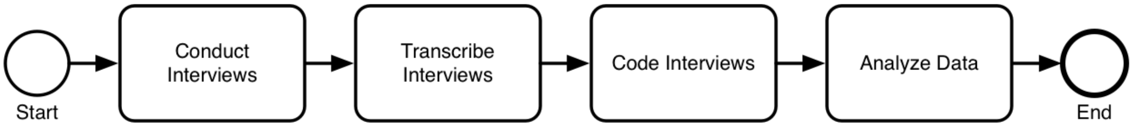


Figure 7 - Interviews Overall Process

After conducting and recording the interviews, all audio files were fully transcribed and coded using NVivo Suite (QSR International, 2016). Coding the data helps on summarizing and organizing the important findings from the interviews in an easy and retrievable way for later being analyzed. As mentioned, the interviews were conducted with seven domain experts from the medicine field, with different lines of expertise, as shown in Table 4. Additionally, for privacy reasons, the real name of the participants will not be shown.

Participant	*Expertise*
Interviewee 1	Pediatrician, Neonatologist and Medical Researcher at the UMCU – WKZ with more than 10 years of experience in the area.
Interviewee 2	Nurse from the Neonatal Intensive Care Unit from the UMCU – WKZ, working for the past 8 years within the WKZ.
Interviewee 3	Pediatrician, Neonatologist and Medical Researcher at the UMCU – WKZ, with more than 12 years of working experience.
Interviewee 4	Consultant Neonatologist and Medical Researcher at the UMCU – WKZ with more than 30 years of experience in the segment.

Interviewee 5	Anesthesiologist from the pediatric ICU from the UMCU – WKZ, with more than 10 years of working experience.
Interviewee 6	Clinical Health Sciences teacher at Utrecht University, Medical Researcher, and previously a neonatal nurse at the UMCU – WKZ, with 20 years of working experience.
Interviewee 7	Epidemiologist and Medical Researcher at the UMCU, with more than 10 years of experience in the area.

Table 4 - Interviewees Information

Moreover, the interviews were made following an interview protocol (Appendix A) which covered the following main topics:

- *Knowledge Discovery Understanding*: what the interviewee understands about knowledge discovery, their thoughts about its benefits for the organization and patients, and what is the understanding about the process of discovering knowledge from data;
- *Data Preparation and Modeling Understanding*: this aimed to understand if the interviewees have any technical knowledge such as statistical and programming skills, their experience in extracting knowledge out of data and which tools they used (if any), their difficulties, and knowledge over the available data;
- *Expectations and Thoughts over KD*: aiming to understand their expectations of being able to analyze data themselves, if they would be able to do it in their daily work, and their experiences (if any) with third-party data analysts doing data analysis.

Hence, after coding all interviews, the data was then analyzed. Moreover, Figure 8 illustrates when a new concept was introduced (yellow box) by a participant per main subject during the interviews, illustrating why data was considered being saturated after the seventh interview. Finally, all results are described below.

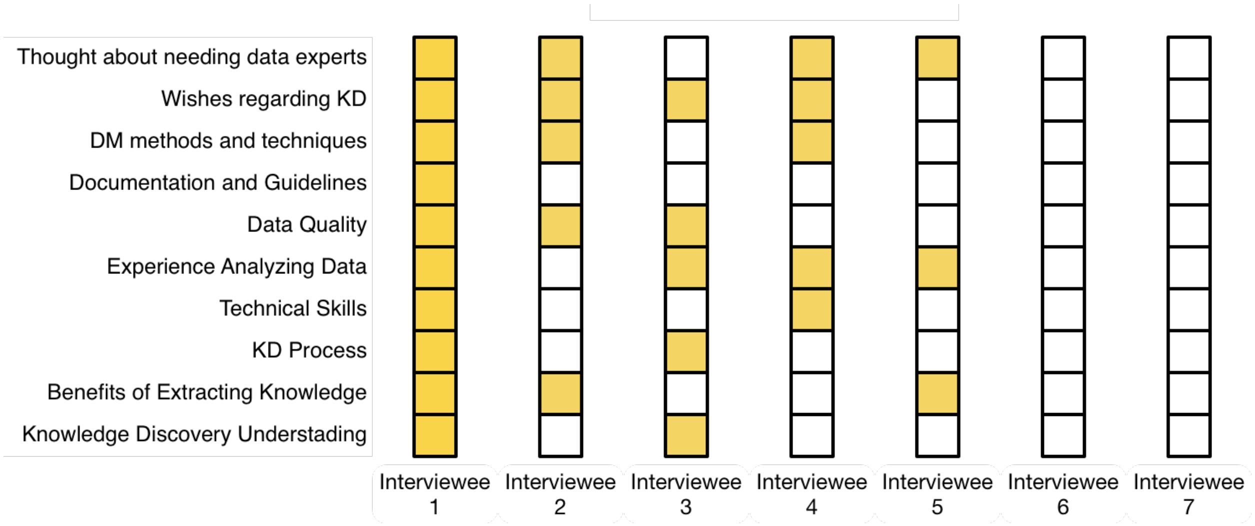


Figure 8 - Data Saturation: New Concepts Introduced by Interviewees

4.1.1 Knowledge Discovery Understanding

As mentioned in Chapter 2, the definition of KD is "*the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data*" (Fayyad, Piatetsky-shapiro, & Smyth, 1996). Based on that, when the participants were asked about their understanding about the term Knowledge Discovery in databases, none of them gave an exact definition of what the term meant. However, all of them (although in different levels of details) knew what the term was referring to and tried to explain it by presenting arguments and ideas of its practical applications and benefits to the healthcare segment. Interviewee 3 for example, said: "*…by analyzing patient's data continuously, the simple software could predict infections a lot sooner than doctors, nurses, or parents could, and thereby saving lives*". Additionally, Interviewee 4 explained Knowledge Discovery as follows: "*You have a bunch of data, and you try to extract from that bunch of data observations that may help you in further analysis".* Moreover, Interviewee 1 stated that by using data and analyzing it "*we can have a better understanding of what happened yesterday and create a pattern, or create assumptions, or create algorithms to help us on doing the care of tomorrow".*

The next question was specifically about the benefits that KD could bring to the healthcare segment, for the hospital and for its patients. All participants emphasized predictive analytics as a huge benefit of KD for the healthcare segment, for their patient's treatment and prevention of diseases. Hence, they all believe that by analyzing historical data/or real time data, health complications can be foreseen allowing doctors and nurses to act quicker, and therefore, save lives. Moreover, Interviewee 5 also explained how reactive all activities within today's NICU are, and how data analytics could help them to improve on that matter: "*We see a lot of continuous flow of data coming of our patients. It is hardly impossible, for people, nurses, and doctors, to see trends […] within the measurements that they see, and on this moment, they are very reactive. So, you cross a boundary and there will be an alarm […] and I think that the only way of getting less alarms […] is trying to understand how to stay within the margin you wish to be. […] We are not proactive in that perspective. And data analytics needs to help us to get more proactive"* Lastly, some participants also mentioned monetary savings as a benefit of KD, as can be seen in the following fragment from Interviewee 1: "*I think, in our case is best to prevent things from happening. If you can prevent sepsis, or if you can prevent inter-ventricular breathing, or if you can prevent preterm birth, because of the data you collected earlier, that will be a major saving in the national healthcare".*

The last question from this first part was about how well domain experts understand the process of extracting knowledge from data, and what steps are in between of starting a project and collecting its results. From all the participants, just one could not give a reasonable answer and deviated a lot from what was asked. The rest of them provided similar answers, where was possible to see and derive, for example, CRISP-DM's phases from them, such as business understanding, data understanding, data preparation and modeling. Three examples of that can be seen in the answers given by Interviewee 1, 5 and 7. Interviewee 1 said: "*You are wondering on how something works, or why something is being done in a specific way […], and then you start generating hypothesis and start wondering what the outcome would be if you do a study, and […] you are either using historical data or you are planning to collect data from patients[…], you start investigating what information is in there, and if the information you have gathered also answers the questions, or if it is in line with the hypothesis you generated earlier".* Additionally, Interviewee 5 states: "*So, what we are doing is to see and understand the data we have, […] and trying to see which data tells us what, and then start creating an ideal database structure, where we make our own database copy, and we will do data mining"*. In this process he also states that data cleaning is being performed while creating this "ideal" database structure. Finally, Interviewee 7 said: "*I think all research must start with a very good and focused question. So, depending on the question it kinds depends how you will handle your statistical analysis of course, but I think is very important to start getting a basic grip on the data that you have. So, do some kind of general analysis on the data to see what […] kind of information I have in my database. So, you really have to know your data first. So, what kind of information is missing for subjects in my databases? How many missing values I have. are there some groups where I have everything missing? […] Then I start usually by describing some kind of baseline characteristics of my subjects, […] general important variables for the questions that you have, and then only then we actually start with the data analysis*". Then, she complements it: "*First you have to handle your missing data […] and then, when you have your complete dataset, you can start your analysis*". Moreover, Interviewee 3, argued about the challenges and obstacles that exists within the KD process, especially in understanding and preparing the data for further use: *"… because you have to realize that sometimes the electrodes are loose, and then you don't have the recording of the meantime […] and then what do you do with that? And then what is the output format? How is the sampling frequency? Is it measured once every hour, or how is it stored? So, there are a lot of things that I thought in the begging, when I first heard about big data, that I thought it was very simple, which were actually very complex"*.

4.1.2 Data Preparation and Modeling

Moving forward, next the participants were asked about their overall technical skillset, that includes statistical knowledge, and if they have the ability to code in any programming language. For the latter, all interviewees consider their programming abilities as non-existent or very limited. Those who consider having limited programming skills said that they can do some minor things in tools like R and Excel, such as applying functions or developing macros using VBA (Visual Basic for Applications), respectively. On the other hand, most of them consider having a good statistical knowledge, at least sufficient for doing research and using SPSS, for example.

The next question was about the hospital's data and its quality. When asked if they knew what data was being collected and stored within the hospital's databases, all of them said that they knew in part what was being collected. That knowledge, most of the times, was visibly limited accordingly to which tasks the participant was involved. Thus, it was clear that each interviewee knows the data that he/she is working with and have an overall idea of what data was supposed to be stored and be retrievable in the hospital's databases. Illustrating what has just been said, Interviewee 4, for example mentioned: *"...I have some basic knowledge of the data obtained and the quality of the data…*". Then, he complements it: "*...for my own research, I know what sources we use"*. Additionally, Interviewee 6 said: "*I have an insight at microbiology [referring to the data], because I am an infection preventionist here in our ward […] and that is the data I have to work with*". In addition, most participants did not know whether there is a document mapping and explaining the database structure, which could provide them a guideline for understanding and retrieving the data. Furthermore, in some cases, participants are collecting and using their own data for research purposes, and that data is either coming from the hospital's databases or some specific data collection activity.

Regarding the quality of the data that is being collected, many interviewees complained about it. From the answers obtained, it is possible to see that, although most data are automatically generated from equipment and machines connected to the patients, the quality of that information still depends a lot on human interactions, for example, ensuring that everything is connected properly to the patient and stayed connected during the data gathering process, or own interpretation of the patient's health condition. Interviewee 1 exemplified that by saying: "*It depends a little bit about what you are looking for, because there is machine generated numbers as well, and those should be reliable. But the only thing is that also machine output can be subjected to malfunctioning of its measurement system. The ventilator can produce numbers without needing a baby attached to it, because it is warming up. Or if you are measuring desaturation, so lower oxygen levels in a baby, we know that the way we measure it with a badge around its hand, so if the baby moves, that influences the numbers that comes out of the machine. So there a lot of artifacts that are also available that you have to consider when you start analyzing the data*". Moreover, Interviewee 2 stated: "*...some parameters are not good in quality. We have parameters like the skin temperature. If the thermometer is laying on the bed and not in the diaper, you don't have a good quality for the parameter. But if a colleague of mine is validating the temperature that is not in the diaper [...], you cannot do anything with it. That is not good*". Interviewee 2 and 5 also mentioned that nurses are supposed to look all the data points generated within the NICU every hour and validate them before they are finally added into the databases. However, they said that this process also has flaws since not every professional validates the data at the same level of detail, and therefore wrong values end being added into the databases anyway. Two interviewees define the overall quality of the data as: 'rubbish in, rubbish out'. Interviewee 3 supports that statement by giving an example about types of information that can be found within the databases: "*At four different places, you would find four different birth weights*". Furthermore, Interviewee 2 also mentioned the big amount of data added by the nurses and doctors in free text format without some kind of pattern. As per this participant, this data added may contain typos, personal observations over the patient and abbreviations, which makes things harder when retrieving and analyzing this information.

Next, they were asked about their experience analyzing data (if any), and if they succeeded on doing it. Again, although in different levels of details and complexity, all participants already looked into the data mainly for research, monitoring, or to understand some illness scenario by exploring the data. As mentioned, some interviewees are also medical researchers and therefore data analysis activities are within their daily work. Some examples of activities made by them can be seen in some interview fragments from, for example, interviewees 4 and 1 respectively: "*based on hypothesis we do an analysis mostly in variable analysis, to look whether there is an association within the dependent and independent variables*"; "*…we have a lot of data that is stuck in the SAS system and by doing some queries and filtering you can come to some subsets or specific part which are smaller, and which is easier to investigate using the tools we are normally using. So, if you are just looking [...] if there is a difference between babies with respiratory support and without respiratory support, those are the things that can relatively easily be extracted from the SAS database*". Furthermore, their data analytical toolbox is based on many statistical methods, but also simple techniques, such as plotting and eye-balling the data. Methods such as correlation analysis, linear or logistic regression, etc., are frequently used by them, as can be seen in this fragment from Interviewee 1:

"*We use simple association methods, like a Pearson R, or those things that are easily extracted by attaching two columns with each other and asking the computer to decide if there is any correlation. So, correlations are very frequently used*". Hence, many interviewees mentioned that they use mainly SPSS or Excel for their data driven tasks. The tools that are used that differ from those are: SAS, used by Interviewee 1 and R used by Interviewee 4 and 7. Interviewee 5 also mentioned R and Python but those are used by a data scientist that he is working with. Despite performing some data driven activities, many of the professionals also talked about their difficulties on analyzing the data when, for example, the number of variables and the number of records required in the analysis increase; or how hard it is to clean the data from all quality problems such as outliers and data errors; or even when dealing with big sets of data, spread across different tables and databases, and having to combine and prepare that data for further analysis. This discontentment can be seen, for example, in the fragments extracted from interviewees 6, 1 and 3 respectively: "*In the patient's system nowadays, if I want to know something, is hard to extract data to and to run queries*"; "*…but it gets more difficult if you start adding all different variables together, or if you have to import or export those variables to combine them to other types of information"*; "*There is no one big master file where you just go and do some analytics*". It was visible that the bad quality data interfere when they try to analyze data, and they struggle to overcome these problems. Furthermore, integrating data is also a huge problem for them since the variables are spread across many tables, and they do not know how overcome this situation.

4.1.3 Expectations and Thoughts

The last questions were about their wishes and expectations over an analytical tool, their thoughts about the possibility of analyzing data themselves, thoughts and experiences about hiring a third-party analyst to do the analysis for them. Starting with their wishes and expectations, they all brainstormed over features that they would like an analytical tool to have so they could use it. Their answers differed a little bit, however, one thing that is common to most participants is the wish of having a step-by-step guidance to help them through the process of analyzing data. This can be seen in following fragments from Interviewee 3, 6, and 1: "*I think that a program that would take you by the hand and ask you, okay this is your database, but what are the steps? Almost like a disclaimer, pointing you to the right direction*"; "*First of all, I think it starts knowing what you should do. Your outcome, what do you want to know, and what steps are needed so you can come to that*"; "*…or that you can say that for me that is the most important outcome variable, and then the system would tell you, based on what I've explored, this is the best route to take*".

When asked about their thoughts of being able to analyze data themselves, and if they would like to do it, most interviewees demonstrated excitement over this possibility, and although time is a huge constraint for them, they said they would like to do it. Interviewee 4 for example said: "*…if you do it yourself, it might speed up the whole process. It would speed up research, and thereby speed up the whole pipeline. […] The advantage of doing it yourself is that you have total control of the process, […] the limitations of the process, potential errors in your results. So, personally, I would prefer to do it myself*". Interviewee 1 also demonstrated interest by saying: "*For me, if it would be easier than it is now, I would spend more time on extracting data or data files, because I know that I will be able to do something with it*". Furthermore, when asked about their thoughts of having external data analysts and data scientists instead for doing these activities, some of the participants argued about the costs that hiring and delegating all data driven tasks for third-parties would have for the business, and also about the limited medical knowledge that these professionals usually have in order to understand and extract useful information from the data. In addition, Interviewee 4 also mentioned: "*If you have to go to a data analysis, it may take up to four or six weeks before you get the answers, and there is six weeks of nothing*". Some of the interviewees however, were in favor of having data experts hired, given time constrains and lack of technical skills, where doing data analytics themselves would not be so productive and successful for clinical use (as it may be for research), since is a time-consuming task in a field where everything happens within a blink of an eye. Moreover, interviewees 5 and 7 mixed both scenarios arguing that the best approach would be domain experts working together with data analysts providing them medical knowledge to successfully understand and analyze the data, which is also the scenario that is supported in recent literature. Nevertheless, despite being crucial, transferring knowledge can also be challenging and very time consuming as shown in following fragment from Interviewee 1: "*So, answering the right question […] is the most difficult part on working with data scientist people. But the knowledge that I lack in doing data science myself, the other side of the table does not have the medical knowledge, and that medical knowledge is needed for doing the analysis, because if they do not know what the data is telling, or what data they are using, […] then you can come to all kinds of associations that have no meaning whatsoever*". He complements it: "*I think I spent hours or days just bringing by knowledge to them, so they could use the knowledge to explore the data*".

4.2 Domain Experts vs Knowledge Discovery

Based on the information collected and described above, together with what was found in literature, this section will address the two first sub-research questions from this study.

SQ1 – *What is the current understanding that domain experts have on Knowledge Discovery and Data Mining activities?*

Based on what has been described above, it was clear that domain experts have a clear view and understanding of the purposes, benefits and activities related to KD and DM, as well as the challenges that it consists of. Additionally, they shown excitement to do data analytics themselves, however, they also know that it still an activity that vary in levels of complexity and that it requires a constant interaction between domain experts and data analysts to explore the most complex and hard scenarios. Furthermore, although lacking on having technical skills (e.g. programming), most of them said to have good statistical knowledge which allowed them analyze data by applying statistical methods to test hypothesis while doing research. Thus, elaborating on that, even though the exact definition of KD is not known, the idea of using data to extract information that can be used on the improvement of processes, to better treat the patients, and even preventing and predicting some health condition to happen is well acknowledged by all the interviewees. Additionally, although not mentioning anything about the any KD process, is possible to see that its concepts and phases (from CRISP-DM, for example), such as business understanding, data understanding, data preparation, and modeling can be seen in the answers given by the participants, and therefore most of them have an overall understanding of the activities that exist in between defining a goal and analyzing data. Furthermore, as expected, domain experts lack on programming skills, and therefore, any analysis based on coding activities can be somehow difficult for them. However, most of them said to have a reasonable knowledge of statistical methods, which on the other hand, allows them to, by using other means and techniques, analyze data. Hence, as mentioned, their experience in the matter is mostly based on some basic exploratory analysis or applying statistical methods for testing research hypothesis. Some of them however, went a little further and learned the basics on how to use tools such as R to help them in their analysis. Also, most of them demonstrated excitement when asked if they would like to spend more time analyzing data if that activity was somehow facilitated. Nevertheless, none of them said a word about replacing data scientists by being able to do data analysis themselves, on the contrary, they know the challenges and difficulties of dealing with data and consider these professionals essential for the process. However, they know that most of the times data analysts and scientists don't have the medical background to understand and extract all relevant information from the data, and that therefore, they have to work together with these people, providing them medical knowledge in order to get the "right" answers.

SQ2 – What are the risks and benefits of allowing domain experts to analyze data themselves?

Although facilitating the KD process for domain experts would give them some independence over data scientists, it is clear that this independence would be mainly over tasks that would not require high levels of technical skills, and given the high demand, big variety, and the different levels of complexity that a data analytical problem may have, the interaction between data scientists and domain experts for more complex data problems are still the best route to take. Nevertheless, from the literature review and the interview's results above, it is possible to argue about the benefits and potential risks of allowing domain experts to analyze data themselves, which is illustrated next.

The most obvious and main benefit of letting domain experts doing data analytics is related to the domain knowledge that these people are supposed to have. As mentioned in Chapter 2, domain ubiquitous intelligence (knowledge over the problem domain, environment, and data) is the main aspect needed in order to perform Actionable Knowledge Discovery (AKD). Thus, as domain experts are required to be specialists on their field, they should be able to identify patterns, inconsistencies, outliers, etc., more easily than a non-expert when using the right tools. Hence, by having a deep knowledge over the domain, the KD process can be turned into an AKD process directly. Another benefit (also pointed out by some of the interviewees) is the total control over the analytical process and therefore, over the results. By being able to analyze data, there would not be any miscommunication or misunderstanding between domain experts and technical people, thus, the objective should be well known, the understanding over the variables within the data should be optimal, and what has been done to achieve such outcome should also be clear. Although interactions with data scientists are the best way to go in more advanced and complex projects, avoiding that interaction (by providing means for domain experts to analyze data) can sometimes help saving time, and maybe even achieve better results for a specific objective. To support this statement, Interviewee 7 said: "*The difficulty is that we actually speak a different language* [making reference to medical experts and computer scientists]. *So, my experiences when I talked to a computer scientist, I usually don't understand what they are saying, and vice-versa*". Hence, since transferring knowledge is a very time-consuming activity, for less complex projects, avoiding doing it can result in faster outcomes. On the other hand, and despite all that, the KD process can also be very time-consuming, and domain experts usually have different activities than analyzing data in their daily routine, hence, depending on the project, transferring the required knowledge to a data scientist could be faster than trying to analyze the data alone. Additionally, the KD process is about discovering hidden and unknown information from datasets, which requires sometimes to forget what you already know, be open-minded, try to find patterns where should not be any, etc. Hence, relying on domain experts to explore data (unveiling hidden information) can also have a drawback such as the possible absence of creativity (to think "out-of-the-box") when doing it since, one is looking to a specific scenario based on the current knowledge and understanding that he/she already has over the problem. Additionally, less external knowledge (i.e. data analysts) could also mean less innovation and new ideas being added into the domain.

Thus, risks and benefits exist on allowing domain experts doing data analysis, however, those are directly connected to external variables such as project complexity, analytical mindset, time available to spend doing data analysis, knowledge over the domain, etc. Nevertheless, it is clear that the main benefit described above that supports domain experts, relies over the knowledge that these professionals have over the domain (which usually is an issue for external data analysts), or in other words, the domain ubiquitous intelligence. For that, by being allowed to analyze data, even if the whole process is not completed for any reason, or some major technical skills is needed to move on, domain experts could be able to better communicate, and better understand what is and has to be done by the data analysts, and therefore, together, achieving better results.

5 DATA QUALITY UNDERSTANDING

In this chapter the current data quality situation Dutch (academic) Hospitals, represented by the Wilhelmina Kinderziekenhuis (focusing on the Neonatal ICU) will be discussed. First, a more detailed view on the Data Understanding phase from the CRISP-DM framework will be given (since it illustrates steps that may need to be considered before starting with the data quality assessment), followed by a description of the WKZ's environment and the problems found during the data quality evaluation.

5.1 THE DATA UNDERSTANDING PHASE

Earlier in Chapter 2, all phases from the CRISP-DM were briefly described, but the focus was given only to Data Preparation, which was explained in more detail. Nevertheless, since the purpose of this chapter is to provide quality assessment of the data within the WKZ, a more elaborated explanation of the Data Understanding phase from the CRISP-DM framework will be given, as the quality of the available data, as per its guideline (Pete et al., 2000), should be addressed within this phase.

As mentioned earlier in this study, Data Understanding (as per the CRISP-DM framework) is all about getting familiarized with the data at hand, assimilating its content (e.g. variables, values, etc.), the possible data quality problems that it might have, and even to start addressing the data mining goals of the project. Hence, to guide users on acquiring knowledge over their data, the CRISP-DM presents four sub-steps for this phase, as shown in Figure 9. Below, each activity will be explained, based on the CRISP-DM guideline (Pete et al., 2000).

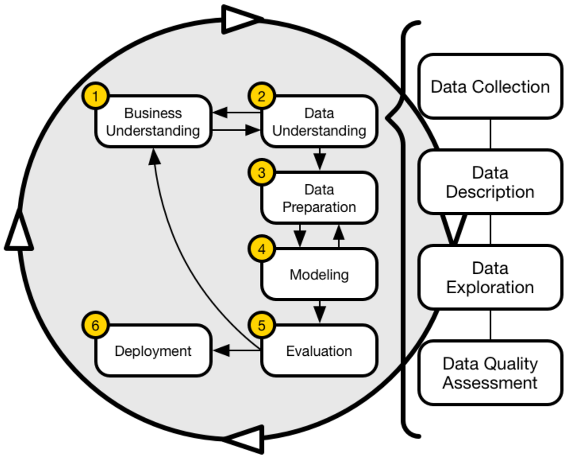


Figure 9 - CRISP-DM's Data Understanding Step

Data Collection: After the Business Understanding phase where the project and its objectives are defined, the data that will be used in the analytical task needs to be identified and then loaded into the tool that will be used to explore and analyze it. For example, a *.csv* file being opened in Excel, or loaded into *R*.

Data Description: This activity is mainly about getting to know the properties of the data that was chosen in the previous step. That is, an overall picture of the data structure and its content, such as variables types, meaning of the attributes, descriptive statistical details about the data, etc.

Data Exploration: The data exploration phase is used to inspect the data more in-depth with the intention of trying to get more acquainted with its content, summarizing it, identifying possible areas that might contain quality issues, etc. These tasks are usually done by means of using simple data manipulation techniques and visualization approaches that would help checking and exploring the data. Thus, although some data mining goals can start being addressed in this sub-step (based on some questions or findings), the main objective is to acquire a good knowledge over the data and start identifying possible data preparation needs for the next KD process phase.

Data Quality Assessment: Understanding the data means not only checking what it contains, and trust in whatever the data is depicting, but also looking for what is missing, or what does make sense and what does not. This step is an extension from the previous one, where the main objective is to examine the data for quality issues, such as, if the dataset is complete; typographical problems, such as misspelled words; if date values are in the right format for all entries; if constraints are being persisted throughout the whole database schema, etc. Moreover, identifying the issues is only the 'tip of the iceberg', once that the problems may continue occurring if the reason of bad data generation is not identified and fixed. Hence, sometimes not only the data has to be checked, but depending of the dirty data occurrences, the whole process has to be revised.

Thus, as part of the KD process, Data Understanding has a great impact on the overall execution and outcome of an analytical project. Without a good understanding of the variables, values, structure, etc., is possible that the main objective of the whole KD process fails to be achieved, simply by the fact that the person who is analyzing the data might not have enough knowledge to distinguish important patterns from other groups of information within it. Additionally, without a good understanding of the data, ensuring its quality can also be hard, since for example, knowing if a value is missing for some random reason, or if there is any external influence causing the missingness, can change completely how to pursue the whole data preparation activity to be done next, and therefore, the overall analytical task.

5.2 Data Quality Assessment

Much has been said about the benefits of using data analytics as a decision-support mechanism in different areas of application, especially in the healthcare. However, these benefits are directly related to the quality of data that is being used during the analysis, and therefore, people have to be really sure that the data is trustworthy. Nowadays most data within hospitals is being generated by means of electronic health records (EHRs), which should be, most of the times, reliable. However, as mentioned earlier, even those mechanisms sometimes depend on human factors, such as an electrode being connected correctly to a patient. Besides that, as shown in the last section, a lot of data from the intensive care units are still being inputted by doctors and nurses as free-text based on observations or comments that can differ for each professional. Thus, this data could contain typographical errors, abbreviations, or other inconsistencies that can hinder data analytic approaches and outcomes. Furthermore, data quality also depends on the database structure, and how it is mapped in relation to the data generation equipment, considering avoiding duplicate values, wrong data insertion, or data integration problems.

As per the interviews' results described earlier, many professionals complained about the quality of the data within the hospital's databases, and therefore, before moving forward with this research, an understanding of these data problems, data generation methods, and database structure can be useful. A study by Batini, Cappiello, Francalanci, & Maurino (2009), which compare and summarize methodologies for data quality assessment and improvement, state that in order to do so, one should start by collecting and understanding contextual and environmental information about how the data is generated and stored, and then applying data quality assessment techniques, and defining improvement strategies. Still according to the authors, data quality can be assessed by the following steps: analyzing the data at hand, analyzing the data quality requirements, identifying critical areas, mapping the processes of generating and storing the information, and measuring the data quality by checking how the data problems affect four main quality dimensions, which although do not represent all dimensions, are considered the center of attention for the majority of researchers (Batini, Cappiello, Francalanci, & Maurino, 2009).

Hence, the dimensions are: *accuracy, completeness, consistency, and timeliness.* The first one refers to syntactic and semantic accuracy for example, which sees if a value is syntactically correct, and if that same value is making reference to what it was supposed to respectively. Completeness makes reference to the amount and impact of missing values within a dataset. Consistency is when values, attributes, and constraints are persisted across the whole database. And lastly, timeliness refers to how current the data is, and whether the it is available when expected and needed to be. Furthermore, as stated by Christoulakis, Spruit, & Van Dijk (2015), data quality is defined by its "fitness to use", and therefore, achieving good results in these dimensions still does not mean that the data quality is in the desired level for a given project. However, evaluating the data based on those dimensions, could help on understanding the quality problems, and what has to be done to fix them. Regarding data quality's improvement, Batini, Cappiello, Francalanci, & Maurino (2009) mention two main approaches: process-driven and data-driven. As its own name says, the process-driven strategy aims to improve the quality of the data focusing in solving problems by redesigning the processes that generates and stores the data. On the other hand, data-driven strategies focus on improving the quality of the data within the databases. Some examples of data-driven approaches are: data normalization, record linkage, data integration, and error identification and correction (Batini, Cappiello, Francalanci, & Maurino, 2009). More about this will be explained in the following sections.

Thus, although a full and extensive data quality assessment would be very helpful for both thesis development and data managers within the hospital, realizing it was not a possibility given time constrains. Nevertheless, in order to understand and have a grasp of the hospital's data quality, meetings were held with people more familiarized with the hospital's data, such as the WKZ's data manager, and external data analysts who are working constantly with this data, in order to extract their viewpoint of the overall data quality. Additionally, an exploratory data analysis was made aiming to identify some of the issues as well. Moreover, it is important to mention that this study was made over the data that is generated within the Neonatal Intensive Care Unit (NICU) of the WKZ.

5.2.1 Data Environment

The data environment around the neonatal ICU within the WKZ is based mainly on two data managing systems: an EHR platform called HiX, and MetaVision. HiX, developed by ChipSoft (ChipSoft, 2014), is the UMC's (which comprises the WKZ) primary system for healthcare, where all patients' personal information (e.g. name, address, admission and discharged dates, billing information, etc.), and all activities that are performed in their treatment are recorded. Additionally, MetaVision, developed by iMDsoft

(iMDsoft, 2017), is the system responsible for managing the data within the NICU. Thus, all patients' health information, measurements, observations and notes that are being generated by monitors and equipment that are connected to the patients, as well as the information added by doctors and nurses manually, are handled by this application. Moreover, both systems interact with each other in order to keep track of which measurement belong to which patient. Additionally, the environment also counts with the Research Data Platform (RDP), that has been set up to make the generic healthcare data available for research purposes. At the RDP (focusing again in the NICU), data from HiX and MetaVision are made available for researchers and medical personnel accordingly to their needs.

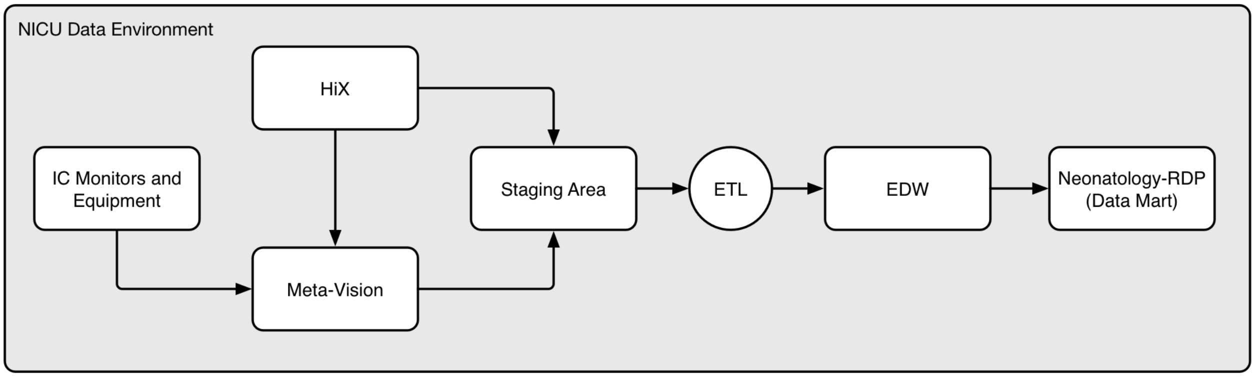


Figure 10 – Neonatal ICU Data Environment

The WKZ's data environment follows an Enterprise Data Warehouse (EDW) architecture, where EDW is defined by Moody & Kortink (2000) as a "*database which provides a single consistent source of management information for reporting and analysis across the organization*". According to these authors, the architecture can be better explained by making use of a supply chain metaphor, where (adapted for the study scenario) the monitors, HiX, and MetaVision are the data suppliers, which provide the data that will be stored into the EDW. The data is then made available to the users via Data Marts (it can be accessed by other means, which are not relevant for this research), which are basically a subset of the EDW's data, specific for one group, subject, or department. Furthermore, the information from HiX and MetaVision, before being added into the EDW, is temporarily stored in a staging area, where it will be validated based on existing business rules. Moreover, MetaVision also communicates directly with HiX, retrieving some of its information (in this case, general patient's admission and discharged date and time). However, all data that comes from MetaVision, when within the staging area, is validated against the information from HiX, as it is the main and most reliable system, which could contain more recent information. Hence, if there is any difference in values (e.g. in the registered dates and times) the value from HiX overwrites the one from MetaVision before flowing into the EDW.

In addition to that, the hospital's data environment follows a Detailed Clinical Models (DCM) modeling design. Dimensional models have the objective of creating interpretable database schemas, which will also require low effort to query upon. Some examples are Flat Schema, Star Schema, and Snowflake Schema. The choice of using a specific design, as per Moody & Kortink (2000), needs to consider a trade-off between redundancy and the complexity to visualize and access the information. According to Goossen (2014) , DCM is defined as "*an information model designed to express one or more clinical concept(s) and their context in a standardized and reusable manner, specifying the requirements for clinical information as a* *discrete set of logical clinical data elements*". Thus, DCM was built as a new way of structuring and organizing clinical data in a consistent and reusable fashion (e.g. data components), and it can be deployed in many technical representations, among them, Electronic Health Records (EHR) and EDW. Hence, one Data Mart can be composed by one or more data components (depending on what information needs to be retrieved), such as the Intensive Care (IC), Laboratory, or Radiology component, for example. This flow can be seen in the schema shown in Figure 10, where rectangles and circles represent data stores and processes respectively. Moreover, the ETL process (Extract, Transform and Load) is responsible for extracting, transforming and loading the data from the staging area that receives the information from different data sources (MetaVision and HiX, for example), in a uniform format into the DWH. Finally, the data within the Data Marts are accessible to the users via software applications (i.e. SAS) available within the UMC.

NEONATOLOGY - RESEARCH DATA PLATFORM

Based on the environment described earlier, the Neonatology-RDP and its content will be described next.

As mentioned earlier, the Neonatology-RDP is a Data Mart built from different data components, where in its core is the IC (Intensive Care) component. Additionally, the Neonatology-RDP also has tables with information coming from different departments and areas within the hospital, such as radiology and laboratory. Nevertheless, the focus of this chapter and quality assessment will be upon the IC data component. Having said that, the IC data component within the Neonatology-RDP combines the information from HiX with the measurements taken from MetaVision, and it currently consists of fifteen tables in the format of a relational database, as shown in the dimensional modeling depicted in Figure 11. Moreover, it contains general IC recording data, such as clinical events, applied catheters, nursing activities, and forms. Furthermore, it has been built to handle both adult's ICU and NICU measurements and information, however, data from the first one mentioned is still not available. A brief explanation of each table will be given below:

- *IC_Opname*: this table contains the ICU admission and discharged information at a patient level.
- *IC_Meting*: four tables with all information about the measurements' taken from the patients
  - *IC_Meting_Metadata:* this table has the metadata of the different types of measurements, such as units, types, quantity descriptions, etc.
  - *IC_Meting_Numeriek*: this table contains all numerical measurements from MetaVision, such as babies' length and weight, quantity of medication given to the patient, number of lines applied to the baby, etc.
  - *IC_Meting_Tekst*: this table consists of all the structured text measurements from MetaVision, such as child's activity, aspect, color, etc.
  - *IC_Meting_VrijeTekst:* this table contains the free text measurements from MetaVision, such as measurements taken by nurses and doctors, that were added into the system.
- *IC_Events:* this table has all the clinical events over a patient, that were registered in MetaVision.
- *IC_Lijnen:* this table contains all information over the lines that have been applied to the patient.
- *IC_VPK_Activiteiten:* four tables with information over the nursing activities ('*verpleegkundig'*) done to the patient

  - *IC_VPK_Activiteiten_Metadata:* this table contains the metadata of the different types of nursing activities.
  - *IC_VPK_Activiteiten_Numeriek:* this table has all numerical nursing activities taken from MetaVision.
  - *IC_VPK_Activiteiten_Tekst*: this table contains all the nursing activities in form of structured text taken from MetaVision.
  - *IC_VPK_Activiteiten_VrijeTekst:* this table consists of all nursing activities stored in the form of free text in MetaVision.
- *IC_Form:* four tables with unlocked and accessible forms information over patients, where three out of the four tables are filled with meta-data.
  - *IC_Form_Warde:* this table contains all values from the forms at a patient level.
  - *IC_Form_Metadata:* this table has the metadata of all forms available.
  - *IC_Form_Parameters:* this table has all the parameters per form version.
  - *IC_Form_Keuzelijstitems:* this table has all the items that can be chosen from the picklists in the forms.

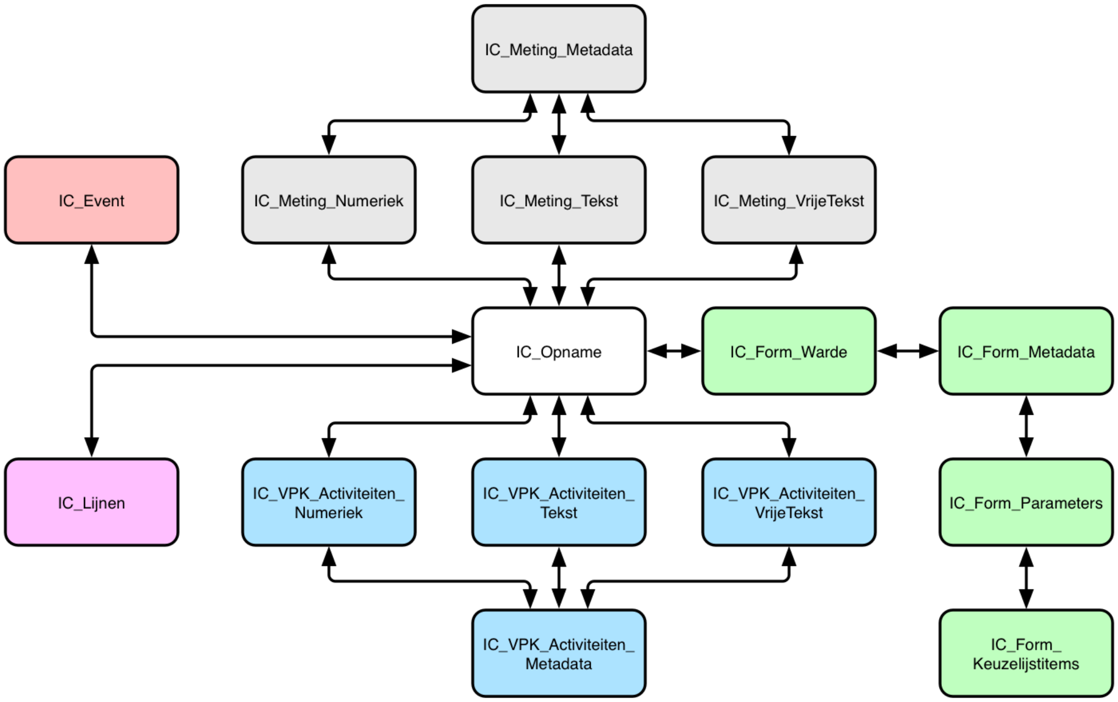


Figure 11 – Neonatology-RDP Intensive Care Component

5.2.2 DATA QUALITY FINDINGS

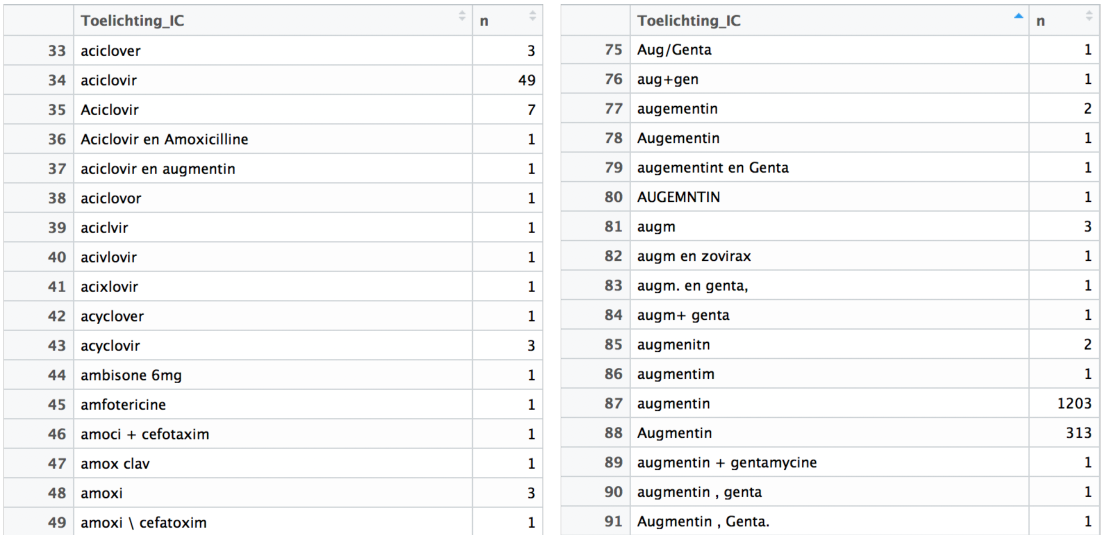
In this section, some examples of the data quality problems found within the Neonatology-RDP will be shown and explained. Again, given time constraints it was not possible to realize an extensive data quality assessment of the environment and data available. Thus, the problems that will be shown next do not represent all the issues that may exist within this Data Mart. Additionally, the data was analyzed using *R*, where by means of an exploratory analysis some problems were discovered. The analysis was made without following any specific procedure. In short, first, the goal was to replicate some of the data problems mentioned during the interviews, such as the statements referring to typographical errors and inconsistent data. Second, based meetings held with external analysts (who at the time of writing were working with the hospital's data), more problems could be identified and later confirmed. Finally, some of the data quality issues were possible to find due to expending time exploring the data and thinking of scenarios that could indeed depict problematic or inconsistent information. Hence, the table below summarizes the problems that were found, per quality dimension.

Quality Dimensions	*Type of Problems Found*
Accuracy	• Syntactic: Typographical Errors, Abbreviations; • Semantic: Wrong Data; • Other: Duplicated Data.
Consistency	• Non-normalized Data; • Problems referring to Integrity Constrains; • Different values of the same entity across the tables.
Completeness	• Quite a lot of missing data throughout the many tables, where the mechanism of missingness is not certain and requires further analysis;
Timeliness	• No explicit problems were found in this dimension.

Table 5 - Quality Issues Overview

ACCURACY

Earlier in this chapter, two types of accuracy measurements were given: syntactic and semantic accuracy. As described above, many of the tables within the Neonatology-RDP have free text data within them, hence, as expected (and as described by most of the interviewees in the previous chapter), sometimes typographical errors, abbreviations, and other syntactic issues can be found within the datasets.



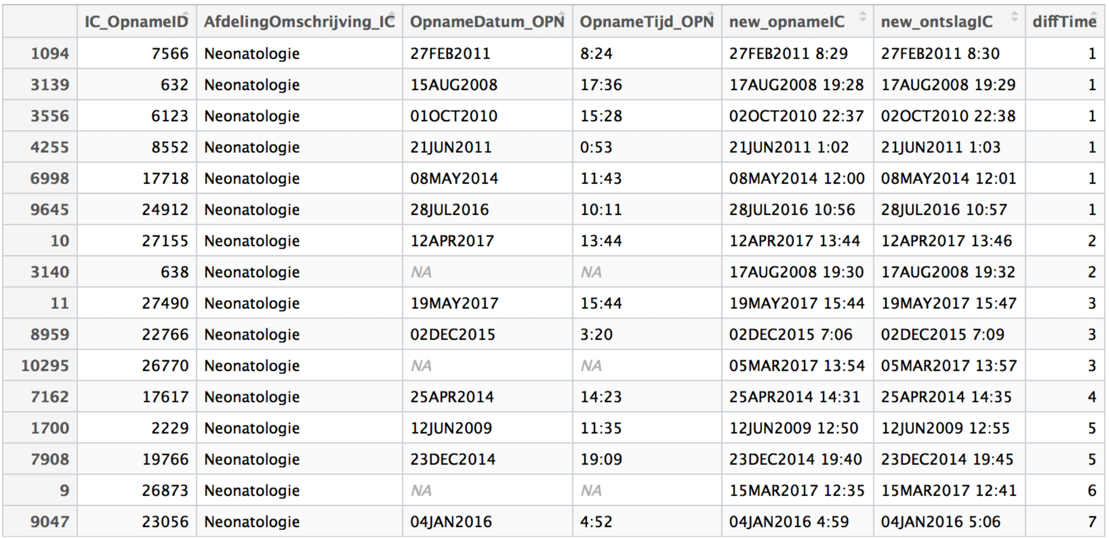
	Toelichting_IC	n
33	aciclover	3
34	aciclovir	49
35	Aciclovir	7
36	Aciclovir en Amoxicilline	1
37	aciclovir en augmentin	1
38	aciclovor	1
39	aciclvir	1
40	acivlovir	1
41	acixlovir	1
42	acyclover	1
43	acyclovir	3
44	ambisone 6mg	1
45	amfotericine	1
46	amoci + cefotaxim	1
47	amox clav	1
48	amoxi	3
49	amoxi \ cefatoxim	1

	Toelichting_IC	n
75	Aug/Genta	1
76	aug+gen	1
77	augementin	2
78	Augementin	1
79	augementint en Genta	1
80	AUGEMNTIN	1
81	augm	3
82	augm en zovirax	1
83	augm. en genta,	1
84	augm+ genta	1
85	augmenitn	2
86	augmentim	1
87	augmentin	1203
88	Augmentin	313
89	augmentin + gentamycine	1
90	augmentin , genta	1
91	Augmentin , Genta.	1

Figure 12 - Syntactic Problem Example: Different Spellings for the Same Type of Medication.

To illustrate that, Figure 12 shows examples of syntactic issues that were found in the *IC_Meting_Numeriek* table, as well as the number of times that they occur in the data sample explored. The examples extracted from the data show many variations of syntax for some medications, such as '*Acyclovir'* (an antiviral medication) and '*Augmentin'* (antibiotic) for example. The image shows that 'Acyclovir' has been written in many different ways, such as '*aciclover'*, '*aciclovor'*, '*acixlovir'*, etc. The same happens to '*Augmentin'*, where there are almost twenty syntax variations for this medication, as well as abbreviations for its name. In addition to that, sometimes, more than one medication is added in the description field, which is not correct. Thus, although this is only a small example, several occurrences of syntactic problems can be seen throughout the tables.

Furthermore, some unexpected values were found when exploring the Neonatology-RDP which could be also examples of accuracy problems in that database. For instance, some patients from the Neonatal ICU, as per the data, stayed in the ICU for less than 5 minutes, which is very unlikely, especially considering that the median of the time spent in the ICU (until the day of writing) is approximately 7000 minutes, or 117 hours. Hence, the record may also have been saved with the wrong date. This possible problem is illustrated in Figure 13, where the difference in minutes (*diffTime*) was taken between the IC discharged date/time (*new_ontslagIC*) and the IC admission date/time (*new_opnameIC*). In addition to that, one patient stayed in the ICU for more than three years, where the highest period of time after that is approximately six months.



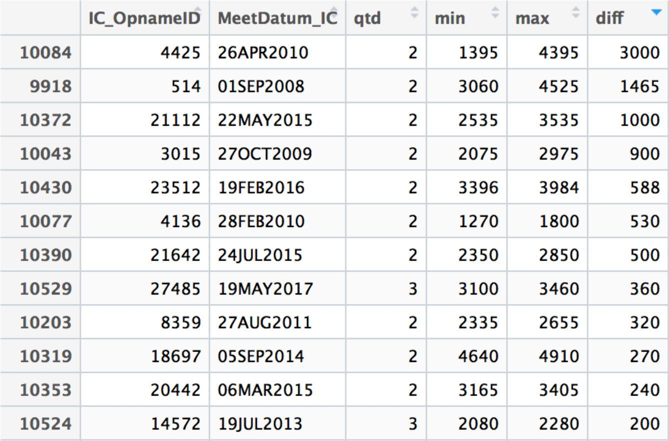
	IC_OpnameID	AfdelingOmschrijving_IC	OpnameDatum_OPN	OpnameTijd_OPN	new_opnameIC	new_ontslagIC	diffTime
1094	7566	Neonatologie	27FEB2011	8:24	27FEB2011 8:29	27FEB2011 8:30	1
3139	632	Neonatologie	15AUG2008	17:36	17AUG2008 19:28	17AUG2008 19:29	1
3556	6123	Neonatologie	01OCT2010	15:28	02OCT2010 22:37	02OCT2010 22:38	1
4255	8552	Neonatologie	21JUN2011	0:53	21JUN2011 1:02	21JUN2011 1:03	1
6998	17718	Neonatologie	08MAY2014	11:43	08MAY2014 12:00	08MAY2014 12:01	1
9645	24912	Neonatologie	28JUL2016	10:11	28JUL2016 10:56	28JUL2016 10:57	1
10	27155	Neonatologie	12APR2017	13:44	12APR2017 13:44	12APR2017 13:46	2
3140	638	Neonatologie	NA	NA	17AUG2008 19:30	17AUG2008 19:32	2
11	27490	Neonatologie	19MAY2017	15:44	19MAY2017 15:44	19MAY2017 15:47	3
8959	22766	Neonatologie	02DEC2015	3:20	02DEC2015 7:06	02DEC2015 7:09	3
10295	26770	Neonatologie	NA	NA	05MAR2017 13:54	05MAR2017 13:57	3
7162	17617	Neonatologie	25APR2014	14:23	25APR2014 14:31	25APR2014 14:35	4
1700	2229	Neonatologie	12JUN2009	11:35	12JUN2009 12:50	12JUN2009 12:55	5
7908	19766	Neonatologie	23DEC2014	19:09	23DEC2014 19:40	23DEC2014 19:45	5
9	26873	Neonatologie	NA	NA	15MAR2017 12:35	15MAR2017 12:41	6
9047	23056	Neonatologie	04JAN2016	4:52	04JAN2016 4:59	04JAN2016 5:06	7

Figure 13 – Semantic Problem Example: Short Stay in the NICU.

Another example is the huge gain of weight that some babies had between two or more measurements taken on the same day. As per Figure 14, which shows the number of measurements realized in the same day per patient (*qtd*), together with the minimum and maximum weight values collected and the difference between the two (all in grams), is possible to see that some patients gain up to three kilograms of weight at the same day, where an regular number should not exceed two decimals places (in grams).

Moreover, accuracy does not only consider single values (i.e. checking whether something was spelled correctly or not). It may also refer to attribute accuracy, relations within tables, or even to a whole database schema (Scannapieco, Missier, & Batini, 2005). Hence, one huge accuracy problem is duplicated data.

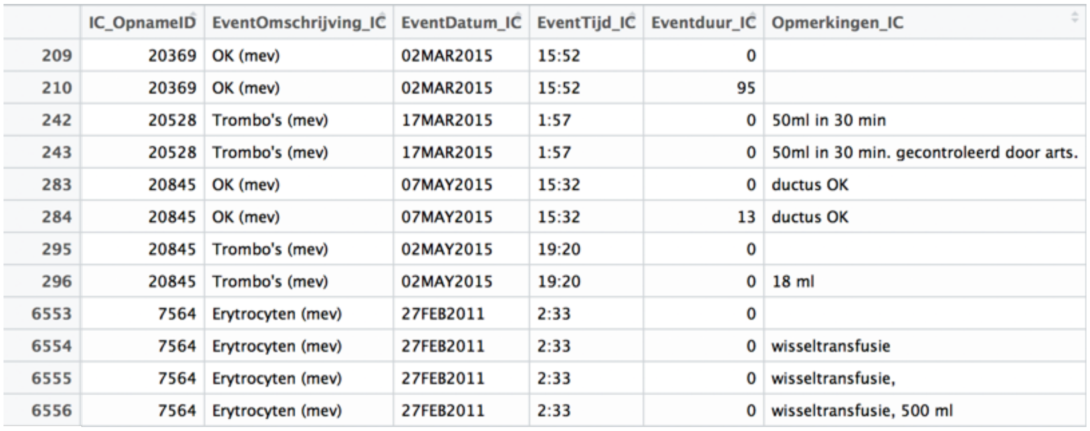
Duplicated data happens when the same entry is replicated two or more times in a dataset, polluting the dataset with needless records. It can occur for many reasons, such as software 'bugs' when storing the data, wrong constraints definitions, or even human errors. As an example, in the *IC_Event* table, from a random sample of roughly sixteen thousand records, approximately three hundred duplicated entries were located. Figure 15 shows a few examples of the duplicated data found within this table. As per the image, for each group of duplicated values, one record has some additional information over the other (even if it is only one character). Thus, it is clear that instead of updating the record, a new one is being created every time that a change is being committed, overpopulating the table with redundant and also incomplete information. Although in this example there is not a great number of duplicated data (only 2% of the values were duplicated), the problem is likely to happen every time an entry is updated, and therefore is worth mentioning since the problem could get bigger over time.



	IC_OpnameID	MeetDatum_IC	qtd	min	max	diff
10084	4425	26APR2010	2	1395	4395	3000
9918	514	01SEP2008	2	3060	4525	1465
10372	21112	22MAY2015	2	2535	3535	1000
10043	3015	27OCT2009	2	2075	2975	900
10430	23512	19FEB2016	2	3396	3984	588
10077	4136	28FEB2010	2	1270	1800	530
10390	21642	24JUL2015	2	2350	2850	500
10529	27485	19MAY2017	3	3100	3460	360
10203	8359	27AUG2011	2	2335	2655	320
10319	18697	05SEP2014	2	4640	4910	270
10353	20442	06MAR2015	2	3165	3405	240
10524	14572	19JUL2013	3	2080	2280	200

Figure 14 - Semantic Problem Example: High Gain of Weight in Last Than One Day.

Also, finding duplicated data can be a tricky activity when dealing with large datasets. One well-known method for finding these types of problems (among other things) is called Record Linkage. Record Linkage is the process of identifying data entries relating to the same person, attribute or entity, by means of common attributes that are used to define true matches. Hence, by using it, is possible to identify two or more records that relate to the same entity but are not absolute identical, as can be seen in the example below.



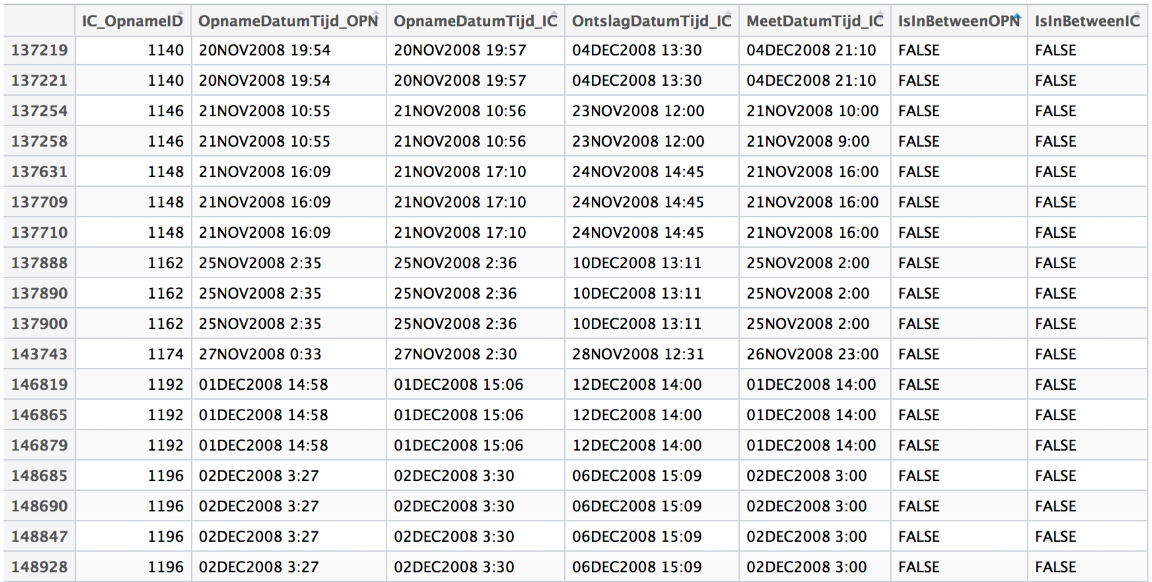
	IC_OpnameID	EventOmschrijving_IC	EventDatum_IC	EventTijd_IC	Eventduur_IC	Opmerkingen_IC
209	20369	OK (mev)	02MAR2015	15:52	0	
210	20369	OK (mev)	02MAR2015	15:52	95	
242	20528	Trombo's (mev)	17MAR2015	1:57	0	50ml in 30 min
243	20528	Trombo's (mev)	17MAR2015	1:57	0	50ml in 30 min. gecontroleerd door arts.
283	20845	OK (mev)	07MAY2015	15:32	0	ductus OK
284	20845	OK (mev)	07MAY2015	15:32	13	ductus OK
295	20845	Trombo's (mev)	02MAY2015	19:20	0	
296	20845	Trombo's (mev)	02MAY2015	19:20	0	18 ml
6553	7564	Erytrocyten (mev)	27FEB2011	2:33	0	
6554	7564	Erytrocyten (mev)	27FEB2011	2:33	0	wisseltransfusie
6555	7564	Erytrocyten (mev)	27FEB2011	2:33	0	wisseltransfusie,
6556	7564	Erytrocyten (mev)	27FEB2011	2:33	0	wisseltransfusie, 500 ml

Figure 15 - Duplicated Data Example: Updates Generate New Records in the Data Tables.

CONSISTENCY

Next, another important dimension that helps understanding and improving the quality of the data is consistency. It evaluates whether a database schema follows a set of integrity constrains in a traditional dependency level, and also constraints that help on maintaining the semantic consistency of the data and must be satisfied by all instances of a dataset. As an example, a possible consistency check would be the violation of an integrity constraint such as: the admission date/time of a patient has to be lower than its discharged date/time for all instances (as it would not make sense the other way around). Additionally, the consistency dimension is very much related to the just mentioned accuracy dimension, since if a constrain is violated, it could cause, for example, semantic problems for the dataset as well, for example, if a date is written in the format *mm/dd/yyyy* instead of *dd/mm/yyyy*.

As an example, different units of measurement (e.g. milliliters, grams, kilograms, etc.) are being utilized within the *IC_Meting_Numeriek* table, where it varies depending on the measurement type. Furthermore, for some specific types of measurements the unit employed is being specified in the its name (i.e. grams for measuring the patient's weight), however, for others it is very hard to identify it. Hence, there is no consistency between these values. Also, in the interview's results it is possible to see a clear consistency problem when Interviewee 3 said: "*At four different places, you would find four different birth weights*", as the same data value should be persisted and be the same in the entire database schema. Hence, although separately all instances may satisfy the integrity constrains, together they fail on it. Additionally, from the meetings with external data scientists working within the UMC, besides the birth weight, other inconsistencies were mentioned, such as a patient having different gestational period prior the birth. Regarding the Neonatology-RDP, the duplicated data example illustrated above, can also be seen as a consistency problem since records are being created instead of being updated.



	IC_OpnameID	OpnameDatumTijd_OPN	OpnameDatumTijd_IC	OntslagDatumTijd_IC	MeetDatumTijd_IC	IsInBetweenOPN	IsInBetweenIC
137219	1140	20NOV2008 19:54	20NOV2008 19:57	04DEC2008 13:30	04DEC2008 21:10	FALSE	FALSE
137221	1140	20NOV2008 19:54	20NOV2008 19:57	04DEC2008 13:30	04DEC2008 21:10	FALSE	FALSE
137254	1146	21NOV2008 10:55	21NOV2008 10:56	23NOV2008 12:00	21NOV2008 10:00	FALSE	FALSE
137258	1146	21NOV2008 10:55	21NOV2008 10:56	23NOV2008 12:00	21NOV2008 9:00	FALSE	FALSE
137631	1148	21NOV2008 16:09	21NOV2008 17:10	24NOV2008 14:45	21NOV2008 16:00	FALSE	FALSE
137709	1148	21NOV2008 16:09	21NOV2008 17:10	24NOV2008 14:45	21NOV2008 16:00	FALSE	FALSE
137710	1148	21NOV2008 16:09	21NOV2008 17:10	24NOV2008 14:45	21NOV2008 16:00	FALSE	FALSE
137888	1162	25NOV2008 2:35	25NOV2008 2:36	10DEC2008 13:11	25NOV2008 2:00	FALSE	FALSE
137890	1162	25NOV2008 2:35	25NOV2008 2:36	10DEC2008 13:11	25NOV2008 2:00	FALSE	FALSE
137900	1162	25NOV2008 2:35	25NOV2008 2:36	10DEC2008 13:11	25NOV2008 2:00	FALSE	FALSE
143743	1174	27NOV2008 0:33	27NOV2008 2:30	28NOV2008 12:31	26NOV2008 23:00	FALSE	FALSE
146819	1192	01DEC2008 14:58	01DEC2008 15:06	12DEC2008 14:00	01DEC2008 14:00	FALSE	FALSE
146865	1192	01DEC2008 14:58	01DEC2008 15:06	12DEC2008 14:00	01DEC2008 14:00	FALSE	FALSE
146879	1192	01DEC2008 14:58	01DEC2008 15:06	12DEC2008 14:00	01DEC2008 14:00	FALSE	FALSE
148685	1196	02DEC2008 3:27	02DEC2008 3:30	06DEC2008 15:09	02DEC2008 3:00	FALSE	FALSE
148690	1196	02DEC2008 3:27	02DEC2008 3:30	06DEC2008 15:09	02DEC2008 3:00	FALSE	FALSE
148847	1196	02DEC2008 3:27	02DEC2008 3:30	06DEC2008 15:09	02DEC2008 3:00	FALSE	FALSE
148928	1196	02DEC2008 3:27	02DEC2008 3:30	06DEC2008 15:09	02DEC2008 3:00	FALSE	FALSE

Figure 16 - Consistency Issue Example: NICU Measurement Outside the Overall Admission and Discharged Date

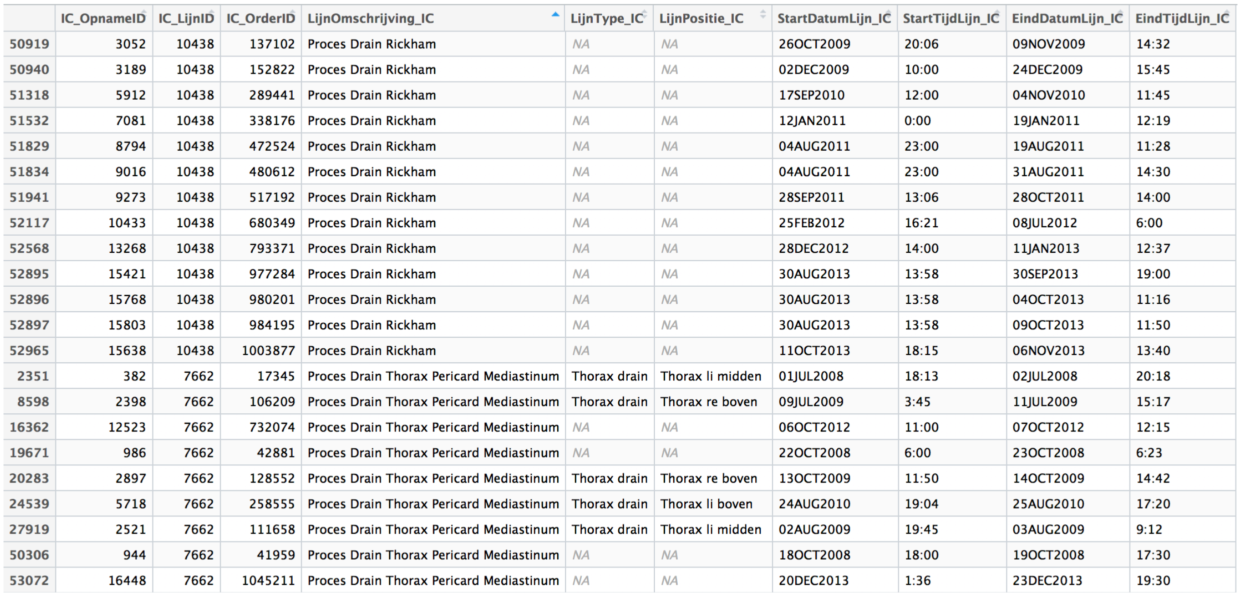
In addition to that, another example is being illustrated in Figure 16, where both *IC_Opname* and *IC_Meting_Numeriek* tables were merged, and as result, several records appear having the measurement time (*MeetDatumTijd_IC*) outside the interval defined by the overall and ICU admission date (*OpnameDatumTijd_OPN* and *OpnameDatumTijd_IC,* respectively) and ICU discharged date (*OntslagDatumTijd_IC*). The analysis was made upon a random sample of around ten thousand patients (*10.647*) in the *IC_Opname* table against more than two million measurements from the *IC_Meting_Numeriek* table, where from that experiment almost two thousand (*1.859*) patients were, as per the data, registered after (or discharged before) a measurement was realized. Thus, almost 20% of the data had problems in that specific scenario. Hence, this is clearly a consistency problem throughout the database schema, where one is allowed to register a measurement for a patient in the ICU that was not officially admitted/discharged into the system.

COMPLETENESS

Moving forward, when assessing data quality problems, one of the main aspects that is related with this topic is missing data. Missing data has been one of the most challenging data quality issues faced by researches during the years. Baraldi & Enders (2010) relates to it as one of the major statistical and design problems in research. For example, an incomplete dataset can make the application of algorithms and data mining techniques very difficult when encountering a not expected blank cell in the dataset. Additionally, a cautious analysis has to be made before trying to fix those types of problems, for example, removing records that contain missing data, once they could contain useful information after all. Furthermore, as defined by Baraldi & Enders (2010), there are three different mechanisms used to define the type of missing data:

- *Missing Completely at Random* (MCAR), when the motive why the variable is missing is independent of the variable itself or any other element, and therefore, considered random. For example, if a record of a given study is lost due to some accident or a rare malfunction of a system.
- *Missing at Random* (MAR), when the motive why the variable is missing is independent of the variable itself, but it may have some relation with other variables. Thus, some authors claim that the right name for this mechanism should be Conditionally Missing at Random instead (Graham, 2009), since, for being correlated to some other variable, it is not random at all. For example, if a survey asks both women and men about their personal information (e.g. age, height and weight) and for all the missing values related to the weight variable the respondents were women (since they are usually more concerned or inhibited about this matter than men), then the reason of missingness would be MAR as it is correlated with the gender of the respondents.
- *Missing not at Random* (MNAR) when the motive why the variable is missing depends on the variable itself among other things. For example, if a monitor cannot measure the heart rate of a child that surpasses a given threshold, the reason of missingness would be MNAR since in this case, the probability of a missing value is directly related with the ability of measure it.

Baraldi & Enders (2010) also state that by knowing which mechanism relates to missingness of the data, one could choose which technique could be used to better handle the missing data. Moreover, the authors also describe several techniques for handling such problem, such as deletion, where records with missing values are discarded; single amputation approach, where the one analyzing the data adds a value replacing the blank cell with a "suitable" value; multiple imputation approach, where, similarly to the single imputation approach, one creates a number of copies of the dataset, and inserts different "suitable" values in each one of them, analyzes all, and then combines the results into the final outcome; and maximum likelihood estimation, where all available data is analyzed in order to identify the values that have the highest chance of completing the sample data. Additionally, the option of using each method, as said, may depend on the missingness mechanism of the data. For example, deletion approaches are mostly recommended for MCAR, and multiple imputation and maximum likelihood estimation approaches for MAR (Baraldi & Enders, 2010). Thus, as can be seen, if one does not have a good understanding of the data, and in the mechanisms and processes involved in the data generation, the reasons why the data is missing can be misinterpreted or biased, and therefore improving the quality of the data (e.g. choosing the more suitable and efficient technique) could be more complicated leading to bad results in the analytical activities.

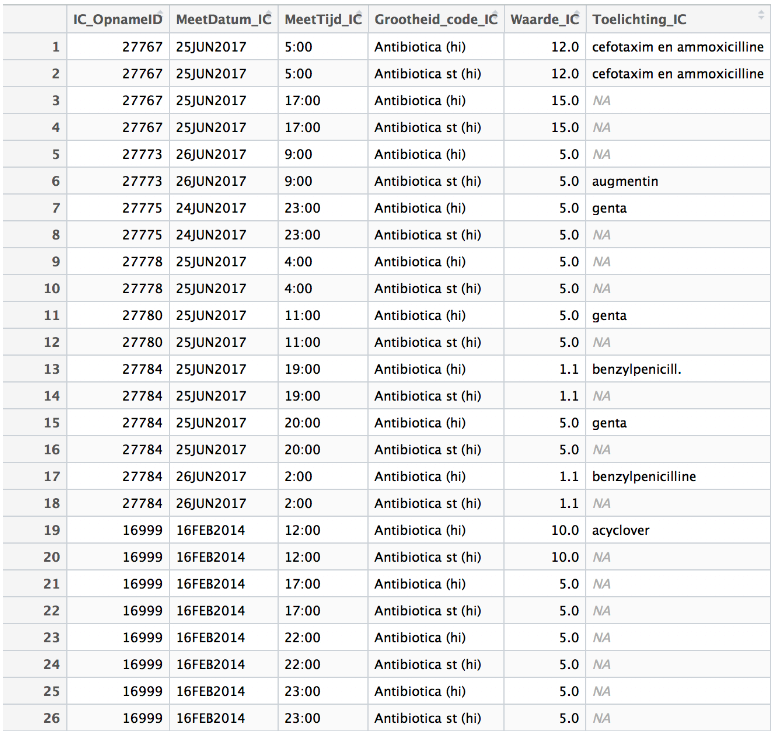


	IC_OpnameID	IC_LijnID	IC_OrderID	LijnOmschrijving_IC	LijnType_IC	LijnPositie_IC	StartDatumLijn_IC	StartTijdLijn_IC	EindDatumLijn_IC	EindTijdLijn_IC
50919	3052	10438	137102	Proces Drain Rickham	NA	NA	26OCT2009	20:06	09NOV2009	14:32
50940	3189	10438	152822	Proces Drain Rickham	NA	NA	02DEC2009	10:00	24DEC2009	15:45
51318	5912	10438	289441	Proces Drain Rickham	NA	NA	17SEP2010	12:00	04NOV2010	11:45
51532	7081	10438	338176	Proces Drain Rickham	NA	NA	12JAN2011	0:00	19JAN2011	12:19
51829	8794	10438	472524	Proces Drain Rickham	NA	NA	04AUG2011	23:00	19AUG2011	11:28
51834	9016	10438	480612	Proces Drain Rickham	NA	NA	04AUG2011	23:00	31AUG2011	14:30
51941	9273	10438	517192	Proces Drain Rickham	NA	NA	28SEP2011	13:06	28OCT2011	14:00
52117	10433	10438	680349	Proces Drain Rickham	NA	NA	25FEB2012	16:21	08JUL2012	6:00
52568	13268	10438	793371	Proces Drain Rickham	NA	NA	28DEC2012	14:00	11JAN2013	12:37
52895	15421	10438	977284	Proces Drain Rickham	NA	NA	30AUG2013	13:58	30SEP2013	19:00
52896	15768	10438	980201	Proces Drain Rickham	NA	NA	30AUG2013	13:58	04OCT2013	11:16
52897	15803	10438	984195	Proces Drain Rickham	NA	NA	30AUG2013	13:58	09OCT2013	11:50
52965	15638	10438	1003877	Proces Drain Rickham	NA	NA	11OCT2013	18:15	06NOV2013	13:40
2351	382	7662	17345	Proces Drain Thorax Pericard Mediastinum	Thorax drain	Thorax li midden	01JUL2008	18:13	02JUL2008	20:18
8598	2398	7662	106209	Proces Drain Thorax Pericard Mediastinum	Thorax drain	Thorax re boven	09JUL2009	3:45	11JUL2009	15:17
16362	12523	7662	732074	Proces Drain Thorax Pericard Mediastinum	NA	NA	06OCT2012	11:00	07OCT2012	12:15
19671	986	7662	42881	Proces Drain Thorax Pericard Mediastinum	NA	NA	22OCT2008	6:00	23OCT2008	6:23
20283	2897	7662	128552	Proces Drain Thorax Pericard Mediastinum	Thorax drain	Thorax re boven	13OCT2009	11:50	14OCT2009	14:42
24539	5718	7662	258555	Proces Drain Thorax Pericard Mediastinum	Thorax drain	Thorax li boven	24AUG2010	19:04	25AUG2010	17:20
27919	2521	7662	111658	Proces Drain Thorax Pericard Mediastinum	Thorax drain	Thorax li midden	02AUG2009	19:45	03AUG2009	9:12
50306	944	7662	41959	Proces Drain Thorax Pericard Mediastinum	NA	NA	18OCT2008	18:00	19OCT2008	17:30
53072	16448	7662	1045211	Proces Drain Thorax Pericard Mediastinum	NA	NA	20DEC2013	1:36	23DEC2013	19:30

Figure 17 - Missing Data Example: Empty Records That Should Contain Lines (Catheters) Information.

As an example of missing data in the Neonatology-RDP, a subset from the table *IC_Lijnen* containing fifty thousand records was collected and examined. From those records, variables such as 'line type' and 'line position' had more than twelve thousand missing values each. Thus, this information is missing in almost 25% of the cases. As can be seen in Figure 17, the field *LijnOmschrijving_IC* (line description) appears to have sometimes an impact on whether *LijnType_IC* (line type) and *LijnPositie* (line position) are filled or not, as it is the case with the value "*Proces Drain Rickham*", where for all cases both variables are empty, which would describe the missingness mechanism of the data as MAR or MNAR. However, for other descriptions (i.e. "*Proces Drain Throrax Pericard Mediastinum*"), there is no observed pattern on the missing values. Hence, it would require a further investigation on the data generation process to understand what may be causing the missingness of the data and the consequences of it for a specific analytical task.

Another example can be seen in Figure 18, using again the *IC_Meting_Numeriek* table, where for a random sample of almost twenty thousand measurements referring to antibiotics treatment, only four thousand have information about what kind of antibiotic was indeed given to the patient. For the rest of the approximately fifteen thousand measurements that were analyzed, that information is missing. The same happens to other types of measurements (besides the antibiotics treatment) where further description is also missing. The sample analyzed (considering all kinds of measurements) have approximately 2.28 million records, where from those, the field *Toelichting_IC*, which refers to the comments mentioned above, are empty in roughly 2.25 million records, that is, this occurs in 98% of the cases.



	IC_OpnameID	MeetDatum_IC	MeetTijd_IC	Grootheid_code_IC	Waarde_IC	Toelichting_IC
1	27767	25JUN2017	5:00	Antibiotica (hi)	12.0	cefotaxim en ammoxicilline
2	27767	25JUN2017	5:00	Antibiotica st (hi)	12.0	cefotaxim en ammoxicilline
3	27767	25JUN2017	17:00	Antibiotica (hi)	15.0	NA
4	27767	25JUN2017	17:00	Antibiotica st (hi)	15.0	NA
5	27773	26JUN2017	9:00	Antibiotica (hi)	5.0	NA
6	27773	26JUN2017	9:00	Antibiotica st (hi)	5.0	augmentin
7	27775	24JUN2017	23:00	Antibiotica (hi)	5.0	genta
8	27775	24JUN2017	23:00	Antibiotica st (hi)	5.0	NA
9	27778	25JUN2017	4:00	Antibiotica (hi)	5.0	NA
10	27778	25JUN2017	4:00	Antibiotica st (hi)	5.0	NA
11	27780	25JUN2017	11:00	Antibiotica (hi)	5.0	genta
12	27780	25JUN2017	11:00	Antibiotica st (hi)	5.0	NA
13	27784	25JUN2017	19:00	Antibiotica (hi)	1.1	benzylpenicill.
14	27784	25JUN2017	19:00	Antibiotica st (hi)	1.1	NA
15	27784	25JUN2017	20:00	Antibiotica (hi)	5.0	genta
16	27784	25JUN2017	20:00	Antibiotica st (hi)	5.0	NA
17	27784	26JUN2017	2:00	Antibiotica (hi)	1.1	benzylpenicilline
18	27784	26JUN2017	2:00	Antibiotica st (hi)	1.1	NA
19	16999	16FEB2014	12:00	Antibiotica (hi)	10.0	acyclover
20	16999	16FEB2014	12:00	Antibiotica st (hi)	10.0	NA
21	16999	16FEB2014	17:00	Antibiotica (hi)	5.0	NA
22	16999	16FEB2014	17:00	Antibiotica st (hi)	5.0	NA
23	16999	16FEB2014	22:00	Antibiotica (hi)	5.0	NA
24	16999	16FEB2014	22:00	Antibiotica st (hi)	5.0	NA
25	16999	16FEB2014	23:00	Antibiotica (hi)	5.0	NA
26	16999	16FEB2014	23:00	Antibiotica st (hi)	5.0	NA

Figure 18 - Missing Data Example: Empty Information about Type of Medication Given to a Patient.

TIMELINESS

Finally, as mentioned, most of the data within the Neonatal ICU is being generated by means of electronic health records (EHRs) and managed by MetaVision, which receives new values every minute. Before the data goes to the DWH though, it has to be validated (again referring to the Neonatal ICU and MetaVision) by the personnel (e.g. nurses), meaning that the parameters represent the truth. This validation occurs every hour, which means that a nurse for example, has to check on the values (one entire hour of measurements) and validate them before they are saved into the DWH. Regarding the Neonatology-RDP, their values, as already mentioned, are based on values from MetaVision and HiX that are stored in the DWH. The Neonatology-RDP is updated in a weekly frequency. Thus, it receives fresh data from MetaVision and HiX once a week, meaning that the most current data could be one day to one week old at the time of the analysis, which does not represent an issue since the update frequency is established and well-known, and if researchers and doctors need real time data, MetaVision can be accessed directly (with the right credentials), as well as other systems within the UMC which could provide real time data for analysis.

5.2.3 QUALITY ISSUES SUMMARY

The main findings described above can be seen in a summarized manner in the table below, where the problem found is briefly described making reference to the quality dimension it refers to, and the data table that it was extracted from.

Dimensions	*Data Table*	*Problem Found*
Accuracy	*IC_Meting_Numeriek*	Medication being written in up to 70 different manners.
	IC_Meting_Numeriek	NICU patients gaining up to three kilograms in the same day.
	IC_Opname	Length of stay in the NICU of less than 5 minutes for some of the patients.
Consistency	*IC_Events*	Around 2% of the random sample of 16.000 records was duplicated.
	IC_Opname merged with *IC_Meting_Numeriek*	Almost 20% of the *10.647* patients analyzed were officially admitted after or discharged before a measurement was realized.
	IC_Meting_Numeriek	different units of measurement (e.g. milliliters, grams, kilograms, etc.) are being utilized, where for some specific types of measurements the unit employed is being specified in the its name (i.e. grams for measuring the patient's weight), however, for others it is very hard to identify it.
Completeness	*IC_Lijnen*	Almost 24% (from a random sample of 50.000 records) of the information about line's type and position was missing from the table
	IC_Meting_Numeriek	From a random sample of 12.000 records referring to antibiotic treatment, the information of which type of medication was given is missing in around 67% of the cases.
	IC_Meting_Numeriek	98% of the variable designated for comments (*Toelichting_IC*) is empty, for a random sample of 2.28 million records. Although for some kinds of measurements no comment is needed, for others the value is missing in great proportions.

Table 6 - Quality Issues Summary

5.3 CURRENT QUALITY OF THE DATA

As stated before, the quality of a dataset is measured based on its "*fitness to use*" (Christoulakis, Spruit, & Van Dijk, 2015). Thus, to say if a dataset is suitable to be utilized in an analytical activity or not, depends on the project at hand and how it will be used on it. Hence, all the problems described above could mean nothing for a specific project, if, for example, the dirty data is somehow unnecessary for that particular problem. Therefore, to answer the third research question, it has to be clear that the overall data quality is different from the quality and applicability of the data in a specific scenario. Nevertheless, in an environment where huge amounts of data are being generated every day, and especially when such delicate matter as medical research is being based upon it, making sure that the data is trustworthy is essential.

However, sometimes it is tricky to identify the problems within the database, which enables wrong or missing data to enter the analysis without people being aware of it. Therefore, ensuring data quality should not only be a periodical activity that comes and goes every few months depending on whether there is a data analytical project or not, but a continuous activity that helps on improving the processes, creating better and consistent data every day. Based on that, quality dimensions (as the ones described above) were introduced to guide people through the aspects that have to be considered when assessing data quality. As mentioned in the beginning of the previous session, the issues and observations described above do not represent all problems that might exist within the Neonatology-RDP. Nevertheless, it was possible to see the many different problems that could be found by means of exploring the tables and by doing some simple data manipulations in *R*. Thus, based on the environment and the issues described above it is possible now to address the third sub-research question of this study.

SQ3: What is the current quality of the available data within Dutch (academic) Hospitals?

To answer this question, interviews results (Chapter 4), meeting with external analysts, and the exploratory data analysis conducted above have to be considered. Thus, initially, from the interviews, many participants complained about the overall quality of the data, and even referred to it as "*rubbish in, rubbish out*". Many problems were described such as different values referring to the same entity, wrong values being inputted manually and even by the machines when they were not even connected to the patient, typographical errors, etc. Later, meeting with external data analysts working within the WKZ were taken, where many data problems in dimensions such as accuracy (syntactic and semantic problems) and consistency (different values for the same data across the tables) were also mentioned. Thus, based on the issues that were described from the exploratory data analysis realized within the Neonatology-RDP, it is was possible to identify problems in three of the four quality dimensions assessed, as well as confirm some of the issues mentioned by both domain experts and data analysts.

In terms of accuracy, many datasets show quite a lot of syntactic issues, since there are many fields that allow free-text input, confirming what has been told by the domain experts and data analysts. For instance, a single type of medication was written in roughly twenty different ways, and many other examples can be seen throughout the tables within the Data Mart. In addition to that, semantic problems (although not as many as syntactic problems) were also observed, such as possible wrong values referring to admission and discharged date/time, which says that a patient stayed in the ICU for one single minute, or the huge gain of weight that babies had (up to three kilograms) in one single day. Moreover, although in small percentage, duplicated records were also found, which overpopulates the dataset with useless and redundant information. Regarding this issue, the amount of duplicated data is not the biggest problem here, but the possibility of duplicated data to be generated within the data, which could lead to a larger problem over time. In terms of consistency, apparently there are also problems regarding nonmatching values for the same entity throughout the database, as well as records that violates integrity constrains. Additionally, one consistency problem found was the ICU measurement events happening outside the admission/discharged date/time period. From the approximately ten thousand patients in the sample data examined, twenty percent had some measurement taken outside the admission/discharged period. Moving forward, in the completeness dimension, missing data is also a problem given that many tables within Neonatology-RDP present quite a lot of missing values. As illustrated above, for a sample of approximately fifty thousand records containing information over the lines added to the patients, the line type and position were missing on around twelve thousand tuples, which is just a small example of the many tables with missing information throughout the Neonatology-RDP. Data tables with many missing values usually displays that something is wrong with either the data generation process (e.g. ensuring that the value is filled in, by making the field obligatory), or the database design.

In conclusion, many data problems related to the quality dimensions were found, depicting flaws that probably extend to both data generation process and technical aspects (e.g. better definition of integrity constrains to avoid human errors). Regarding some examples given above, such as the wrong values concerning the weight of the babies, is hard not to ask further questions such as whether those values were typographical errors or not; or if, in reality, those values belonged to some other patient and were exchanged by mistake; if yes, if that could be happening to other variables as well and how often, etc. Thus, although many data quality problems exist and can be easily seen, some of them open new questions about the whole validity of the available data, which would require a more extensive data quality assessment to be checked. Nevertheless, with the knowledge that was acquired, it is clear that this matter requires more attention and continuously improvement to slowly transforming and creating a more trustworthy and consistent data environment.

6 A KNOWLEDGE DISCOVERY ADAPTATION

The motivation behind this study is to help domain experts on pursuing an analytical activity, allowing them to perform KD tasks, aiming for the extraction of knowledge and insights from datasets that could help the care of their patients (especially the preterm babies from the WKZ). Thus, in this chapter the focus will be given to the artifact development, where, by combining all information acquired and described in the previous chapters, a Meta-Algorithmic Model (MAM) will be devised with the objective of guiding domain experts through a KD process based on the CRISP-DM framework. Additionally, the proposed artifact is immersed within the Applied Data Science (ADS) context (explained in Chapter 3), which combines three disciplines (Data Mining, Engineering, and Domain Expertise), as shown in Figure 18.

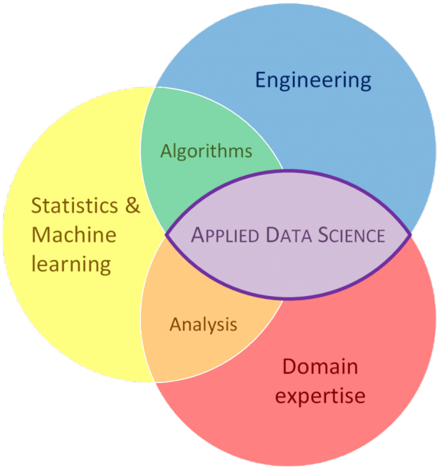


Figure 19 - Applied Data Science (Spruit & Jagesar, 2016)

The concept of MAM was inspired by the Method Engineering discipline, which is defined as a discipline to "*design, construct and adapt methods, techniques and tools for the development of information systems"* (Brinkkemper, 1996). In turn, MAM is given the meaning of the "*engineering discipline where sequences of algorithm selection and configuration activities are specified deterministically for performing analytical tasks based on problem-specific data input characteristics and process preferences*" (Spruit & Jagesar, 2016). Thus, its main objective is to devise a step-by-step guideline, composed by method fragments, used to guide experts from the application domain (without deep technical expertise) in the understanding of some design science's artifact. Furthermore, in the paper *Power to the People!* (Spruit & Jagesar, 2016), with the objective of also helping domain experts to perform data analytical tasks, meta-algorithmic fragments were created (based on the CRISP-DM framework) showing the steps that are recommended to be done in order to have the KD process facilitated. However, the focus was to help domain experts on all the way through Understanding and Preparing the data, as well on applying Machine Learning algorithms for binary classification activities on structured data. Thus, as somehow an extension of that, the focus of this chapter will be providing a guideline over both Data Understanding and Data Preparation phases, where it will contain the main steps that domain experts are expected to follow to go through the CRISP-DM process, as a starting point for their analysis focusing in a descriptive or verification DM task. Those DM tasks are usually done by means of exploratory data analysis, and by applying statistical techniques in order to identify useful patterns from the data for further human interpretation. The choice for these DM mechanisms was based on the domain experts' interviews and their experiences analyzing data, as well as what could be feasible given technical and other constraints. Additionally, why the focus was given to Data Understanding and Data Preparation will be explained later on this chapter. Moreover, in order to devise this method fragments, some important remarks that were found and described in the previous chapters need to be taken in consideration. They are described in Table 7 below.

Aspect	*Comments*
Handling data	• No specific data mining tool is common to all domain experts. Tools vary between Excel, SPSS, R, and SAS. • Large datasets available, which makes things harder for domain experts when preparing the data. • There is no '*Master-file*' integrating all variables required for answering a question in a single view.
Quality of the available data	• Accuracy problems such as typographical errors, abbreviations, wrong data, and duplicated data. • Consistency problems, such as different values of the same entity across the tables and non-normalized data. • Completeness problems such as a lot of missing data throughout the many tables, where the mechanism of missingness is not certain;
Technical skills, and DM methods and tools	• Most available analytical tools do not support less technical people. • Domain experts have limited or non-existent programming skills. • 'Black-box' problem, where many techniques and tools only provide a non-transparent execution of DM algorithms and techniques
External data analysts	• Usually they do not have the domain knowledge to make the most of their analysis. • Domain experts complain about not having control over the process by delegating all data mining tasks to external analysts, and the long time waiting for an outcome.
Wishes	• Step-by-step guidance throughout the KD process • No black-box, thus, more control over the project development and outcomes

Table 7 - Remarks and Problems faced by Domain Experts regarding KD

Some of the topics mentioned above played an essential role when identifying important aspects that had to be considered when developing the MAM, such as: lack of technical (programming) skills; wish to avoid the 'black-box' scenario; different tools for handling data; and no master-file to base the analysis upon. As explained earlier, domain experts know what KD represents, and have an idea of its overall process. Also, based on their knowledge and experience, an analytical task always starts with a hypothesis or question for which they seek answers. Then, data is selected and filtered, and a statistical analysis is made (such as regression and correlation analysis for example) to achieve the outcome and to answer the research question. However, this process is filled with problems and difficulties that hinder their analytical power. Starting with the tools available, as stated in Chapter 2, only few of them support citizen data scientists, thus, these tools do not make things easier for people with less technical skills to do data analysis. Additionally, given time constrains (taking in consideration that these professionals have their day filled with different activities others than Data Analysis) is hard to stop and learn how to use a tool from scratch, as it can be very time consuming. Moreover, making use of the tools available (which they are already used to) domain experts described their difficulties when trying to analyze data mentioning problems such as handling large sets of data; integrating it, and therefore, not having a master file where they can find all variables that they need to perform the analysis; handling the dirty data (e.g. missing data, wrong values, etc.) within the datasets; non-transparency when applying DM methods within the existing tools, therefore, not being sure of the outcomes' validity, etc. Thus, most of the difficulties described were related to preparing the data prior to the analysis.

Moreover, in order to devise the MAM, the next section will have the purpose of addressing sub-research questions ***SQ4***, ***SQ5*** and ***SQ6***, which will provide basis for the artifact construction.

6.1 Data Analytics for Domain Experts

As per the CRISP-DM framework, all main phases of the KD process and their respective outcomes are very well defined. However, there is no distinction of how phases should be pursued (and what outcomes are expected) depending of the type of user who is following the guideline. For example, the majority of data driven tasks are being done mainly by data analysts and scientists, who spend hours, days, and even weeks, understanding and mapping inconsistencies and potential problems, and applying DM methods on the data. In other words, those professionals are usually hired to work entirely and intensively with data and in extracting knowledge from it. On the other hand, domain experts usually have different priorities where unfortunately the focus is not KD. Hence, they do not have the same amount of time to spend on data analysis, and therefore, not all KD phases will be conducted with the same level of details when compared with data experts. Thus, as the focus is different, the way of pursuing the phases from CRISP-DM should be different as well. Phases that consume quite a long time such as Data Understanding and Data Preparation, should focus only in tasks that would indeed facilitate the KD process for these professionals, instead of, for instance, mapping and removing every single inconsistency of the dataset. Additionally, regarding DM orientations to be followed, given time and technical constrains, not all of them are feasible to be done by domain experts without a guidance or help from data experts, such as predictive data mining. Therefore, domain experts should focus on either Verification or Descriptive Data Mining (DDM), where, as explained in Chapter 2, most techniques rely on exploratory data analytics.

The CRISP-DM starts with the Business Understanding phase, in which the goal is to acquire knowledge over the business itself and data mining goals. Considering the great expertise that the audience of this project has over the domain, the only activity left (which still should be really straightforward for the domain experts), should be translating the hypothesis or research questions that they might have into an analytical project. Thus, the more complex activities in the KD process for domain experts, that are focus of this research, is the Data Understanding and Data Preparation phases. Hence, below both phases will be addressed considering what has just been said, and how those activities should be adapted for being used by domain experts.

6.1.1 Data Understanding

Data Understanding, as per the CRISP-DM framework, consists of four sub-steps (Data Collection, Data Description, Data Exploration, and Data Quality Assessment) which guides the user on how to fully comprehend in detail the data that will be used in the analysis. Understanding the data includes not only its content, but also its structure, as well as identifying data quality problems within it. Additionally, the CRISP-DM guideline states that it can also be the start of an exploratory data analysis, where one starts looking for hidden patterns and useful information within the data. Furthermore, most of the tasks involved in this phase are time consuming, and sometimes some tasks require some technical skills, as it is the case of Data Quality Assessment, in which data has to be (most of the times) manipulated, constructed and transformed in order to investigate all possible scenarios. However, as mentioned above, it is not accurate to assume that different types of users (i.e. data scientists and doctors) will follow the activity with the same level of detail and produce the same outcomes, hence that leads us to the following sub-research question:

SQ4: What aspects have to be considered, regarding the Data Understanding phase of the CRISP-DM, so it can be adapted for the domain experts needs?

Based on the steps from the Data Understanding phase defined by the CRISP-DM, the following table illustrates the aspects that can be handled by the domain experts, as well as what could be more difficult for them, in the format of pros and cons.

Pros	• Knowledge over the data content, e.g. understanding of the attributes and its values • Identification of data quality issues (especially in the accuracy and completeness dimensions). • Statistical knowledge to perform exploratory data analysis.
Cons	• Scattered data (e.g. different tables and/or databases) • Large sets of data, which makes it harder to visualize data problems • Data structure and how tables are related may be unknown, and the documentation of the environment can be unavailable.

Table 8 - Data Understanding vs Domain Experts

Since the beginning of this essay, it is stated that the main aspect that differentiates domain experts from the majority of external analysts, is the huge domain knowledge that these professionals have (e.g. knowledge over the organization, its environment, its people, and its data content). While data scientists would spend quite some time mapping and understanding the meaning of the attributes and its values, domain experts should in theory already be familiarized with them. On the other hand, the database environment in which the information is stored, and of course, where and how the data is stored, is something that most domain experts would have difficulties to explain, which was also said during the interviews described in Chapter 4. Although domain experts already have a deep understanding of the variables and information within a dataset, is not always easy for them to know where to find (and how to access) such information. Also, data might be spread across many tables within the database, and examining the datasets one by one, trying to find the right information to be used in the analysis, can be very time consuming and demotivating. Hence, one of the main center of attention for this phase when dealing with domain experts should be on understanding the data environment, and by doing this, facilitating the access to the data. Thus, by focusing on first understanding the data tables arrangement, data tables relationships, attributes within each table, etc., before focusing in the content of each data table, the user can have a better idea of where each information is actually located, and how each table is related to the others. Moreover, as mentioned earlier, based on the DM orientations in which domain experts are expected to follow, the second topic on how this phase should be adapted for domain experts relies on allowing domain experts to start and complete the exploratory data analysis within this phase, till the point of achieving the analytical project goal.

Therefore, to complete, two main topics should be highlighted to facilitate the Data Understanding stage for the domain experts: *understanding the data environment* and *focusing on the exploratory data analysis*. Moreover, regarding this phase, after addressing the two subjects mentioned above, it is possible that domain experts would have an advantage over data analysts and scientists. It is true that some data manipulation has to be made to extract all important and relevant information, such as outliers and wrong data, however, the knowledge over the content facilitates the other tasks considerably.

6.1.2 Data Preparation

Moving forward, as it should be clear now, preparing data can be very time consuming depending on the data quality level one wants to achieve. For some data mining methods, ensuring that only valid and clean data enters into the analysis is mandatory for a good outcome, as it is the case for classification and predictive DM methods for example. However, that does not mean that for the other types of DM activities data quality is less important, yet, by means of exploratory data analysis and knowledge over the quality issues, problems can be considered and avoided during the analysis, and the quality improvements, if required, made on demand. Moreover, as stated before, most data preparation activities are highly situational as they are most of the times realized specifically for one analytical task, based on the data available and the project's objectives. The CRISP-DM's Data Preparation phase depicts five main activities to be conducted in order to prepare the data (data selection, data cleaning, data construction, data integration, and data formatting), where among them, many activities are possible, such as data deduplication, data transformation, filtering, data subseting, data appending, aggregation, data merging, data normalization, etc. Additionally, CRISP-DM does not make very clear (especially for less technical people), for example, what 'Data Construction' means only by looking at the model. Furthermore, as there are too many options of how to approach a preparation task, it is quite hard to map all of them into a model. Hence, that lead us to the next research question:

SQ5: How and which data preparation step's activities should be included in the model, and how they can be adapted for the domain expert's needs?

The Data Preparation goal for domain experts, based on their technical skills, time constraints, and type of DM orientation to be pursued, should only focus in making the dataset simpler and smaller for further analysis. Based on all that, any proposed solution has to be straightforward, since, besides the time constraint aspect, people tend to get demotivated if stuck into something for too long without much progress. Nevertheless, as it is hard to specify the exact activities that should be pursued in this phase (given that many activities are highly situational), two aspects can be highlighted, as ones that fit domain experts' needs (based on the difficulties stated by them during the interviews) and constraints, and at the same time, aligned with the phase's goal stated above: (1) creating a unified view for the data, (2) constructing the dataset with the purpose of making any analytical task easier afterwards.

One of the main problems stated by domain experts when trying to analyze data, was the fact that the information was most of the times scattered in different places, and there was no 'master-file' containing all variables and information needed for answering the analytical project goal. Thus, the Data Integration sub-activity (part of the CRISP-DM) should be highlighted within the model and pursued whenever possible, since it is important to create a simpler and unified dataset in which domain experts can base their analysis upon, rather than allowing them to struggle on analyzing many single datasets. The second aspect stated above is the construction of a simpler dataset. Thus, some of the activities depicted by CRISP-DM should be pursued here as well, such as Data Selection, Data Formatting, Data Cleaning, and Data Construction (renamed to Data Engineering not to confuse with the second major aspect mentioned above). Regarding the first, there is no point in having a huge dataset, with lots of variables if one is only interested in some small subset of it. Also, by reducing the size of a dataset, statistical methods can be easier applied, and the data can be better described and understood. Formatting data in the right way, can also help those professionals on better analyzing and visualizing the data content. Moreover, domain experts also complained about data quality problems that they found or knew that existed, such as missing data for example. So, cleaning the data and removing some of the major "easy to fix" problems, would impact greatly the size and simplicity of the dataset (such as removing the missing data). Finally, sometimes engineering new attributes based on already existing ones, and adding them in the dataset, would help on speeding up the analysis, such as calculating the body max index out of weight and height attributes. Hence, the activity of engineering data is also an important task aligned with the objectives mentioned above.

To conclude, all sub-phases from the CRISP-DM are applicable for domain experts as well, of course, adapted for domain experts and aligned with the new objectives of this phase stated earlier. Nevertheless, the focus should first be first given to the Data Integration sub-step, where once the "master-file" created (if feasible), the other activities such as Data Formatting, Data Engineering, Data Selection and Data Cleaning can be pursued. All those activities should be done with the only objective of creating a simpler and smaller dataset. For example, Data Cleaning for domain experts will suggest cleaning out just easy to see and straightforward problems, and not to fix every single error or inconsistency out of the data.

6.1.3 The Framework Adaptation

Considering everything that has been said during this chapter, when focusing in domain experts, there is no point in spending time pursuing the CRISP-DM's phases exactly as they were designed. The overall process has to be simplified. First, the Business Understanding phase remains important, however, domain experts do not require a guideline for it since they are already immersed into the business, and most data analytical tasks are based on questions and hypothesis extracted from their daily routine. Phases four to six (Modeling, Evaluation, and Deployment) from the original CRISP-DM are also not entirely the focus here, considering the DM methods suggested (Verification and Descriptive), and the fact that no complex model would be created for deployment (e.g. prediction and classification models). Of course, these phases could still be totally feasible if a more complex project has to be pursued and should still be accessible in the model. However, those would require acting together with data external analysts. Thus, that leave us with the both Data Understanding and Data Preparation phases, and the last sub-research question:

SQ6: What is the best way to guide domain experts throughout the Knowledge Discovery process, based on the CRISP-DM, so they can most likely successfully accomplish the analysis?

Based on the type user who is following the CRISP-DM model, the way to navigating though it should be different given all reasons already explained during this chapter. Focusing on domain experts, only the three first phases should be highlighted for them to go through alone (without the help of a data analyst). Business Understanding, as stated earlier, should be the starting point still, however, with the objective of only translating the research question or hypothesis into an analytical project goal. Next, considering all that has been said above about the Data Understanding and Data Preparation phases, the certainty that preparing the data is most of the times needed to the full comprehension of the data content and to perform a full data quality assessment, and the fact that it was suggested for domain experts to in fact pursue the whole data analysis within the Data Understanding phase, two alterations are proposed in the original CRISP-DM model, as can be seen in Figure 20.

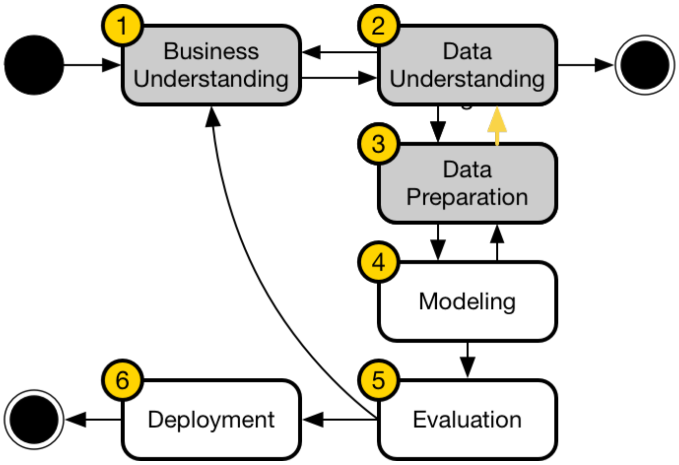


Figure 20 - CRISP-DM for Domain Experts

First, a two-way relationship was added between Data Understanding and the Data Preparation phases (arrow in yellow). That way, it is possible to prepare and manipulate the data prior or during the exploratory analysis, as well as (if desired) to fully examine data quality problems within the data, similarly to what has done in Chapter 5. That was an unexpected non-existent relationship in the original CRISP-DM, given the fact that even data analysts in order to fully understand the data, take advantage of some data preparation tasks in order to the explore it. Second, a new ending point was added after the Data Understanding phase. Thus, the process now has two ending points depending on the activity to be done, and the type of user who is conducting the analysis. The ending point after the Data Understanding phase would mean that domain experts would have concluded the exploratory data analysis, answered their research questions, and no further interactions are needed. Finally, the steps in gray are the ones suggested for domain experts.

7 A CRISP-DM BASED META-ALGORITHMIC MODEL

In this section the MAM, which is the main objective from this thesis, will be shown and explained. The models were created using two different modeling notations: The Process-Deliverable Diagram (PDD) and a Business Process Modeling Notation (BPMN) based syntax. A PDD consists of two integrated diagrams, which express both process-view and deliverable-view of an artifact construction (van de Weerd & Brinkkemper, 2008). The process-view, also called meta-process modeling, consists of a model showing the activity flow of a specific process, while the deliverable-view, also called meta-deliverable modeling, comprises the expected outcome of an activity, in the form of a concept diagram (van de Weerd & Brinkkemper, 2008). On the other hand, the BPMN depicts the overall process, better illustrating the main activities, conditional events, and their disposal and flows within the three highlighted phases from Figure 20.

The choice for having both modeling notations was made since it was important to highlight that the whole process was indeed making use of what was proposed regarding the CRISP-DM framework adaptation in the previous section, and also to ensure the comprehension of the activities that are suggested to be performed, as well as what outcomes are expected from each task. Thus, the model then focuses on the three first phases of the CRISP-DM framework, and it is composed of six main activities, as can be seen in Figure 21, where the overall process is depicted. The whole PDD can be found below, where it will be followed by the description and further explanation of each of its activities.

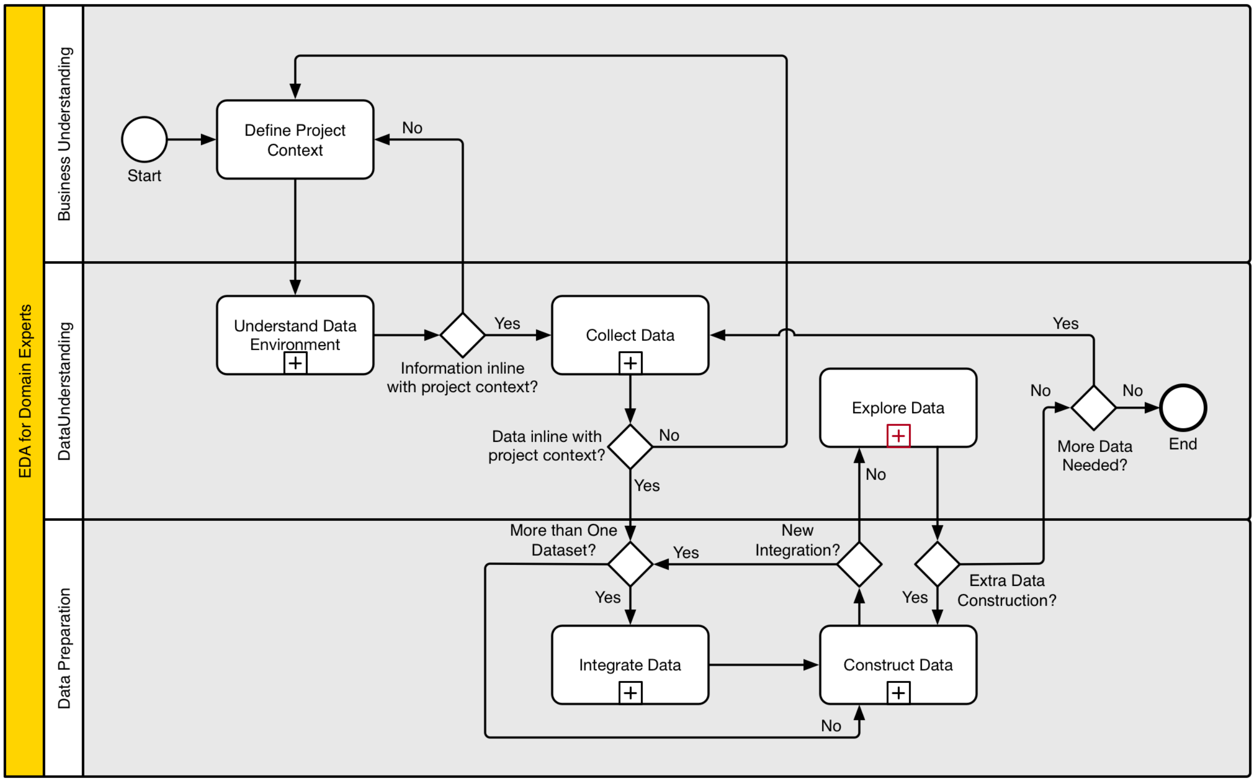


Figure 21 – EDA for Domain Experts

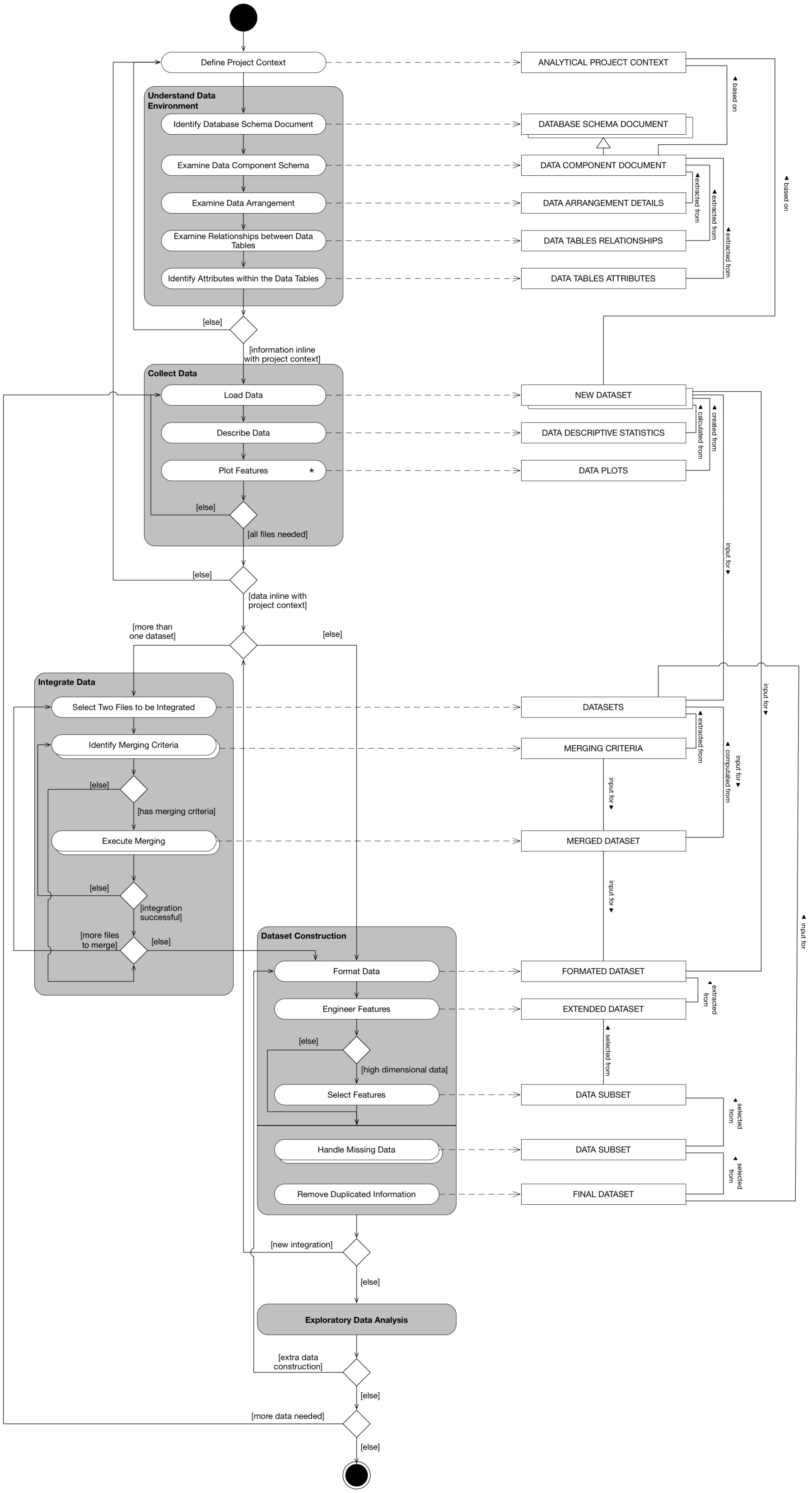


Figure 22 - Process-Deliverable Diagram Overview

7.1.1 CRISP-DM FOR DOMAIN EXPERTS: BUSINESS UNDERSTANDING

The Business Understanding phase is where all KD activities should begin with, since it has the purpose of contextualizing the people who are involved in the analytical task with the company's environment, vision, and on defining a project that correlates with the business goals. Furthermore, when the focus is on domain experts, this phase should be smoothened, once that these professionals already have a deeper knowledge over the business, processes, environment, and goals when compared with external data analysts. Thus, the MAM shows only one main activity for this phase, called *'Define Project Context'*.

ACTIVITY: DEFINE PROJECT CONTEXT

This activity does not require much explanation, and it has the purpose of being a starting point for the analytical task to be pursued. Although domain experts have most of the times a hypothesis or research question already defined, this activity has the goal of making sure that domain experts (before diving into an analytical task) will organize their ideas and translate the research question into an analytical project by defining a project goal about what they want to analyze and achieve when performing KD, and what would be required to do in order to achieve the expected results.

7.1.2 CRISP-DM FOR DOMAIN EXPERTS: DATA UNDERSTANDING

Next, considering that the hypothesis and research questions about a given situation are already established, and the goal of the analytical task defined, the process should move on to the Data Understanding phase, which focusing in the domain experts, as described in the previous section, should be dedicated extensively to understand the database schema, understand how to access its content, understand how the data is distributed, and collect the required information for the analysis. In addition to that, this phase should also be where the exploratory data analysis is performed, as described in Section 6.1.1.

Differently from CRISP-DM where the Data Understanding phase already starts with collecting and loading the data, in this research it is believed that understanding the data environment and how the information is stored is of high importance for the comprehension and further data manipulation by the domain experts, and therefore deserves a separate activity in the beginning of this phase (*"Understand Data Environment")*. Thus, this task focuses on providing means towards a good understanding of the data environment, that is, helping the user to comprehend how the data is scattered within the database, how each table relates to the other, where and how the data is stored, etc. After the fulfillment of this activity, the following step, called *"Collect Data'* is used for, as its own name says, collecting and describing the data that will be used in the analysis. Furthermore, by first understanding the environment in which the data is stored, the task of collecting the right dataset (accordingly to the project scope) and understanding its content should be easier and straightforward. Furthermore, those two activities can flow back to the Business Understanding phase if the project context need some refinement based on the data information that was retrieved, or if some constraint is found that requires a new project definition. Both activities will be explained in detail next.

The third main activity that belongs to the Data Understanding phase of this model is called *'Explore Data'*, which represents the actual exploratory data analysis activity. Since the tasks that are comprised within this activity are highly situational and depends entirely on the data and project at hand, and once this task is out of the scope of this project, no sub steps will be provided for it. Nevertheless, once in the *'Explore Data'* activity, flows to both *'Collect Data'* and *'Construct Data'* are allowed if one needs new data to analyze, or maybe further preparation over the existing dataset to continue the analysis. More about the Data Preparation phase will be explained along this chapter.

ACTIVITY: UNDERSTAND DATA ENVIRONMENT

The *'Understand Data Environment'* activity contains five sub-steps as depicted in Figure 23. It starts from the assumption that documents that describes in detail the database schema from a given business are updated and available for checking. Hence, the first proposed activity called '*Identify Database Schema Document*' refers exactly to the task of identifying and retrieving those documents that describe and help on understanding the company's database schema. This document should contain explanations about the data tables within the database, how they relate to each other, which features are comprised within each table, a brief explanation over the attributes that are being stored into them, etc. The next step is find and examine the information just mentioned focusing in the data component in which the analysis should be based upon. For example, the WKZ's data environment is composed by many data components where the Neonatal-RDP is one of them. Hence, if one is planning to perform an analytical task using neonatology data, the focus should be given in the understanding of the database environment for that specific data component.

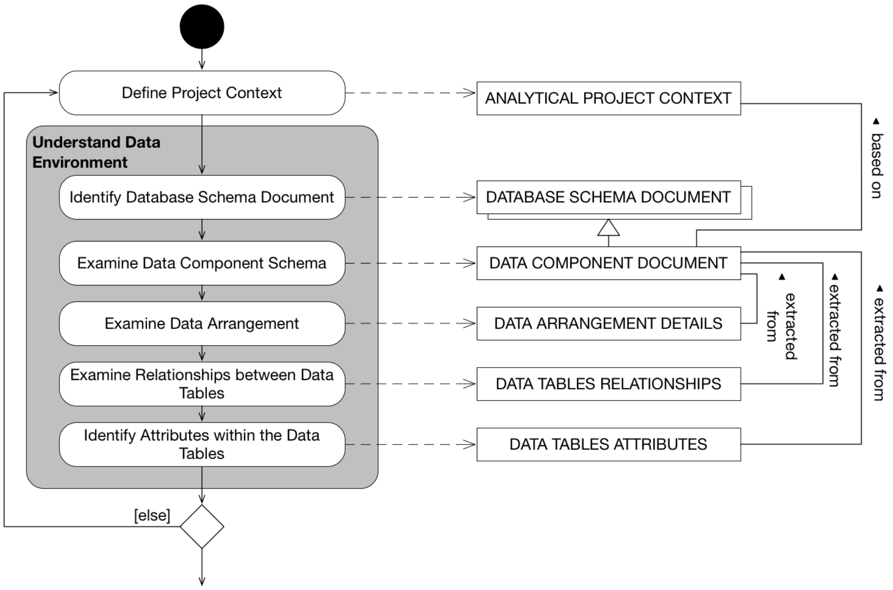

Define Project Context
ANALYTICAL PROJECT CONTEXT
Understand Data Environment
Identify Database Schema Document
DATABASE SCHEMA DOCUMENT
Examine Data Component Schema
DATA COMPONENT DOCUMENT
Examine Data Arrangement
DATA ARRANGEMENT DETAILS
Examine Relationships between Data Tables
DATA TABLES RELATIONSHIPS
Identify Attributes within the Data Tables
DATA TABLES ATTRIBUTES
◂ based on
◂ extracted from
◂ extracted from
◂ extracted from
[else]

Figure 23 - Data Environment Understanding and Data Collection

The third sub-activity, *"Examine Data Arrangement"* focuses on examining and identifying which data tables are available, their meanings, their purposes, and how they are arranged. By doing this, the user could already have a better understanding about which type of information is available and where they are being stored, which will be useful for further activities. The next task to be performed is to identify how data tables relate to each other, that means, which variables and attributes allow the establishment of a relationship between two tables. In technical terms, this kind of relationship is usually expressed by a Primary Key and/or Foreign Key relation, which determines the attribute that represents the linkage factor between two data tables. This task should provide insights about integrating data and what makes sense integrating and what does not. Finally, the last sub-activity called *'Identify Attributes within the Data Tables'* has the purpose of understanding which variables are being stored within each table, and how to refer to them. No content is being analyzed so far, only attributes and their meanings.

ACTIVITY: COLLECT DATA

After acquiring a reasonable knowledge over the data environment, the *"Collect Data"* activity has the purpose of loading the data files that are required for the analytical task and getting familiarized with them. Thus, this activity has three components, as shown in Figure 24.

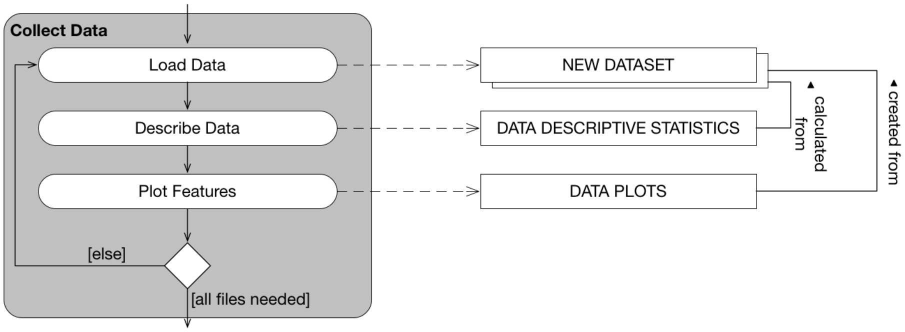


Figure 24 - Collect Data

The process starts by loading the data file into a tool of choice, such as loading a .CSV file into Excel or R. Next, it is recommended to describe the data and then plot its features to start exploring the data and the relationship between its variables. As mentioned in the previous section, domain experts should in theory not have difficulties to understand the meaning and values of a variable within a dataset that is related to their domain. Thus, this activity follows the same idea and goals from the Data Description sub-step provided by the CRISP-DM within the Data Understanding phase, therefore, describing the data has the purpose of being a straightforward activity that aims to provide an overall picture of the data's content, such as some descriptive statistics, how the data is distributed, and some of its quality problems such as quantity of missing data related to a given variable. Ways of pursuing this activity vary from eyeballing to applying statistical functions to extract that information. Furthermore, the Plot Features activity refers to the development of graphs that could help on better visualizing relationships between variables or how the data is distributed in a graphical manner. Lastly, if more data files are needed, the process can be repeated.

7.1.3 CRISP-DM FOR DOMAIN EXPERTS: DATA PREPARATION

Moving forward, after collecting and understanding the data, the acquired datasets should be prepared for the analysis accordingly to the user needs. Based on what has been explained in previous sections, the main objective of the Data Preparation phase for domain experts should be on creating a simplified and smaller dataset for an exploratory data analysis. Thus, considering the user group to whom this MAM is being created, two main activities are suggested within this phase: *'Integrate Data'* and '*Construct Data'*.

The Data Preparation phase starts at a conditional event where, if during the previous activities more than one dataset was collected, the user has the option (if suitable and feasible) to merge those datasets into one master-file, thus, bringing all information into one place and reducing the amount of data files to be analyzed. However, if just one dataset was collected, the user should go straight to the *'Construct Data'* task, where the objective is to clean, format, select and engineer the features from the given dataset, similarly as proposed originally by the CRISP-DM framework. Additionally, if integrating the data is not feasible, the user should follow the steps from the *'Construct Data'* activity for all datasets available, and afterwards, depending on the changes that were made, go back and try to merge them again. Finally, as proposed in Figure 20, after completing these activities and starting the exploratory data analysis, the user has the option and liberty to come back to the Data Preparation phase if needed.

ACTIVITY: INTEGRATE DATA

Integrating datasets can be a tricky activity for those who do not have experience doing it. First, one has to know what can be integrated and what makes sense integrating. For domain experts, most of the information needed about this matter should have been acquired during the *'Understand Data Environment'* activity, where the data tables and their relationship were examined. Thus, the *'Integrate Data'* activity, depicted in Figure 25, consists of three tasks: *'Select Two Files to be Merged'*, *'Identify Merging Criteria'*, and *'Execute Merging'*. The first activity, as its own name says, is the selection of the two convenient datasets to be integrated. After that, the merging criteria between those files have to be identified, where the correct and successful identification of such criteria is mandatory for a successful integration between the two datasets. Thus, the last activity, *'Execute Merging'* should only be pursued if the merging criteria is indeed found.

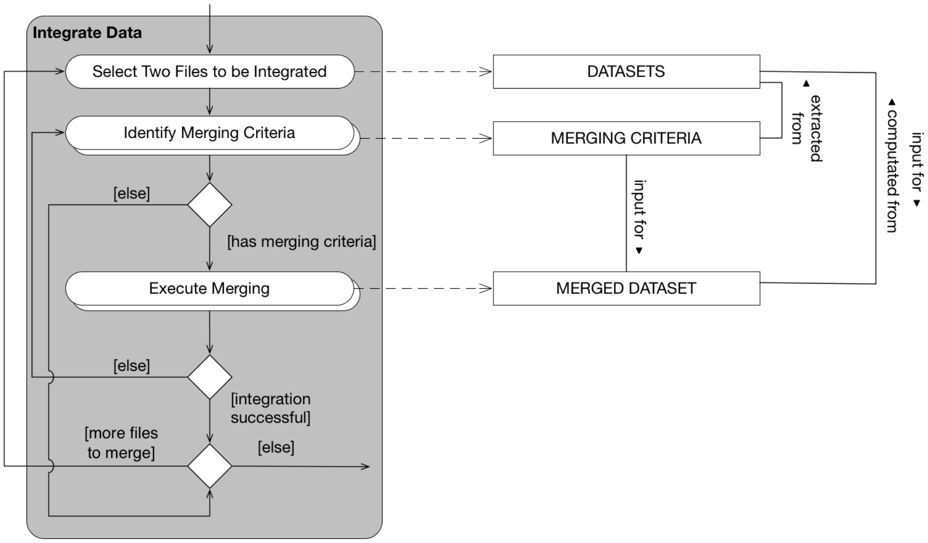


Figure 25 - Integrate Data

As mentioned earlier, the merging criteria is usually defined by a primary and foreign key relationship between two data tables, that is, a common attribute that allows to identify matching records between two datasets. In order to help domain experts to successfully identify the merging criteria four sub-tasks are suggested as shown in Figure 26. First, a user has to identify common variables (even if with different names) between both datasets, with the goal of reducing the number of variables that could be defined as merging criteria. The next step is to select from those variables the one(s) that uniquely identify single observations for each dataset, that is, the attribute(s) that permits to differentiate one record from another.

For example, the weight attribute could exist in two datasets, however, is not possible to uniquely identify each record only by the weight value. On the other hand, if instead of weight, a patient identification code is common in both datasets, it is more likely to be a good choice for the merging criteria since that attribute uniquely identify each record from the table, knowing that it belongs to that specific patient. Furthermore, if both activities are successfully fulfilled, the last step is to make sure that the integration will be effective with the choices previously made.

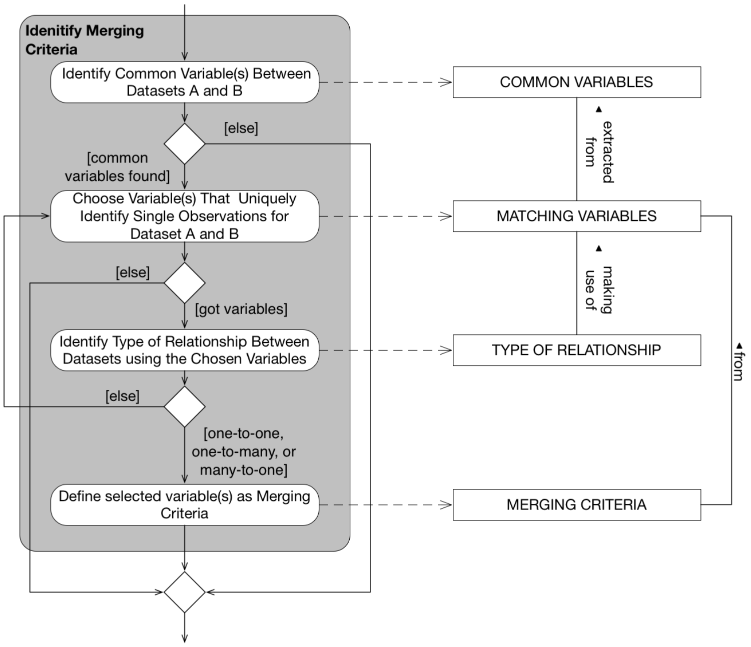

Idenitify Merging Criteria
Identify Common Variable(s) Between Datasets A and B
[else]
[common variables found]
Choose Variable(s) That Uniquely Identify Single Observations for Dataset A and B
[else]
[got variables]
Identify Type of Relationship Between Datasets using the Chosen Variables
[else]
[one-to-one, one-to-many, or many-to-one]
Define selected variable(s) as Merging Criteria
COMMON VARIABLES
extracted from
MATCHING VARIABLES
making use of
TYPE OF RELATIONSHIP
from
MERGING CRITERIA

Figure 26 - Identify Merging Criteria

Thus, the final sub-activity is to identify the type of relationship (in terms of cardinality) between the two datasets considering the selected attributes as the merging criteria. Four distinct types of relationship cardinality exist, they are: One-to-One (1:1), One-to-Many (1:n), Many-to-One (n:1), and Many-to-Many (n:n). A One-to-One relationship between two tables exists when one record from Dataset A relates to exactly one record from Dataset B. On the other hand, the One-to-Many relationship happens when one record from Dataset A relates to more than one record from Dataset B, as the same way that Many-to-One is when more than one record from Dataset A relates to exactly one record from Dataset B. Finally, the Many-to-Many relationship in when many records from Dataset A relate to many records from Dataset B. Although in theory datasets could be merged despite the type of relationship, for the domain experts only the first three mentioned are suggested to be used, since the Many-to-Many could create very complex datasets, with several duplicate records, and even wrong information. Hence, if the data relationship cardinality using the merging criteria extracted from previous steps happens to be one-to-one, one-to-many or many-to-one, the rest of the *"Integrate Data"* activity should be pursued, and the integration is most likely to be successful. However, again, if the cardinality happens to be many-to-many, it is suggested for the domain expert to choose other variables as merging criteria. Additionally, if there are no variables left to choose from, the user should then drop the integration activity and move on to the *"Construct Data"* task.

The last step for integrating the data is the actual execution of the merging task. It can be pursued in several ways, using different tools and notations. Moreover, the goal is not enforcing the user to choose one tool, and teach how to execute such task, but to provide the knowledge of what is needed in order integrate datasets. Thus, despite the means, the parameters which are required doing so are basically the same in any tool available. Two of them were already defined in the previous activities: the datasets to be merged, and the merging criteria. In order to conduct the '*Execute Merging'* activity, those parameters have to be known, since they are now going to be used. Thus, the only missing parameter is the merging type, which represents the definition of the content that should be returned after the conclusion of the merging task. Figure 28 illustrates the *'Execute Merging'* activity, where four merging types (the most commonly used and known) are suggested, they are: Inner Join, Left Join, Right Join, and Outer Join.

To better understand each merging type given above, consider Figure 27 and the explanations below:

- *Inner Join:* probably the most commonly used merging type, it returns all records from Dataset A which have a corresponding matching record in Dataset B (illustrated in the upper-left corner from the image below);
- *Left Join:* it returns all records from Dataset A regardless if that record has a match or not in Dataset B, together with the matching records (if any) from Dataset B (illustrated in the lower-left corner from the image below);
- *Right Join:* similar to the *Left Join*, it returns all records from Dataset B regardless if that record has a match or not in Dataset A, together with the matching records (if any) from Dataset A (depicted in the lower-right corner from the image below);
- *Outer Join:* the last merging type returns all records from both tables, matches and un-matches (depicted in the upper-right corner from the image below);

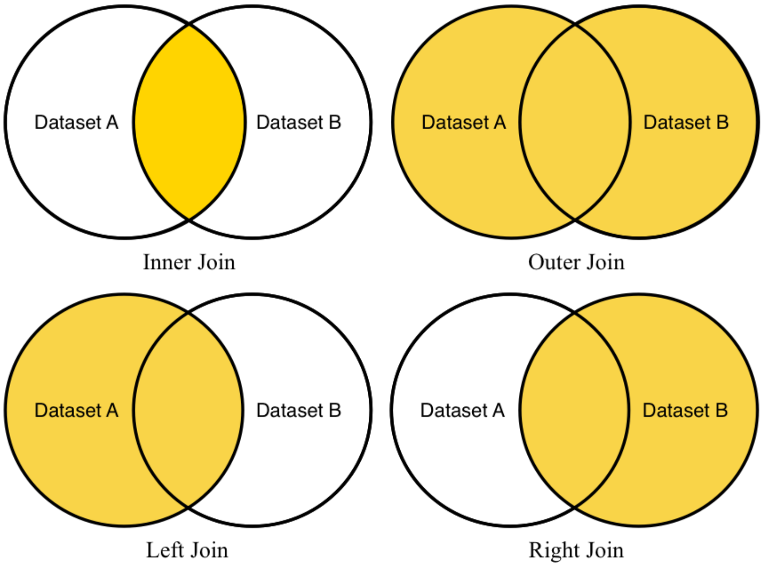


Figure 27 – Data Joins

After choosing the right merging type for the given situation, all three main components for integrating two datasets were identified, what allows the merging task execution to be done, using the most convenient method and tool for the user. In *R* for example, one could use the *merge* function and add the correct parameters based on the MAM above. The whole data integration process can be repeated if the merging was not successfully done (which could happen due to a bad judgment of the merging criteria) or if there are more files collected in previous steps to be merged.

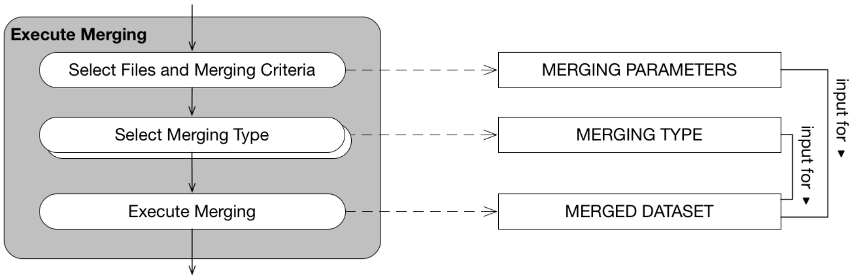


Figure 28 - Execute Merging

ACTIVITY: CONSTRUCT DATASET

The last main activity within the Data Preparation phase is called *"Construct Dataset"*. It comprises most of the tasks defined by the original CRISP-DM framework for the Data Preparation phase, such as data selection, formatting, construction, and data cleaning. This activity is suggested to be pursued with the dataset resulted from the *'Integrate Data'* task or with the datasets initially collected. The activity is illustrated in detail in Figure 29.

It starts with the *"Format Data"* activity which is basically the same as the Data Formatting step from the CRISP-DM. Examples of tasks that can be done within this activity are: rearranging attributes, changing text from upper to lower case, etc., with the goal of formatting the variables without changing their meaning, building that way a better visualization (based on the user's interpretation) of the dataset to be analyzed. The next proposed activity is called *"Engineer Features"*, where new attributes can be constructed if needed. For example, if the dataset has the weight and height information from a given person, a new variable could be their body max index, calculated based on the existing variables. Next, as proposed by Spruit & Jagesar (2016), if the dataset is high dimensional, that is, if it has a high number of variables and records, a feature selection should be done, first to reduce the size of the dataset which will facilitate the analysis, and second, to remove variables and records that may not be relevant to the project goal and analytical task. Thus, the feature selection can be done both horizontally and vertically, where horizontally means applying feature selection techniques (like a simple filtering) to the attributes (columns) of a given dataset. On the other hand, vertically means applying those techniques upon the records (rows) from the dataset. Finally, the last two activities are dedicated to handle the missing data and removing duplicate information. Thus, by cleaning the data building a simpler dataset, it should be easier to achieve and produce better results during the exploratory data analysis.

Additionally, missing data, as seen already in previous chapters, if not identified and considered during the analytical task, can heavily interfere in the outcome by making the analysis biased due to the incomplete information. Thus, identifying and handling missing data is of high importance for any analytical activity. Moreover, when focusing in domain experts, suggesting them to only handle missing data can be a very broad and hard activity, hence it deserved further explanation, as seen in Figure 30. There are several ways of pursuing this activity, however, in the end the idea of reducing the size of the dataset aiming a simpler analysis should be kept, and therefore, the suggestion is to remove missing records from the dataset, as explained next.

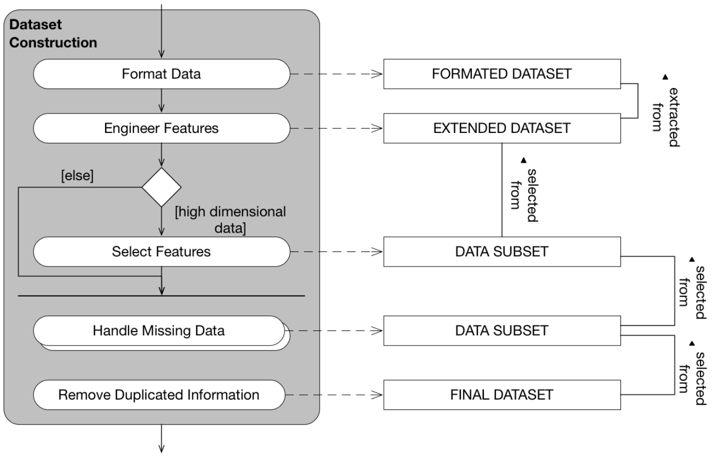


Figure 29 - Data Construction

Thus, it starts suggesting that the user examine the missingness patterns of the given dataset. Two main aspects should be noticed when examining the patterns: the proportion of missing data when compared with the content of the dataset, and if it is scattered throughout the many variables or concentrated in only few attributes. In the end, the main suggestion still is to remove all missing information, since even if the mechanisms of missingness (explained in the previous chapter) are well known, it is not guaranteed that by using the existing methods and techniques (which require time and technical skills) to fix that specific issue will result in an optimal dataset. Thus, what is being suggested is to evaluate the missingness scenario, and to try removing as few as possible the number of records. The method fragment then has three conditional events one after the other. The first one suggests removing the records that contain missing values if those represent a low proportion of the data content available. Additionally, the term 'low proportion' is quite subjective, so it varies based on the situation and interpretation of the user. However, if a high proportion of the data contains missing information, the next step is to check whether the missing values are scattered or concentrated within some few variables. If the latter, the domain expert should examine whether that specific attributes are indeed important for the analysis, once that by removing them most of the missing values would be gone, and the records preserved. If the attributes are important for the analytical project, then an optional analysis are suggested to be made, in order to better understand the reasons why the missingness is happening, such as explained in the previous chapter, when describing the missingness mechanisms (MCAR, MAR, and MNAR). Nevertheless, one way or another, the next activity suggests reporting the findings (which variables are missing, how many of them, and any information over why it might be missing) to data specialists, or database managers, who can explore the problem and fix the data generation process if needed. In the end, as already mentioned, the suggestion is to remove the records that contain missing information in order to facilitate the exploratory analysis. However, one has to keep in mind that, since information is missing, the findings might be biased given incomplete data. Hence, a close interaction with data experts and database managers should be required to extract better and reliable results.

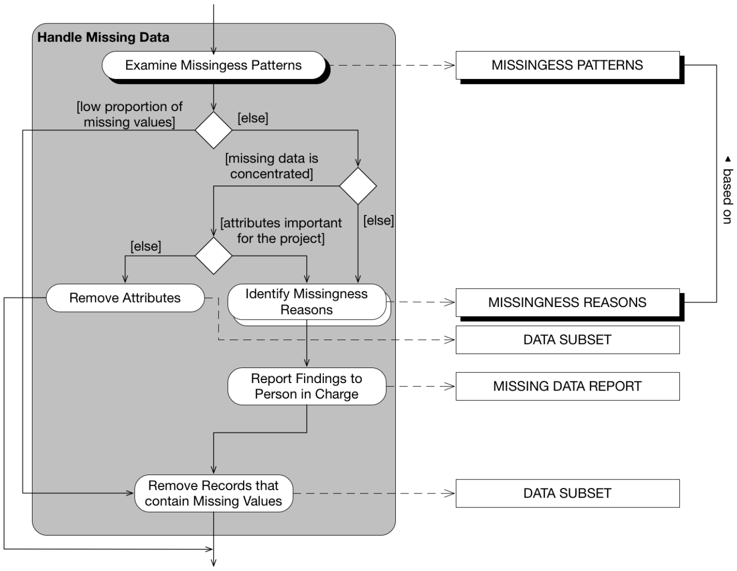

Handle Missing Data
Examine Missingess Patterns
MISSINGESS PATTERNS
[low proportion of missing values]
[else]
[missing data is concentrated]
[else]
[attributes important for the project]
[else]
Remove Attributes
Identify Missingness Reasons
MISSINGNESS REASONS
◂ based on
DATA SUBSET
Report Findings to Person in Charge
MISSING DATA REPORT
Remove Records that contain Missing Values
DATA SUBSET

Figure 30 - Handle Missing Data

8 EVALUATION DESIGN

As per Pries-Heje, Baskerville, & Venable (2008), evaluating both theoretical and practical DSR outcomes is a crucial task for showing the artifact's usefulness, benefits and qualities. Additionally, the authors also state that evaluation tasks have also to consider the behavior of systems, people and organization that the produced artifact interacts with. Thus, DSR evaluation tasks are performed by "*testing the developed solution against its requirements and by identifying its impact to the real world*" (Mettler, Eurich, & Winter, 2014). Additionally, two types of activities have to be considered, they are: artificial evaluation and naturalistic evaluation. Artificial evaluation, as its own name says, is not applied in a real-world scenario and most likely not applied to real users; it includes laboratory tests, simulation activities, mathematical proofs, etc. On the other hand, naturalistic evaluation check the impact of the proposed solution in its real environment of application, interacting with real systems, real people, and real data (Venable, Pries-Heje, & Baskerville, 2012). Moreover, evaluations in DSR can be categorized in two types: Ex Ante (e.g. evaluation of a model or design) and Ex Post (evaluation of an instantiated artifact) perspectives. The Ex Ante evaluation is usually used to decide whether to acquire, develop, or implement a specific solution. In the case of an artifact built using DSR, the Ex Ante evaluation method aims to provide theoretical ways of testing the solution without the need of implementing it. Thus, the produced artifact is "*evaluated based on its design specifications alone*" (Pries-Heje, Baskerville, & Venable, 2008). On the other hand, the Ex Post is usually used to evaluate a given solution after its acquisition, development, or implementation, hence, in its real application environment. Regarding the MAM, its evaluation required collecting feedback from its real audience, that is, domain experts for whom the model was actually developed for. Thus, it was done using a Naturalistic Ex Post setting, since the goal was to evaluate it in a real-world scenario. Moreover, structured walk-throughs, case studies, and a survey were used as strategies to evaluate the artifact, as described further in this chapter. Furthermore, a prototype tool was also developed to assist on the evaluation procedure, with the purpose of illustrate and exemplify the expected outcomes for the activities depicted within the model in an interactive and visual manner.

Participant	*Expertise*
Participant 1	Pediatrician, Neonatologist and Medical Researcher at the UMCU – WKZ
Participant 2	Consultant Neonatologist and Medical Researcher at the UMCU – WKZ
Participant 3	Anesthesiologist from the pediatric ICU from the UMCU – WKZ
Participant 4	Clinical Health Sciences teacher at Utrecht University, Medical Researcher, and previously a neonatal nurse at the UMCU – WKZ
Interviewee 5	Epidemiologist and Medical Researcher at the UMCU

Table 9 - Experiment Participants

The participants included in the validation process were domain experts from the medical domain (as illustrated in Table 4) which also were part in the interviews described back in Chapter 4, where again, for privacy reasons, their real names will not be shown.

Moreover, each participant was asked to evaluate the following topics:

- *Interpretability:* accordingly to Bibal & Frénay (2016), it can be explained by the following three connected subjects: *understandability*, *accuracy*, and *efficiency*. The first one means that a model is only interpretable when it can be understood. Accuracy refers to how accurate the model is to the data in hand since a model can be rather simple and easily understood without having any relationship with the data. Finally, efficiency, refers to the time and effort it takes to understand the model.
- *Perceived Usefulness:* refers to the degree to which the participant considers the artifact effective for structuring and preparing the data for an analytical project.
- *Ease of Use:* measures the degree to which the participant considers following the guideline free of effort
- *Intention to Use*: like its own name says, whether the participants intent to use the guideline for future analytical projects.

Moreover, the last three topics follow the so-called Technology Acceptance Model (TAM), which was and still is one of the most important models for measuring user acceptance regarding some technological artifact (D. L. Moody, 2003). In addition to that, there is no consensus in literature on the best way to measure interpretability, and therefore, based on the use case scenarios, the user was asked to evaluate the interpretability of the model based on its own experience and knowledge. Next, the evaluation setup will be described, followed by the prototype tool, and the case scenarios.

8.1 EVALUATION SETUP

Several approaches for evaluating DSR projects exist, where all of them have their pros and cons for a given scenario and research purpose. Venable, Pries-Heje, & Baskerville (2012) developed a framework for positioning different evaluation methods based on the environment and purpose of the evaluation. For example, in a Naturalistic Ex-Post scenario, as per the authors, seven evaluation methods are highlighted, among them, action research, case studies, focus groups, participant observation, and both qualitative and quantitative surveys. Moreover, Rozanski & Woods (2005) illustrate a different perspective of validation purposes and present few more techniques, and although their purpose is on validating a designed software architecture, the following concepts can also be applied to the DSR artifact evaluation for this project. First, the authors state that the validation procedure (among other things) should not only be used to check technical aspects and collect feedback, but also to fully explain the model and to ‘sell’ the solution to the stakeholders. Thus, it is mandatory to show why and how the proposed artifact will indeed be of value for the domain experts, meeting their needs, and making them understand the importance of using it. Hence, the authors highlight some validation techniques that allows one to present the architecture, collect feedback, validate its technical aspects, etc., such as simple presentation, use case scenarios, prototyping and proof-of-concept, and formal reviews and structured walkthroughs, each of them with their own level of complexity and benefits.

Given the purpose of this research, what it aims to achieve, and the setting where the evaluation was made, the procedure has been identified with six main steps, as shown in Figure 31.

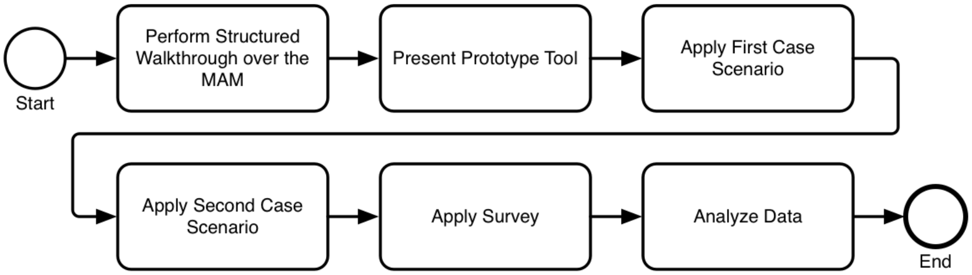

Start
Perform Structured Walkthrough over the MAM
Present Prototype Tool
Apply First Case Scenario
Apply Second Case Scenario
Apply Survey
Analyze Data
End

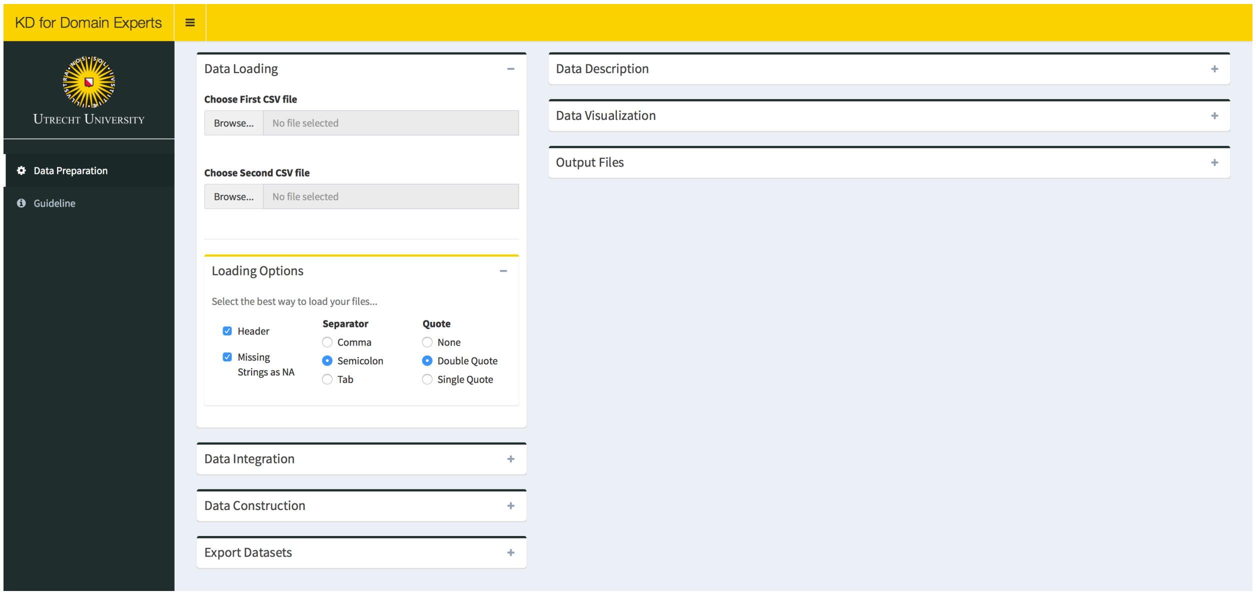
Figure 31 - Evaluation Overall Process

The process starts with a Structured Walkthrough, which is the process of explaining in detail every aspect of the (in this case) MAM with the purpose of ensuring comprehension from the domain experts over the model, as well as guide them through why such activities exist, their expected outcomes, decisions that were made in order to create a given activity, what should be the benefits of using the guideline, and answer any questions they had. The choice of using a Structured Walkthrough approach was made to provide a more valuable insight over the MAM to the participants, and as mentioned earlier, given domain experts' time and technical constraints, providing guidance has always been of huge importance for this project. Thus, it would not be optimal to develop a model and simply give it to the domain experts to interpret, when they do not have the absolute knowledge over the subject, risking that, in the end, they would still struggle on how to start a KD project.

The next step of the evaluation procedure was to present the prototype tool to the participant, which although not being part of the final artifact and main deliverable of this thesis, it was developed with the objective of facilitating comprehension over the MAM, regarding its activities and their expected outcomes. Moving forward, with the purpose of exemplifying the model usage in the real-world, two use case scenarios were created, where an analytical project goal was defined in which domain experts had to go through the model by using real data from the WKZ's databases, with the aid of the prototype tool. After concluding both case studies, a questionnaire was applied where questions regarding the four topics mentioned earlier were presented to them. Finally, the data was analyzed, and the results are displayed in the next section. In addition to that, both the tool and the use case scenarios will be further explained in the next sub-sections.

8.1.1 PROTOTYPE TOOL

A prototype tool, as defined by Rozanski & Woods (2005), is a functional subset of a system for feedback collection and validation purposes. For this thesis, the prototype tool was developed in such a way that examples of activities and ways of pursuing each task described in the model could be seen and experienced by the domain experts, with the purpose of increasing their understanding of what had to be done, and what should be a suitable outcome from each task. The tool was developed using *Shiny*, a R package for creating web-applications directly from the R suite ("Shiny," n.d.). From Figure 32, which illustrates the home screen from the tool, is possible to see seven main modules: *(1) Data Loading, (2) Data Description, (3) Data Visualization, (4) Data Integration, (5) Data Construction, (6) Output Files, and (7) Export Datasets*. Each module has a set of functionalities that exemplify the suggested tasks from the MAM, which will be further explained below.



Data Loading –

Choose First CSV file

Browse... No file selected

Choose Second CSV file

Browse... No file selected

Loading Options –

Select the best way to load your files...

☑ Header

☑ Missing Strings as NA

Separator
○ Comma
● Semicolon
○ Tab

Quote
○ None
● Double Quote
○ Single Quote

Figure 32 - Prototype Tool Main Interface

Collect Data Activity

The *Collect Data* activity from the MAM, has three suggested sub-activities: *Load Data, Describe Data, and Plot Features*. Those tasks can be experienced from the tool using modules 1 to 3 respectively.

Within the *Data Loading* module, users have the possibility to initially load up to two different CSV files. Moreover, few loading options can be chosen based on the users' preferences and file structure, such as whether the file should be loaded with its original headers, if the missing strings should be replaced by '*NA*' (Not Applicable) for further missing data manipulation, if the column separators in the CSV file are configured as commas, semicolon or tab, and the correct type of quotation that the CSV has. All these options can be changed, and the data will reload automatically with the new settings, so the user can test the combinations until reach the optimal one. Next, in the *Data Description* module, the user has the possibility to preview the loaded file and check some of its descriptive statistical information, such as mean, maximum and minimum values for numerical variables, times of occurrences for categorical attributes, number of missing values per attribute, etc. Thus, acquiring practical knowledge over the *Describe Data* activity from the MAM. Finally, the *Data Visualization* module allows the user to create some graphical visualization over the loaded datasets where some few charts, such as Histogram, Scatter Plot Matrix, and Missing Information Visualization charts (to visualize missingness patterns, as stated by the guideline), can be constructed based on the data, thus, representing the *Plot Feature* sub-activity from the MAM.

Figure 33, illustrates the different tasks described above. In the top-left corner, the *Load Data* activity is being represented, where two CSV files were loaded into the tool. Both top-right and bottom-left corner screenshots represent the *Describe Data* activity, where the first illustrates the data previewing, and the second the data descriptive statistics information, as described above. The last corner (bottom-right), represents the *Plot Features* activity, where few charts can be created to illustrate graphically the data content.

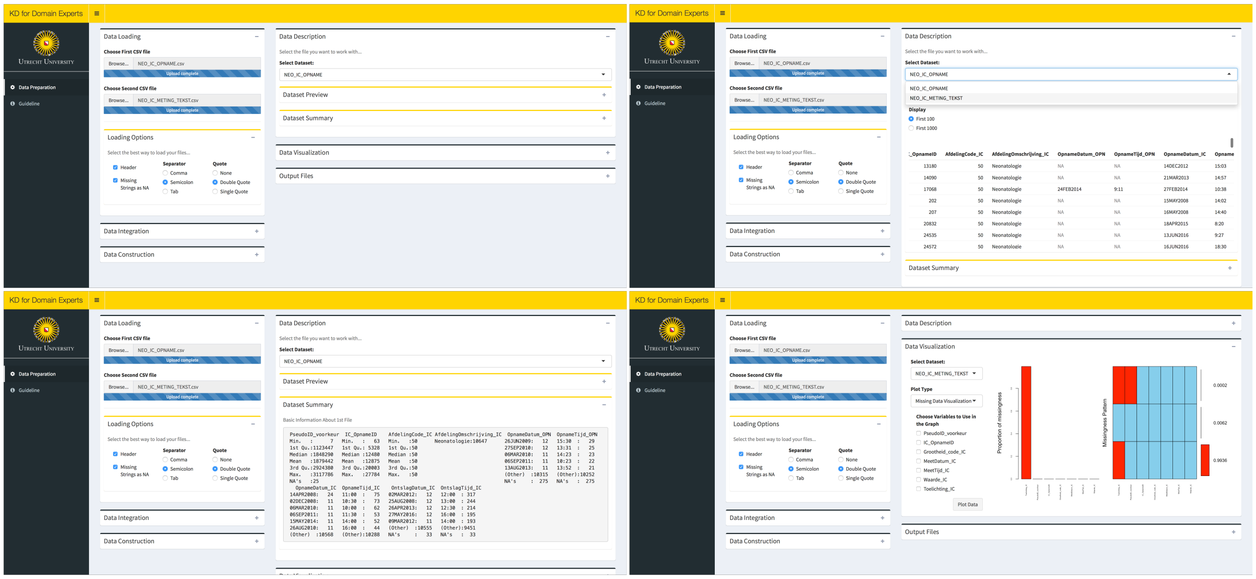


Figure 33 - Prototype Tool - Collect Data

Integrate Data Activity

The integration task can also be experienced within the prototype tool using the *Data Integration* module. Figure 34 illustrates the activity being pursued within the tool, in which the user starts by selecting the two datasets that should be merged. The next step is to identify the merging criteria for both datasets. Finally, the merging type (Inner Join, Outer Join, Left Join and Right Join) should be selected, as specified in the MAM. Although the model identifies a set of conditions before the user can actually merge the datasets, the tool does not implement them (it does not treat errors or help in choosing the right attributes). Thus, if the user selects the wrong parameters, the tool will still try to merge the two files, creating a wrong merged file, which should be noticed by the user by examining the dataset. Nevertheless, the outcome will be shown in the sixth module *(Output Files)*, where it will depict a small preview of the merged information and some descriptive statistical measurements, as shown in the image below.

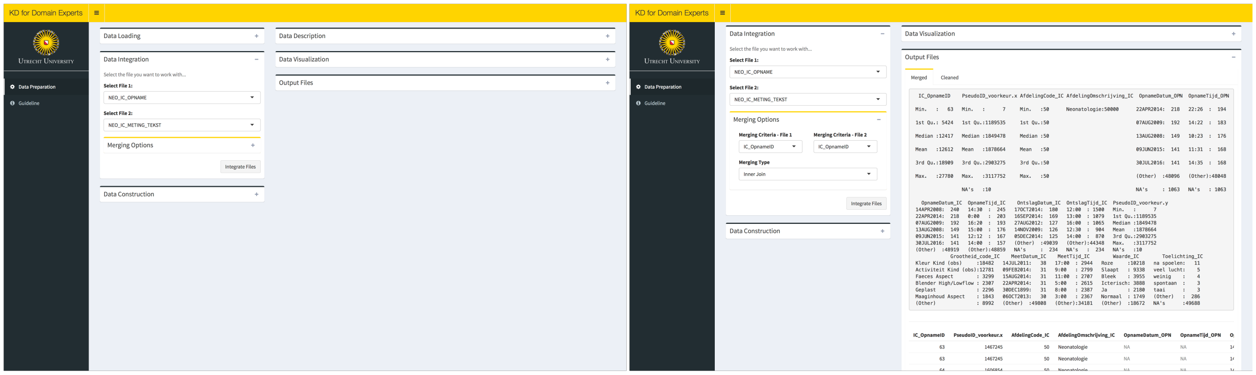
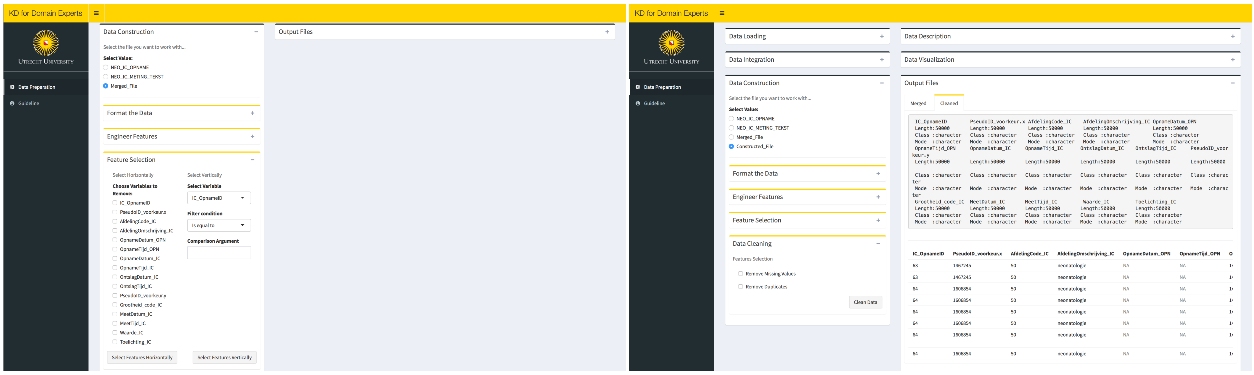


Figure 34 - Prototype Tool – Integrate Data

Construct Data Activity

The last main activity within the Data Preparation phase proposed by the MAM is the *Construct Data* activity, which is represented by the *Data Construction* module of the prototype tool. First, the user has to choose the file to be "constructed", and as shown in the guideline, the possibility of "constructing" the initially loaded data files exist (if those could not be merged for any reason), as well as constructing the new merged dataset. After selecting a suitable file, there is the possibility to perform few example tasks for each sub-activity depicted in the MAM, with the purpose of illustrating what is meant by each task and what is expected from their outcomes. Thus, the tool allows the user to format all data into lower case, as an example of the *Format Data* activity. Additionally, domain experts can, as an example of the *Engineer Features* activity, join two columns into a new one (e.g. date and time columns into a single variable). Furthermore, the *Select Features* activity was developed to allow the user to filter the data both horizontally and vertically. That is, the user can select which columns/variables to remove from the dataset, as well as filter its records by selecting only rows that meats some specific condition added by the user, such as maintain records where a specific attribute either starts or has a given string, or even is exactly equals to that string. Finally, the user can clean the data by removing all missing data and duplicate information, as examples of *Handle Missing Data* and *Remove Duplicate Information*. If any of those activities are performed, a new dataset is created (Constructed_File), which will be shown in the *Output Files* module in the *Cleaned* tab. Figure 35 illustrates the tool interface for this scenario.



Select the file you want to work with...

Select Value:
○ NEO_IC_OPNAME
○ NEO_IC_METING_TEKST
● Merged_File

Choose Variables to Remove:
□ IC_OpnameID
□ PseudoID_voorkeur.x
□ AfdelingCode_IC
□ AfdelingOmschrijving_IC
□ OpnameDatum_OPN
□ OpnameTijd_OPN
□ OpnameDatum_IC
□ OpnameTijd_IC
□ OntslagDatum_IC
□ OntslagTijd_IC
□ PseudoID_voorkeur.y
□ Grootheid_code_IC
□ MeetDatum_IC
□ MeetTijd_IC
□ Waarde_IC
□ Toelichting_IC

Select Variable: IC_OpnameID
Filter condition: Is equal to
Comparison Argument:

Select the file you want to work with...

Select Value:
○ NEO_IC_OPNAME
○ NEO_IC_METING_TEKST
○ Merged_File
● Constructed_File

□ Remove Missing Values
□ Remove Duplicates

```
 IC_OpnameID        PseudoID_voorkeur.x AfdelingCode_IC    AfdelingOmschrijving_IC OpnameDatum_OPN
 Length:50000       Length:50000        Length:50000       Length:50000            Length:50000
 Class :character   Class :character    Class :character   Class :character        Class :character
 Mode  :character   Mode  :character    Mode  :character   Mode  :character        Mode  :character
 OpnameTijd_OPN     OpnameDatum_IC      OpnameTijd_IC      OntslagDatum_IC     OntslagTijd_IC      PseudoID_voor
keur.y
 Length:50000       Length:50000        Length:50000       Length:50000        Length:50000        Length:50000

 Class :character   Class :character    Class :character   Class :character    Class :character    Class :charac
ter
 Mode  :character   Mode  :character    Mode  :character   Mode  :character    Mode  :character    Mode  :charac
ter
 Grootheid_code_IC  MeetDatum_IC        MeetTijd_IC         Waarde_IC          Toelichting_IC
 Length:50000       Length:50000        Length:50000       Length:50000        Length:50000
 Class :character   Class :character    Class :character   Class :character    Class :character
 Mode  :character   Mode  :character    Mode  :character   Mode  :character    Mode  :character
```

IC_OpnameID	PseudoID_voorkeur.x	AfdelingCode_IC	AfdelingOmschrijving_IC	OpnameDatum_OPN	OpnameTijd_OPN	O
63	1467245	50	neonatologie	NA	NA	14
63	1467245	50	neonatologie	NA	NA	14
64	1606854	50	neonatologie	NA	NA	14
64	1606854	50	neonatologie	NA	NA	14
64	1606854	50	neonatologie	NA	NA	14
64	1606854	50	neonatologie	NA	NA	14
64	1606854	50	neonatologie	NA	NA	14
64	1606854	50	neonatologie	NA	NA	14

Figure 35 - Prototype Tool - Construct Data

Finally, the user has the option of exporting the produced dataset, in the *Export Datasets* module, and saving it (as a CSV file) on its own local environment to later analyze and explore it in any tool of preference. The full tool can be accessed in: https://github.com/Dedding/R-Shiny-Prototype-Tool

8.1.2 Case studies

After performing the structured walkthrough over the model and presenting the prototype tool that has been explained above, two use case scenarios were used to exemplify how the model should behave in a real-world scenario, and to help evaluating the understanding that the participants had over the model. Hence, the scenarios that were used are described below:

Case Scenario 1

This scenario had the objective to observe whether the participant could successfully follow all steps from the guideline in order to pre-process the data focusing in one specific analytical project goal. Thus, in this scenario, the analytical project objective (part of the project context definition step) was to identify whether more active patients (e.g. registered differently than sleeping) are discharged faster from the IC than the ones that are less active (e.g. registered as sleeping). Furthermore, the purpose was actually to check if they could understand and follow the guideline's steps accordingly with the project goal, reaching the exploratory data analysis activity with a suitable pre-processed dataset. The data needed for pursuing this case scenario are available in the *IC_Opname* and *IC_Meting_Tekst* tables from the Neonatology-RPD, explained earlier in this essay. Although identifying the right tables are part of the model activities, since the experiment time is limited, those tables were provided to the participants in order to save time. Moreover, the data component document was also given and explained to the participant at the beginning of the experiment, for the same reason just mentioned.

Additionally, although the prototype tool has a limited number of features and functionalities, the objective was to use it to validate the understanding of the MAM by illustrating and letting the participants experience and interact with the data by performing some examples of activities for each task depicted in the model and chosen to be pursued during the case study. Hence, the overall quality of the final dataset was not evaluated, but if the participant managed to go through the tasks and understands what has to be done to create a desirable dataset to explore. Thus, Table 10, shown below, depicts the scenario specifications, where it consists of acquiring, integrating, and constructing two datasets available within the Neonatology-RDP.

Subject	*Description*
Case Outline	Consider that an analytical project exists with the objective of identifying whether a patient with a more active behavior gets released earlier from the ICU. The task is to follow the guideline using the data accordingly to the project scope, to get familiarized with its activities and outcomes.
Data to be used	IC_Opname (containing IC admission and discharge date/time of patients) and IC_Meting_Tekst (containing the observations about patient's behavior, events, and physical aspects, taken by nurses and doctors).
Tool to be used	Prototype Tool developed by the researcher.
Pre-Condition	Data Component document and the Datasets to be used will be made available to the participants.
Success Condition	Participant can follow the guideline accordingly to the analytical project goal and notice that data integration can and should be done between the two datasets.
Failed Condition	Participant still cannot pre-process the data following the model and showed major difficulties on trying to do so.

Table 10 - Case Scenario 1 Specifications

CASE SCENARIO 2

The second scenario shares the same objective of the first one, however, the analytical project goal was different; it had the objective of identifying whether there was a correlation between medical events that a patient was submitted to, with its staining aspects registered by the doctors and nurses. Thus, again, the actual achievement of an answer to that research question was not the goal, but if the participant was able to understand and follow the guideline steps to in order to create a simpler data set to be analyzed. The data that should be used for conducting such activity were in the *IC_Events* and *IC_Meting_Tekst* tables from the Neonatology-RPD, which again will be already provided to the participants, together with the data component document. Differently than the first case scenario, the participant should notice that those tables, if merged, will result in a many-to-many relationship, and therefore, they should not be integrated directly.

In addition to that, again, the quality of the achieved dataset after pre-processing the data was not evaluated, as the prototype tool was developed with the purpose of simply exemplifying how each specific outcome from the model should look like. Thus, as already mentioned, that means that only a few features were implemented, and some functionalities may not work properly depending on the parameters selected. Hence, the full data preparation cannot be currently done using the tool alone. Nevertheless, Table 11 summarizes the use case scenario specifications.

Subject	*Description*
Case Outline	Consider that an analytical project exists with the objective of identifying whether there is a correlation between medical events that a patient was submitted to, with its staining aspects registered by the doctors and nurses. The task is to follow the guideline using the data accordingly to the project scope, to get familiarized with its activities and outcomes.
Data to be used	IC_Events (containing all medical events realized on the patients) and IC_Meting_Tekst (containing the observations taken by nurses and doctors).
Tool to be used	Prototype Tool developed by the researcher.
Pre-Condition	Data Component document and the Datasets to be used will be made available to the participants.
Success Condition	Participant can follow the guideline accordingly to the analytical project goal and notice that merging those two datasets will result in a many-to-many relationship, and therefore, the merging is not suggested.
Failed Condition	Participant still cannot pre-process the data following the model and showed major difficulties on trying to do so.

Table 11 - Case Scenario 2 Specifications

9 RESULTS

After conducting both case scenarios, a survey was given to each participant containing questions about the interpretability of the guideline, as well as intention to use it, its perceived usefulness, and ease of use. The full questionnaire can be seen in Appendix C. Furthermore, the answers acquired were analyzed and the results are depicted in detail below in a descriptive statistical format.

9.1 DEMOGRAPHICS AND EVALUATION MEETINGS

Five domain experts working at the UMC were involved in the evaluation procedure as described in the previous chapter. In terms of technical knowledge and KD experience, all the information about that matter was described in Chapter 4, since the participants were also involved in the interviews conducted earlier in this project. Hence, in short, from Table 9, participants 1, 2, 3 and 4, based on the interviews conducted earlier, had considerable good theoretical knowledge over KD, however, they either do not have any technical skills (i.e. programming), or it is very limited to daily routine tasks. Participant 5 on the other hand, as per the interview, have some experience working with R, where although not “fluent” on programming, she can still do some data analysis within the tool. The table below shows which interviewee from Chapter 4, each participant from the evaluation procedure is.

Evaluation Respondents	*Interviews' Participants*
Participant 1	Interviewee 1
Participant 2	Interviewee 4
Participant 3	Interviewee 5
Participant 4	Interviewee 6
Interviewee 5	Interviewee 7

Table 12 - Evaluation vs Interviews' Participants

Regarding the actual evaluation procedure, each meeting was one hour long, where the MAM and the Prototype tool were explained, and the two case scenarios applied. The survey was sent by email, as its fulfillment was not mandatory to be done at the same moment as the meetings were held. Starting with the Structure Walkthroughs, it took around 20 minutes to fully explain the model in detail. Since all participants were also part of the interviews performed earlier in this study, they were already contextualized with the thesis topic and goals. During the walkthrough, the domain experts demonstrated being interested on the model, as they interacted by asking questions about the activities, as well as stating their viewpoints about the MAM. After describing the guideline, the Prototype Tool was presented for about 10 minutes, showing how it related to the model, what was its purpose, and what could be done within it and how. Domain experts looked very enthusiastic about the tool, by the fact that it created an easy and interactive interface for start doing KD, with a lot of room for improvement.

Next, the case scenarios were shown, and the files and documents needed for their fulfillment given to the participant. For the first case scenario, the objectives were mainly for them to follow the guideline, understand how both data tables related to each other, notice that the two datasets could actually be merged into one “master-file”, and to correctly identify the required parameters to do so. Gladly, all participants were able to achieve those goals, and by using the tool, examine and actually merge the datasets. Moreover, since not all features were implemented into the tool, when the domain experts reached the Dataset Construction activity, only few examples of tasks were available to be done, such as data selection, and data cleaning. Thus, since the data’s full preparation was not possible to be realized, the participants ended up discussing about what should be actually done, what should remain and be left out from the dataset, etc., in order to make it simpler and smaller, with the purpose of facilitating further analysis. The second scenario was smoother since they had already acquired knowledge from the first one. This time the goal was to notice that by merging the two datasets, they would end up with a cardinality of many-to-many, which, as per the model, would stop the integration activity. As expected, all of them were able to visualize it, and then argue about ways of transforming that many-to-many relationship into a one-to-many or many-to-one relationship within the Dataset Construction activity. Again, since the tool was not built to allow a full data preparation, it was not possible to continue the activity as they wanted in practice. However, by the end of the experiment, it was possible to see that they acquired a good understanding of the suggested activities and were able to better structure their thoughts of what had to be done in order to achieve a suitable dataset for answering the analytical project research question.

9.2 Survey Results

9.2.1 Case Scenarios and Prototype Tool

As stated above, the case studies were discussed and explained to them during the evaluation procedure, so no miss-understandings or miss-interpretation would lead them in the wrong direction when performing the designed activities. Hence, in the survey, they were asked whether both case studies were indeed clear for them. Gladly, all participants confirmed the full comprehension of both scenarios. Furthermore, having their statement that the case studies and its objectives were indeed well known, help on ensuring the validity of the other answers.

The development of the prototype tool, as well as its usage, was entirely related to the evaluation procedure, as an attempt to exemplify each activity and outcome from the MAM, by using real functions and techniques over the actual data collected from the Neonatology-RDP. Despite the fact that the tool is only a prototype, therefore not part of the main outcome of this research, it was important to know whether it was fulfilling its purpose of facilitating the comprehension over the MAM. Thus, first, participants were asked to rate the effort it took in order to understand the prototype tool, in a Likert Scale from 1 to 5 (1 being very hard, and 5 very easy). Eighty percent of the participants rated the effort with a 4, considering quite easy to understand it; just one participant rate it with a 3. Moreover, in the same Likert Scale format, participants were asked to rate, based on their opinion, if the prototype tool indeed helped them on understanding the MAM, where 1 meant that it did not make any difference in the understanding of the MAM, and 5 that it definitely helped. Again, 80% of the respondents answered the question with a 4, as shown in Figure 36. Based on their answers, it was clear that, although not perfect, the tool was of value for the evaluation procedure, since most participants answered the questions positively. However, during the evaluation procedure, functionalities that would have helped to better prepare the data were missing, once that only few were actually implemented (since the purpose of the tool was to exemplify some of the possible activities, and not to allow a full data preparation). Thus, some tasks that would have been required to achieve a 'better' dataset accordingly to the analytical project goal, could not be done, which could have influenced the not optimal results in the survey.

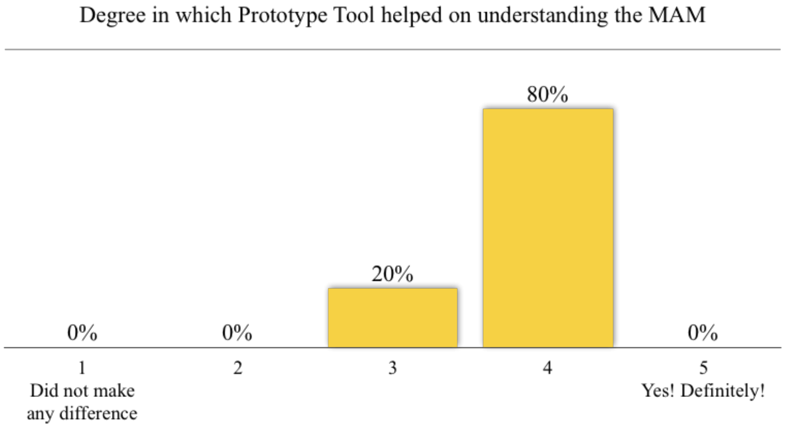


Figure 36 - Prototype Tool Benefit

9.2.2 Interpretability of the MAM

To evaluate how interpretable the guideline was for domain experts a few questions were asked about both understandability and accuracy of the MAM. Again, most of the questions were either in Likert Scale format (ordinal scale) or categorical format (e.g. yes or no), as described next.

First, when asked about their opinion, in a scale from 1 to 5, about the overall understandability of the model (1 being very hard to understand, and 5 very easy to understand), all participants rated the model with a 4, showing a quite good outcome for its overall understandability. In addition to that, all participants also answered positively that all activities and outcomes depicted within the MAM were well acknowledged and understood.

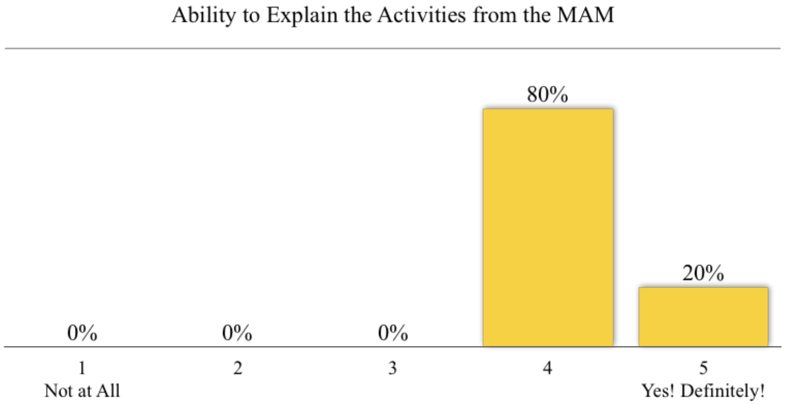


Figure 37 - Understandability Evaluation 1

Furthermore, participants were asked to rate, in a scale from 1 to 5 whether they would be able to explain all tasks and outcomes depicted in the model, 1 being not at all, and 5, definitely. This time, both fours and fives appeared in the answers, in a proportion of 4:1, as can be seen in Figure 37, illustrating a very good result for the model's overall understandability.

Next, two more specific questions about the MAM's understandability were asked, this time about its specific activities and outcomes. First, regarding the Understand Data Environment activity (prior to data collection, within the Data Understanding phase), this task was proposed within the MAM with the purpose of addressing one major difficulty stated by domain experts in the interviews, which was the lack of knowledge over 'where the data was stored, and how to access it'. By understanding the environment that surrounds that data, one could be able to identify the information needed way quicker than searching for it in every single data table available. Thus, participants were asked to rate how important they considered (after what was shown and explained to them) understanding the data environment to be, before actually starting to manipulate the data. The question was designed in a Likert Scale format from 1 to 5, where 1 meant total indifference regarding the importance of the activity, and 5 that it was indeed very important for the process. Gladly, the answers were concentrated within 4 and 5, as can be seen in Figure 38, which meant that they comprehended that this activity could help them on different stages from the KD process, as examined during the case scenarios.

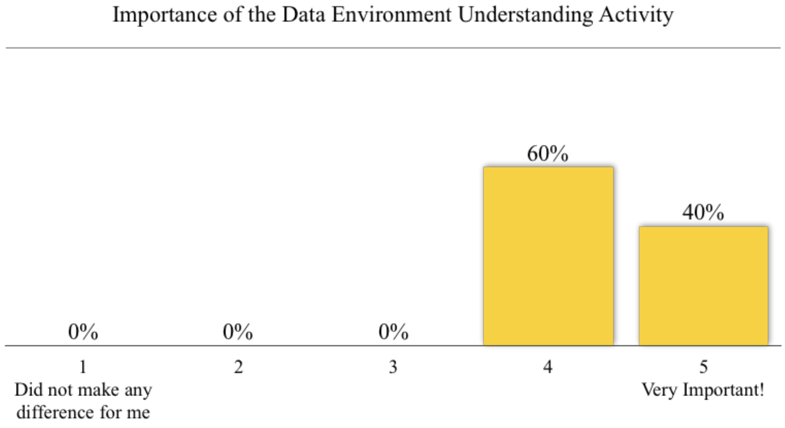


Figure 38 - Understandability Evaluation 2

Second, regarding the Integrate Data activity, as mentioned earlier, the purpose was not to teach the user how to perform the merging by following some specific syntax within some specific tool. However, the goal was to indicate which parameters and which attributes would always be required in order to perform such task, how to identify them, and what information is important when addressing such task. This activity, again, addressed one major difficulty as stated by domain experts during the interviews, which was not having a "master-file" when handling data concentrating all variables in one single view. Thus, a question was asked about whether they indeed understood all the attributes needed in order to merge two datasets, that is, the files selection, merging criteria identification, and merging type. Again, gladly, all participants answered that they indeed understood what was required, and based on the evaluation process, they now know where to look for such information, and how to retrieve it.

Moving forward, regarding the accuracy of the model, a question was asked about how accurate they thought that the model was when using a real dataset. Again, a Likert Scale was used, from 1 to 5, 1 being not accurate, and 5 very accurate. All answers were concentrated within 4 and 5, with a proportion of 3:2 people respectively, as shown in Figure 39. That demonstrated high acceptance on the way that the model was designed, in terms of how well it fitted to the hospital's datasets, as well as to their needs.

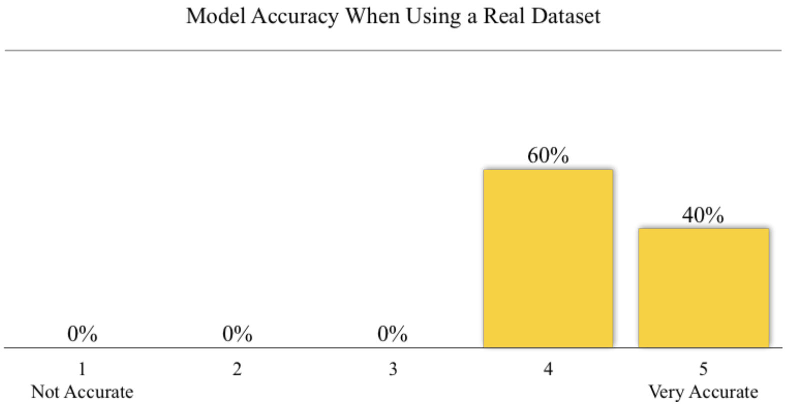


Figure 39 - Accuracy Evaluation

Furthermore, still regarding accuracy, questions were asked about whether the participants were able to successfully identify the right activities (regarding both case scenarios), and the right outcomes, when handling the given datasets and based on the project goal. All participants answered that they were indeed able to identify all suitable activities based on the data that they had, which was indeed the case during the evaluation procedure. Moreover, for the second case scenario however, one participant added more information to the answer by saying that the many-to-many relationship (see Table 11) when merging the two datasets, although more difficult to pursue, was indeed possible to accomplish. That statement is correct, however, although possible, it would make things harder for domain experts to interpret the new dataset, which as explained during this essay, is not the purpose of the model. The model suggests a dataset construction to be made trying to change this cardinality into either one-to-one, many-to-one or one-to-many, and then flow back to the merging activity for trying to merge the files again. Nevertheless, all the scenarios were discussed with the professionals, who demonstrated (at the time of evaluation) acknowledgement over the given statements, assumptions, and design options.

9.2.3 Usefulness, Ease of Use, and Intention to Use

The perceived ease of use of the model measures the degree to which the participant considers following the guideline free of effort. When building the model, all attention was concentrated on making it very straightforward, that at the same time, could concentrate just enough technical aspects to help domain experts to pre-process the data as part of the KD process. Thus, when asked about the effort it took to understand the model in scale from 1 to 5 (1 being too much effort, and 5 effortless), the respondents were divided between 3 and 4 in a proportion of 3:2 participants, respectively, as can be seen in Figure 40. Since domain experts are not technically involved most of the times on any analytical activity, the effort to comprehend the model after one structure walkthrough was expected not to be optimal. Nevertheless, the ratings are still good considering that most of these professionals never saw nor experienced such thing before. Moreover, once understood, the respondents rated, again in a Likert Scale from 1 to 5 (1 being very hard to use, and 5 very straightforward to use) their opinion about how easy was to actually follow the guideline through both case scenarios. This time, 40% of the participants rated the model as 5, that is, very straightforward to use, and 60% rated it with a score of 4. Thus, although the effort to initially understand the MAM was not considered optimal, once the model was understood, the effort to follow its activities and comprehend each expected outcome was considered, as per the domain experts, quite uncomplicated and simple.

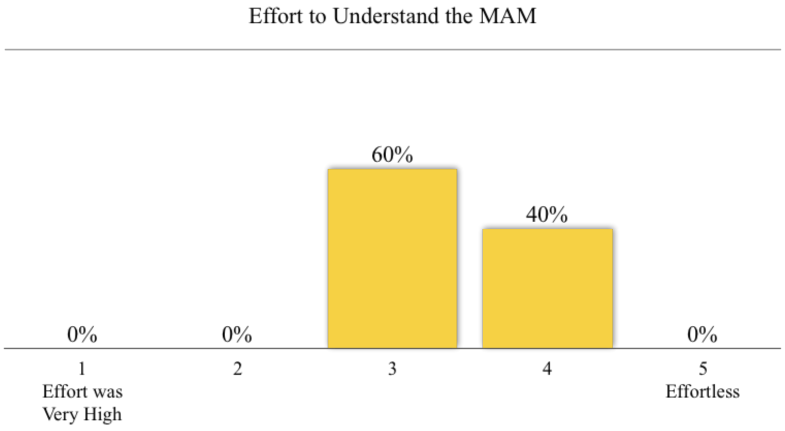


Figure 40 - Ease of Use

Next, the perceived usefulness of the model can be translated to which degree the participant considers the artifact effective for structuring and preparing the data for an analytical project. Again, a question in the format of a Likert Scale from 1 to 5 (1 being not useful, and 5 very useful) was asked. As can be seen in Figure 41, 80% of the respondents rated the model as either 4 or 5, indicating that they considered it quite useful. Only participant 5 rated it as 3, indicating some degree of indifference regarding the usability of the guideline. Since participant 5 has more experience in KD and DM than the others, is comprehensible that the model appears not to be so useful for her in certain aspects. However, the main goals and activities still apply even for more experienced users, however, they can be enhanced and extended as needed.

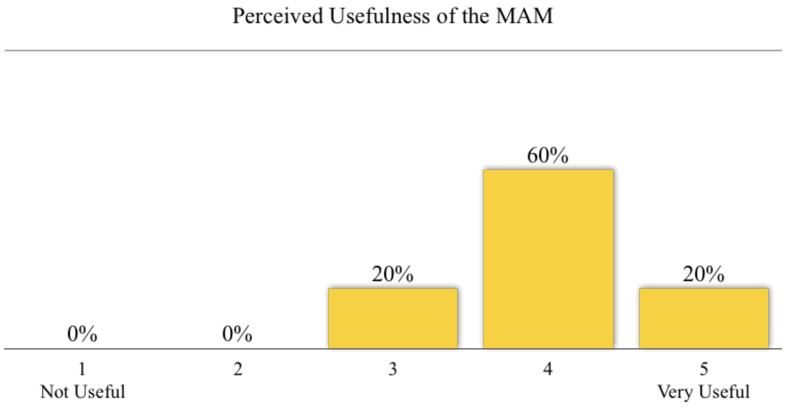


Figure 41 - Perceived Usefulness

The final subject was to check whether the participants intended to use the guideline for future analytical projects. First, the respondents were asked whether they were confident to follow the guideline in practice, and therefore, start doing some KD tasks. The answers were collected in a scale from 1 to 5, where 1 meant that the respondent would not be confident, and 5 that (s)he would be very confident. The answers were concentrated between score 3 and 4, in a 40:60 percent ratio, respectively. Although most of the previous answers demonstrated that domain experts were able to comprehend and to follow the guideline in a practical experiment, is understandable that some of them do not fell entirely comfortable to start pursuing an analytical task after just one interaction with the guideline.

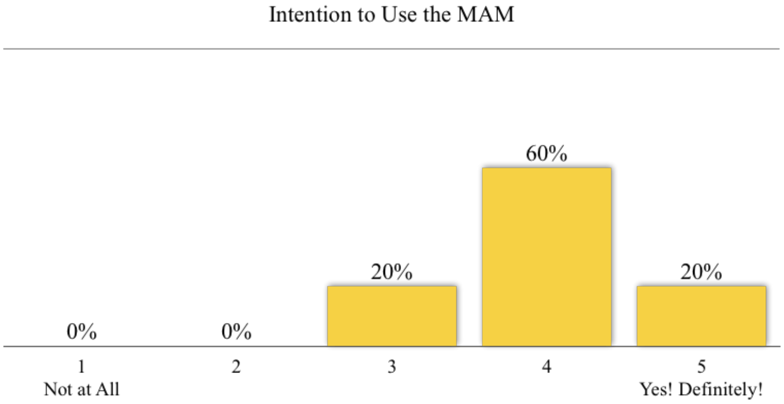


Figure 42 - Intention to Use

Additionally, the same format of question was made to see if the professionals would like to use the tool in future activities. Figure 42 illustrates the answers obtained, in which 80% of the respondents either choose score 4 or 5, meaning that they intended to use the guideline. Participant 5 however, choose a score of 3, mostly given the fact, as per her, of the approach chosen on how to handle missing data, avoiding pursuing more complex scenarios such as data imputation, as discussed in Chapter 5, and also, the fact of her being already more familiarized with the KD process, being a R user already, influenced her in choosing a lower score on this answer.

9.2.4 Survey Conclusions

In this section a summary from the results is illustrated, making reference to the topics that were evaluated.

Interpretability

As shown above, to evaluate interpretability two sub-topics (understandability and accuracy) were assessed. In terms of understandability, the ratings, as per the respondents, were considered very good, as they declared being able to fully comprehend the activities and outcomes from the guideline, and the importance of specific activities such as Understand Data Environment and Integrate Data, which were designed to facilitate not only the current step in the process, but its following activities. In terms of accuracy, it was clear that the participants felt confident on how real datasets from their domain fitted the MAM, and how the tasks and problems could be represented and assessed by following it. Therefore, the overall interpretability from the model, after conducting the evaluation, was considered high and fulfilled the expectations.

Ease of Use

To evaluate the overall ease of use of the MAM, three topics had to be considered: how much effort it took to understand the guideline, how much effort it took to follow it, and if the tool influenced positively (or not) in the evaluation of that matter. First, the effort to understand the model was not optimal nor high, it was rated to be between an average level of effort and almost effortless. As domain experts are not used to perform such activities, hence, it was expected for them to have some difficulties interpreting all activities and seeing the big picture immediately. Thus, that supported the choice of pursuing a structured walkthrough technique in the evaluation procedure. On the other hand, after the understanding of the MAM, the participants rated as almost effortless to follow its activities. In addition to that, most participants had good opinions about whether the tool helped on understanding the model and how easy was to use it. However, as it was limited to some small number of functionalities, the data preparation activity was limited to a pre-defined set of possibilities to be performed. Hence, that may have influenced negatively some ratings. Nevertheless, although a little bit of effort was needed to fully comprehend the model, it was possible to see a good evaluation of its ease of use.

Perceived Usefulness

Regarding the perceived usefulness, the majority of the respondents declared that they perceived the MAM to be indeed useful. Only one participant (Participant 5) rated it as average. However, the level of usefulness of the guideline is directly influenced by the level of experience that one might have in the subject. Regarding this specific participant, as she had a little bit more experience on the subject, she did not need guidance for all activities depicted within the model. Moreover, she also stated that she missed some activities within the model, such as missing data imputation. However, as mentioned earlier, the model was built for an audience without prior experiences with data analytics and on preparing data, hence it had to be kept simple to what was feasible to the majority of this professionals. Therefore, considering the target audience and the problems that they face when trying to do KD, the overall perceived usefulness of the model fulfill the expectations.

Intention to Use

Last but not least, most domain experts declared that they indeed intent to use the model in future activities. In the same way to what happened on the perceived usefulness evaluation, only one participant (Participant 5) rated her intention to use the MAM as average, which again relates to her level of experience on the matter and which tasks she intent to perform, as explained above. Therefore, as 80% of the respondents declared that they have the intention to use the model, it also achieved the expectation on this matter.

10 DISCUSSION

In this thesis, a Meta-Algorithmic Model was developed, with the objective of facilitating the KD process (focusing in the Data Preparation phase) for domain experts. By using the guideline, these professionals would better visualize and structure a KD process, going through the definition of an analytical project, the understanding of the data environment and the information available within it, and finally, preparing the data with a simple objective of creating a smaller and simpler dataset for an exploratory data analysis to be conducted later on.

The model created was based on the CRISP-DM framework, which has been adapted to the needs of domain experts, therefore, less technical people. Moreover, the process of developing the MAM was not trivial, as several studies had to be conducted to gather enough information and requirements in order to create a feasible and usable solution. Among those studies, interviews were conducted with professionals from the medical segment (e.g. doctors and nurses, experts in their domain) to better understand what they knew about KD, and what they thought about the possibility of having this process facilitated for them. Also, a data quality assessment was performed on a small subset of available data within the WKZ, in order to identify problems and address them in the MAM.

Data Analytics has been growing in importance, and every day more businesses are implementing and adapting their processes to use knowledge extracted from data to improve the way work is being done. The UMC is not different, as it has been investing heavily on creating a data environment that was capable of supporting data analytics. Moreover, based on the studies performed, it was discovered that domain experts know the benefits of using data to the better care of their patients, and most of them would spend more time analyzing data if that process was somehow facilitated. Additionally, although the hospital was busy creating the data environment, several problems were found during the quality assessment, where many of them related to the data generation process. Furthermore, as mentioned above, the CRISP-DM framework, which provides a guidance for doing KD, had to be adapted as it originally did not differentiate activities and processes based on different type of users, and therefore, it was somehow unsuitable for domain experts' needs.

Regarding the evaluation of the proposed solution, meetings with domain experts were held, where the model has introduced and extensively explained, and two case studies representing a real analytical project (using real data) were performed. The goal was to follow the MAM, understanding what has to be done in order to identifying the required data, and what tasks of the Data Preparation phase were important and suitable for the given project and available information. A prototype tool was also developed (in the format of a Web Application) which permitted the participants of the evaluation procedure to visualize the outcomes for each activity depicted within the MAM. The findings of this study were acquired by means of a survey, which extracted their opinions about the interpretability (understandability and accuracy), ease of use, perceived usefulness, and intention to use the MAM. The results (described in the previous section) showed that domain experts were satisfied about both understandability and accuracy of the model, as all participants rated it as either 4 or 5 in the Likert Scale, where 5 represented the maximum satisfaction for a given topic. Additionally, regarding the model's ease of use, that is the effort it took to both understand it and to follow it, regarding the first, opinions were divided between 3 and 4, out of a scale of 5, where 5 represented the task to be effortless. Although not optimal that was expected since most of these professionals are not used to perform such activities, and they were seeing it for the first time. However, they opinion of the effort it took to follow the MAM when performing the case studies was very good, as all respondents graded it as either 4 or 5. The perceived usefulness of the model and the intention that participants had on using it were similar as 80% of the participants were very positive about those two topics, and graded it as either 4 or 5 in a scale from 1 to 5, where the latter meant maximum acceptance for the topics. Just one participant rated both topics with a 3 since she already had some technical experience doing KD, and thus, for her some of the activities and outcomes were already known.

Summarizing, Chapter 4 depicts the interviews' findings, Chapter 5 the data quality assessment, Chapter 6 the CRISP-DM's adaptation, Chapter 7 the final MAM, and finally in Chapter 9, the results are described and explained. Below, each sub-research question will be shortly commented, followed by the answer to the main research question, limitations of this study and future research.

10.1 Research Questions

SQ1 – What is the current understanding that domain experts have on Knowledge Discovery and Data Mining activities?

As discussed in Chapter 4, based on the interviews conducted, it was clear that domain experts have a clear view and understanding of the purposes, benefits and tasks related to KD, as well as the challenges that it consists of, such as ensuring data quality. Additionally, by knowing the power that analyzing data has, and the benefits that it can bring to the organization and to the people involved with the business, most of domain experts shown excitement about the possibility of being able to do data analytics themselves. On the other hand, they also know that it still an activity that varies in levels of complexity and that it still requires constant interaction with data analysts to explore the most complex and hard scenarios. Furthermore, their practical experience with KD relies mostly in previous interactions with data analysts, or by applying some statistical methods to test hypothesis for research purposes, since it is an activity that usually requires more technical expertise, such as programming skills, which they do not have.

SQ2 – What are the risks and benefits of allowing domain experts to analyze data themselves?

Based on the literature review described in Chapter 2, and the interviews held with domain experts described in Chapter 4, it is possible to state that both risks and benefits exist on letting domain experts conduct data analytical tasks, and those are directly related to external variables such as project complexity, time constraints, domain knowledge, etc. Thus, as per the analysis made, it is clear that the knowledge that domain experts have over the domain, that is, the extensive knowledge over the business, environment, people, and data, is what can be considered the main existing benefit which supports those professionals to do KD. Moreover, domain ubiquitous intelligence, that is, the extensive domain knowledge mentioned above, is the main factor that allows KD to be transformed into AKD. On the other hand, domain experts most of the times start an analytical task already with a research question or hypothesis, which also represent a drawback due to the possible lack of creativity when looking for hidden information. Additionally, less external knowledge (i.e. data analysts) also means less innovation added into the domain. Hence, in the end, allowing domain experts to do KD does not exclude the need of external data analysts, and vice-versa.

Depending on the scenario one side can thrive and the other struggle, however, in the end, most analytical activities still would need interaction between the two parties.

SQ3: What is the current quality of the available data within Dutch (academic) Hospitals?

Four quality dimensions were chosen to address the current data quality situation of the WKZ, as shown in Chapter 5. They are: *Accuracy, Consistency, Completeness, and Timeliness.* Based on those quality dimensions and the problems found (which can be easily seen in Table 6), such as typographical errors, wrong data, abbreviations, problems in integrity constrains, and lots of missing data, some of the issues most likely extend to the data generation process, and although many quality problems were found in low proportions and could be easily seen, some of them open new questions about the whole validity of the available data, which would require a more extensive analysis to be ensured. Additionally, the data quality assessment was only performed with a small subset of the complete data available within the Neonatology-RDP. Hence, the findings do not represent all problems that might exist within the data environment. Nevertheless, quite a lot of problems were found in a very short time, and as mentioned earlier, it is clear that this subject need more attention and continuously improvement to ensure a more trustworthy and consistent data environment.

SQ4 – What aspects have to be considered, regarding the Data Understanding phase of the CRISP-DM, so it can be adapted for the domain experts needs?

As stated in section 6.1.1, two main topics were highlighted to facilitate the Data Understanding stage for domain experts: *understanding the data environment* and *focusing on the exploratory data analysis*. Despite the fact that domain experts already have a good understanding of the attributes and overall content of a dataset, they lack on identifying where to find and how to access the information needed. Thus, the first main activity on this phase should be on understanding the data environment in which the information needed is stored, that is the data tables arrangement, data tables relationships, which attributes are within each table, etc., before focusing on loading and start exploring the data. Thus, that way the user can have a better idea of where the required information actually is, and to access it. Additionally, as explained in section 6.1.1, based on the DM orientations suggested to be followed by domain experts (Verification and Descriptive), the second adaptation of this phase to the needs of domain experts reckon on permitting users to start and complete the whole exploratory data analysis within it, till the point of achieving the analytical project goal. The other activities suggested by CRISP-DM such as Data Description and Data Quality Assessment are still valid, however, the focus and adaptation should really be on the two topics mentioned above.

SQ5: How and which data preparation step's activities should be included in the model, and how they can be adapted for the domain expert's needs?

As explained in section 6.1.2, all the sub-activities depicted within the CRISP-DM's Data Preparation phase are applicable in the model, however based technical and time constraints, and DM orientation to be pursued, the whole objective of this task should on creating a simplified and smaller dataset. Thus, two aspects can be highlighted, based on domain experts' difficulties mentioned during the interviews and the given goal for this phase: creating a unified view for the data and constructing the dataset with the purpose of making any analytical task easier afterwards. The first aspect has to do with the fact that most of the times the data to be used is scattered in different places, and there is no 'master-file' containing all information needed for answering the analytical project goal. Thus, the Data Integration is suggested to be pursued whenever possible, since that would indeed facilitate and concentrate the analysis in one single file. The second aspect stated above is the construction of a simpler dataset, which involves the other activities introduced in the CRISP-DM: Data Selection, Data Formatting, Data Cleaning, and Data Construction (renamed to Data Engineering). Summarizing, all sub-phases from the CRISP-DM are also applicable for domain experts, however, aligned with the "new" objective for this phase. Nonetheless, Data Integration must receive a special attention, as it targets one of the major complaints from domain experts. Moreover, all the activities should be done with the only objective to creating a simpler and smaller dataset, trying to limit the time spent and technical knowledge needed to do so.

SQ6: What is the best way to guide domain experts throughout the Knowledge Discovery process, based on the CRISP-DM, so they can most likely successfully accomplish the analysis?

Figure 20 introduced a CRISP-DM adaptation that would allow users to follow the guideline as both the original proposes, as well as (considering less technical users) following the topics mentioned in ***SQ4*** and ***SQ5***. Thus, for domain experts, only the three first phases should be indeed the focus, where Business Understanding still is the starting point, having the objective translating any research question or hypothesis into an analytical project goal; Data Understanding is where the whole exploratory analysis is suggested to be done, as explained in section 6.1.1; and Data Preparation should directly support the analytical activity. Thus, two alterations are proposed in the original CRISP-DM model: a two-way relationship between Data Understanding and the Data Preparation phases, that would permit to prepare and manipulate the data prior or during the exploratory analysis, as well as (if desired) to fully examine data quality problems within the data, and a new ending point for the framework within the Data Understanding phase, meaning that the exploratory data analysis reached an end and not further interactions are needed. Thus, the model depicted in Figure 20, of course, together with the method fragments introduced in Chapter 7, represents what was proposed to best guide domain experts through the KD process.

10.2 CONCLUSION

In this research project it was found that domain experts share interest on using data to enhance the way business is being conducted. Also, the huge importance of domain knowledge for the KD process was explained, as only by using it, it is possible to transform KD into AKD, which supported the idea of facilitating the KD process for domain experts. Moreover, one of the most problematic tasks within KD process still is to prepare the data for it to be analyzed, given that it is a very time-consuming task, and which still usually requires a minimum technical knowledge. Finally, most of the existing tools nowadays do not support less technical people, that is, do not provide an easy interface and an easy way of doing KD. Hence, throughout this project means for allowing and facilitating such tasks for domain experts were researched. Based on the findings and results achieved throughout this research, the overarching research question for this thesis (shown below) can now be answered.

How can the data preparation phase, embedded within the knowledge discovery process, in an applied data science context, be facilitated so domain experts can explore an analytical problem more easily and intuitively?

First, the original CRISP-DM (which was chosen as basis for the entire research) was identified as being indifferent regarding the type of professional who is following it. Hence, it does not differentiate nor change the way an activity is supposed to be pursued based on a person's knowledge over the domain, technical skills, and experience doing KD. Thus, the type of user who is conducting the analysis, in conjunction with the type of analytical project and data available, should determine how to pursue an activity, and which tasks to actually perform. As explained in section 6.1.3, an adaptation of the CRISP-DM was proposed, aligning the objectives of the framework with what is believed to be indeed important for domain experts (based on the interviews, data quality assessment, and literature review), where only the activities (as well as their inner tasks) that would add some value into the analysis, and at the same time, would be feasible considering all the mentioned constraints, were suggested to be followed by domain experts.

Second, regarding the Data Preparation phase, one cannot prepare any data without first defining a project context and going through the Data Understanding phase. It was not possible to focus only in the Data Preparation task, without providing domain experts the means and the goals for preparing the data. Thus, to facilitate the Data Preparation phase the Business Understanding and Data Understanding phases had to be addressed and simplified as well, as explained in section 6.1.2.

Third, as mentioned throughout the entire research, Data Preparation is considered to be even more time consuming and complicated than DM itself. Defining how to pursue this activity, depends most of the times to the project at hand and information available. Thus, in order to facilitate it, the goals of this phase had to be limited to only making the dataset simpler and smaller (and not fixing and cleaning all possible scenarios), given domain experts' time and technical constraints. Additionally, based on the difficulties mentioned by domain experts during the interviews and the quality of the data that they would be dealing with, some activities within the Data Preparation phase were highlighted, such as Data Integration and Data Construction, focusing on allowing those professionals to prepare the data, and at the same time, to not spend more time than required on this task. Therefore, Data Preparation for domain experts should not have the purpose of creating a perfect dataset, but to create a simpler and smaller one for further exploring it.

Therefore, in order to facilitate the Data Preparation for domain experts, three aspects had to be considered. First, the way of pursuing the KD process had to be different. Domain experts do not share the same goals and knowledge as data scientists, so the activities that compose the KD process have to adapted and take advantage of their specific qualities. Second, is not possible to focus only on the Data Preparation phase, as the former phases are mandatory for one succeeding on preparing the data. Finally, the third aspect, as explained in section 6.1.2 and in section 7.1.3, regards the objectives of the data preparation phase, which should be aligned to the DM orientation, as well as with the constraints that involve the user who is performing the activity. In the case of domain experts, time is a huge constraint, as well as technical skills, thus, the goals of this phase were aligned accordingly.

10.3 CHALLENGES AND LIMITATIONS

This project, as being part of a master's thesis, was restricted by a few factors such as time and number of resources. In this section, some of the topics that somehow limited the research (as well as might have influenced its outcomes) will be described and discussed.

As could be seen in the previous chapters, the decisions that lead to the development of the models (that is, the CRISP-DM adaptation depicted in Figure 20, the BPMN model in Figure 21, and the method fragments in Chapter 7) were based on several activities and findings, such as interviews, data quality assessment, literature study, etc. Thus, first, back in Chapter 4, although the information gathered during the qualitative study was very insightful, only domain experts from the medical segment were interviewed. Despite the fact that most of them had different areas of expertise, they all had similar experiences and had to handle the same problems regarding the data and its environment. The difficulties describe by them were addressed in the model, however, others that might affect domain experts from different segments may have been "ignored" in the MAM, simply by the fact that they were not known. In addition to that, only seven people were interviewed, and however the findings demonstrated to reach data saturation, that is, no new ideas or relevant thoughts were being introduced, it does not mean that different problems or insights do not exist. Nevertheless, as mentioned earlier, given a close schedule (and following the data saturation theory), seven interviewees was considered a good number.

Next, the data quality assessment described Chapter 5 was made only upon some few tables from the Neonatology-RDP. The database environment from the WKZ (extending it to the UMC) is huge, hence, the problems that were found, although interesting, do not represent all the problems that may exist within their complete data environment. Thus, even though it would be very interesting to examine a larger portion of the database, it was simply not possible given time constraints. In addition to that, the MAM was also built considering the problems found during this activity, and despite the fact that the guideline was designed to be as generic as possible, some major drawback, due to some scenario that could not be observed during the data quality assessment, might exist.

Furthermore, the evaluation procedure illustrated in Chapter 8 consisted of only five participants (due to a tied schedule, and difficulties to find a spot in their agendas). Although the result shown to be similar for all respondents and indicated satisfaction with the proposed guideline, a larger number of domain experts could emphasize how good the model indeed is, or in which aspects it should be improved. Moving forward, a prototype tool was developed, as described in section 8.1.1, where its purpose was to help domain experts to understand the MAM, and to experience some activities that are suggested by it. However, it was not implemented to allow a full data preparation to be made, as it contained only a few functionalities for each task. Hence, despite the fact that it aimed to help on the evaluation, at the same time, the limitations might have been a problem for the some aspects evaluated (as described in Chapter 9), when the participants wanted to pursue some task, and the same was not available, hence, not given them much freedom to examine different options for the case scenarios.

10.4 FUTURE RESEARCH

In this last section, few topics that were left open during and after the model's construction will be discussed as opportunities of future research related to this thesis project.

Data analytics, as could be seen throughout this document, is a very promising and important field nowadays, as it is still growing and being adapted within many companies around the globe. The paper *Power to the People!* (Spruit & Jagesar, 2016) represented a starting point for spreading the power of KD, of technology, to people who are no experts in the area, who have other qualities that could indeed help on extracting information as good as (or sometimes better) data analysts or scientists. This thesis followed the same line of research, focusing in the applied data science area of study.

First, all activities that were made within this thesis were based on the medical segment. The interviews, data quality assessment, and even the evaluation was made with people and data from the UMC. Hence, it would be interesting to investigate how this guideline would apply to other segments in the market, and if it would indeed (although is believed that it would) help domain experts on pursuing KD based on their daily routines and data, achieving the same results.

Next, the model focus on activities that help domain experts on pre-processing the data prior to the analysis. However, it does not guide them on what tasks should be pursue next. One main wish stated by all professionals was to have a step-by-step guidance throughout the KD process, helping them to make the best decisions based on the data and situation encountered by them. Hence, pre-process the data is only the beginning. With the advancements of Artificial Intelligence (AI) and Machine Learning (ML), it would be interesting to provide them a more situational guidance (to both pre-process data and to explore it afterwards), that would indeed take them by the hand and advise them to take the best decisions, while still avoiding the black box scenario.

Last but not least, although providing the steps that are needed or suggested to be done during a KD project is helpful, and it already helps on structuring the thoughts and the ideas on how to pursue an analytical project, the means of how to achieve the results depicted by the model can still be hard to see or use. Hence, it would be interesting to investigate specifically how domain experts should use the model (tools, notations, syntax, etc.), since no tools exists that allows KD to be done by those professionals efficiently. The MAM development in this thesis depicted activities which in theory can be pursued in most tools mentioned by them during the interviews (e.g. Excel, SAS, R, etc.), however, the exact way of pursuing it is nowhere to be found, and it would be interesting to have it.

BIBLIOGRAPHY

10 Key Marketing Trends for 2017. (n.d.). Retrieved December 20, 2017, from https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, (January), 182–185. Retrieved from http://recipp.ipp.pt/handle/10400.22/136

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5–37. https://doi.org/10.1016/j.jsp.2009.10.001

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big Data Analytics in Healthcare. *Hindawi Publishing Corporation*, *2015*, 1–16. https://doi.org/10.1155/2015/370194

Bibal, A., & Frénay, B. (2016). Interpretability of Machine Learning Models and Representations : an Introduction. *ESANN European Symposium on Artificial Neural Networks*, (April), 27–29.

Blencowe, H., Cousens, S., Chou, D., Oestergaard, M., Say, L., Moller, A.-B., … Lawn, J. (2013). Born Too Soon: The global epidemiology of 15 million preterm births. *Reproductive Health*, *10*(Suppl 1), S2. https://doi.org/10.1186/1742-4755-10-S1-S2

Brinkkemper, S. (1996). Method engineering: Engineering of information systems development methods and tools. *Information and Software Technology*, *38*(4 SPEC. ISS.), 275–280. https://doi.org/10.1016/0950-5849(95)01059-9

Cao, L. (2012). Actionable knowledge discovery and delivery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*, 149–163. https://doi.org/10.1002/widm.1044

CAO, L., & ZHANG, C. (2007). THE EVOLUTION OF KDD: TOWARDS DOMAIN-DRIVEN DATA MINING. *International Journal of Pattern Recognition and Artificial Intelligence*, *21*(04), 677–692. https://doi.org/10.1142/S0218001407005612

ChipSoft. (2014). ChipSoft - HiX: eenduidigheid in een modern jasje. Retrieved February 28, 2018, from https://www.chipsoft.nl/oplossingen/84

Christoulakis, M., Spruit, M., & Van Dijk, J. (2015). Data quality management in the public domain: A case study within the Dutch justice system. *International Journal of Information Quality*, *4*(1), 1–17. https://doi.org/10.1504/IJIQ.2015.071672

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence, AI Magazine*, *17*(3), 37–54.

https://doi.org/10.1145/240455.240463

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health*, *25*(10), 1229–1245. https://doi.org/10.1080/08870440903194015

Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. *ACM SIGKDD Explorations Newsletter*, *1*(1), 20–33. https://doi.org/10.1145/846170.846172

Goossen, W. T. F. (2014). Detailed clinical models: Representing knowledge, data and semantics in healthcare information technology. *Healthcare Informatics Research*, *20*(3), 163–172. https://doi.org/10.4258/hir.2014.20.3.163

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Gualtieri, M. (2017). The Forrester Wave^{TM}: Predictive Analytics And Machine Learning Solutions, Q1 2017. *Forrester Research*.

Hackbarth, A. D. (2012). Eliminating Waste in US Health Care. *JAMA*. https://doi.org/10.1001/jama.2012.362

Harpaz, R., Dumouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics*, *91*(6), 1010–1021. https://doi.org/10.1038/clpt.2012.50

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2), 87–92. https://doi.org/http://aisel.aisnet.org/sjis/vol19/iss2/4

Hmamouche, Y., Ernst, C., & Casali, A. (2015). Automatic KDD Data Preparation Using Multi- criteria Features. In *IMMM 2015 : The Fifth International Conference on Advances in Information Mining and Management* (pp. 33–38).

Howson, C. P., Kinney, M. V, McDougall, L., & Lawn, J. E. (2013). Born too soon: preterm birth matters. *Reproductive Health*, *10 Suppl 1*(Suppl 1), S1. https://doi.org/10.1186/1742-4755-10-S1-S1

iMDsoft. (2017). MetaVision ICU - clinical information system for critical care. Retrieved February 28, 2018, from http://www.imd-soft.com/products/intensive-care

Kantardzic, M. (2011). Data-Mining Concepts. In *Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition* (pp. 1–25).

Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. *International Journal of Computer Science and Information Technology*, *2*(2), 55–66. https://doi.org/10.5121/ijcseit.2012.2206

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2003). A Taxonomy of Dirty Data. *Data Mining and*

Knowledge Discovery. https://doi.org/10.1023/A:1021564703268

King, J., & Magoulas, R. (2016). 2016 Data Science Salary Survey. *O'Reilly Strata*, 23.

Linden, A., Krensky, P., Hare, J., Idoine, C. J., Sicular, S., & Vashisth, S. (2017). Magic Quadrant for Data Science Platforms, 48. Retrieved from https://www.gartner.com/doc/reprints?id=1-3TKR16P&ct=170215&st=sg

Livari, J., & Venable, J. (2009). Action Research and Design Science Research - Seemingly similar but decisively dissimilar. *Prooceedings of the ECIS 2009*, *Paper 73*, 1–13.

Luo, Q. L. Q. (2008). Advancing Knowledge Discovery and Data Mining. *First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008)*, 7–9. https://doi.org/10.1109/WKDD.2008.153

Makary, M., & Daniel, M. (2016). Medical Error - the third leading cause of death in the US. *British Medical Journal*.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*, *90*(10), 60–68. https://doi.org/00475394

McGregor, C. (2013). Big data in neonatal intensive care. *Computer*, *46*(6), 54–59. https://doi.org/10.1109/MC.2013.157

Mettler, T., Eurich, M., & Winter, R. (2014). On the Use of Experiments in Design Science Research: A Proposition of an Evaluation Framework On the Use of Experiments in Design Science Research: A Proposition of an Evaluation Framework. *Communications of the Association for Information Systems*, *34*(10), 223–240. Retrieved from http://aisel.aisnet.org/cais%0Ahttp://aisel.aisnet.org/cais/vol34/iss1/10

Moody, D., & Kortink, M. A. . (2000). From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design. *Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000)*, *2000*, 5–16.

Moody, D. L. (2003). The Method Evaluation Model : A Theoretical Model for Validating Information Systems Design Methods. *Information Systems Journal*, 1327–1336. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.3682&amp;rep=rep1&amp;type=pdf

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data minning guide. *CRISP-DM Consortium*.

Pries-Heje, J., Baskerville, R. L., & Venable, J. R. (2008). Strategies for Design Science Research Evaluation. *European Conference on Information Systems (ECIS)*, *Paper 87*, 1–13. https://doi.org/10.1177/1933719108329095

QSR International. (2016). What is NVivo? | QSR International. Retrieved January 25, 2018, from http://www.qsrinternational.com/nvivo/what-is-nvivo

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

Rani, B. K., Govrdhan, A., Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering*, *2*(January 2010), 250–255. https://doi.org/10.1.1.163.4924

Rostin, A., Albrecht, O., Bauckmann, J., Naumann, F., & Leser, U. (2009). A machine learning approach to foreign key discovery. *12th International Workshop on the Web and Databases (WebDB), Providence, Rhode Island*, (WebDB), 1–6. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.150.2150&amp;rep=rep1&amp;type=pdf

Rozanski, N., & Woods, E. (2005). *Software Systems Architecture: Working With Stakeholders Using Viewpoints and Perspectives*. *Addison-Wesley*.

Russom, P. (2011). BigData Analytics. *TDWI Research*, *TDWI BEST*.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. https://doi.org/10.1109/CTS.2013.6567202

Saigal, S., & Doyle, L. W. (2008). An overview of mortality and sequelae of preterm birth from infancy to adulthood. *The Lancet*. https://doi.org/10.1016/S0140-6736(08)60136-1

Scannapieco, M., Missier, P., & Batini, C. (2005). Data Quality at a Glance. *Datenbank-Spektrum*, *14*(January), 6–14. https://doi.org/10.1.1.106.8628

Shahri, H., & Barforush, A. (2004). A flexible fuzzy expert system for fuzzy duplicate elimination in data cleaning. *Database and Expert Systems Applications*, (May), 161–170. https://doi.org/10.1007/978-3-540-30075-5

Shiny. (n.d.). Retrieved April 30, 2018, from https://shiny.rstudio.com/

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, *17*(8), 43–48. https://doi.org/10.5120/2237-2860

Spruit, M., & Jagesar, R. (2016). Power to the People! - Meta-Algorithmic Modelling in Applied Data Science. *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, *1*(January 2016), 400–406. https://doi.org/10.5220/0006081604000406

Spruit, M., & Lytras, M. (2018). Applied data science in patient-centric healthcare: Adaptive analytic systems for empowering physicians and patients. *Telematics and Informatics*, *35*(4), 643–653. https://doi.org/10.1016/j.tele.2018.04.002

Tsai, C.-W., Lai, C.-F., Chao, H.-C., & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big Data*, *2*(1), 21. https://doi.org/10.1186/s40537-015-0030-3

Tsipouras, M. G., Exarchos, T. P., Fotiadis, D. I., Kotsia, A. P., Vakalis, K. V., Naka, K. K., & Michalis, L. K. (2008). Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. *IEEE Transactions on Information Technology in Biomedicine*, *12*(4), 447–458. https://doi.org/10.1109/TITB.2007.907985

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). *Data Mining and Knowledge Discovery Handbook*. *Journal of Chemical Information and Modeling*. https://doi.org/10.1017/CBO9781107415324.004

van de Weerd, I., & Brinkkemper, S. (2008). Meta-Modeling for Situational Analysis and Design Methods. *Handbook of Research on Modern Systems Analysis and Design Technologies and Applications*, 35–54. https://doi.org/10.4018/978-1-59904-887-1

Venable, J. R., Pries-Heje, J., & Baskerville, R. L. (2012). A comprehensive framework for evaluation in design science research. In *International Conference on Design Science Research in Information Systems* (pp. 423–438). https://doi.org/10.1007/978-3-642-29863-9

Wang, Y., & Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, *70*(August), 287–299. https://doi.org/10.1016/j.jbusres.2016.08.002

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), 29–39. https://doi.org/10.1.1.198.5133

Zhang, M., Hadjieleftheriou, M., Ooi, B. C., Procopiuc, C. M., & Srivastava, D. (2010). On multi-column foreign key discovery. *Proceedings of the VLDB Endowment*, *3*(1–2), 805–814. https://doi.org/10.14778/1920841.1920944

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180

APPENDICES

A INTERVIEW PROTOCOL

Interview protocol form

Thomas Dedding

Introduction

- Introduce myself
- Explain the purposes of the interview and my research project

Notes to the interviewee

Before we start, we would like your permission for recording this interview for later transcribing it. All the information collected here will be used only for scientific research purposes and therefore, will be held confidential and it will not be shared anywhere outside the university.

Start recording…

Background

First things first, could you tell me a little bit about yourself? (Name, education, how long you've been working here, what do you do, etc.)

Knowledge Discovery Understanding

The definition of knowledge discovery that I'm using is: "the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data" (Fayyad, Piatetsky-shapiro, & Smyth, 1996).

1) What is your understanding of the term Knowledge Discovery?
2) What are the potential benefits of extracting knowledge from data for healthcare?
3) How do you (would you) perform the Knowledge Discovery process? Which tasks you consider important?

The benefits of extracting knowledge can be financial (by creating better ways of diagnosing and treating patients, detecting frauds, etc.) and also in the patient's care (predicting outcomes based on symptoms, allowing doctors to act before is too late, etc.).

Data Preparation and Modeling

4) Do you have any programming or statistical skills?

   IF YES – can you elaborate on it? For what do you use it?

5) Did you try to analyze the data via SAS or any other tool?

IF YES – Did you succeed? How? Was it hard? What the problems were? How hard is to understand analytical outputs given by SAS or another tool?

IF NO – Why not?

6) Did you use any documentation as a guideline?

    IF YES – How was it? Was it accurate? Was it understandable? Did it help?

    IF NO – why not?

7) Data Understanding and Data Preparation can be sometimes more time consuming and difficult than data mining itself. It is the process of cleaning, reducing, filtering the data, so the useless, duplicated, fuzzy, empty information is removed before further analysis. Regarding this…

    a. Do you know what data is collected and stored in the hospital's database?

        IF YES -- What you think of it regarding its quality and how reliable it is? Have you seen the hospital's raw data already?

    b. There is any documentation where this information is available? Do you know if they keep it updated?

8) Regarding descriptive analytics (which the objective is to find patterns, associations, modifications, peculiarities and noteworthy structures, for further human interpretation). Did you already explore data?

    IF YES -- Did you know and/or use any specific method(s) for describing the data? (Clustering, Association Rules, etc.)?

        IF YES -- How come you used this specific method? Have you ever tried other methods, and if not, why not?

    IF NO – Why not?

9) Regarding DM methods. Do you know any of these methods: classification, regression, clustering, association rules, etc.?

    IF YES -- If DM methods and their specific techniques and algorithms were presented to you, would you know which one to use?

    IF NO – What you think the tool should offer so you could use a DM method to play with the data?

Expectations and Thoughts about Knowledge Discovery

10) What you think an analytical tool should have so you could easily use it? (Interactive, good documentation, step-by-step guidance, good exploration features, etc.)

11) How would you respond if you could analyze the data yourself?

12) What's your vision/thoughts on how the business might benefit if domain experts (doctors) were able to analyze the data themselves?

13) How’s your vision if this task was delegated to a third party, i.e. an external analyst?

14) In the past or present, did or do you have contact with data analysts in your department or know someone who did?

    IF YES - Did they solve what was needed?

    What’s your opinion on the solution they delivered, i.e. did they deliver what was expected?

    Can you elaborate if the solution was/is feasible for the business? Did they understand the business goals, or how did you translated the business goals to them to make this project successful? Do you think they fully understood the available data? The solution they provided, is it still in use? What has changed or not?

Closure

…Stop recording…

- Talk about the further steps of our research project
- Reassure confidentiality

~End~

B EVALUATION FORM

Evaluation Form - Knowledge Discovery for Domain Experts

This form should be filled after the conclusion of both observational case studies applied to collect feedback over the Meta-Algorithmic Model developed for facilitating data pre-processing activities of a knowledge discovery project performed by domain experts.

Email address: _________ Name: __________ Background: ___________

1) Both use case scenarios were clear to you?

   ○*Yes* ○*No*

   a. If no, why?

2) What is your opinion about the model understandability?

   very hard to understand ○ - ○ - ○ - ○ - ○ *very easy to understand*

3) Did you understand all activities and expected outcomes from the model?

   ○*Yes* ○*No*

   a. If no, what was unclear?

4) Would you now be able to explain the tasks that are suggested to be done in order to prepare the data for an exploratory analytical activity?

   not at all! ○ - ○ - ○ - ○ - ○ *yes! definitely!*

5) What is your opinion about the effort it took to understand the prototype tool?

   very hard to understand ○ - ○ - ○ - ○ - ○ *very easy to understand*

6) Were you able to follow the activities depicted within the model by using the tool?

   ○*Yes* ○*No*

7) Did the prototype tool help you on understanding the MAM?

   did not make any difference ○ - ○ - ○ - ○ - ○ *yes! definitely!*

8) Did you manage to successfully identify all the suggested activities when performing case scenario 1?

   ○*Yes* ○*No*

   a. If No, what happened? (case scenario 1)

9) Did you manage to successfully identify all the suggested activities when performing case scenario 2?

   ○*Yes* ○*No*

   a. If No, what happened? (case scenario 2)

10) What is your opinion on the effort it took to understand how to pre-process the data by following the model guideline?

effort has very high ○ - ○ - ○ - ○ - ○ *effortless!*

11) What is your opinion on the importance of understanding the data environment, as suggested by the model, prior to loading the data files?

did not make any difference ○ - ○ - ○ - ○ - ○ *very important!*

12) Did you understand the attributes needed for integrating datasets?

○*Yes* ○*No*

a. If No, what was unclear?

13) In your opinion, how accurate the model was when using a real dataset?

not accurate ○ - ○ - ○ - ○ - ○ *very accurate*

14) What is your perceived usefulness for the model?

not useful ○ - ○ - ○ - ○ - ○ *very useful*

15) In your opinion, how would you "rate" the guideline?

very hard to use ○ - ○ - ○ - ○ - ○ *very straightforward to use*

16) How confident are you in following that guideline in practice?

not confident ○ - ○ - ○ - ○ - ○ *very confident*

17) Would you use this model in future activities?

not at all ○ - ○ - ○ - ○ - ○ *yes! definitely!*

a. If no, why?

18) Is any suggested activity still confusing or unclear to you?

○*Yes* ○*No*

a. If Yes, which one(s) and why?