

A Standardized Data Mining Method in Healthcare:

a pediatric intensive care unit case study

Utrecht, 27 August 2018

Author:

Eldin Karisik – 5701686

Master Business Informatics Student

Utrecht University, Information & Computing Science

e.karisik2@students.uu.nl

1st Supervisor

Dr. Marco R. Spruit

Utrecht University

2nd Supervisor

Dr. M.J.S Brinkhuis

Utrecht University

1st External Supervisor

Dr. Erik Koomen

Wilhelmina Kinderziekenhuis

2nd External Supervisor

Dr. Teus Kappen

Wilhelmina Kinderziekenhuis



Wilhelmina Kinderziekenhuis



Universiteit Utrecht

Abstract

The growth of available data in the healthcare led to numerous data mining projects being launched over the years, that revolves around knowledge discovery. In spite of this, the medicine domain experiences several challenges in their quest of extracting useful and implicit knowledge due to its inherent complexity and unique characteristics, as well as the lack of standards for data mining projects. Hence, the aim of this research is to bring some standardization in data mining processes in the healthcare based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) method. The CRISP-DM is widely adopted in various industries and is suitable as a base method on which enhancements can be made in order to bring domain specific standardizations.

This proposed method which is named MSP-DM was evaluated by domain experts from the UMC and UU. Additionally, these expert interviews were conducted in identifying any missed method fragments that were not captured during the case study or mentioned in the literature, as well as evaluating the found method fragments. During the course of the case study, one of the provided projects was successfully completed and implemented, as for the second project insight was gained about the possibilities of predictive modeling.

Moreover, during the expert sessions and the case study, a high emphasis was given to more involvement of clinical professionals and domain experts during a data mining project, i.e. in the selection of parameters, modeling, and evaluation. The clinical staff is usually unacquainted with the concept of data mining, which can create a gap between the researcher performing the analysis and the (medical) domain experts. Similarly, not involving clinical practitioners in the data mining project could lead to a failure to adopt a certain technology or analysis result, because the clinical practitioners could feel surpassed not being consulted or involved in the process. In addition, for researchers that are unfamiliar with the medicine domain it is essential to interact with clinical professionals in order to attain a sufficient understanding of the domain, which will eventually help in comprehending the problem, data, and objectives. Hence, a collaboration is required in mitigating this problem through their (clinical practitioners) provided input that can determine relevant outcomes and issues, which will lead to better analysis and easy implementation of the outcomes that are found. Likewise, practical activities and concepts were found that were missing in the original method.

For this reason, these and other findings were incorporated in the MSP-DM, which proved to be viable during the case study. In consideration with the results, the created method provides an extension of the CRISP-DM tailored for the healthcare that includes the current challenges of data mining projects which may be extended to comprise processes relevant to other domains as well.

Acknowledgement

First, I would like to thank my parents and family in supporting and motivating me over the years through my academic life. Without them, I would not have pursued in further education. Moreover, I would like to thank my wife Imane for her support and patience through every challenge faced while studying throughout my master business informatics program. Hence, my gratitude to them all.

Furthermore, I would like to thank Erik Koomen, Teus Kappen, and Joppe Nijman from the WKZ and UMC for their effort in guiding and helping me through my thesis. They did a wonderful job in welcoming me at the WKZ, as well as making me feel comfortable being one of their team members during the projects. It would not have been possible in completing this project if they did not provide the necessary means, advice and time. Hereby, I would like to show my appreciation for their effort.

Besides, I would like to thank Vincent Menger and Sebastiaan Broekema for their time and expertise by providing me with valuable insights for achieving the objectives for this research.

Finally, I would like to thank Marco Spruit as my first supervisor who has given me the chance to be part of this project at the WKZ and the connections he has provide for me to successfully write this thesis. In addition, I would like to thank him for his insights and trust during this research project.

Contents

Abstract.....	ii
Acknowledgement.....	iii
Chapter 1: Introduction.....	1
1.1 Research context.....	1
1.1.1 Data Mining and Knowledge Discovery.....	1
1.1.2 Data Mining in Medicine.....	2
1.1.3 Case study: The VDS project.....	5
1.2 Scientific & Societal Relevance.....	5
1.2.1 Scientific relevance.....	5
1.2.2 Societal relevance.....	6
1.3 Problem statement & Research objective.....	7
1.4 Research questions.....	8
Chapter 2: Design science research.....	10
2.1 Research methodology.....	10
2.1.1 Research framework.....	10
2.1.2 Literature review.....	11
2.1.3 Design science approach.....	11
2.1.4 Meetings.....	13
2.1.5 Evaluation.....	13
Chapter 3: Data Mining Process Methods.....	15
3.1 Data Mining Methods.....	15
3.1.1 CRISP-DM.....	16
3.1.2 SEMMA.....	18
3.1.3 KDD.....	19
3.2 Method Selection.....	20
3.2.1 Limitations of the Selected Method.....	20
Chapter 4: Overcoming the Challenges.....	22
4.1 Data integration.....	22
4.2 Data quality.....	23
4.3 Causal inference in observational data sets.....	23
4.4 Legal issues.....	25
4.4.1 The legal applications in the Healthcare.....	25

4.5 Identified method fragments	27
Chapter 5: Case study: The VDS project	28
5.1 Business Understanding	28
5.1.1 VDS1: Business Understanding	29
5.1.2 VDS2: Business Understanding	29
5.2 Data Understanding	30
5.2.1 VDS1: Data Understanding	30
5.2.2 VDS2: Data Understanding	31
5.3 Data Preparation.....	31
5.3.1 VDS1: Data Preparation.....	31
5.3.2 VDS2: Data Preparation.....	31
5.4 Modeling.....	32
5.4.1 VDS1: Modeling.....	32
5.4.2 VDS2: Modeling.....	33
5.5 Evaluation	33
5.5.1 VDS1: Evaluation.....	34
5.5.2 VDS2: Evaluation.....	35
5.6 Deployment.....	37
5.7 Identified method fragments	38
Chapter 6: Method Fragments	39
6.1 Method Fragments.....	39
6.2.1 Domain Understanding	42
6.2.2 Data Understanding.....	46
6.2.3 Data Preparation	50
6.2.4 Modeling.....	52
6.2.5 Evaluation	54
6.1.6 Deployment.....	57
Chapter 7: Method evaluation	59
7.1 Qualitative: Expert Interviews.....	59
Chapter 8: Conclusion & Discussion.....	68
8.1 Conclusion of the sub-questions	68
8.2 Conclusion of the main research question	70
8.3 Discussion.....	71
8.3.1 Limitations	71

8.3.2 Reflection.....	72
References.....	74
Appendices.....	81
A. Process-deliverable-diagram CRISP-DM.....	81
Business Understanding	81
Data Understanding.....	84
Data Preparation.....	85
Modeling.....	87
Evaluation	88
Deployment.....	90
B. Experiment notebook – VDS1.....	91
B.1 VDS1 code.....	91
C. Experiment notebook – VDS2.....	97
C.1 VDS2 code.....	97
D. Expert interviews.....	102
D.1 Interview protocol form.....	102
Interviewee A1:.....	104
Interviewee A2.....	107
Interviewee A3.....	109
Interviewee A4.....	113
Interviewee A5.....	117

List of Figures

Figure 1.1: Frequency of DM applications in the top 10 industries based on the KDnuggets poll.....	3
Figure 2.1: Research Framework based and edited on the Verschuren & Doorewaard (2010)	10
Figure 2.2: Design Science Approach based and edited on the Hevner et al. (2004)	12
Figure 3.1: The CRISP-DM process method (Chapman et al., 2000)	17
Figure 3.2: Four-level breakdown of the CRISP-DM method (Chapman et al., 2000).....	18
Figure 3.3: SEMMA based on Mariscal et al. (2010).....	18
Figure 3.4: KDD method originally retrieved from Fayyad et al. (1996)	19
Figure 5.1: Regression lines and confidence limits (95%) of minute volume to weight versus etCO2 per ventilator mode	35
Figure 5.2: The creation of the Neural Network.....	35
Figure 5.3: Results of the Neural Network prediction of one patient on a small dataset.....	36
Figure 5.4: Results of the Neural Network prediction with all patients on a big dataset.....	36
Figure 6.1: Process-deliverable-diagram of the Domain Understanding	43
Figure 6.2: Process-deliverable-diagram of the Data Understanding	47
Figure 6.3: Process-deliverable-diagram of handling legal issues	49
Figure 6.4: Process-deliverable-diagram of the Data Preparation	51
Figure 6.5: Process-deliverable-diagram of the Modeling	53
Figure 6.6: Process-deliverable-diagram of the Evaluation	55
Figure 6.7: Process-deliverable-diagram of the Deployment	57

List of Tables

Table 1.1: Challenges of data mining in healthcare	4
Table 1.2: An overview of the related topics and types of analysis within the case study	5
Table 2.1: An overview of the participants of the interviews.....	14
Table 3.1: Poll on data mining process method usage by KDnuggets.com	20
Table 4.1: An overview of identified method fragments in chapter 4.....	27
Table 5.1: Baseline characteristics	34
Table 5.2: Linear mixed-effects models fit by maximum likelihood	34
Table 5.3: MSE results of different node parameters	36
Table 5.4: An overview of identified method fragments in chapter 6.....	38
Table 6.1: An overview of identified method fragments	40
Table 6.2: The Generic tasks (bold) and Outputs (<i>Italic</i>) of the CRISP-DM retrieved from Chapman et al. (2000), page 12	41
Table 6.3: Activity table for Domain Understanding	44
Table 6.4: Concept table for Domain Understanding	44
Table 6.5: Activity table for Data Understanding	47
Table 6.6: Concept table for Data Understanding	48
Table 6.7: Activity table for handling legal issues	49
Table 6.8: Concept table for handling legal issues	50
Table 6.9: Activity table for Data Preparation	51
Table 6.10: Concept table for Data Preparation	52

Table 6.11: Activity able for Modeling.....	53
Table 6.12: Concept table for Modeling	54
Table 6.13: Activity table for Modeling	55
Table 6.14: Concept table for Evaluation.....	56
Table 6.15: Activity table for Deployment.....	58
Table 6.16: Concept table for Deployment	58
Table 7.1: Results of the expert interviews	60

List of Abbreviations

WKZ	Wilhelmina Children’s Hospital
DM	Data mining
MSP-DM	Medicine Standardized Process – for Data Mining
EHR	Electronic health records
KDD	Knowledge discovery in databases
CDM	clinical data mining
VDS	ventilation decision support
M.A.M	Meta-Algorithmic-Modeling
PDD	process-deliverable-diagram
CRISP-DM	Cross-Industry Standard Process for Data Mining
SEMMA	Sample, Explore, Modify, Model, and Asses
EM	Enterprise Miner
GDPR	General Data Protection Regulation
EU	European Union
PICU	Pediatric intensive care unit
NICU	Neonatal intensive care unit
ATPD	Ambient temperature and pressure dry gas
BTPS	Body temperature and pressure saturated gas
VT	Tidal volume
EMR	Electronical Medical Record
PRVC	Pressure-regulated volume control
UMC	Universitair Medisch Centrum
UU	Utrecht University
SD	Interquartile range
IQR	Standard deviation
3PM	Three-phases method

Chapter 1: Introduction

The 21st century is an age of big data that encompasses all aspects of life in which humans are involved with, including biology and medicine (Z. Zhang, 2014). With the rapid development in metabolomics, genomics, proteomics, and other types of omics technologies throughout the past decades, an incredible amount of data related to molecular biology has been created (Li, Kang, & Zhao, 2014). This shifting environment involves large amount of data that range from clinical records and numeric laboratory values to video, photo or audio files in which data mining can be used as an important tool to transform these data into information (Rivo et al., 2012). Additionally, the transition from medical records to electronic health records (EHR) has steered to a rapid change in growth of data (Sessler, 2014). Sessler (2014) further explains that this growth in big data provides a wonderful opportunity for health policy experts, physicians, and epidemiologists to make data-drive decisions that will eventually improve patient care. This big data is not only applicable for the biomedical scientist, but a necessity that must be understood and used well in the search for new knowledge (Margolis et al., 2014). Ketchersid (2014) explains that accumulating large data sets is of no value if the data cannot be analyzed in a way that generates insights that can be acted upon.

1.1 Research context

1.1.1 Data Mining and Knowledge Discovery

Data mining (DM) and Knowledge discovery in databases (KDD) are popular terms that are used interchangeably within projects that are involved with (big) data (Wasan, Bhatnagar, & Kaur, 2006). KDD refers to the process of finding useful knowledge from data, and DM refers to the process of extracting knowledge (patterns and information) derived by the KDD process in which algorithms are applied to extract knowledge (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). However, over the years, DM is used as a term that encompasses the whole of KDD process and therefore both terms can be used as synonyms when referring to this area (Mariscal, Marbán, & Fernández, 2010). Along with this line, DM can be defined as the process of knowledge discovery (Witten, Frank, & Hall, 2011). Lavecchia (2015) provides a more extended definition of DM as an automatic extraction of useful previously unknown information from data sets or databases by using advanced search algorithms and techniques in order to discover correlations and patterns. Hence, DM will be used in this research to reference the whole process of knowledge discovery.

Furthermore, in industries such as retailing, insurance, banking, and medicine usually use DM to increase sales, reduce costs, and enhance research (Seifert, 2004). In addition, DM applications are used as a means to detect waste and fraud, along with improving and measuring program performance or search for trends in data within various domains in order to achieve organizational goals (Pal, 2011). DM can

be classified as a multidisciplinary field that combines statistics, machine learning, database technology, data visualization, pattern recognition, and expert systems (Obenshain, 2004). Hence, the tasks of DM can be summarized as tasks of description and prediction in finding human-interpretable patterns and associations, after considering the whole data and creating prediction models that seek to foretell some response of interest (Bellazzi & Zupan, 2008).

In the literature, there are three main categories of data mining strategies described as: supervised, unsupervised, and semi-supervised learning (Obenshain, 2004; Zhu, 2008). In a supervised learning setting, a set of input variables are used in building statistical models for predicting or estimating an output variable. In an unsupervised learning setting, the output variable does not exist and DM techniques are used in discovering relationships, clusters and patterns from data sets. In semi-supervised learning, a restricted amount of output variables are provided and with DM techniques the missing values can be predicted or relationships, clusters, or patterns can be extracted in the data set (Iavindrasana et al., 2009; James, G., Witten, D., Hastie, T., Tibshirani, 2013).

1.1.2 Data Mining in Medicine

The medical domain is known for its ontological constraints and complexity in regards with healthcare process computerization and medical data analysis (Cios & Moore, 2002). The origin of this complexity of healthcare is derived from the heterogeneity of treatments and outcomes, the diversity of health-related ailments (disorders), the subtle intricacies of study designs, analytical methods and approaches for collecting, processing and interpreting healthcare data (Dinov, 2016). The result of this increasing complexity carries a growing demand for more personalized medicine (Hingorani et al., 2013; Joyner & Paneth, 2015; Van Giessen et al., 2015). Hence, the main vocation of a healthcare institute is to provide individualized patient care rather than collecting data to fit for mining, which makes it a challenge in modernizing clinical data mining (CDM) in order to discover exciting and valid knowledge from the gathered clinical data (Iavindrasana et al., 2009).

The increase of analytical capabilities, data availability and the demanding need to improve the healthcare quality and patient outcome are the drivers of the big data era in healthcare (Rumsfeld, Joynt, & Maddox, 2016). With its current development in technology, large amount of data is produced that require appropriate analytical techniques and technology, as well as systems for extracting knowledge in order to optimize decision making in their treatment (Milovic, 2012). There are ranges of different sources in obtaining medical data such as large clinical trials, biomarker data, administrative claim records, prospective cohort studies, patient reported data, electronic health records, clinical registries, medical imaging and the internet (Rumsfeld et al., 2016; Slobogean et al., 2015) This abounds in various sources in the healthcare, which causes the need of DM applications. Figure 1.1 illustrates an increasing trend in the usage of DM in the healthcare and other industries. In fact,

there was an increase of 62.1% from 2015 to 2016 based on the KDnuggets polls (Piatetsky, 2016, 2017).

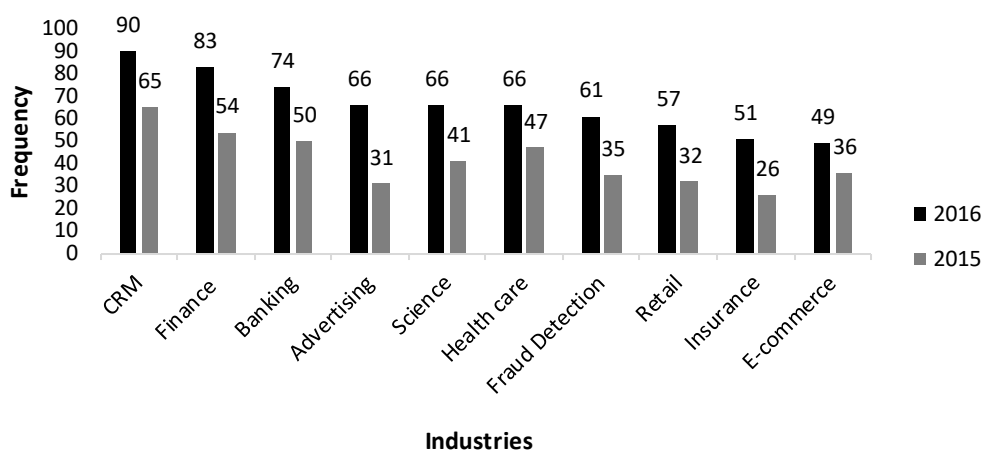


Figure 1.1: Frequency of DM applications in the top 10 industries based on the KDnuggets poll

Resulting that DM is becoming an indispensable tool for clinical practitioners and researchers in medicine in which treatments, diagnoses and prognosis’ can benefit from (Bellazzi & Zupan, 2008). The potential uses of data mining techniques in the healthcare have been successfully applied to help predicting healthcare cost, the state of health of the patients, under-diagnosed patients and health insurance fraud (Yoo et al., 2012). Hence, Esfandiari, Babavalian, Moghadam, & Tabar (2014) define DM in medicine as “*Extraction of implicit, potentially useful and novel information from medical data to improve accuracy, decrease time and cost, construct decision support system with the aim of health promotion*”. The authors explain that this definition contains three parts:

- I. Data mining: the extraction of useful and implicit knowledge
- II. Medical nature: the use of medical data and applying the extracted models to the medical domain.
- III. Goals: medical DM aims to improve efficiency and decreasing human error, decreasing time and cost, enhancing medical support systems and knowledge extraction.

This enables health institutions to use DM applications for a variety of areas such as physicians using patterns by measuring economic indicators, customer satisfaction, quality indicators and clinical indicators; optimizing healthcare, proactively intervening, identifying high-risk patients and clinical performance from multiple perspectives in order to optimize the use of resources, cost effectiveness, and decision-making based on evidence (Koyuncugil & Ozgulbas, 2010). Likewise, with other informatics disciplines, the analytic goals in medicine when faced analyzing large data sets are; prediction, modeling and inference in which regression, clustering, and classification are the commonly used methods (Sinha, Hripcsak, & Markatou, 2009). The key of success to improve the quality within healthcare is in

having the ability to use data with the purpose of extracting useful information (Eapen, 2004).

However, DM in the medical domain comes with various challenges that have been covered in multiple works (Bellazzi & Zupan, 2008; Bhoj Raj Sharma, Kaur, & Mnanju, 2013; Cios & Moore, 2002; Esfandiari et al., 2014; Hosseinkhah, Ashktorab, Veen, & Owrang O, 2009; Kleinberg J, Ludwig J, Mullainathan S, 2014; Lee & Yoon, 2017; Niakšu, 2015; Obermeyer & Emanuel, 2016; Rumsfeld et al., 2016). The challenges are described in the following Table 1.1.

Table 1.1: Challenges of data mining in healthcare

Challenge	Description
Data integration	Data accessibility in the healthcare can be restricted for DM because data sources that can serve as input for DM are frequently scattered in different systems and settings such as clinics, administrations, and laboratories (Cios & Moore, 2002; Milovic, 2012). In addition, the lack of data standards in the medicine can result in heterogeneous data issues that can cause inconsistency and instability in data sources (Rumsfeld et al., 2016).
Data quality	The data quality can be affected by inaccurate measurements, human or equipment errors (Niakšu, 2015). This can cause biases in the data collection, which can significantly affect both the generalizability and performance of future developed predictive models. Essentially, having higher-quality clinical data may result in more clinically useful, valid, and stable DM projects (Altman & Ashley, 2015; Shah et al., 2015).
Causal inference in observational data sets	The availability of large amounts of data does not remove the inherent limitation of observational data (Rumsfeld et al., 2016). It is possible to apply analytics on large amount of data, although having fully comprehensive data sources is very unlikely. Herein, issues of sampling bias, which mainly comes from the influences of the state of a patient, are not always measured nor observed within data sources (Rumsfeld et al., 2016). Algorithms can provide good predicting outcomes, but this does not necessary mean that these predictors are causes of something in order for physicians to intervene (Obermeyer & Emanuel, 2016).
Validation & analytical problems	Validation of DM projects can influence the overall performance when applied in clinical care. Therefore, it is important to examine the data set for missing data, noisy data, risk of false-positive associations, multiple comparisons, and the potential of overfitting of prediction models (Rumsfeld et al., 2016).
Legal issues	Data sources are becoming more available to assist DM projects in knowledge discovery (Rumsfeld et al., 2016). Hence, factors such as data security, patient consent and privacy and other legal issues related to electronic health information need to be considered (Gray & Thorpe, 2015; Murdoch & Detsky, 2013). There is legislation that protect personal privacy and prohibits the use of patient's clinical information without the consent which may complicate the use of such information for research purposes (Niakšu, 2015).
User-friendliness	Currently, it may require experts to understand the results of a DM model (Bellazzi & Zupan, 2008). However, this can be meaningless if models or the outcomes of a DM project are intended for an average database user. Hence, appropriate selections of models need to be made when considering who the end users will be. This mainly applies if those models are deployed within a systematical setting that assist decision support for medical professionals (Bellazzi & Zupan, 2008). Herein, the technological

1.1.3 Case study: The VDS project

The ventilation decision support (VDS) project is an initiative from the Wilhelmina's Children's Hospital. The project focuses on specific patient groups within the knowledge domain that requires the attention where medical data can be used for providing person-oriented advice, person-oriented diagnostics, treatment and person oriented signaling. In Table 1.2, an overview is presented of the related topics and types of analysis that are conducted within the case study.

Table 1.2: An overview of the related topics and types of analysis within the case study

Case ID	Topic	Type of analysis
VDS1	Did the software update of the vendors of ventilators impact the tidal volume (V_T)?	Comparative analysis
VDS2	What model is suitable in predicting the respiratory deterioration and improvement or show inferences between the various predictors, which may require adjustment of ventilator settings?	Machine/deep learning modeling

The outcomes of this project need to be preferable scalable and should be able to contribute to the care continuum and be directly applicable in the healthcare. Eventually, using DM is a first step in realizing a smarter way of making data understandable for medical professionals, that will lead in discovering novel insights in order to improve patient care. Therefore, this requires a DM process method in guiding such projects and ambitions, which is currently missing.

1.2 Scientific & Societal Relevance

This section explains the relevancy of the thesis and discusses the scientific and social perspective.

1.2.1 Scientific relevance

Developing a tailored DM process method should not only benefit the VDS project but science in general because of its usefulness within the medicine domain and DM projects. Thus, this research project aims to uncover the unique characteristics and challenges within DM in medicine and integrate them into a standardized method that can be used by researchers in order to make the process of data mining reliable and usable. Those researchers and clinicians can have limited skills in the field, but with a high degree of knowledge within the domain or for those who are unfamiliar with DM in the medicine domain. Therefore, this method can be used as a guide in the DM process in medicine.

Lastly, the process method with its underlying method fragments can be used for further research and creation of adapted method fragments that can be applied in other industries or research domains.

1.2.2 Societal relevance

This thesis project is part of the VDS initiative as earlier described. By providing a tailored process method for DM projects within their department it will enable them to implement future projects related to DM, in order to further their insight in respiratory improvement, diagnosis and treatment of their patients with a breathing condition. In addition, this method can be used in other departments for other disorders or conditions, in order to improve the diagnosis and treatment of patients in general.

1.3 Problem statement & Research objective

As previously pointed out, over the decade the medicine domain and other industries experienced an exponential growth in their quest for knowledge discovery. This led to numerous DM projects being launched. However, the medicine domain encountered several challenges in DM due to its inherent complexity and unique characteristics. Additionally, there have been methodological issues that limit DM projects, such as data inconsistency and instability, data quality, legal concerns, validation & analytical issues and limited observational studies (Lee & Yoon, 2017). This is mainly attributed by the absence of an universal protocol to model, compare or benchmark the performance of various data analysis strategies (Dinov, 2016).

Likewise, the increasingly complex data sources, types and structures such as time-series, multi-relational and object data types and natural language texts require the development of new methodologies as well as algorithms, grid services and tools (Bhoj Raj Sharma et al., 2013). Similarly, Esfandiari et al. (2014) are confirming this and elaborating that the application of DM in medicine specially lack standards in the overall knowledge discovery process. Patel et al. (2009) further explain that building (predictive) models come with challenges because of the lack of standards and confidential issues.

Thus, considering the recent challenges in the medicine domain, the unique characteristics and the lack of standards for DM projects, the aim of this thesis is to develop a standardized data mining process method based on an existing DM method amplified with the literature, the case study, and domain experts' opinion for the medicine domain. Herein, the fragmented methods will be brought together and complemented with the current developments within the medicine domain in order to establish some standardization.

Eventually, this newly created method is named the medicine standardized process for data mining (MSP-DM) must be capable of guiding DM projects in the healthcare. Hence, the VDS case study serves to guide this thesis in acquiring the necessary method fragments and validating the performance and usefulness with experts in the field of DM and medicine. Therefore, the MSP-DM will be to guide researchers that begin practicing DM techniques in the medicine or for those who lack domain knowledge within medicine.

1.4 Research questions

This research will focus on taking the problem statement and research objective into account as well as the need of the VDS project and its common problem space within the medicine domain. Thus, the main research question is formulated as follow:

RQ: How to develop a standard-based and enhanced data mining process method for researchers within the medicine domain to better guide them in the process of data mining projects considering the domain's specific challenges and unique characteristics?

This thesis project is centered in developing a standardized data mining process method for medical and IT researchers within the medicine domain by considering various unique characteristics and challenges within this field that can generate insight and suitable (predictive) models to be deployed in real-life scenarios. In addition, the MSP-DM will align with the basic requirements of the VDS project. Hence, the project will focus on developing suitable (predictive) models that can be used within its domain.

The creation of the MSP-DM should not only benefit this project, but science in general because of its usefulness within the medicine domain and DM projects. In addition, improving DM projects within the healthcare can lead to new insights and better medical care for patients. The following sub-questions are formulated to structure and guide the research objective and to provide a suitable answer for the main research question:

RSQ1: What are the existing data mining process methods for guiding data mining projects in the healthcare in order to support the development of a standardized process method for the healthcare?

With the aim of developing a standardized process method, it is required to set the initial direction for the literature review and to examine theoretical models and frameworks that relate to DM projects, specifically in the medicine domain. A selection will be made of the most used DM process methods within the healthcare for laying the foundations of a more standardized process method.

RSQ2: What are the main concepts and activities involved in medical data mining for constructing a standardized process method for the healthcare?

An outline will be created of all the concepts and activities that are related in the process of developing a standardized process method. This is important because having a good understanding of these concepts and activities will determine the legend interface. Herein, the literature, expert interviews and the VDS case study will be consulted in order to find scattered activities and concepts that are relevant for the medicine domain and overcoming the challenges of data mining in the healthcare.

RSQO3: How can the main concepts and activities be modelled into a standardized method for DM projects in the healthcare?

For the development of the MSP-DM, it is necessary to make use of Meta-Algorithmic-Modeling (M.A.M) (Spruit & Jagesar, 2016; Spruit & Lytras, 2018). M.A.M is an inspired by the method engineering discipline to construct, design and adapt tools, techniques and methods for the development of information systems (Brinkkemper, 1996). Herein, a meta-modelling approach will be applied proposed by Weerd & Brinkkemper (2009) in which a *process-deliverable diagram* (PDD) is produced.

Chapter 2: Design science research

2.1 Research methodology

2.1.1 Research framework

With the aim of providing an appropriate result for the given research objective in developing a standardized data mining process method, a research framework is needed to be set. Herein, a schematic and highly visualized representation is provided of the steps that are needed to be taken in order to complete one's research objective (Verschuren & Doorewaard, 2010). In Figure 2.1 the framework is illustrated based on the suggested representation of Verschuren & Doorewaard (2010).

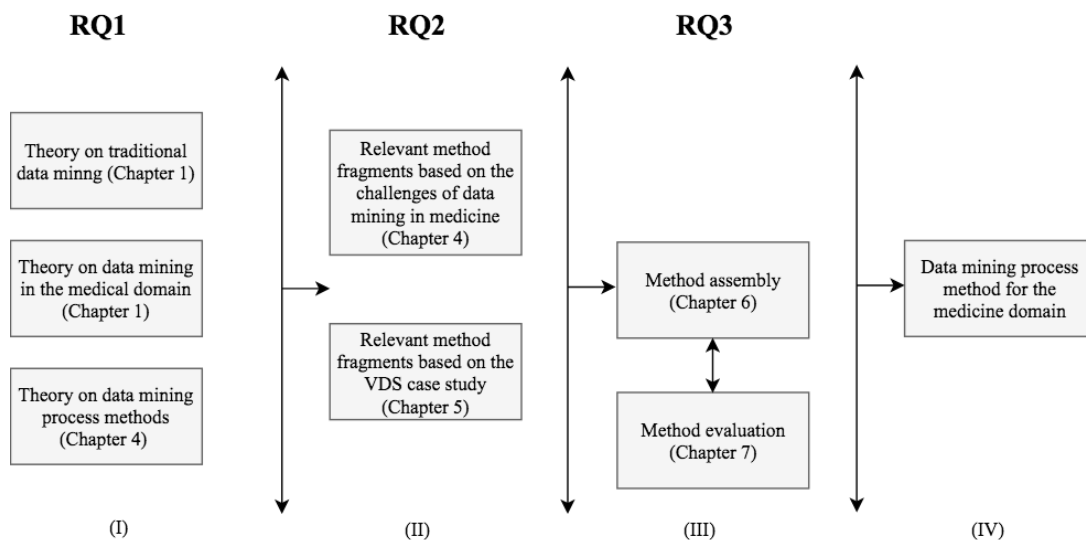


Figure 2.1: Research Framework based and edited on the Verschuren & Doorewaard (2010)

Initially, a theoretical framework is established on the related subjects of data mining, data mining process methods, and its application in the medical domain (I). This phase (I) is based on literature review that enables answering the first research question (RQ1) of selecting a suitable DM process method as foundation layer for a more standardized method. Moreover, based on the gathered information, method fragments are collected in a general framework combined with the method fragments from the VDS case study and research on overcoming the challenges of DM in Medicine. In this phase (II) the second research question (RQ2) will be answered, wherein an overview is provided of the main concepts and activities related to DM in the healthcare.

The general framework of all these method fragments collected will become a part of the method assembly (III). Once the method base is established with the relevant collected method fragments, useful method fragments are then selected and assembled into a newly created method (MSP-DM). This assembled method will be

evaluated through expert interviews on the practicality in order to ensure that it meets the research goals, as well as identifying overlooked method fragments. This approach is based on the generic steps developed for situational method engineering by Weerd, I. van de, Brinkkemper, S., Souer, J. & Versendaal (2006). Finally, the results will be a DM process method within the medical domain (IV).

2.1.2 Literature review

Within the theoretical framework a literature study is performed. Webster & Watson (2002) recommend and propose to use snowballing as the main method in finding relevant literature in the field of information systems instead of the systematic literature study. Moreover, since this research is related to information systems because of the relation with M.A.M, this method is used. The authors highlighted the backward (from the reference list) and forward (finding citations to the papers) snowballing techniques in the usage of this method. This snowballing approach starts with a set of papers that are based on identifying a set of papers from leading journals in the area (Jalali & Wohlin, 2012).

In addition, in order to access and find relevant academic journals, books and papers the search engine Google Scholar is used in combination with the usage of the Utrecht University library proxy link: scholar.google.com.proxy.library.uu.nl. Therein, various terms are used related to DM within the medical domain. This includes the different synonyms of DM such as knowledge discovery. Furthermore, books and other materials are used provided by the Wilhelmina's Children's Hospital such as ventilation manuals and pediatric books related to ventilations in order to understand the pediatric domain and to familiarize with the case study.

2.1.3 Design science approach

Furthermore, the design science approach is based on the Hevner, March, Park, & Ram (2004) framework that emphasizes on a construction-oriented view of information systems research, wherein innovative IT artifacts can be designed and build. The authors have indicated that with this approach models or methods can be constructed and evaluated with scientific rigor as artifacts and this is in line with the development of the MSP-DM in the healthcare. In Figure 2.2, the design science approach is illustrated in the context of this research project.

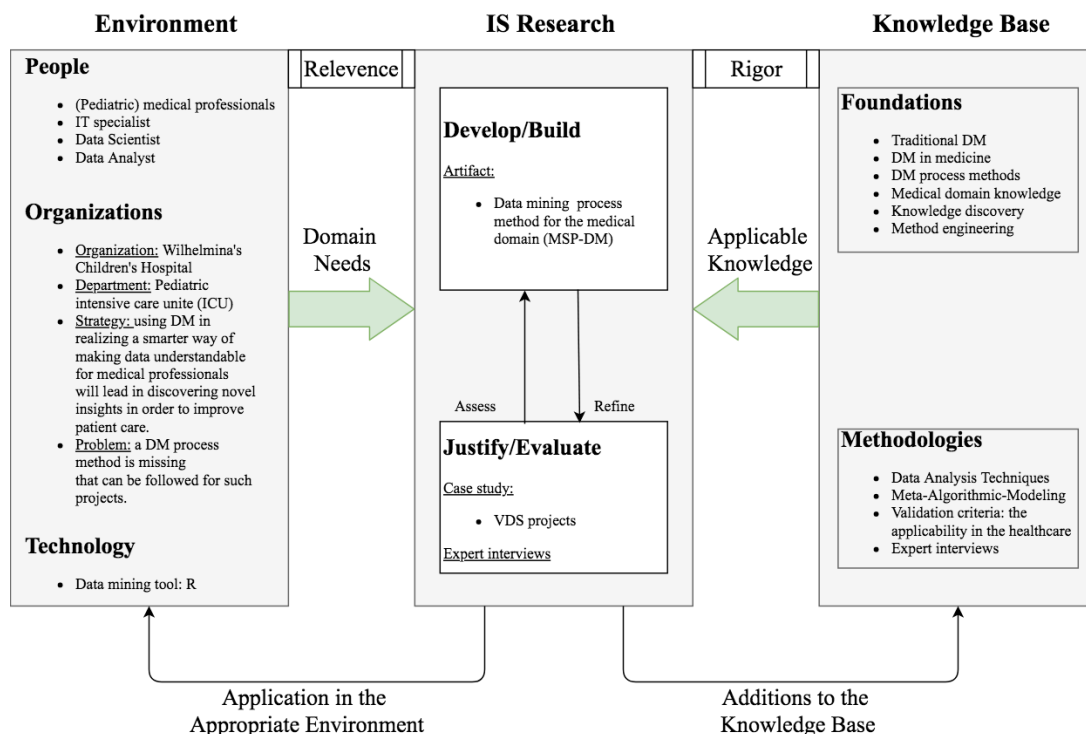


Figure 2.2: Design Science Approach based and edited on the Hevner et al. (2004)

The design science approach starts with the relevance cycle that clarifies the domain needs within a certain environment and project context. In this research, the environment or research context is explained in chapter 1. In addition, the problem statement, research objectives, and research questions are addressed in section 1.3 and 1.4. Moreover, this cycle is incorporated in the whole process of development of the artifact due to the continuous interaction with application domain and the evaluation outcomes, wherein new requirements can be discovered.

Furthermore, the rigor cycle provides the required knowledge base in both the construction and evaluation of the design artifact (Hevner et al., 2004). Herein, scientific methods and theories are selected and applied during the research project. This knowledge base is explained in this chapter 2, as well as indicated in each chapter what the theories and methods are that are used.

Lastly, within the construction of the artifact, a cycle is performed together with the requirements provided from the applicable domain and knowledge base. In addition, the artifact is assessed, justified, refined, and evaluated with the selected methods (case study & expert interviews). In this research, the case study was used in discovering and justifying the acquired method fragments and the expert interviews were conducted in evaluating the MSP-DM and its findings, as well as extracting overlooked method fragments. In addition, the case study of the VDS projects underwent the phases of the selected DM process method (see chapter 3) that assisted in extracting method fragments.

2.1.4 Meetings

In the course of the research various types of meetings were conducted with domain experts that guided the research project.

2.1.4.1 Weekly meetings

During the process of the research project and the case study weekly meetings were organized at the WKZ. In these meetings updates were provided about the ongoing research and case study, as well as other related researches were discussed from other researchers. These regular settings provided insights on the domain along with the problematics of the case study. Likewise, valuable implicit knowledge was gained from the involved domain experts during these meetings that enabled in understanding the requirements and possible solutions based on their experience within the medicine domain and data mining.

2.1.4.2 Scheduled meetings

In the case study projects, there was a clinical practitioner closely involved in the analysis and the technical aspect. Herein, a close cooperation was performed in better understanding the domain and the needs along with finding technical solutions for certain analytical problems. These cooperation sessions were regularly scheduled during the week and decreased over time while finalizing this thesis.

Moreover, other scheduled meetings were conducted with other domain experts that were needed for certain encountered problems during the research. For example, meetings were scheduled with a statistical expert who explained some aspects of certain models such as the multi-level model or there were sessions with more participants, wherein causality was explained from a medical and statistical viewpoint.

These scheduled meetings allowed in attaining a better understanding of various concepts within DM and the medicine domain.

2.1.5 Evaluation

As previously mentioned in the section 2.1.3, evaluation of the MSP-DM is essential in assessing and refining it. The evaluation was performed through expert interviews.

2.1.5.1 Expert interviews

Once the method fragments were assembled from the research and case study as explained in section 2.1.1, these findings were put forward to experts through interviews for evaluation and their feedback was collected. The interviews were conducted with five domain experts from the medicine field, with diverse lines of proficiency, as shown in Table 2.1. Due to privacy reasons, the participants were identified with a unique IDs. The selection of these experts was based on their involvement of DM in the medicine field and expertise in a particular line of business. Three expert participants (A1, 2, and 3) were closely involved with the

case study and the other participants were experts from the UMC that expressed their experience and knowledge of DM in the medicine.

Table 2.1: An overview of the participants of the interviews

IDENTIFIER	DOMAIN	ORGANIZATION	EXPERIENCE
A1	Anesthesiology, Statistics	UMC	9+ years
A2	Anesthesiology, Pediatrics	UMC	12+ years
A3	Pediatrics, ICT	UMC	9+ years
A4	Data Science, Psychiatry	UMC / UU	4+ years
A5	Data Science, ICT	UMC	3+ years

The approach of conducting expert interviews is a specialty within semi-structured interviews as the experts are determined deliberately (Muskat, Blackman, & Muskat, 2012). Hence, semi-structured interviews were conducted with medical professionals and domain experts in order to evaluate the assembled method fragments and to become acquainted with the pediatric domain, as well as extracting overlooked method fragments. These semi-structured in-depth interviews are mainly used for qualitative research which can be conducted with an individual or groups of people (DiCicco-Bloom & Crabtree, 2006). In addition, semi-structured interviews are widely employed by various healthcare professionals in their research, wherein the respondents have to answer predetermined open-ended questions (Jamshed, 2014). Herein, a schematic presentation of questions or topics are explored by the interviewer (DiCicco-Bloom & Crabtree, 2006). In addition, a protocol form was developed for the interviews as shown in the appendix: D.1 Interview protocol form, that enabled to retrieve information from domain experts through the means of interview content analysis. These interviews, which typically lasted for at least an hour, were recorded and transcribed subsequently. Then, the transcriptions were examined and interpreted in respect to the evaluation and the MSP-DM was refined accordingly with the overlooked method fragments.

Chapter 3: Data Mining Process Methods

In this chapter, an outline is provided of the various used DM process methods in the industries. The aim of this chapter is to select a fitting DM process method that will provide a foundation for the creation of a more standardized process method in healthcare. The criteria of selection are as follow:

- **Applicability in various industries**
The process method needs to have the ability to be used in various other industries and therefore be generalizable. This creates a possibility to facilitate and construct a more tailored method in which it can be built upon and modifications can easily be made.
- **Comprehensibility**
It is important that the process method is comprehensible of its phases, activities and outcomes in order to understand its place and relevance within the method.
- **Relevance in data mining**
The process method needs to be solely designed for the purpose of data mining projects. This means, that it considers not only the domain environment, but also the technical part and other disciplines. Therefore, it should have the ability to distinguish between the various related disciplinary tasks within the phases.
- **Popularity**
The popularity of a specific process method in data mining indicate its usefulness and effectiveness of the method. Thus, the most popular methods will be mainly considered as candidates.

In this research, the following process methods will be explored: CRISP-DM, SEMMA, and the KDD.

3.1 Data Mining Methods

DM techniques and methods have become an essential research area, because of their presence in the data analysis process, which makes it possible to reveal behavioral patterns, hidden relations, similar regularities, and entity profiles in data that are stored in large warehouses or databases (Bošnjak, Grljević, & Bošnjak, 2009). Ever since the emerging of DM projects in the early 90s, several industry standards, application methodologies, domain independent process models have been proposed such as the prominent of them CRISP-DM, SEMMA and PMML (Niakšu, 2015).

For the purposes of avoiding ambiguity, a definition needs to be provided for data mining method. As defined in chapter 1, data mining can be defined as the process

of mining valuable and hidden knowledge from various sources of data such as data warehouses, data bases and other information storage locations (Sun & Li, 2008). Furthermore, in the Oxford Dictionaries (2018), method can be defined as “*a particular procedure for accomplishing or approaching something, especially a systematic or established one: the quality of being well organized and systematic in thought or action*”. Hence, in this thesis the data mining method will be defined as the set of needed activities and outcomes in order to complete a DM project in which useful and implicit knowledge is being extracted from data.

3.1.1 CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a methodology that was developed in the early 2000s by Chapman et al. (2000). In 1996, the initial phases started in creating the CRISP-DM by Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR for the data mining market due to the need for a data mining process model (Chapman et al., 2000).

The purpose of the method is described as to carry out data mining projects by following the provided processes of the model in order to complete projects successfully, which is the main goal of the method (Wirth & Hipp, 2000). In addition, the method can be used in different situations to plan, document and communicate within and outside the team's projects. In other words, the method provides guidelines and a uniform framework for data mining projects. The process can be described as a cyclic, in which iterations are performed before reaching the final results and business goals (Moro, Laureano, & Cortez, 2011).

The method provides an overview of the life cycle of a data mining project that contains the corresponding phases of a project, their tasks, and relationships between these tasks (Mariscal et al., 2010). The life cycle consists of six main phases as shown in Figure 3.1. The arrows show the frequent and important dependencies between the different phases, although moving back and forth between different phases is always required as explained by the authors (Chapman et al., 2000). Hence, the cycle or order of the phases is not rigid.

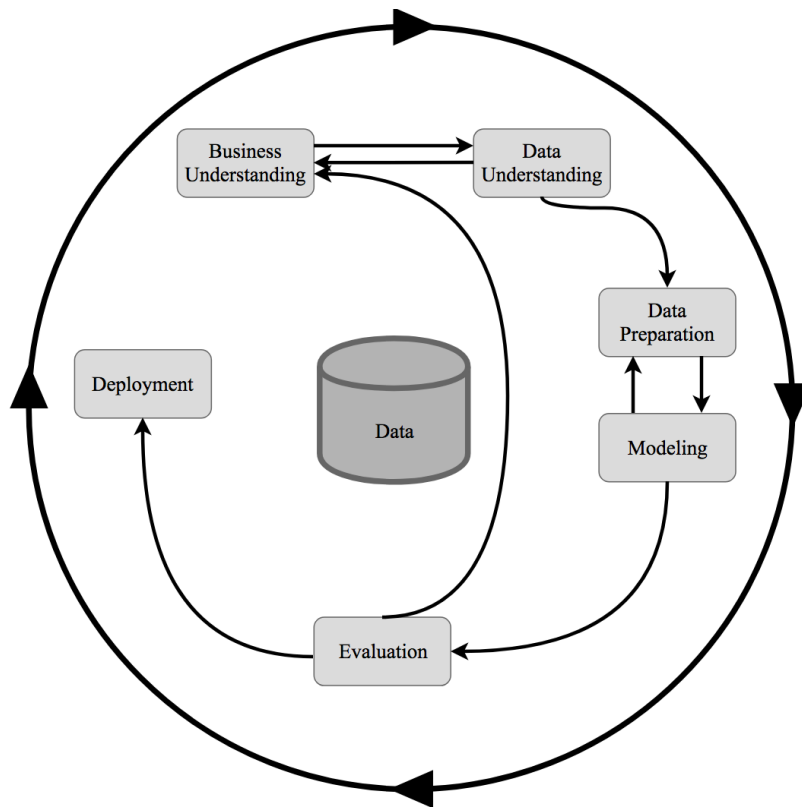


Figure 3.1: The CRISP-DM process method (Chapman et al., 2000)

Furthermore, Chapman et al. (2000) describe the method in terms of a hierarchical process model that consists of sets of tasks described at four levels of abstraction from general to specific as illustrated Figure 3.2. At the top level, the phases are shown in which each phase consists of several second-level generic tasks. Hence, the second level are the generic tasks that belong to a corresponding phase. These generic tasks are defined in a general sense, in order to be able to cover all possible data mining circumstances. Furthermore, the third level shows the specialized tasks where the actions in the generic tasks are explained on how they should be carried out. The last level is defined as process instance, that represent a record of the decisions, actions, and results of an actual data mining project in which deviations and particularities from the process are documented. In addition, the authors explain that in practice previous actions need to be repeated in which the flow of the tasks should not be taken as a rigid process similar with the life cycle.

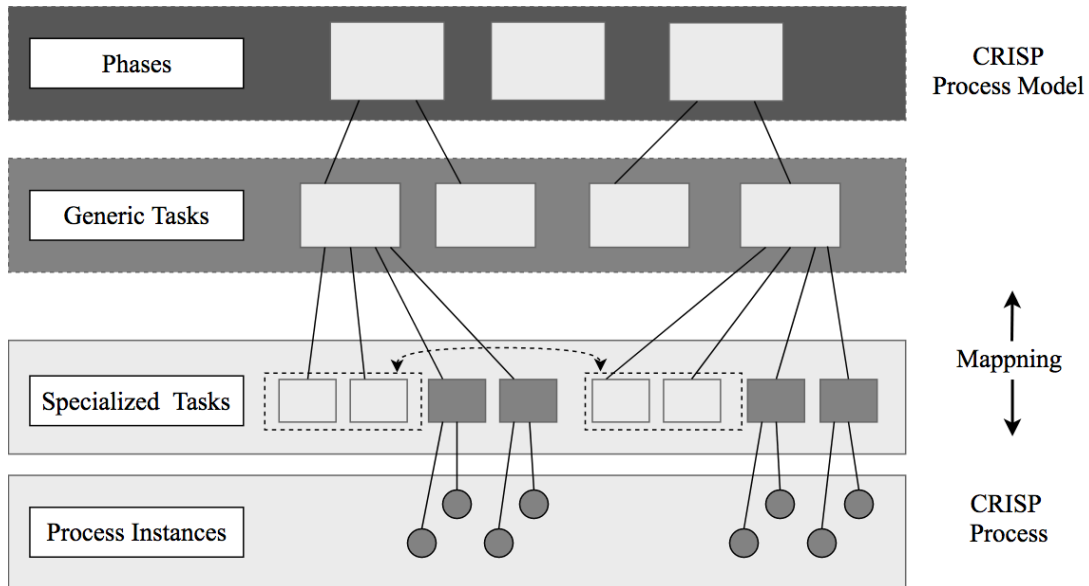


Figure 3.2: Four-level breakdown of the CRISP-DM method (Chapman et al., 2000)

The authors make a distinction between the CRISP-DM method and the user guide Chapman et al. (2000). Herein, the method presents a brief overview of the phases, tasks, and their outputs that describes what to do in a DM project. As for the user guide, more detailed hints and tips are provided for each phase and each task within a phase that depicts on how to do a DM project.

Moreover, in 2008 there was an attempt to develop a 2.0 version of the CRISP-DM, but was put at standstill wherein the status still remains unknown nor is the website active any longer (Marban O, Mariscal G, 2009).

3.1.2 SEMMA

The SEMMA method was developed by the SAS Institute that specializes in the development of analytical and processing software. The method was built for in-house usage and later adopted in other industries. The name of the method was initially proposed by Bulkley in 1991, but it was not commercially adopted until 2008 (Marban O, Mariscal G, 2009). The SEMMA is an acronym to describe the SAS data mining process which stand for Sample, Explore, Modify, Model, and Asses (SAS Institute, 2018) as shown in Figure 3.3.



Figure 3.3: SEMMA based on Mariscal et al. (2010)

Furthermore, the method is based on the technical part of the project, for example; it solely aims in solving a DM problem and ignores the managerial side (Marbán, Segovia, Menasalvas, & Fernández-Baizán, 2009). Hence, the method focuses mainly on the application of exploratory statistical and visualization-based data mining techniques (Bellazzi & Zupan, 2008).

Moreover, the SEMMA sets out a waterfall life cycle during the course of the project until the end and takes the same approach similar to the CRISP-DM in being iterative once a stage is fully completed (Marban O, Mariscal G, 2009; Mariscal et al., 2010). In addition, the method is most effectively used along with the Enterprise Miner (EM) software (Rogalewicz & Sika, 2016).

Furthermore, Mariscal et al. (2010) indicate that the method differentiates from most other process methods because it skips steps related to understanding a particular domain as well as exploring and evaluating business goals which are considered to be essential in carrying out a successful DM project.

3.1.3 KDD

The Knowledge discovery in databases (KDD) entered the stage in the early 90s to highlight that knowledge is the product of a discovery process guided by data, and is a place of assembly of various research areas that focus on data analysis and knowledge extraction from different perspectives such as artificial intelligence, statistics, logic, mathematics, and data bases (Mariscal et al., 2010; Piatetsky-Shapiro, 1990). The KDD is defined as a process that uses DM methods to extract useful and understandable patterns in data (Fayyad et al., 1996). The authors explain the importance for proper transformation and preprocessing procedure such as data selection and cleaning before commencing in analysis. Furthermore, the authors mention that the starting point of a DM project is developing an understanding of the application domain and other relevant prior knowledge and goals, although not illustrated in their visualized method as shown in Figure 3.4.

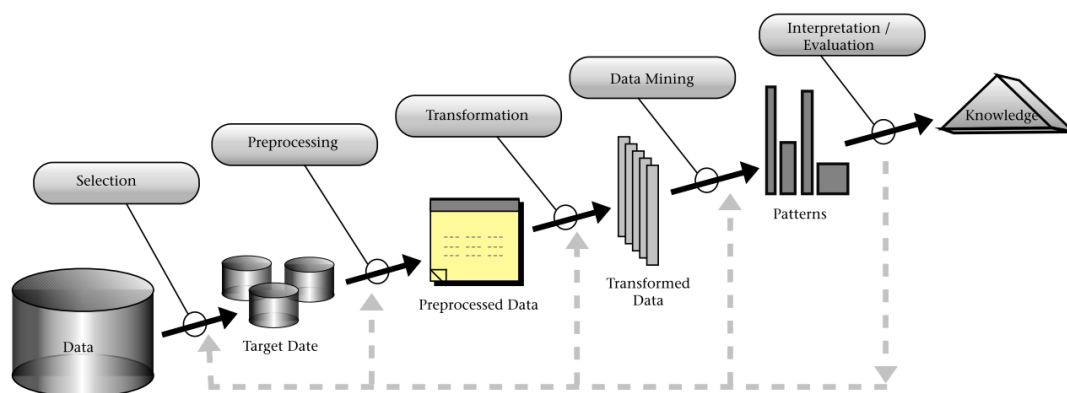


Figure 3.4: KDD method originally retrieved from Fayyad et al. (1996)

Hence the correct sequence of the phases of the method are: domain understanding, selection, preprocessing, transformation, data mining, and interpretation / evaluation.

3.2 Method Selection

This section discusses the selection criteria earlier mentioned at the beginning of chapter in relation to the above-mentioned methods.

The CRISP-DM is by far the most commonly used data mining methodology (Sharma, Osei-Bryson, & Kasper, 2012) and widely adopted in the industry (Onwubolu, 2009). In addition, the CRISP-DM method is considered to be technology neutral, industry independent, and the de facto standard for DM (Azevedo & Santos, 2008; Shearer, 2000). Likewise, in a relevant online poll by KDnuggets in 2014, 43% of the respondents as shown in Table 3.1 choose the CRISP-DM as their main method for analytics, data mining, or data science projects (Piatetsky, 2014).

	2007	2014
CRISP-DM	42%	43%
My own	19%	27,5%
SEMMA	13%	8,5%
KDD Process	7,3	7,5%

Table 3.1: Poll on data mining process method usage by KDnuggets.com

Moreover, the SEMMA and the KKD score below the 10% and custom methods indicated as “my own” score noteworthy high with an increase over the years, although it is unknown how the self-made methods are constructed. Furthermore, the CRISP-DM shows a greater extent of depth containing multiple layers of concepts and activities compared to the other methods. However, all three methods are uniquely developed for DM projects and share equally the relevance, although one is more comprehensibly described than the other.

In conclusion, from the three selected methods the CRISP-DM scores the best in regard to the criteria’s aforementioned. The CRISP-DM is the widely used and adopted method in the industries exclusively made for DM projects. Additionally, the method provides a great layer of foundations with a level of depth that enables to develop a method based on method fragments. Hence, the CRISP-DM will be used in this thesis as the base for a more domain specific DM method.

3.2.1 Limitations of the Selected Method

Despite the extensive use and popularity within DM projects, applying its phases remains a challenge for the CRISP-DM in certain domains such as the healthcare. The process model is still not mature enough to deal with the complex problems it needs to address, which results that it does not produce the expected results in which the effectiveness of its deployment is reduced (Marbán et al., 2009). Within the modeling phase of CRISP-DM method, it includes the application of various knowledge discovery and DM methods, with an extensive scale of tunable parameters (Bošnjak et al., 2009).

However, Bošnjak et al. (2009) further explain the challenges that exist with this, is that no method dominates the other methods all the time. This makes it difficult in selecting the most suitable way of modeling or knowing if the best method is chosen. Many developers of models leave the parameters to their default values, which leads to that many select algorithms which are based on intuitive appeal or reputation (Thornton, Hutter, Hoos, & Leyton-Brown, 2012). This may lead to that underperforming algorithms and/or models are selected with suboptimal results, that are deployed or not used at all.

Moreover, there have been attempts in improving the CRISP-DM methodology in general and tailoring it for the specific needs in the healthcare. In the medical domain, there have been few known attempts to provide a specialized process model for application or DM methodology (Li J, Zhang Y, 2016). Speckauskiene & Lukosevicius (2009) have proposed a generic workflow of taking care of medical DM applications. However, the authors did not cover some of the aspect within a practical DM application, such as the deployment of the modeling results, data understanding, mining non-structured data and data preparation.

Catley, Smith, Mcgregor, & Tracy (2009), introduced an extension of the CRISP-DM for temporal medical multidimensional streaming data of intensive care unit (ICU) equipment. Therein, the authors provided an example in which the activities were mapped out with the technical aspect, DM problem type, and defined application domain in order to support the researchers of intensive care unit temporal data. However, this was specifically tailored for this goal, which made it not directly applicable for other DM application goals or medical data types. In addition, the evaluation phase was excluded from the research.

Furthermore, Niakšu (2015) introduced a unique method named CRISP-MED-DM based on the original CRISP-DM method. The author purposes to resolve the challenges of the medical domain such as clinical data quality and completeness, patient data privacy, heterogeneous data, and variety of data formats and representations. However, some recent challenges were missed and most of the modifications made are already mentioned in the original CRISP-DM in generic terms. Hence, no considerable changes were made except for a few exceptions that focus more on specialized tasks instead of the generic tasks.

Lastly, Menger, Spruit, Hagoort, & Scheepers (2016) presented the CRISP-IDM, which is mainly focused in involving medical professionals or domain experts in every phase of the CRISP-DM in which modifications were implemented in order to achieve this goal. Due to the main focus being specific to a certain setting, it is not directly applicable for other DM objectives or settings. Hence, this research intends in achieving that.

Chapter 4: Overcoming the Challenges

In this chapter, the challenges that come with DM in the healthcare are discussed and how to overcome them. As mentioned in Chapter 1, the focus will be on relevant concepts and activities that are related in overcoming these challenges except validation & analytical problems, as well as user-friendliness due to the time constraints of this research will not be explored.

4.1 Data integration

Data accessibility in the healthcare can be restricted for DM purposes because data sources that can serve as input for DM are frequently scattered in different systems and settings such as clinics, administrations, and laboratories, without any common format or principles (Cios & Moore, 2002; Milovic, 2012). Therefore, data integration is required to provide a unified framework to access voluminous and diverse data sources while preserving and improving the veracity of data in order to explain the various conditions of patients (Brazhnik & Jones, 2007; Louie, Mork, Martin-Sanchez, Halevy, & Tarczy-Hornoch, 2007). This is an important step in improving data analysis (Sinha et al., 2009).

However, the aspect of clinical integration and utility is largely overlooked (Lee & Yoon, 2017; Neff, 2013; Rumsfeld et al., 2016). One of the main challenges of healthcare institutions is the provision of a unique medical knowledge extraction framework due to the lack of appropriate collection and transmission standards (Esfandiari et al., 2014). Frequently, in healthcare institutions clinical information systems are not integrated across the various departments (Niakšu, 2015). In addition, according to the survey conducted by Niakšu & Kurasova (2012) medical information systems that are frequently being used in hospitals do not support data exchange standards such as HL7, CDA, DICOM. The lack of data standards in the medicine aggravates the instability and inconsistency in medical terminology in data sources that can result in heterogeneous data issues (Rumsfeld et al., 2016). Raw medical data are heterogeneous and huge that can be collected from various sources such as interviews with the patient, images, and physician's notes and interpretations (Cios & Moore, 2002; Wasan et al., 2006).

Therefore, prior when conducting DM and knowledge discovery data must be preprocessed and transformed, in order to be able to collect the data and use it accordingly (Milovic, 2012). Big data approaches could enable such integration of the various data sources related to the omics and personal data of patients (Murdoch & Detsky, 2013). Through the establishment of data responsive data warehouses or adequate infrastructures will enable better accessibility and flow of data, although this can be an expensive and time consuming undertaking (Milovic, 2012). Thus, healthcare organizations that desire to broaden their scope in DM must consider investing resources such as time, effort and money to enable such endeavors (Koh & Tan, 2005). Without the managerial or organizational support, such projects can fail

or be delayed. Hence, DM in the healthcare demands a close collaboration between the DM experts and the management in achieving the desired objectives (Stühlinger, Hogl, Stoyan, & Müller, 2000) Therefore, integration of the information systems ranges from data exchange architecture, data transformation methods and ending with semantic data integrity (Niakšu, 2015).

4.2 Data quality

The healthcare industry has been cautious in embracing big data, because of the additional cost of adding analytical functions to the existing EHRs, poor-quality data, privacy issues, and the lack of willingness to share data (Hansen, Miron-Shatz, Lau, & Paton, 2014). However, there is an increase in data assets, while major gaps remain in the quantity and quality of data (Behrns, 2015). Rumsfeld et al. (2016), affirms in their research the limitations of data quality in the healthcare. This can heavily effect DM results and applications, since it depends on the quality of the data (Koh & Tan, 2005). In general medical datasets are large, heterogeneous and hierarchical, complex and may vary in quality (Milovic, 2012). The data quality can be affected by inaccurate measurements, human or equipment errors (Niakšu, 2015). This can cause biases in the data collection, which can significantly affect both the generalizability and performance of future developed predictive models. Lee & Yoon (2017) emphasize the necessity of improving the data quality of EHRs in which many of the technical issues are remained to be solved.

Nonetheless, several quality measures can be used in evaluating the quality of data, although accuracy and consistency are considered the most important measures that decide the quality of data (Patil, Joshi, & Toshniwal, 2010). In addition, for large amount of data the four “Vs” of big data analytics in healthcare should be taken into account: volume, variety, velocity and veracity (Raghupathi & Raghupathi, 2014). Essentially, having higher-quality clinical data may result in more clinically useful, valid, and stable DM projects (Altman & Ashley, 2015; Shah et al., 2015).

Hence, it is crucial when conducting DM in the healthcare to consider larger samples of clinical data and to perform data preprocessing, in which techniques could be used such as feature selection where outliers can be identified and ruled out (Niakšu, 2015; Rumsfeld et al., 2016). Moreover, having a good understanding of the available data is crucial in order to identify what can be used or not. In addition, being creative in identifying what can be used can result in finding ways on conducting a desired research, for example; if the quality of the data is inadequate perhaps other sources of data can be sought such as text data.

4.3 Causal inference in observational data sets

The availability of large amounts of data does not remove the inherent limitation of observational data (Rumsfeld et al., 2016). It is possible to apply analytics on large amount of data, although having fully comprehensive data sources is very unlikely. Herein, issues of sampling bias, which mainly comes from the influences of the state of a patient, are not always measured nor observed within data sources (Rumsfeld et

al., 2016). Algorithms can provide good predicting outcomes, but this does not necessarily mean that these predictors are causes of something in order for physicians to intervene (Obermeyer & Emanuel, 2016). Therefore, it is immensely important for data analysts to differentiate between correlation and causation and how those two can be applied within a DM projects in medicine.

Moreover, observational studies are unable to test for causality and therefore should be considered as hypothesis-generating (Lee & Yoon, 2017). However, the potential of big and heterogeneous data to help in understanding causal relations should be taken into account, for example; research could support comparative studies of outcomes when applying different treatments or using better characterizations of patient profiles that could provide researchers' the ability to control factors that can misperceive studies leading to false conclusions (Behrns, 2015).

Furthermore, it is recommended that the outcomes from the observational studies should not influence clinical practice until the related hypotheses are tested in a passably randomized controlled trails (Tai, Grey, & Bolland, 2014). In addition, the effect size of observational studies is frequently exaggerated because of confounding, selection bias, and methodological weaknesses such as measurement error (Lee & Yoon, 2017). Large observational studies have the ability to produce improbably precise estimates that are highly statistically significant but clinically unimportant (Slobogean et al., 2015; Tai et al., 2014).

Lee & Yoon (2017) recommend to adopt specific, scientific best practices, such as generation of a priori hypotheses in a written protocol, transparent reporting with justification of any changes in plans, and detailed analytical plans stating the specific method and precautions against bias in order to minimize the obstacles in getting valid inferences. In addition, Slobogean et al. (2015) suggests further that potential clinically important effects should be defined a priori and the outcomes discussed accordingly.

Furthermore, there are several analytical techniques that address the issue of confounding in observational studies: instrumental variable analysis, multivariable and propensity score analysis (Laborde-Castérot, Agrinier, & Thilly, 2015; Stel, Dekker, Zoccali, & Jager, 2013). In addition, there is an another variant of the instrumental variable analysis that is being used more frequently lately the Mendelian randomization study in which genetic variants of instrumental variables are used to bypass the issues of reverse causation and unmeasured confounding in observational studies (Boef, Dekkers, & Le Cessie, 2015). These techniques aim to attain a better understanding in unmeasured and unknown confounders in observational studies.

Nevertheless, DM can provide some insight or suggestions about causality but in order to discovery causal influences prospective studies can be also an option, in which a group of similar individuals (cohorts) that differ with respect to certain factors under study are tracked over time in which the effects of a certain outcome

can be measured on the basis of these factors (NCI Dictionary of Cancer Terms, 2018). Hence, prospective studies are able to provide some light on cause and effect in a given experimental setting (Straughan, P., & Seow, 2000). Additionally, as previously mentioned randomized controlled trials is also highly recommended to be conducted before implementing any outcome.

4.4 Legal issues

On May 25, 2018 the General Data Protection Regulation (GDPR) privacy law became effective in the European Union (EU) (European Commission, 2018). This new regulation is intended to place a high level of protection to personal data of European citizens. In other words, companies or organizations around the world need to establish ownership and transparency of individuals' data, in which a clear declaration of consent from them is needed to process and save their personal data.

This new privacy law includes the healthcare sector, in which obligations and responsibilities designated for data controllers and processors are established. Controllers will have the responsibility to ensure that processing of personal data fully complies with the GDPR requirements wherein technical and organizational measures needs to be established. Herein, data protection policies will need to be set and implemented.

Processors will be obliged to maintain records of all processing activities performed and to uphold disclosure readiness in order to show compliance. In addition, processing on behalf of a controller needs to be set out in a contract, which is in compliance under the GDPR.

4.4.1 The legal applications in the Healthcare

In the GDPR (2018), the personal data related to health is defined as “*information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*”. The GDPR further defines three forms of health data that are subject to higher standards of data protection:

- **Data concerning health:** defined as “*personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status*” (GDPR, 2018).
- **Biometric data:** defined as “*personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data*” (GDPR, 2018).

- **Generic data:** defined as “*personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question*” (GDPR, 2018).

In the GDPR, the process of one of these three forms of health data is prohibited, except if one of the following conditions apply:

1. The data subject must give “*explicit consent*” to the processing of the data (GDPR, 2018b).
2. That processing the data is necessary for the purpose of preventing or occupational medicine, wherein the working capacity of the employee, provision of health, medical diagnosis are needed for treatment, the management of health, social care or social care systems and services (GDPR, 2018b).
3. Processing the data is essential for public interest in the area of public health, such as safeguarding high standards of safety, quality medical devices, and medicinal products, as well as protecting against serious threats to health (GDPR, 2018b).
4. Using data for achieving purposes related to the public interests, along with statistical, scientific, and historical research purposes (GDPR, 2018b).

These conditions can be adjusted in the future by the member states (GDPR, 2018c). Thus, monitoring these legal proceedings is essential, in order not to break the law and to be aware what the possibilities there are in using personal data.

The healthcare sector will need to carry out a more holistic approach to data management. This may lead if conducted correctly to better compliance practices and reduced risks because of the awareness where the data is stored and what it is used for.

Moreover, from these new legal issues related to personal data it is important for researchers that intend to begin with a DM project to investigate with his/her associates or colleagues the availability of data that comply with these new privacy regulations and what the constraints can be. These constraints can entail depersonalizing or anonymizing the personal data by those who have the access to do so before receiving the data. In addition, if the researcher has the authority to access the data him or herself and works with others that do not have the right to see the personal data, then the data needs to be depersonalized or anonymized if cooperation is intended. Thus, the implications are that researchers that are subject to the GDPR must build robust anonymization into data science and data engineering processes.

4.5 Identified method fragments

The findings of this part of the research are mainly related to specified issues and specialized tasks that can be performed in overcoming the addressed challenges. This thesis mainly focuses on the generic tasks within a DM projects as illustrated in Figure 3.2 of chapter 3 and further explained in chapter 6, wherein the specialized tasks can be a part of a generic task. Hence, two method fragments are found that can be used, as shown in Table 4.1.

Table 4.1: An overview of identified method fragments in chapter 4

Identified method fragments	Type method fragment	Designed phase	Section
Legal constraints	Concept	Domain Understanding	4.4
Anonymize dataset	Activity	Data Preparation	4.4

Chapter 5: Case study: The VDS project

In this chapter, the insights and experience gained from the VDS case study are presented. In addition, both the VDS1 and VDS2 projects are provided by the Wilhelmina's Children's Hospital (WKZ) and are discussed within the same sections, in which the experiences are documented. The sections of this chapter will be structured in a way that represents the phases of the CRISP-DM, in which each phase will have its own section. Therein, the steps undertaken and extracted method fragments will be captured that are unique within the medical domain or not currently used within the CRISP-DM.

In the literature it was indicated, that the clinical staff that are involved in a DM projects are usually unfamiliar with its concepts, in which a gap can be created between the (medical) domain experts and the researcher performing the DM tasks (Menger et al., 2016; Meulendijk et al., 2013). In this case study, the involved (clinical) practitioners were all well aware of the DM concepts, although some were more acquainted than others. Nevertheless, not involving clinical practitioners in this process could lead to failure to adopt a certain technology or DM outcome, because the clinical practitioners could feel surpassed, not being consulted or involved in the process (Menger et al., 2016). Hence, a collaboration is required in mitigating this problem through their (clinical practitioners) provided input that can determine relevant outcomes and issues, which will lead to better analysis and easy implementation of the outcomes that are found (Brennan & Bakken, 2015; Menger et al., 2016). During the whole course of the case study (medical) domain experts were involved in guiding and assisting the projects. In addition, these experts were also interviewed and are indicated as interviewees A1, A2, and A3. Furthermore, the programming codes of the case study can be found in the appendix B.1 VDS1 code and C.1 VDS2 code.

5.1 Business Understanding

The case study was conducted at the pediatric intensive care unit (PICU) at the Wilhelmina's Children's Hospital, which is one of the four pediatric cardiac centers in the Netherlands and a university teaching hospital with a PICU and a level of 3 neonatal intensive care unit (NICU). Neonates with congenital heart diseases that require cardiac surgery are usually admitted to the PICU, while other neonates and those with non-cardiac congenital malformations are admitted to the NICU, although there could be exceptions if there are shortages of beds at one of the units (Snoep, Jansen, & Groenendaal, 2018). The relevant clinical staff consist of various practitioners involved in the pediatric, anesthesiology care, as well as nurses and other additional staff that directly support treatment.

Firstly, in order to become acquainted with the PICU department and the environment of the healthcare a process of initiation was accompanied that consisted of spending time and getting introduced in the field of nursing and pediatric care

which aided in a better understanding of the domain and purpose of the case study. Herein, the first two weeks were spent shadowing nurses and doctors in their day-to-day work. In addition, there have been multiple sessions with ventilation practitioners that explained the technical aspects of the ventilators, as well as the biological aspect of ventilation.

In addition, regular meetings with clinical practitioners were held in understanding the issues at hand in regard to the provided projects. This included extracting the necessary requirements and needs for the projects as well as developing a sufficient understanding of the medical domain. Moreover, this resulted in a clear and specific mapping of the requirements of the projects as illustrated in Table 1.2 in chapter 1 that were needed for acquiring and selecting the relevant data sources, and type of analysis. Furthermore, scientific literature was consulted in getting a better understanding related to the project issues such as literature about causality and how to work with it.

5.1.1 VDS1: Business Understanding

The VDS1 project was primarily set to be a training exercise before conducting the VDS2. However, it did manifest itself to be more than that because it provided an initial challenge in the analysis and the outcomes had a direct impact on the current understanding of software updates of clinical ventilators. Vendors of these ventilators regularly provide software updates, with the intention of optimizing the clinical support. This was also the case in 2013, in which Gettinge, former Maquet a vendor of such ventilators changed the inspiratory tidal volume measurement in their update from ambient temperature and pressure dry gas (ATPD) to body temperature and pressure saturated gas (BTPS) in order to enhance the display of tidal volume (V_T). Hence, the aim of this project was to investigate the impact of this software update. Significant changes due to the update will not only have clinical implications but managerial too wherein the supplies of these ventilators will be involved in discussing these impacts.

In this study, it was required to get familiar with the domain of respiratory ventilation and the related biological implications. Thus, regular meetings were scheduled with experienced clinical practitioners, wherein sessions were held in attaining a sufficient knowledge of the related subjects with the aim of being able to conduct what needs to be researched and to understand the various provided parameters from the ventilators. This research required a comparative analysis in order to provide a sufficient answer to the given research question of the project.

5.1.2 VDS2: Business Understanding

During the process of the VDS1 project a vast amount of domain knowledge was developed, which would be applicable for the second project because similar data sources and parameters were used, although with another goal and data. Similarly, the regular meetings continued with (clinical) practitioners discussing the project, goals, processes, analysis and models. In these meetings, insights were gained in

regard to what models would fit and what to consider and what not. Herein, the goal was to find a suitable predicting model that can provide good predictions in respiratory deterioration or improvement, wherein the inferences is shown between the various predictors. Respiratory deterioration may be caused by progression of the underlying illness (e.g. pneumonia or viral bronchiolitis) and/or suboptimal ventilator settings. Likewise, it is important in the setting of respiratory improvement, to optimize ventilator settings to reduce patient ventilator-induced injury. Similarly, reducing or ending ventilation of patients is also considered as an improvement. Hence, this project required a machine learning model that is able to learn from large amount of data and make accurate predictions.

5.2 Data Understanding

The data sources for both VDS1 and VDS2 were provided anonymized by the WKZ which made it not possible to track back the patients. The data sources were based on historical data from various Electronical Medical Record (EMR) sources (Metavision, GLIMS and HiX). These include minute-by-minute ventilator settings and monitoring parameters (SpO₂, EtCO₂), lab values (blood gas analysis and infectious parameters), and patient demographics and disease characteristics. In addition, a selection of parameters was conducted by the practitioners that were involved with the projects and a list was made of all parameters with their meaning which were explained during meetings. Herein, a description was provided of the parameters and its origin, as well as which to consider or not in the analysis. All the datasets were initially depersonalized by one of the practitioners at the WKZ and provided for usage.

5.2.1 VDS1: Data Understanding

The data of all children that were admitted to the PICU of the University Medical Center Utrecht during the period of 2012-2014 that were ventilated with a Maquet SERVO-i® ventilator on Pressure Regulated Volume Controlled mode were collected. In addition, retrospectively per minute ventilator data was acquired. This included in total 454 patients (1,063,901 observations) which were divided by the year before the change of the software update with 221 patients (532,930 observations) versus 235 patients (551,937 observations) the year after the software update (two patients received both). In this phase a close collaboration was established with the clinical practitioners that were involved with the project, in order attain a good understanding of the data and its parameters that were included.

Moreover, all variables were initially included from the datasets which enabled exploratory data analysis to be conducted. The main purpose of this analysis was to find unexpected and new patterns or relations. Herein, scatterplots as well as bar plots were used for visualization purposes. These visualizations provided some indications that there is a difference in the measurement of the V_T , which needed further research in the following phases.

5.2.2 VDS2: Data Understanding

Similarly, with the VDS1 project retrospective ventilator data was used from 2008 – 2017 with a larger data set and with more variables. However, due to technical issues in processing such an amount of data at once, a selection was made only to use the pressure-regulated volume control (PRVC) setting mode as a starting point in doing the research and developing an appropriate predictive model. Herein, the dataset contained 4+ million observations with 150 patients and 38 parameters initially that were later adjusted. Here as well, exploratory analysis was performed in order to get familiar with the dataset and meetings were scheduled in discussing the dataset and the initial findings. These meetings were required because the provided parameters in the dataset are related to medical terminology, wherein clarification was essential as well as how the registrations were recorded and calculated.

5.3 Data Preparation

In this phase for both projects the selected data was first formatted into a desired format such as date's and time. Afterwards, the necessary merges were performed as well as feature engineering, in which derived attributes were made to make new variables. Then, the data was cleaned for noisy or missing data or parameters were omitted due to their irrelevance or lack of registered data. During the data preparation the most important tasks are; integrating, transforming, cleaning, reducing, and discretizing (S. Zhang, Zhang, & Yang, 2003).

Moreover, there was an iterative process between data preparation and modeling during the case study for both projects in which data was prepared for a certain model technique and if changes occurred in the model technique such as change of parameters or choosing another model technique the dataset was adjusted for that.

5.3.1 VDS1: Data Preparation

The VDS1, underwent the typical data preparation procedure once the dataset was selected by formatting first then merging, constructing and finally cleaning the dataset. No abnormalities occurred during the process except that it did not follow the exact same order of activities illustrated in the CRISP-DM, wherein the dataset needs to be cleaned first after the data selection, then constructed, merged, and finally formatted. This deviation was performed due to the convenience of formatting the dataset first and performing the cleaning at the end because this brought some initial structure within the dataset that enabled to get a better understanding of the dataset overall.

5.3.2 VDS2: Data Preparation

Similarly, the same structure was followed as with the VDS1 project during the data preparation. In addition, the dataset needed to be normalized because this was required for the selected modeling technique which could be categorized as part of the construction activity. This was an iterative process between the data preparation and modeling phase once a specific model was selected.

5.4 Modeling

During the modeling phase, both case study projects did not specifically deviate much from the CRISP-DM. However, for researchers who are not vested in the medicine domain it can be challenging in developing appropriate models due to the deficiency of knowledge in regard to the provided parameters. Likewise, this issue happened during the case study, wherein the domain knowledge was lacking in understanding the parameters and their relation with each other from a medical perspective. Hence, meetings were scheduled with the practitioners that were able to provide the needed understanding in selecting or using the appropriate parameters. Menger, Spruit, Hagoort, & Scheepers (2016), explain that usually the researcher or data analyst receive a specific problem by domain experts, wherein statistical or technical modeling is used in answering the problem and then collaborate with the domain experts again to possibly refine and evaluate the models. During the case study there was a close collaboration with the domain experts in regard to selecting and assessing the parameters of the models, as well as evaluating the model. Through some basic visualization such as histograms, scatter plots, bar charts, and trend lines the performance of the models were presented and discussed with the domain experts. The clinical professionals helped in filtering and selecting relevant parameters for a specific output variable on which the models were made for. This interaction was helpful because in depth domain knowledge was essential in understanding the results and its relations between the various variables.

Moreover, the same order was followed provided by the CRISP-DM for both of the projects with not much deviations in regard to the activities and deliverables except that more emphasis was pointed out in respect to the involvement of domain experts. In addition, scientific literature was sought in understanding the domain along with the possibilities of prediction models that could fit the desired objectives and available data. Herein, a list of models was researched such as tree, linear, probabilistic, and rule models. Moreover, it was important to differentiate between correlation and causality in assessing the results of the models because the clinical practitioners involved in this case study indicated that correlation does not automatically mean that intervention is necessary or that there is a real causal effect. In addition, the causal inference in observational datasets is discussed in section 4.3. wherein this was also highlighted.

5.4.1 VDS1: Modeling

During this project it was quite difficult in finding the appropriate model for comparative analysis because of the complexity of not equally having the same sample size as well as the issue of not having the same subjects before and after the software update. Nevertheless, through the weekly meetings wherein practitioners and experts were present in discussing the overall project, ideas and suggestions were provided that made it possible in discovering appropriate models and experimenting with them. Here, the mixed effect model or multilevel model was proposed that has the ability in handling random effects, that was applied in this

case. This helped in overcoming the problem of not having an equal sample size. In addition, the parameters were selected in collaboration with the domain experts, wherein relations between parameters or variables and the output variable were statistically measured that enabled in making conclusions.

5.4.2 VDS2: Modeling

Similarly, with the VDS1 project domain experts were consulted in the assessment of the models and fine-tuning of the parameters. However, the selection of the models was a bit different in which various models were initially tried such as the ARIMA model in predicting trends and finally selecting neural networks for this project. Models were dismissed due to their limited possibilities they provided in respect to the requirements needed for the project. For example, the ARIMA model was initially tried out which in itself is a univariate model but did not perform that well in predicting future outcomes, as well as explaining the influences of its predictions. Hence, a multivariate model was selected that was able in handling large amount of data and providing some explanation of its influence of its predictions. Here, the neural network model was selected as a prediction model for this project because of its ability in predicting accurately by learning from the data. Hence, during this project regular meetings were scheduled in discussing the model selections, as well as the model performance and parameters settings.

Moreover, a smaller dataset of one patient with 2.200 was selected to try the different models and test the performance. This happened iterative with the data preparation phase in adjusting the datasets to fit the model techniques. Moreover, three analyses were performed in this setting: on a smaller dataset of one patient with 2.200 observations, on a bigger dataset of one patient with 68.506 observations, and on the whole dataset with 834 patients and 4.115.249 observations. The reason for such mapping was initially to test the model first on how it performs on a smaller dataset, then on a somewhat larger dataset and later on a big dataset. Another reason was that processing the neural network on the larger dataset is time-consuming and in order to measure the performance of the model a smaller dataset was a good starting point.

5.5 Evaluation

In this phase both of the VDS projects, the results were reviewed and evaluated with healthcare professionals and domain or technical experts. Herein, the healthcare professionals would provide the expertise of their domain and guide the analysis by finding new relevant insights. The experts would provide their feedback concerning the performance of the models and if they meet the requirements, along with improving the technical part if needed. There were instances that models were changed because there were other better models that could have been used or that the dataset needed some adjustment in improving the models or that fine-tuning of the models was required. Note that some of the health care professionals were also technical or data mining experts, in which the exploration and assessment of the project was directed upon.

5.5.1 VDS1: Evaluation

The models were reviewed in collaboration with health care professionals and domain or technical experts. Herein, the results of the models were assessed and adjusted with the experts in order to enable correct conclusions in answering the problem question. Table 5.1 illustrate the baseline characteristics, wherein the minute volume per kg was found to be different in the ATPD versus the BTPS (mean (SD) 259 (83) ml/kg versus 267 (66) ml/kg, $p < 0.001$).

Table 5.1: Baseline characteristics

	All patients	ATPD ventilated patients	BTPS ventilated patients	p-value
	<i>n = 454 ** (1,063,901 observations)</i>	<i>n = 221 (532,930 observations)</i>	<i>n = 235 (551,937 observations)</i>	
Age, median (IQR), months	5.6 (28.1)	11.3 (63.0)	4.3 (9.6)	<0.001
Weight, median (IQR), kg	6.1 (8.3)	8.4 (15.3)	5.45 (4.7)	<0.001
Tidal volume per kg, mean (SD), ml/kg *	7.29 (1.36)	7.31 (1.32)	7.27 (1.39)	<0.001
Respiratory rate, mean (SD), /min *	36 (7)	35 (8)	37 (6)	<0.001
Peak pressure, mean (SD), cm H2O *	21.0 (5.6)	21.8 (5.4)	20.2 (5.6)	<0.001
Minute volume per kg, mean (SD), ml/kg *	264 (75)	259 (83)	267 (66)	<0.001

* analysis of all observations per ventilator

** some patients were admitted in both ATPD and BTPS periods

In addition, in a simple linear regression and a random effect model, as shown in Table 5.2 and Figure 5.1, the interaction between the ventilation mode and minute volume per kg indicated a statistical significant effect on etCO₂.

Table 5.2: Linear mixed-effects models fit by maximum likelihood

	Value	SE	DF	t-value	p-value
Intercept	50.117	0.847	1063444	59.173	< 0.001
Minute volume to weight	-0.022	< 0.001	1063444	-148.548	< 0.001
Mode (ATPD vs BTPS)	-2.446	0.643	450	-3.806	< 0.001
Peak pressure	-0.168	0.032	1063444	-5.258	< 0.001
Weight	-0.128	0.074	450	-1.740	0.083
Age	0.001	0.001	450	1.043	0.298
Minute volume to weight x Mode (ATPD vs BTPS)	0.005	< 0.001	1063444	26.321	< 0.001

Random effects: peak pressure and patient (observations per patient)

Fixed effects: etCO₂ ~ Minute volume to weight x Mode + Peak pressure + Weight + Age

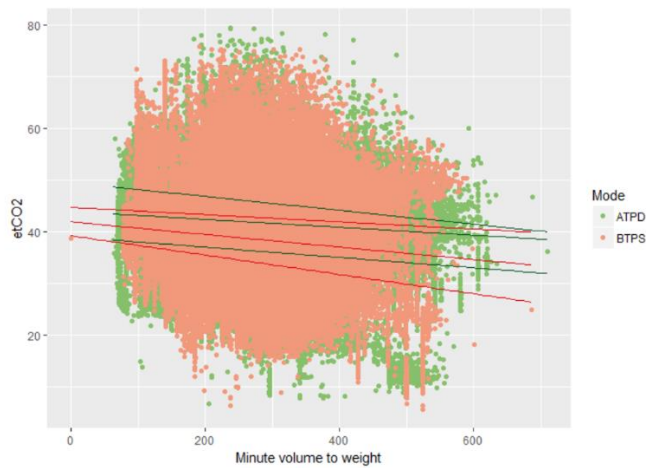


Figure 5.1: Regression lines and confidence limits (95%) of minute volume to weight versus etCO₂ per ventilator mode

The results of the project were reviewed by making sure that everything was correctly performed on the analytical side, as well as the technical side. Then from that point on, the project was taken over by the people involved with the project due to the impact of the findings. The outcome of the comparative analysis indicated that there was a statistical significant effect between the ventilation mode and the minute volume per kg on the etCO₂ that was caused by the software update. Henceforth, presentations were given to the ventilator suppliers, as well other events in which a paper was written referring to the research and its findings.

5.5.2 VDS2: Evaluation

In this phase, the VDS2 project started with a pilot model before being reviewed. Herein, the selected model was tested with a larger dataset since a machine learning technique was used in (deep) learning from the data in order to make accurate predictions. The initial test of the model was tested on a smaller dataset of one patient, which provided good predictions, as shown in Figure 5.2 and Figure 5.3.

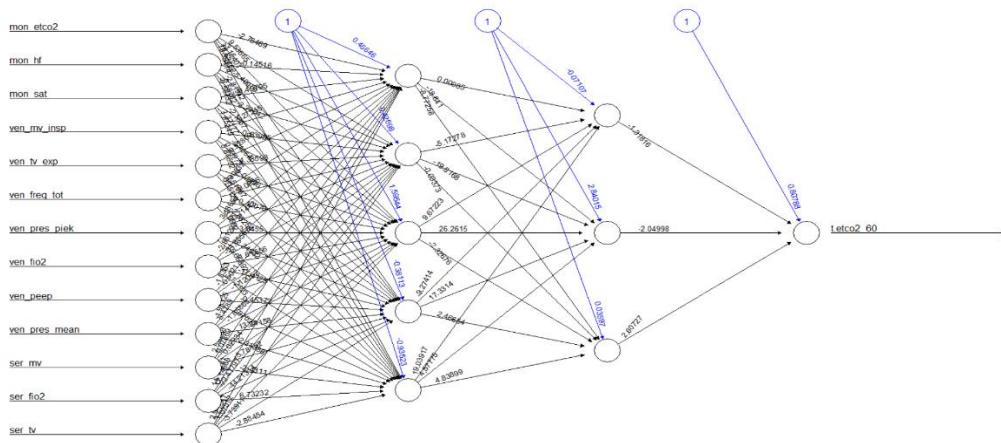


Figure 5.2: The creation of the Neural Network

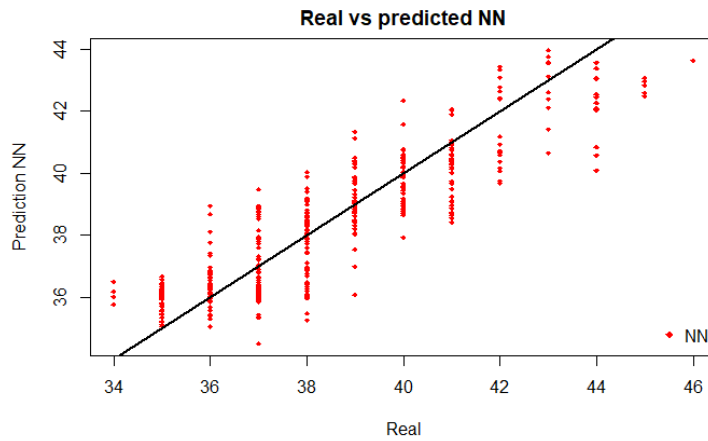


Figure 5.3: Results of the Neural Network prediction of one patient on a small dataset

The neural network was created with two hidden layers (5, 3) with a mean squared error (MSE) of 1.32. Similar positive results were noted when a larger dataset of one patient was used, in which the MSE decreased when the hidden node parameters were adjusted, as shown in Table 5. 3.

Table 5. 3: MSE results of different node parameters

Nodes	MSE
2, 1	4.17
3, 2	3.85
5, 3	3.48
6, 4	3.47

However, when a larger dataset was used to test this pilot model that included more patients, predictions of the model were dramatically less accurate, as illustrated in Figure 5.4.

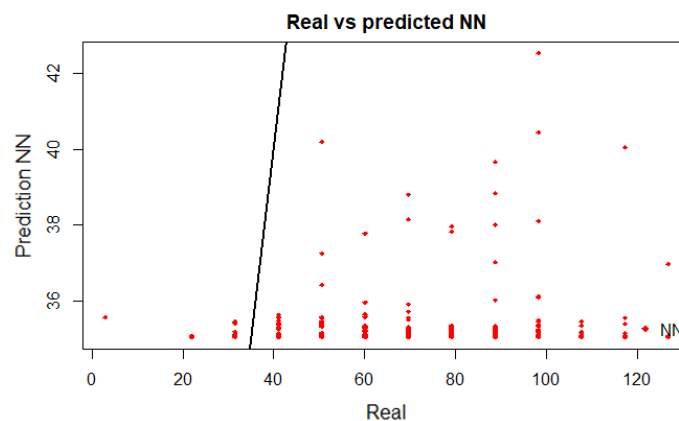


Figure 5.4: Results of the Neural Network prediction with all patients on a big dataset

The results indicated a MSE of 28.97, which is dramatically higher compared to the previous predictions with the same hidden layers (5, 3). This brought some new

implications to the project and possible reasons why the results were disappointing. For example; signal-to-noise ratio that favors noise (unwanted data) instead of the signal (useful information), specific patients that hugely influence the results, heterogeneity between the patients, missing variables that were not included such as clinical data, currently unavailable data such as text data, interpretations or undocumented (prior) knowledge, and another reason can be the overall quality of the dataset in which better selection is required. These issues were brought forth to the (medical) domain experts and were discussed, wherein possible actions were proposed. There were three options put forward; acquiring better data, creating more specified models for this scenario or halt the project until one of the first two options is realized. Due to the time constraint of this thesis it was not possible to work on solving the problem. However, the project revealed significant results such as that neural networks work but need adjustments or that the dataset is incomplete which needs to be worked on before implementing a model into a system.

5.6 Deployment

The deployment phase usually emphasizes on implementing the findings attained from the modeling phase on the work floor (Menger et al., 2016). This phase does not distinctly differ much from typical data mining projects, since the findings obtained need to be deployed and transformed to daily work practice. In this case study the VDS1 project only managed to successfully pass the reviews in the evaluation phase, in which actions and implications of the results are still under review. In addition, the purpose of the VDS1 project was not a new model that would be implemented on the work floor, but more a comparative analysis that could influence the current understanding of software updates of ventilators. The analysis was performed with new insights, wherein the involved parties can determine and assign further actions in a follow-up with the management of the relevant unit(s). Concerning the VDS2 project, it requires further research as mentioned previously in the evaluation phase.

Furthermore, the deployment phase was very limited for the case study due to the previously mentioned matters. Therefore, there are not many particularities to be mentioned in this phase.

5.7 Identified method fragments

In the case study, multiple method fragments were identified that deviated from the CRISP-DM during the provided projects, as illustrated in Table 5.4.

Table 5.4: An overview of identified method fragments in chapter 6

Identified method fragments	Type method fragment	Designed phase	Section
Conduct regular meetings with (clinical) practitioners.	Activity	Business Understanding	5.1
The available scientific literature	Concept		5.1
Assess the data lineage with a domain expert	Activity	Data Understanding	5.2
Anonymize dataset	Activity	Data Preparation	5.3
Order in the preparation of the data	Activity		5.3
(Clinical) practitioners opinion during model assessment	Concept	Modeling	5.4
Involve health care professionals and other experts in the review of the results	Concept	Evaluation	5.5
Possible other actions after evaluating the results: moving back to data preparation, modeling or approving the model	Activity		5.5.2
Produce pilot model	Activity		5.5.2

Chapter 6: Method Fragments

This chapter will assemble the method fragments collected from Chapter 4, the literature, the case study and the interviews conducted with the domain experts. The CRISP-DM will be the foundation on how the method fragments are structured. In addition, the activities and concepts found will be grouped within one of the phases (sections) of the CRISP-DM: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This research will focus mainly on the non-technical aspects of DM and will partially include technical tasks that needs to be performed during data understanding, data preparation, and modeling phases. This means that the technical activities will be explained that need to be executed without going much into depth on the programming aspect, the available tools to be used or functions. Additionally, this research will mainly focus on the generic activities and its deliverables presented in the CRISP-DM as shown in Table 6.2, except in cases that need further explanation for a specific activity.

Moreover, for the modeling of the MSP-DM, a Meta-Algorithmic-Modelling approach will be applied as proposed by Weerd & Brinkkemper (2009). This method engineering technique provides a visualization of relations between the activities (the process), its transitions, and the corresponding deliverables defined as concepts. The outcome of this visualization will be a process-deliverable-diagram (PDD), which shows on the left-hand side an UML activity diagram of the processes and on the right-hand side an UML class diagram with its deliverables/concepts. These diagrams are conjoined with each other and display how the activities are tied to its corresponding deliverable. Additionally, each activity and concept is separately explained in a table that is depicted in the diagrams.

In appendix A. Process-deliverable-diagram CRISP-DM, the method fragments originally from the CRISP-DM are presented that form the foundation for DM projects. Therein, the general method fragments are illustrated and explained with the corresponding activity and concept tables.

Section 6.1 Method Fragments, provides an insight on the method fragments extracted from the VDS case study and conducted research from the literature and expert interviews.

6.1 Method Fragments

In this section, a comprehensive overview is provided with the collected method fragments from the case study, literature and from chapter 4 and amplified with the current CRISP-DM method with the PDDs that are created along with its corresponding activity and concept table. This technique for meta-models consist of a process view on the left-hand side of the diagram, whereas the activities are modeled and on the right-hand side the deliverables and their relations with each other are shown (Weerd & Brinkkemper, 2009). In Table 6.1, an overview is illustrated of the identified method fragments that have been collected from various

sources in which the designation and type is provided. In addition, the case study is also provided as source, wherein domain experts A1, A2, and A3 played a major role in guiding the project that enabled in identifying method fragments. Moreover, the interviewees mentioned as source of the identified method fragments have explicitly emphasized on that particular method fragment, although others have expressed their opinion about those fragments but not in depth. In addition, Table 7.1 from chapter 7 provides the results of the interviews.

Table 6.1: An overview of identified method fragments

Identified method fragments	Type method fragment	Designed phase	Source
Rename Business Understanding to Domain Understanding	Phase name	Domain Understanding	(Menger et al., 2016); Interviewee A4
Clinical and Managerial objectives	Concept		(Niakšu, 2015); Interviewee A1, A4 and A5
Legal constraints	Concept		Section: 4.4; Interviewee A3
Stakeholders	Concept		Interviewee A3
Privacy Risk Assessment	Concept		Interviewee A4 and A5
Conduct regular meetings with (clinical) practitioners.	Activity		Case study: Business Understanding
The available scientific literature	Concept		Interviewee A2, A4, and A5; Case study: Business Understanding
Assess the data lineage with a domain expert	Activity	Data Understanding	Interviewee A3, A4, and A5; Case study: Data Understanding
Anonymize dataset	Activity	Data Preparation	Case study: Data Preparation; Section: 4.4; Interviewee A1 and A3
Order in the preparation of the data	Activity		Interviewee A3, A4 and A5; Case study: Data Preparation
(Clinical) practitioners opinion during model assessment	Concept	Modeling	Interviewee A1, A3, A4, and A5; Case study: Modeling
Involve health care professionals and other experts in the review of the results	Concept	Evaluation	Interviewee A1, A2, A3 A4 and A5; Case study: Evaluation
Possible other actions after evaluating the results: moving back to data preparation, modeling or approving the model	Activity		Interviewee A1, A2, and A5; Case Study: Evaluation
Present the results to stakeholders and write a report/paper	Activity		Interviewee A5

Produce pilot model	Activity		(Tai et al., 2014); Interviewee A4 and A5; Case Study: Evaluation
Delete produce final report	Activity	Deployment	Interviewee A5;
Review with end user	Activity		Interviewee A5;
A follow-up action if feedback is related to the domain understanding	Action		Interviewee A3 and A5;

Moreover, in Table 6.2 the phases of the original CRISP-DM method are presented that will be used as a foundation layer for other method fragments collected. The phases are indicated as an open activity with sub-activities. In addition, the generic tasks are the sub-activities within an open activity and the outputs are the concepts in the right-hand side in the deliverables.

Table 6.2: The Generic tasks (**bold**) and Outputs (*Italic*) of the CRISP-DM retrieved from Chapman et al. (2000), page 12

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives	Collect Initial Data	Select Data	Select Modeling Techniques	Evaluate results	Plan Deployment
<i>Background</i>	<i>Initial Data Collection Report</i>	<i>Rationale for Inclusion/Exclusion</i>	<i>Modeling Technique</i>	<i>Assesment of Data Mining Results w.r.t. Business Success Criteria</i>	<i>Deployment plan</i>
<i>Business Objectives</i> <i>Business Success Criteria</i>	Describe Data	Clean Data	<i>Modeling Assumptions</i>	<i>Approved Models</i>	
Asses Situation	<i>Data Description Report</i>	<i>Data Cleaning Report</i>	Generate Test Design	Review Process	Plan Monitoring and Maintenance
<i>Inventory of Resources</i>	Explore Data	Construct Data	<i>Test Design</i>	<i>Reviews of Process</i>	<i>Monitoring and Maintenance plan</i>
<i>Requirements, Assumptions, and Constraints</i>	<i>Data Exploration report</i>	<i>Derived Attributes</i>	Build Model	Determine Next Steps	Produce Final Report
<i>Risks and Contingencies</i> <i>Terminology</i> <i>Costs and Benefits</i>	Verify Data Quality	<i>Generated Records</i>	<i>Parameter Settings</i> <i>Models</i> <i>Model description</i>	<i>List of Possible Actions</i> <i>Decision</i>	<i>Final Report</i> <i>Final Presentation</i>
Determine Data Mining Goals	<i>Data Quality Report</i>	Integrate Data	<i>Asses Model</i>		Review Project Experience
<i>Data Mining Goals</i> <i>Data Mining Success Criteria</i>		<i>Merged Data</i>	<i>Model Assessment</i> <i>Revised Parameter Settings</i>		<i>Documentation</i>
Produce Project Plan		Format Data			
<i>Project Plan</i> <i>Initial Assesment of Tools and Techniques</i>		<i>Reformatted Data</i> <i>Dataset</i> <i>Dataset Description</i>			

However, due to the complexity of connecting the relations with other concepts between the phases, the phases will be discussed and visualized separately in order to avoid ambiguity. In addition, the connecting relations will be further explained in the activity and concept tables. Moreover, all of the sub-activities are closed because on the deliverable side, those outputs explain what the activity does and going more into depth would be irrelevant. The same applies to some closed concepts, in which it would increase the complexity that are not relevant to the specific context such as a sum up list of a specific concept. In addition, other closed

concepts require separate explanation but due to the visual complexity that can cause further ambiguity separate PDDs were made that mainly explain a certain activity or concept.

6.2.1 Domain Understanding

The phase name “Business Understanding” is renamed to “Domain understanding” to avoid ambiguous meaning from other industries in the private sector to a more appropriate name that fits the clinical and managerial nature in hospitals (Menger et al., 2016). In addition, the task “Determine business objectives” was renamed to “Determine objective” in order to be more generic in determining the objectives, because in the healthcare there could be two separately perspectives in defining goals, i.e. the management and clinical perspective (Niakšu, 2015). This was also highly supported by interviewee A1, A4 and A5.

In chapter 4, legal issues were mentioned in which a new privacy law is being enforced in the EU since May 2018. This will have consequences in regard to collecting personal data from patients. Thus, addressing the issue of patient data privacy is essential before starting with a DM project. Currently, the CRISP-DM contains legal issues under the concept of requirement that checks if the data is allowed to be used. However, the legal constraints are not included in the current method that provide what restrictions a particular dataset has or restrictions of other datasets that are required. In the interview with A3, it was indicated that data needs to be depersonalized for usage which can be a legal constraint. In addition, during the interviews with the experts A4 and A5 it was clearly point out that a privacy risk assessment is included when a project is proposed. Therein, risk factors of privacy data are documented, for example; what the risk can be by using a particular data in respect to where it will be used and by whom and the overall setting around it.

Another finding was that from the case study it was indicated by the responsible for the data collection as well as being an interviewee A3 that collecting the different data from different systems that identifying the stakeholders is highly important in order to obtain on a timely basis the right permissions and access of data. Menger et al. (2016) explain that obtaining relevant data in hospitals can be difficult because the staff that supplies the data may not be accustomed with the idea DM. Thus, the stakeholders are added in the inventory of resources as a property that needs to be identified in order to approach the right people wherein the intentions of the project can be explained. In addition, it was mentioned by the interviewee A2, A4, and A5 that scientific literature is consulted during the assessment of the situation in regard to the problem and/or objective of the project. Figure 6.1, illustrates the changes made in the CRISP-DM, Table 6.3 and Table 6.4 provide further explanation of the activities and concepts.

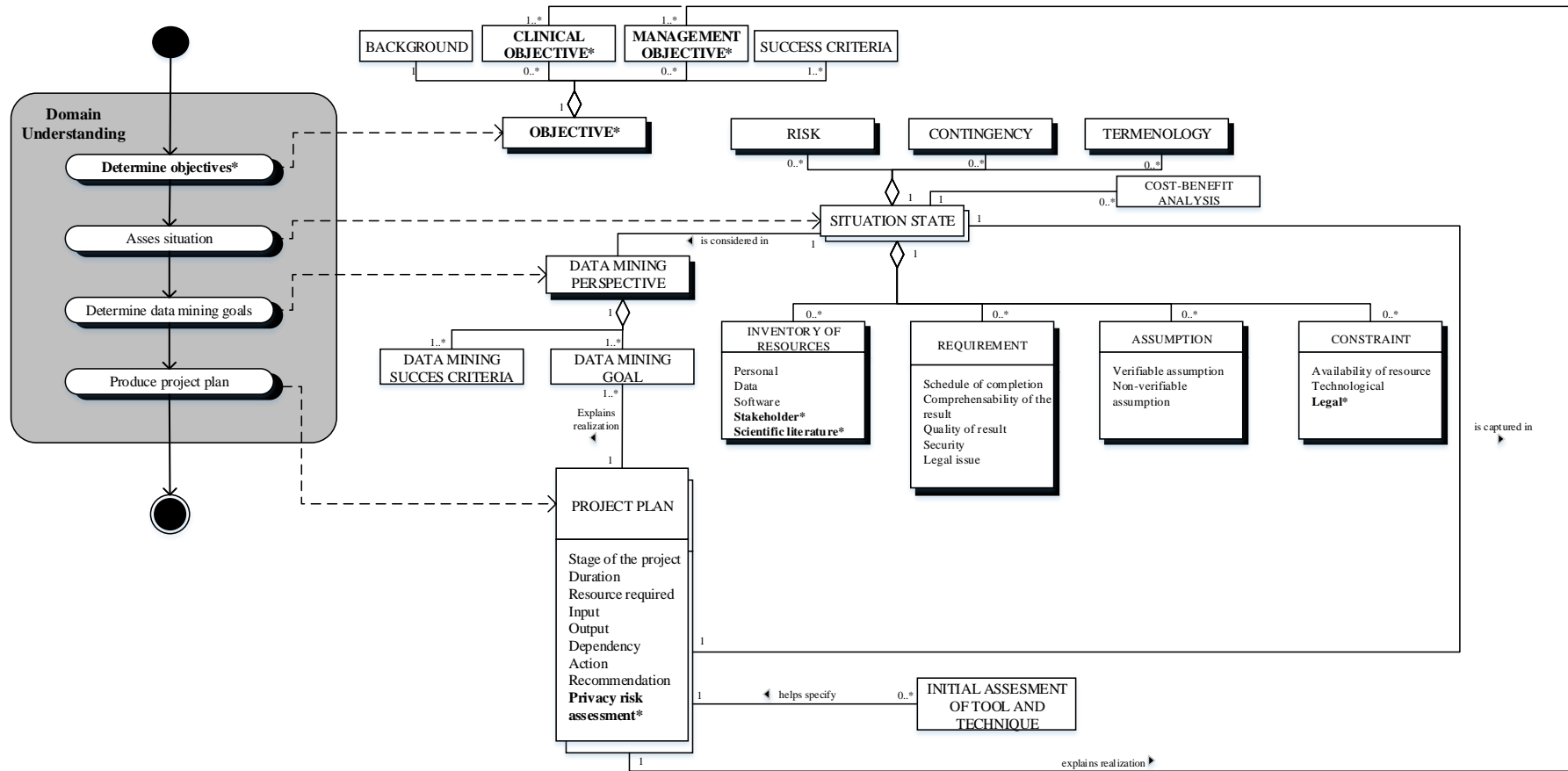


Figure 6.1: Process-deliverable-diagram of the Domain Understanding

Table 6.3: Activity table for Domain Understanding

Activity	Sub-activity	Description
Domain Understanding	Determine objectives	It is important that the data analyst thoroughly understands what the stakeholders really want to accomplish from a clinical as well as managerial perspective. Thus, identifying the CLINICAL OBJECTIVE or MANAGEMENT OBJECTIVE is highly important as well as understanding the BACKGROUND and SUCCESS CRITERIA that needs to be determined.
	Assess situation	This activity involves more detailed fact-findings of the INVENTORY OF RESOURCE, ASSUMPTION, CONSTRAINT, REQUIREMENT and other factors such as the RISK, CONTINGENCY, TERMINOLOGY and COST-BENEFIT ANALYSIS, that need to be considered in shaping the data analysis goal and project plan within a specific SITUATION STATE.
	Determine data mining goals	Likewise, within the BUSINESS PERSPECTIVE the DATA MINING GOAL and DATA MINING SUCCESS CRITERIA are stated in more technical terms within a DATA MINING PERSPECTIVE.
	Produce project plan	The intended plan for achieving the DATA MINING GOAL and BUSINESS GOAL are described in a PROJECT PLAN. In addition, an INITIAL ASSESSMENT OF TOOLS AND TECHNIQUES are performed to help specify the project.

Table 6.4: Concept table for Domain Understanding

Concept	Description
OBJECTIVE	This entails a perspective from a clinical or management viewpoint on what needs to be achieved. This consist of the BACKGROUND information, CLINICAL OBJECTIVE, MANAGEMENT OBJECTIVE and SUCCESS CRITERIA.
BACKGROUND	Consist of information about the organization's situation at the beginning of the project (Chapman et al., 2000).
CLINICAL OBJECTIVE	A description about the clinical objectives or implications related to the project (Niakšu, 2015; interviewee A4 and A5).
MANAGEMENT OBJECTIVE	A description about the management objectives provided by the organization of the hospital that does not have clinical implications such as analyzing budgeting (Niakšu, 2015; interviewee A4 and A5).
SUCCESS CRITERIA	A description of the criteria of successful or useful outcomes to the project from a clinical or management point of view (Chapman et al., 2000).

SITUATION STATE	This provides an explanation of the state of the project that should be considered in defining the data analysis goal and project plan. It consists of INVENTORY OF RESOURCE, REQUIREMENT, ASSUMPTION, CONSTRAINT, RISK, CONTINGENCY, TERMINOLOGY and COST - BENEFITS ANALYSIS.
INVENTORY OF RESOURCE	A list of available resources to the project including personnel, data, and software (Chapman et al., 2000). Additionally, the stakeholders need to be identified and the scientific literature consulted related to the project.
REQUIREMENT	A list of all requirements of the projects that includes schedule of completion, comprehensibility and quality of results, security and legal issues (Chapman et al., 2000).
ASSUMPTION	A list of assumptions that is made by the project (Chapman et al., 2000). This may include verifiable assumptions and non-verifiable assumptions.
CONSTRAINT	A list of constraints on the project (Chapman et al., 2000). These may be on the availability of resources or technological constraints. Additionally, there are legal constraints related to datasets that can be used or other datasets that will be needed.
RISK	A list of the risks or events that may influence the project in a negative way such as delay or cause it to fail (Chapman et al., 2000).
CONTINGENCY	A list of corresponding contingency plans, on how to react if such risks or events take place (Chapman et al., 2000).
TERMINOLOGY	This is a compilation of glossary that is relevant to the project. This may consist of glossary that is relevant to business or data mining terminology (Chapman et al., 2000).
COST AND BENEFIT	This is a COST-BENEFIT ANALYSIS that compares the costs of the project with the possible benefits to the business if it succeeds (Chapman et al., 2000).
DATA MINING PERSPECTIVE	This entails a perspective from technical terms in data mining that consist of a DATA MINING GOAL and DATA MINING SUCCESS CRITERIA.
DATA MINING GOAL	A description of intended outputs of the projects in order to achieve the business objectives (Chapman et al., 2000).
DATA MINING SUCCESS CRITERIA	A description of the criteria of successful outcomes to the project in technical terms (Chapman et al., 2000).
PROJECT PLAN	A list of stages that need to be executed in the project. This includes the duration, resources required, inputs, outputs, dependencies, actions, and recommendations. In addition, within the project plan each phases are discussed in detail and what evaluation strategy will be used in the evaluation phase (Chapman et al., 2000). Moreover, a privacy risk assessment will be considered to be documented in regards of the use of patient data.
INITIAL ASSESSMENT OF TOOLS AND TECHNIQUES	An initial assessment of the tools and techniques available should be performed that will be used during the different phases of the process. (Chapman et al., 2000).

6.2.2 Data Understanding

In chapter 4, the obstacle of data integration was mentioned where data is scattered across various systems and frameworks causing difficulty in gathering the desired data quality. Resolving these issues is only possible with the cooperation of the management of the hospital in improving the data infrastructure and implementing standards for collecting data. However, this research is restricted in the process of undertaking a DM project in the healthcare and this may differ from hospital to hospital, in which the data infrastructures are better accessible and data quality better managed. However, it is important for researchers to be aware of how data infrastructure is ordered within the hospital and how data can be obtained and is structured.

Moreover, in the “Data Understanding” phase a new general task “Assess data lineage” is introduced. Indicated by the interviewees A3, A4 and A5 that assessing the data lineage or tracing the data to its origin may provide better understanding on how the data was collected, registered and what the various inputs mean. Herein, the source of the data needs to be identified and those who are responsible in monitoring and collecting this data to provide explanations when needed, i.e. labelling of parameters or how certain inputs are calculated. Additionally, in section 5.2 of the case study, it was indicated that having sessions with practitioners that could explain the data and its characteristics was crucial in comprehending the data in general. Therein, the data lineage was also discussed. Figure 6.2, depicts the changes made in the CRISP-DM including the activity Table 6.5 and concept Table 6.6.

Moreover, a PDD was created for handling legal issues in Figure 6.3 with its corresponding tables Table 6.7 and Table 6.8. This belongs to a specialized activity within the generic task of “Collect initial data” in the “Data Understanding” phase. Herein, the legal constraints as well as privacy sensitivity are considered during this generic task.

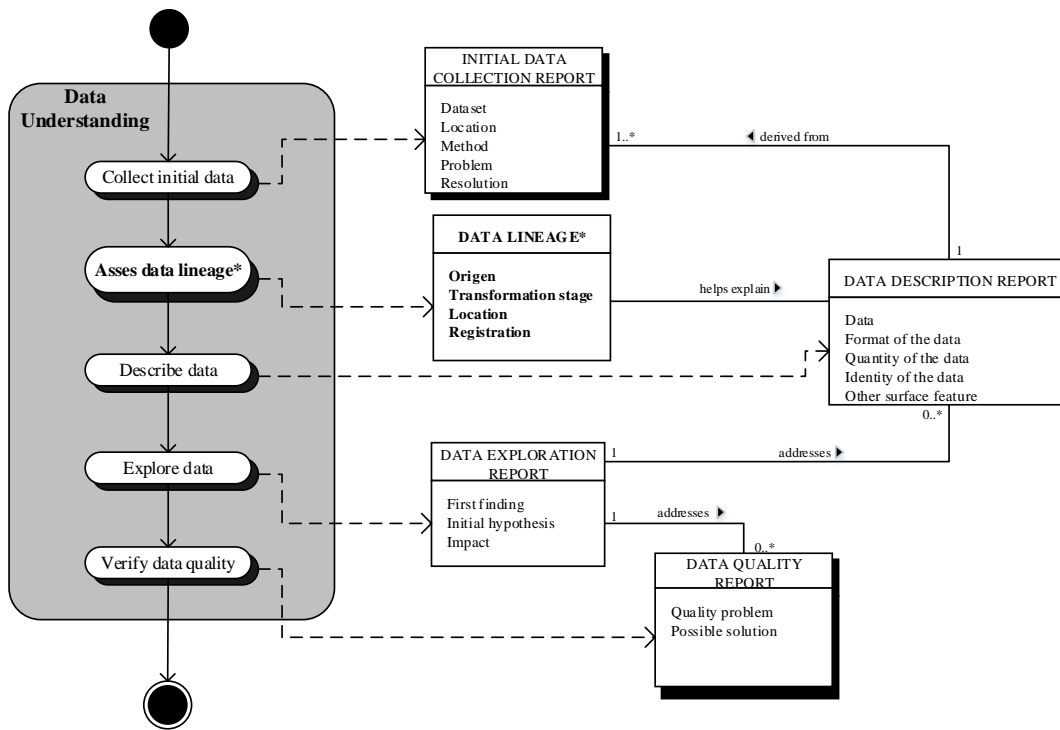


Figure 6.2: Process-deliverable-diagram of the Data Understanding

Table 6.5: Activity table for Data Understanding

Activity	Sub-activity	Description
Data Understanding	Collect initial data	The data needs to be acquired from the project resources, which is listed in the PROJECT PLAN in order to create an INITIAL DATA COLLECTION REPORT. This includes data loading, if necessary for data understanding. In addition, it has a dataset, location, legal status, method, problems, and resolutions.
	Asses data lineage	This activity involves tracing the data source to its origin.
	Describe data	The gross or surface properties of the acquired data needs to be examined and the results need to be reported in a DATA DESCRIPTION REPORT. The data is evaluated if it satisfies the relevant REQUIREMENT.
	Explore data	The data mining questions are addressed in the DATA EXPLORATION REPORT by using visualization, querying and reporting techniques. This process may address the DATA DESCRIPTION REPORT, DATA QUALITY REPORT and the DATA MINING GOAL by contributing or redefining them.
	Verify data quality	The quality of the data needs to be examined by addressing questions such as: if the data is

		complete or correct and if there are errors and how common they are. These kinds of questions need to be addressed in the DATA QUALITY REPORT by verifying the data quality.
--	--	--

Table 6.6: Concept table for Data Understanding

Concept	Description
INITIAL DATA COLLECTION REPORT	A description of the various data that is used for the project identifying whether some attributes are more important than others (Chapman et al., 2000). Besides, it contains a list of the datasets, locations, methods and problems.
DATA LINEAGE	Detecting the transmission history of data across the multiple entities or sources starting from the origin also known as data provenance and source tracing (Backes, Grimm, & Kate, 2016). It has an origin source, transformation stages and registration behavior.
DATA DESCRIPTION REPORT	Herein the acquired data needs to be described including the format of the data, the quantity of data, the identities of the fields and other surface features, which are found (Chapman et al., 2000).
DATA EXPLORATION REPORT	The results of the exploration are described that includes the first findings or initial hypothesis and the impact on the project (Chapman et al., 2000). This report can be illustrated by graphs and plots that indicate the data characteristics or other interesting initial findings.
DATA QUALITY REPORT	A list of the results of the data quality verification. In addition, a list of quality problems and their possible solutions are provided (Chapman et al., 2000). Moreover, the four “Vs” of big data analytics should be taken into account when dealing with large amount of data (Raghupathi & Raghupathi, 2014).

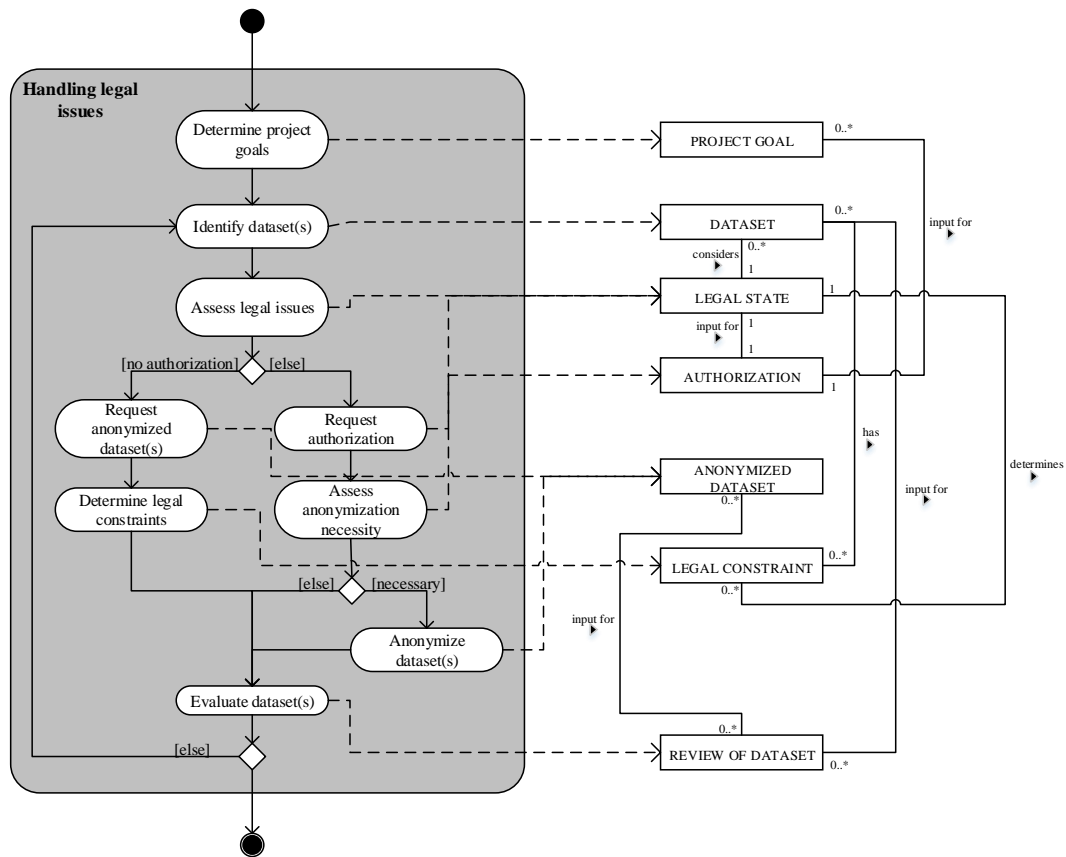


Figure 6.3: Process-deliverable-diagram of handling legal issues

Table 6.7: Activity table for handling legal issues

Activity	Sub-activity	Description
Handling legal issues	Determine project goals	The reason need to be defined why a particular dataset is required for the project in order to be classified within the conditions mentioned in section 4.5.1 to be able to use patient data.
	Identify dataset	The gross or surface properties of the data needs to be identified in respect to what is needed.
	Assess legal issues	This activity involves in assessing the LEGAL STATE in regard to accessibility of data and if authorization or anonymization is required before attaining a desired dataset.
	Request anonymized dataset	An ANONYMIZED DATASET is being requested.
	Request authorization	Requesting AUTHORIZATION in accessing the dataset.
	Determine legal constraints	Here, the LEGAL CONSTRAINT is described on what can be done with the dataset.
	Assess anonymization necessity	This activity involves in assessing the LEGAL STATE in regards depersonalizing or anonymizing the data.

	Anonymize dataset	The data needs to be anonymized that there is not trace to be found that can lead to the identity of the patient.
	Evaluate dataset	The dataset needs to be reviewed if it meets the desired conditions to be used. Otherwise, another data needs to be sought.

Table 6.8: Concept table for handling legal issues

Concept	Description
PROJECT GOAL	Here the goal is being defined in regard to what the purpose is of the project and the reasons why data is needed.
DATASET	Description of the required dataset.
LEGAL STATE	Assessing the legal environment of a particular data source and its constraints.
AUTHORIZATION	Permission is obtained of accessing and using a particular data source.
ANONYMIZED DATASET	A depersonalized dataset that cannot be traced back to identity of patients.
LEGAL CONSTRAINT	Restrictions set on the usage of the desired data.
REVIEW OF DATASET	The data is reviewed if its meets the conditions for data mining purposes.

6.2.3 Data Preparation

In the “Data Understanding” phase, a new general task “Anonymize dataset” has been added, because of the possibility that a researcher may access patient data with authorization but will work with other stakeholders or colleagues that do not have the permission in personal data of patients. Therefore, there needs to be a task after the dataset is been selected to anonymize it for further use if that is the case. This also occurred during the case study, in which the data was received anonymized from one of the researchers and colleagues who had access to it and only could share it with others until it was anonymized. Additionally, this was also supported by interviewee A1 and A3. Furthermore, it was indicated during the interviews with the experts A3, A4 and A5 that once the selection of the dataset took place they would start formatting the data instead of doing it at the end of this phase, due to convenient reasons and for better understanding of the dataset before the clearing and constructing tasks. The same applies after formatting that other data would be integrated with each other before cleaning starts. In addition, this order was also practice during the case study as indicated in section 5.3. In Figure 6.4, the changes of the CRISP-DM are illustrated with their corresponding activity Table 6.9 and concept Table 6.10.

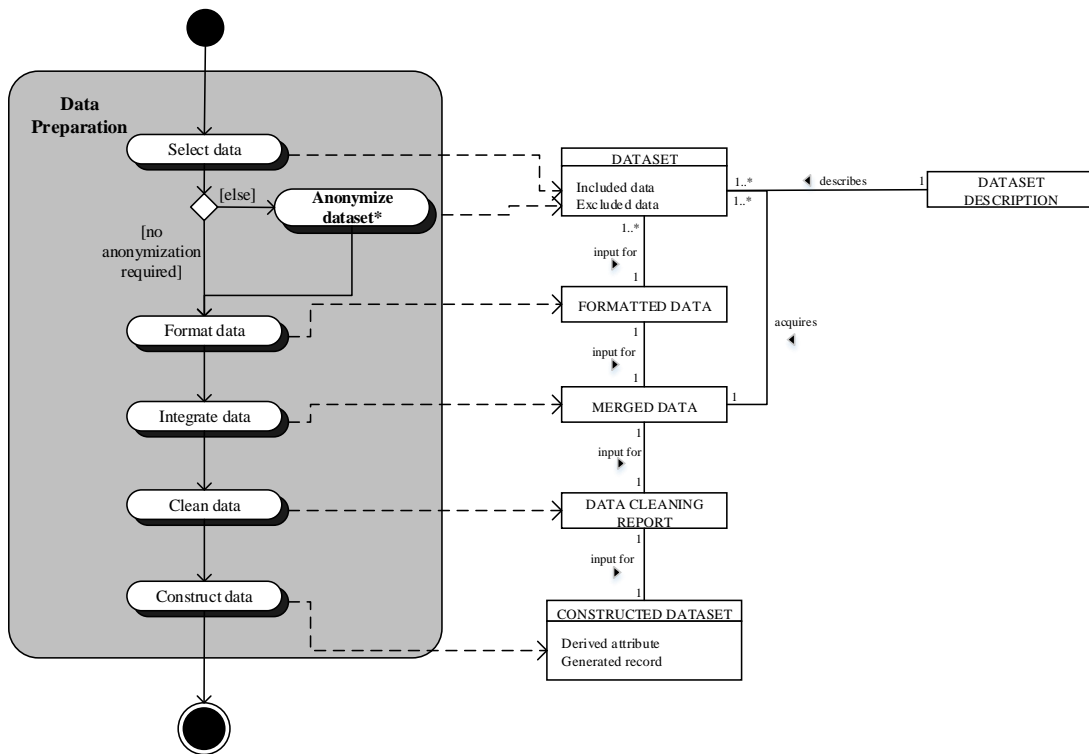


Figure 6.4: Process-deliverable-diagram of the Data Preparation

Table 6.9: Activity table for Data Preparation

Activity	Sub-activity	Description
Data Preparation	Select data	The DATASET that will be used for analysis is selected. In addition, the DATA DESCRIPTION helps to specify the DATASET that will be used for modeling and other analysis activities.
	Anonymize dataset	The activity of depersonalizing the dataset for sharing purposes with others involved with the project.
	Format data	The REFORMATTED DATA is syntactic modified that does not change the meaning in order to be used by the MODELING TECHNIQUE in the modeling phase.
	Integrate data	In the MERGED DATA, information is combined from multiple records or tables of the DATASET in order to create new values or records.
	Clean data	The data needs to be cleaned in the DATA CLEANING REPORT and raised to the level that is required by the selected analysis techniques. This can be done by inserting suitable defaults or selecting clean subsets of the data and by using other techniques that

		help in preparing the data for the modeling phase. Moreover, the data quality problems need to be handled from the DATA QUALITY REPORT, in which decisions and actions need to be taken.
	Construct data	This activity involves constructive data preparation operations such as creating entire new records in the GENERATED RECORD, or the production of DERIVED ATTRIBUTE or transformed values for existing attributes.

Table 6.10: Concept table for Data Preparation

Concept	Description
DATASET	These are datasets that are produced during the data preparation phase that are being prepared for modeling or other major analysis work (Chapman et al., 2000). In addition, a list of included and excluded data is provided with its reasons for these decisions.
DATASET DESCRIPTION	A description of the datasets that will be used for the modeling phase (Chapman et al., 2000).
FORMATTED DATA	The data is accustomed in accordance to the tools requirements that will be used in the modeling phase (Chapman et al., 2000).
MERGED DATA	The datasets are merged with other relevant data that has similar information about a particular object (Chapman et al., 2000).
DATA CLEANING REPORT	A description of the decisions and actions taken that address the data quality problems from the DATA QUALITY REPORT (Chapman et al., 2000). The datasets need to be cleaned from irrelevant fields that create noise in the data that could have an effect on the results.
CONSTRUCTED DATASET	The CONSTRUCTED DATASET contains the derived attributes that are constructed from one or more existing attributes in the identical record. In addition, the creation of new generated records in the datasets are described (Chapman et al., 2000).

6.2.4 Modeling

The activities and deliverables of the original CRISP-DM modeling phase are covering well in the medical domain that can be used for variety of other research and projects purposes. However, during the case study in section 5.4 and expert interviews (A1, A3, A4 and A5) it was emphasized to have regular meetings and assessments in this phase in regard to developing models with (clinical) practitioners to discuss if the models meet the desired requirements and if the correct parameters are used. Thus, a property is added in the “MODEL ASSESSMENT” deliverable that takes the practitioners opinion into consideration. In addition, as indicated in chapter 4 in respect to causal inference in observational datasets researchers need to be aware between the difference between correlation and causality in assessing a result or finding. In Figure 6.5, the changes of the CRISP-DM are illustrated with their corresponding activity Table 6.11 and concept Table 6.12.

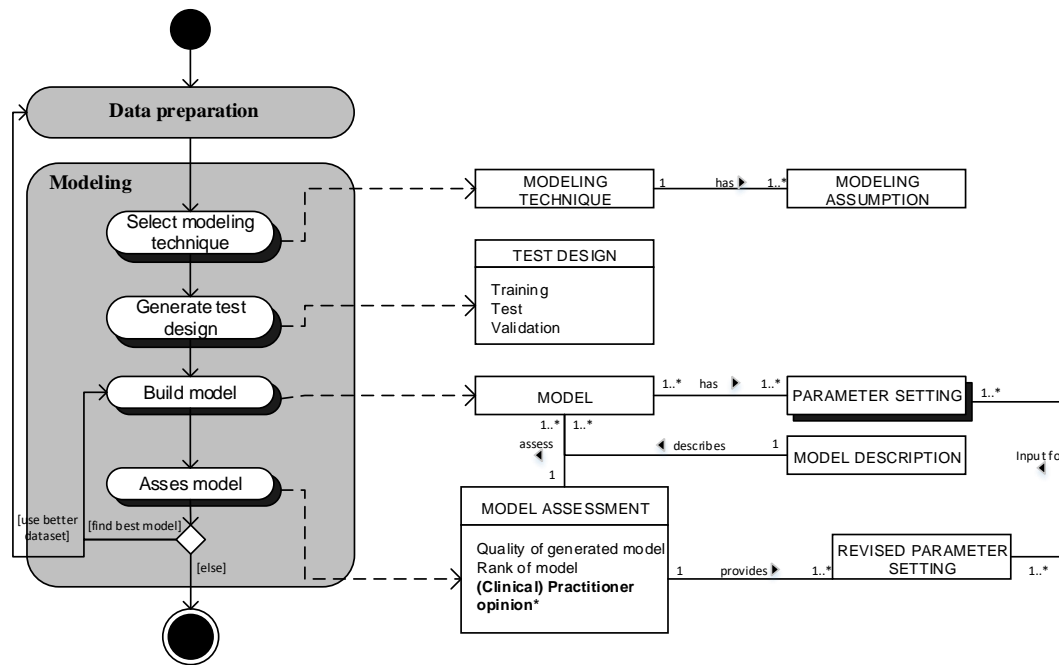


Figure 6.5: Process-deliverable-diagram of the Modeling

Table 6.11: Activity able for Modeling

Activity	Sub-activity	Description
Modeling	Select modeling techniques	The first step in this phase is the selection of the actual MODELING TECHNIQUE, which describes how MODEL needs to be build. This technique makes a MODELING ASSUMPTION about the suitability of the data that is going to be used for a certain MODELING TECHNIQUE.
	Generate test design	Before the actual model is build, a procedure or mechanism is generated in order to TEST DESIGN the model in regard to quality and validity. Herein, the dataset can be divided in three components to test the model: training, testing and validating.
	Build model	The modeling tool is applied to the prepared dataset in order to produce a MODEL. It includes the PARAMETER SETTING that are selected for the modeling tool and the results of the MODEL are described in the MODEL DESCRIPTION.
	Asses model	In the MODEL ASSESSMENT, the models need to be ranked on how they perform. The BUSINESS OBJECTIVE and the BUSINESS SUCCESS CRITERIA are taken into account as well as the DATA MINING SUCCESS CRITERIA and the results of the TEST DESIGN during the assessment. Moreover, the

		parameters that are used can be revised in the REVISED PARAMETER SETTING and iterated between the modeling building and assessment until the best model is found.
--	--	---

Table 6.12: Concept table for Modeling

Concept	Description
MODELING TECHNIQUE	This is the selection of the actual modeling technique that has to be used for modeling.(Chapman et al., 2000).
MODELING ASSUMPTION	A list of assumptions that are made for the modeling technique about the data.(Chapman et al., 2000).
TEST DESIGN	This describes the intended plan for testing, evaluating and training the models (Chapman et al., 2000).
MODEL	These are the actual models that are produced by the MODELING TECHNIQUE (Chapman et al., 2000).
PARAMETER SETTING	A list of parameters and their chosen values that can be adjusted by the modeling tool (Chapman et al., 2000).
MODEL DESCRIPTION	Here the models are described and assessed with their expected robustness, accuracy, and possible shortcomings (Chapman et al., 2000).
MODEL ASSESSMENT	The outcomes are summarized including with a list of qualities of the generated models and their quality rank in relation to each other (Chapman et al., 2000). In addition, (clinical) practitioners' feedback is sought in assessing the models and if revision of the parameters is required.
REVISED PARAMETER SETTING	The PARAMETER SETTING is revised in accordance to the MODEL ASSESSMENT. Herein, the iteration with the model building and assessment is performed until the best model is found (Chapman et al., 2000).

6.2.5 Evaluation

In chapter 4 in regards to causal inference in observational datasets, it was recommended that the outcomes from the observational studies should not influence clinical practice until the related hypotheses are tested in a passably randomized controlled trials (Tai et al., 2014). Hence, a new general task “Produce pilot model” was introduced in the “Evaluation” phase with a new deliverable “PILOT MODEL” in order to test the model in the appropriate setting before implementing the model. In addition, this was highlighted by interviewee A4 & A5 and mentioned in section 5.5 in which similarly a model was tested on a larger dataset once the tests on the smaller datasets were positive.

Moreover, a condition has been added between the task “Evaluate result” and the “Review process”, wherein there could be three possibilities in moving forward: modeling being approved, adjusted or another model needs to be made from scratch again. After evaluating the results, the model can be adjusted due to the feedback of experts and needs to return to the “Modeling” phase. It is also possible that the

model can be rejected because it does not perform well or there is a better model available and therefore needs to return to the “Data Preparation” phase in order to prepare the dataset for another model. This happened in the case study as described in section 5.5 and was highlighted by interviewee A1, A2, and A5.

Moreover, a new general task “Produce final report” with its corresponding deliverables “FINAL REPORT” and “FINAL PRESENTATION” is added. It was indicated during the interview with A5 that this task which was originally located in the “Deployment” phase should be moved to the “Evaluation” phase because to get a project implemented, the stakeholders need to agree on it and this can be done by making a report of the project and presenting it. In Figure 6.6, the changes of the CRISP-DM are illustrated with their corresponding activity Table 6.13 and concept Table 6.14.

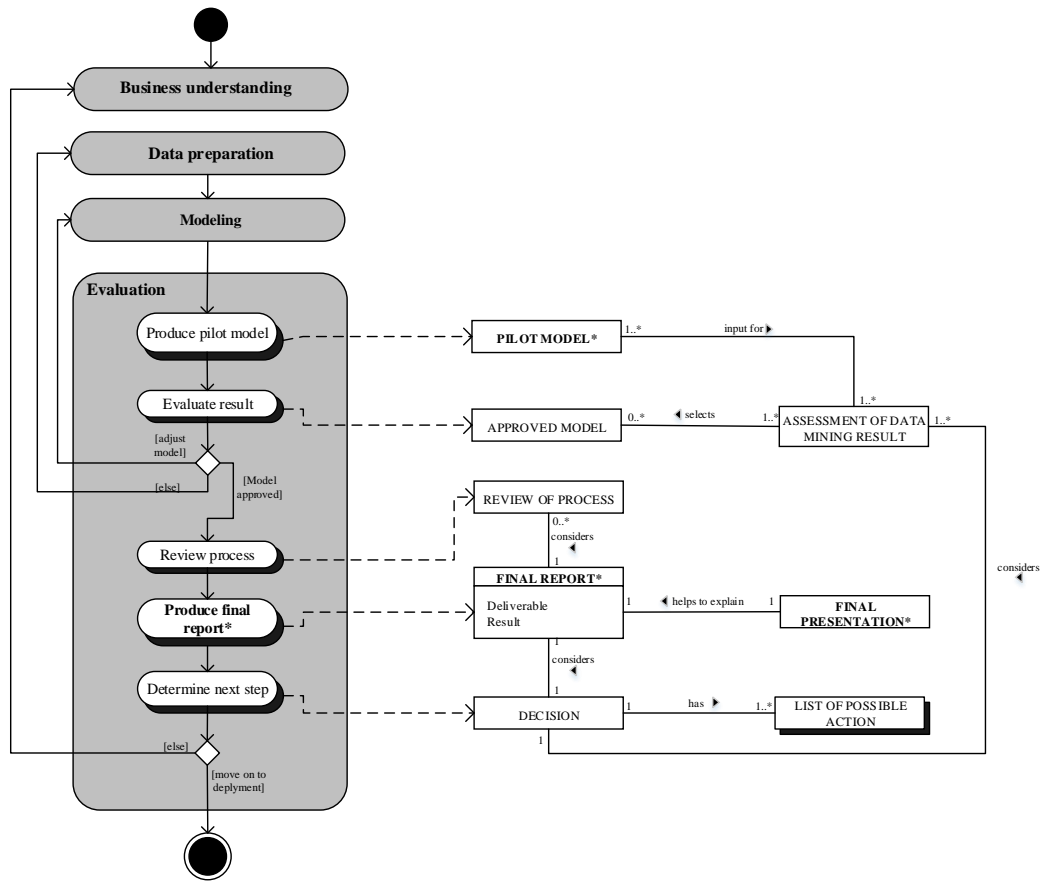


Figure 6.6: Process-deliverable-diagram of the Evaluation

Table 6.13: Activity table for Modeling

Activity	Sub-activity	Description
Evaluation	Produce pilot model	In this activity a test environment is created in order to test the model in a certain required setting.

	Evaluation results	The ASSESSMENT OF DATA MINING RESULT is evaluated to what extent a particular model meets the SUCCESS CRITERIA and CLINICAL or MANAGEMENT OBJECTIVE and if there is a reason, why this model might be deficient. After this assessment an APPROVED MODEL is selected that meets the needed requirements. In addition, the clinical and data science experts are consulted in which feedback is provided in regard to the findings from the model and technical aspect that can result in adjusting the model or choosing a different one. This was highly emphasized by all interviewees (A1, A2, A3, A4, and A5)
	Review results	A thorough REVIEW OF PROCESS is performed of the data mining engagement in order to check if an important task or factor has somehow been overlooked.
	Determine next step	This activity depends on the results of the ASSESSMENT OF THE RESULT and REVIEW OF PROCESS in order to make a DECISION on how to proceed. The team can decide to move on to the next phase or iterate between the phase again or quit and set up a new data mining project according to the LIST OF POSSIBLE ACTION

Table 6.14: Concept table for Evaluation

Concept	Description
PILOT MODEL	A setting in which the proposed models are tested.
APPROVED MODEL	The models that meets the SUCCESS CRITERIA are selected and approved (Chapman et al., 2000).
ASSESSMENT OF DATA MINING RESULT	The assessment results are summarized in terms of the SUCCESS CRITERIA and whether the project already meets the original CLINICAL or MANAGEMENT OBJECTIVE (Chapman et al., 2000). In addition, the feedback of the clinical and data science experts is consulted in assessing the model.
REVIEW OF PROCESS	The process review is summarized and the activities that have been missed and those that should be repeated are highlighted (Chapman et al., 2000).
DECISION	The decisions that are made are described along with the rational for them (Chapman et al., 2000).
LIST OF POSSIBLE ACCTION	A list of potential further actions with the reasons behind it is provided for each option (Chapman et al., 2000).

6.1.6 Deployment

In the “Deployment” phase the task “Produce Final” with its deliverables has been removed and moved to the “Evaluation” phase because as explained in the previous section that it fits better in the “Evaluation” phase.

Moreover, a new general task “Review with end user” has been added as it was indicated in the interview with A3 and A5 because this enables once the model is implemented to receive the feedback of its end users which can provide further insight for new related DM projects or adjustment of the current one by looking at a newer perspective. Hence, the possibility is provided once new requirements are acquired from the end user to return to the “Domain Understanding” and iterate with the whole process if needed. In Figure 6.7 the changes of the CRISP-DM are illustrated with their corresponding activity Table 6.15 and concept Table 6.16.

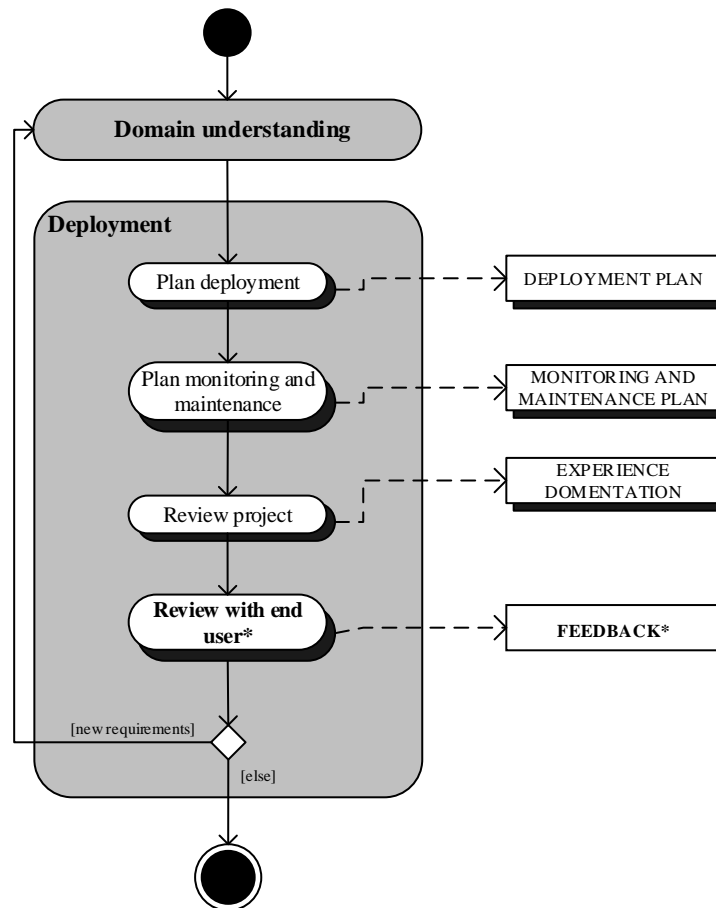


Figure 6.7: Process-deliverable-diagram of the Deployment

Table 6.15: Activity table for Deployment

Activity	Sub-activity	Description
Deployment	Plan deployment	The evaluation results are taken and the strategy for deployment is determined. This procedure is documented in the DEPLOYMENT PLAN and can be used for later deployment.
	Plan monitoring and maintenance	If the data mining results become a part of the day-to-day business and its environment a detailed MONITORING AND MAINTENANCE PLAN needs to be constructed. Therein, a maintenance strategy is developed in order to avoid incorrect usage of the data mining results and the specific type of the deployment are taken into account.
	Review project	A final assessment is performed in an EXPERIENCE DOCUMENTATION report in what went well or what went wrong during the project and what needs to be improved in the future.
	Review with end user	Once the modeling is running in an actual environment, the end users will interact with the new system or model. Their interaction can provide great value and therefore feedback should be sought in regard to the interaction with the new setting and insight for further developed requirements.

Table 6.16: Concept table for Deployment

Concept	Description
DEPLOYMENT PLAN	The deployment strategy is defined including the necessary steps and how to perform them (Chapman et al., 2000).
MONITORING AND MAINTENANCE PLAN	The monitoring and maintenance strategy is summarized and how to perform the required steps (Chapman et al., 2000).
EXPERIENCE DOCUMENTATION	This is a summary of the important gained experience during the project (Chapman et al., 2000).
FEEDBACK	Herein, feedback is provided from the end users of the newly implemented system in which new insights and requirements can be developed.

Chapter 7: Method evaluation

During the course of the design cycle and research, two projects were provided by the WKZ that made it possible to experience the medical domain in respect to DM projects and to acquire relevant method fragments. There have been frequent meetings with experts during this course of the research in discussing the case studies, as well as finding new method fragments that can be used. In addition, interviews were conducted with experts in dissecting the MSP-DM and the missing elements were identified that are applicable within the medicine domain in respect to DM projects. These expert interviews served as a mean of evaluating the extracted method fragments from the case study and the research performed. In addition, overlooked method fragments were indicated and later refined in the MSP-DM as shown in chapter 6.

7.1 Qualitative: Expert Interviews

At the ending of the research, expert interviews were conducted with 5 domain experts from the Universitair Medisch Centrum (UMC) and Utrecht University (UU). In Table 2.1 of chapter 2 an overview was provided of the participants of the interviews that operate in a specific domain and organization. Moreover, the interviews followed a particular protocol that covered the following topics:

- Introduction of the research: a brief introduction was provided of the research, explaining its importance and goals.
- Background of the interviewee: the background of the interviewee was inquired and their experience in data mining projects was documented. Likewise, questions were asked in respect to how data mining projects were conducted, as well as which challenges they faced and how they overcame it.
- Evaluation of the process method: the MSP-DM was initially discussed on how it can be improved and feedback was gained on missed method fragments that were overlooked during this research. In addition, findings from this research and case study were discussed.

Moreover, in these interviews the MSP-DM was discussed in respect to its generic activities and deliverables. Herein, the missing elements were pinpointed and noted. The experts provided their expertise and experience in proposing new method fragments or by adjusting the current elements within the method. This enabled the modification of the current CRISP-DM to a more tailored process method (MSP-DM) for in the healthcare. In addition, there were two interviews conducted with A4 in which the latter one was in assessing the tailored method for the second time with the final adjustments made. The interviewee A4, agreed in all changes that were made with no further comments. Moreover, interviewee A2 was only interviewed on the fragments related to the domain, evaluation and deployment phase because A2 was not adequate in the technical aspect of DM. In addition, with A1 some background questions were not asked because of the lack of time for the interview

which more emphasis was given to DM activities or processes and evaluating the fragments. Nevertheless, A1 was a colleague during the case study, and therefore the background information was already known of this participant. In Table 7.1 the method fragments acquired from these interviews are shown and in the appendix D. Expert interviews the transcriptions can be reviewed.

Table 7.1: Results of the expert interviews

RESPONDENT	QUESTIONS	SUMMARY	COMMENT	CORRESPONDING INTERVIEW OUTCOME	SUPPORT
Business Understanding					
A4	Remark from interviewee*	It is for the healthcare domain more suitable to change the first phase to Domain Understanding because of its clinical nature.	“perhaps change the name of the phase to domain understanding which fits the healthcare better because it also has a clinical perspective and provides a broader view within analyzing a domain.”	Rename Business Understanding to Domain Understanding	
A1	I have found in the literature that there could be two separate objectives in a data mining project in the healthcare; clinical and managerial objectives. Do you agree with this?	The objectives in the healthcare can vary from clinical objectives or managerial objectives for data mining projects.	“Well firstly, a hospital is a company, we have clinical and business processes. The business process means that products need to be bought and archived. Information needs to be shared, communication and technology needs to be arranged. The clinical process contains medical decisions and patient care. These two generally overlap yet we separate them in our mind.”	There are clinical and Managerial objectives	
A2			“Yes, there could be different objectives for data mining projects such as those you mentioned but I am not that familiar with them. Nevertheless, I can imagine that there could be multiple various objectives in a data mining project besides the clinical aspect.”		
A3			“As you know, I have just started in doing some data mining projects and I am still figuring out how things should be done. However, I can imagine that there could be different kind of objectives but currently I am just familiar with the VDS project which has a clinical purpose.”		
A4			“there could be clinical as well as managerial objectives within a data mining project in the healthcare.”		
A5			“there are various types of projects in respect to data mining activities such as performing an analysis on administrative matters within the hospital, as well as clinical research.”		
A3	Remark from interviewee*	Identifying the right stakeholders is highly	“the collection of data in hospitals is quite challenging because the hospital is not set to this kind of analysis.”	Stakeholders needs to be identified	

		important in order to get the necessary means of what is needed during a data mining project in the healthcare.	<p>“Identifying stakeholders, understanding the culture of the hospital is really important in order to get things done. There have been some obstacles in attaining the required data for the project, wherein several request were done to the responsible for retrieving the data but it was very difficult in attaining it until I found the right person who provided me the necessary means to retrieve and get the data. There is a lot of politics within a hospital and it quite important to get familiar with it in order to get things done.”</p> <p>“The issues of collecting data was overcome by identifying the stakeholders and confronting them in order to be able to access the data.”</p>		
A1	I have found that there could be some legal constraints during a data mining project and I have added it in the concepts of constraints, what do you think about this?	Taking into account the legal constraints in working with colleagues that do not have the same authorization in using certain data, wherein anonymizing the data is required.	“There could be legal constraints although there is not a big issues here because frequently the data can be acquired with not that much problem.” (Did not show much concern or emphasis)	Legal constraints needs to be pin pointed	
A2			“There are not much legal issues or constraints that needs to be taken into account. I did not really encounter them.”		
A3			“As you know that during the case study there were some privacy issues in sharing the data with you because you did not have the authorization in overseeing patient data and therefore I had to anonymize the data for you in order to make it possible for you to access it and to use it as well.”		
A4			“Personally, we don’t have that much problem with privacy issues since most of the time our data acquired is already anonymized and we don’t use much poor raw data on which patient information is available. However, I can imagine that perhaps for other projects there could be some legal constraints.”		
A5			“Yes, there could be legal constraints although as earlier mentioned anonymized data can be easily requested here which makes it easier to avoid such constraints.”		
A4	Remark from interviewees*	It is required for getting a data mining project approved, a privacy risk assessment needs to be filled. This can be done for each project or there could be a general policy in handling sensitive data for each project	<p>“we have form called privacy risk assessment that is filled for the every project. However, because there can be multiple projects in a short run, we have created in our department a general privacy risk assessment that applies to other cases as well in which we use that as our form. Usually, this needs to be done for each project separately although practically this is not achievable. Therein, it explains how you handle the data and this needs to be approved.”</p> <p>“the privacy risk assessment is much work but it is important and</p>	Privacy Risk Assessment needs to be included	




		in which a general assessment is made.	should be more highlighted here perhaps in the assess situation.”		
A5			“The privacy risk assessment needs to be added here which is needed for starting a project. It is required for each project that uses patient data to fill in this assessment before getting approval in using certain data.”		
A1	I have found that scientific literature can benefit the projects, what is your opinion about this?	Scientific literature should be consulted in finding relevant cases that are done before and proven when assessing the situation of a data mining project.	“Yes, there is much written related to predictive modeling or other data mining topics. There can be matters that are related to a given data mining project. However, there is also much discussion what is the correct way to do things. So, you may find relevant papers but you will still need to do some research by yourself.”	Check the available scientific literature for a particular related case.	
A2			“Consulting scientific literature which needs to be done always when conducting a research by yourself. So, including this in getting a better understanding of the project can be helpful indeed.”		
A3			“There are multiple ways of finding certain information or answers that are applicable in your project. I think using scientific literature can be a part of this.”		
A4			“Yes, I agree that scientific literature is missing here which can be consulted during a project.”		
A5			“I miss here [in the generic activity, assess situation] the scientific literature when assessing the situation, because it is highly possible that for certain cases it can be scientifically written or done before. Hence, consulting such literature can be beneficial when conducting a data mining project when assessing the situation.”		
Data Understanding					
A1	During my case study I have found that discussing the origin of the data with a domain expert to be very beneficial in understanding the data, what do you think about this and how this fits in this phase as an activity?	Assessing the data lineage can have various benefits in understanding the data that should be formed which is currently missing in the original CRISP-DM.	“When it comes to data understanding, we rarely do it, we usually just do a pilot or a sample size. We do this to see if there is an effect or to see the feasibility.” “The step verifying data quality is done during the process, the steps describe and explore data we do during the analysis. Data understanding in general is for export in a pilot or sample size, it’s not data driven. However, I can understand that for newcomers in this domain can be useful and following this structure.”	Assess the data lineage with a domain expert	
A3			“I think this helped you in better understanding the data, as well as the project overall. For people that are not familiar within the medicine domain I can see that this can be necessary.”		
A4			“knowing the process how data is stored and collected will help you in getting a better understanding of the data or finding interesting		

			insights that can help in the analysis. This can be in a form of retrieving the data lineage of a particular dataset.”		
A5			“the activity of assessing the data lineage needs to be added in order to get to know the data and its origin as well how the registrations were done for better insight.”		
Data Preparation					
A1	I have found that it is important to anonymize the data when working with others that do not have permission in accessing it, what do you think about this?	When working with others that are not allowed in accessing a particular data unless it is depersonalized, the data then needs to be anonymized in order to enable cooperation.	“This is true, like in your case. You received the data from A3 anonymized because you did not have access to personal data from patients.”	Anonymize dataset	
A3			“it is important when working with others that don’t have access to patients’ information to see what the privacy issues are or ethical issues in order to depersonalize the data for further use with others.”		
A4			“Yes, if you do not have permission in accessing the data you can only retrieve it if its anonymized. However, as mentioned earlier we do not encounter such issues since such data is most of the time already anonymized.”		
A5			“Here we really do not encounter it, although yes if you have no authorization in accessing directly patient data then indeed it needs to be anonymized first upon receiving it.”		
A1	I have noticed that the order of preparing the dataset for modeling is a bit different, in which you first select, format, integrate, clean, and then construct the data. Is this correct?	During the data preparation phase, a different order is followed, wherein the data is selected first, then formatted, integrated, cleaned and last constructed.	“In the data preparation phase, is mostly collecting data and finding the right data. There is not a particular structure to follow. This mainly depends what kind of order you prefer.”	Different ordering of the data preparation	
A3			“As you know I am still learning how data mining should be done and currently I am learning from mistakes and seeing how others do it as well as following courses online. So, I am not sure what order is the best to do, although I have followed the same order you described until now and it works fine with me.”		
A4			“The order is first selecting then formatting, integrating, cleaning and then constructing.”		
A5			“The order is a bit different here compared to the CRISP -DM. Here, we start with select data, format data, integrate data, clean data and then construct data.”		
Modeling					
A1	Do you find it useful in involving (clinical) practitioners during the model assessment?	Involving clinical experts in the model assessment is recommended in attaining a better understanding of the parameters	“It depends on your own expertise of the domain. I can imagine after working for a long time in a certain domain setting that you will be able to understand the most of the domain related data. However, checking your findings or decisions with a more experienced expert will not be a harm. I think this will apply more for newcomers in the	Considering the (Clinical) practitioners opinion during model assessment	

		selected and refining it accordingly.	medicine domain than experienced researchers.”		
A3			“Yes, as you know I am a clinical practitioner and do have a good understanding of the medicine domain. However, I as you have seen I still need some help from others in understanding what I am exactly doing and if it’s right. So, consulting with other experts during this process is import if you do not have that much experience or domain knowledge.”		
A4			“In the model assessment, the clinical experts should be involved in evaluating the selected parameters or doing it together. This can be beneficial because those clinical experts have a better understanding of these parameters and its relations and therefore can assist in selecting or revising the parameters that will be used for the model.”		
A5			“involve experts in checking what the important factors are or variables within a model. This is important because they have greater insight within the domain which can help in identifying the right parameters.”		
Evaluation					
A1	Why is it important to involve both the health care professionals and other experts in reviewing the results?	Reviewing the results of the data mining project with various other domain experts can be very beneficial in attaining further insights which may lead to improvement of the model.	“When evaluating the models, there should be other (clinical) experts involved in this process because of their insight that can benefit in finding mistakes or even improving the model because of the suggestions they can make.”	Involve health care professionals and other experts in the review of the results	
A2			“It’s important to me that it’s physiological correct, otherwise I wouldn’t trust it. By simulating it, it’s possible to control whether the data is correct. This is only possible when you involve the practitioners when evaluating the model in order to find the physiological correctness or in convincing the practitioners of the model that would ease the adoption and trust of the model.”		
A3			“In order to overcome some difficulties in implementation of analysis of models, it was mainly done by consulting with experts that can explain or help in a specific situation and also making goals more concrete and tangible.”		
A4			“Assessing with various experts is needed be it with clinical or technical experts or other data scientists if the model is suited for a particular case.”		
A5			“Assessing the data mining results should be done by involving experts as well clinical practitioners in evaluating the model in order to receive feedback that can help in improving the model or finding faults.”		

A1	Do you think that after evaluating the results that there are other possible actions such as moving back to the preparation or modeling phase?	After evaluating the results of the data mining project with other experts can lead that the results can be accepted or that the model needs to be adjusted or a different model needs to be selected.	“Yes, there could be other possible action because during the assessment of the results you would probably get some feedback which can influence your progression. So, those suggestion can be possible.”	Possible other actions after evaluating the results: moving back to data preparation, modeling or approving the model	
A2			“Yes, during the evaluation of the results the actions can be discussed with others and this may entail the actions you suggested.”		
A3			“I think that those are good possibilities, although I am not sure about it because I do not have that much experience in other data mining projects.”		
A4			“Yes, those actions can be taken if needed.”		
A5			“Yes, there could be more actions after evaluating the model. And I agree that you can move back to the modeling phase or to the data preparation phase if other models needs to be constructed after the feedback or adjusted or the model can be accepted.”		
A5	Remark from interviewee*	Once the model is accepted, the results needs to be presented to other stakeholders that can use the findings or in case for implementation desires.	“The activity of produce final report of the deployment phase fits better in this [evaluation] phase than in the deployment phase because during a the deployment in which the results are already implemented it is not necessary to make a final report. However, doing it before it is more appropriate after reviewing the process.”	Present the results to stakeholders and write a report/paper	
A1	I have found that before evaluating the whole results, to use a pilot to test the model on a smaller scale or environment to measure its effectiveness, what do you think about this?	Testing you developed model in a certain environment before evaluating the results can provide you some initial understanding of the effectiveness or usefulness of the model.	“As mentioned before, we do skip some steps and go right to pilot or sample size to see if there is an effect or feasibility. So, it can be placed before evaluating the results.”	Produce pilot model	
A2			“I don’t know what the standard procedure is during this phase but this seems logical.”		
A3			“This should be a good idea and probably wise. However, I am not sure if that’s the standard procedure here.”		
A4			“This is probably correct. Testing the model on a certain setting seems logical. In our cases, we do not have really implementation assignments or projects but are more exploratory driven. So, we do not test our models in certain environments but use only the data that has been provided or retrieved.”		
A5			“Before evaluating the model, there should be a setting in which the model is tested. And as you mentioned this can be similar to a pilot setting in a particular environment. This is recommended because this will provide you some initial understanding in regards to the effectiveness or usefulness of the model.”		
Deployment					

A5	Remark from interviewee*	Moving the activity of producing a final report to the evaluation phase, as earlier mentioned.	“The activity of produce final report of the deployment phase fits better in this [evaluation] phase than in the deployment phase because during a the deployment in which the results are already implemented it is not necessary to make a final report. However, doing it before it is more appropriate after reviewing the process.”	Delete produce final report	
A5	Remark from interviewee*	After the implementation of a data mining project in a real life scenario it is important to have reviews with end users depending on how frequent the results are used during a day-to-day routine.	“at the end of this phase reviewing with the end users is missing. It is important to get their feedback after the results are implemented. This feedback can be asked every month or half year or week. This depends of course how often the results of the data mining projects are used by the end users. “	Review with end user’s	
A5	Remark from interviewees*	The feedback received from the end users can lead to changes or adjustments of the first phase due to the newly acquired insights from the feedback.	“This could mean that the feedback received can mean that more insight is gained about the initial problem of the data mining project which could mean that from the deployment to the business understanding an action can be performed in which changes are necessary in this phase.”	A follow-up action if feedback is related to the domain understanding	
A3			“if the success rate of the implementation is not high then it should be returned back to the first phase. Hence there should be are link between deployment and business understanding.”		

-  High support
-  Neutral
-  No support
- *Remarks**

Lastly, these results were incorporated in Table 6.1 of chapter 6 in which the MSP-DM is based upon. Moreover, the acquired method fragments of the case study and from the research were discussed with domain experts during the evaluation, wherein the proposed suggestions or changes were accepted and supported. These fragments are indicated in Table 6.1 of chapter 6, wherein multiple sources were provided for particular method fragments supported with the opinion of the interviewees. The opinion of the interviewees can be found in Table 7.1. Moreover, remarks and additions from the interviewees about the method can be found under remarks, and these were also incorporated in chapter 6. In addition, all the proposed method fragments were supported and none were rejected, although some interviewees have neutrally supported some fragments with less provided importances than others.

Furthermore, other method fragments were considered yet later removed because it was discovered that they were already in the original CRISP-DM as a specialized

activity or output, i.e. the aspect of funding or legal issues which were already mentioned but not emphasized in the original method.

Chapter 8: Conclusion & Discussion

In this chapter the results will be summarized and the conclusion to the main research question presented:

RQ: How to develop a standard-based and enhanced data mining process method for researchers within the medicine domain to better guide them in the process of data mining projects considering the domain's specific challenges and unique characteristics?

To provide an appropriate answer to the main research question, the conclusions of a number of sub-questions are discussed first that guided the research in order to conclude with the main research question.

8.1 Conclusion of the sub-questions

The first sub-question is based on a theoretical foundation, which is derived from the literature in respect to the various existing data mining process methods. Sub-question two is based on the findings from the literature, case study and experts' interviews that explain the main concepts and activities of data mining in the healthcare. The final sub-question, is grounded on the theory of Meta-Algorithmic-Modeling that enables the realization of the main deliverable of this thesis: a standardized data mining process method in the healthcare (see chapter 6).

Sub-question one:

RSQ1: What are the existing data mining process methods for guiding data mining projects in the healthcare in order to support the development of a standardized process method for the healthcare?

The top three selected data mining process methods were reviewed and explored on the basis of applicability in various industries, comprehensibility, relevance in data mining, and popularity. The research revealed that the CRISP-DM process method is the most suitable and extended process method to support a standardized method for the healthcare. Hence, the CRISP-DM was selected to be used as the laying foundation on which can be built upon and modification can be made.

Sub-question two:

RSQ2: What are the main concepts and activities involved in medical data mining for constructing a standardized process method for the healthcare?

The main concepts and activities related to data mining in the medicine domain, as well as overcoming the unique challenges in the healthcare were explored. In this study, relevant method fragments were discovered that were missing in the original CRISP-DM as illustrated in section 6.1. Moreover, the findings presented that a great emphasis was indicated in involving clinical professionals, as well as (domain) experts during the data mining project because of their (hidden) insight and required support in successfully finishing projects. Likewise, practical activities and concepts were found that were missing in the original method. Hence, modifications within the CRISP-DM were performed that accommodate such settings for the creation of the MSP-DM.

Sub-question three:

RSQO3: How can the main concepts and activities be modelled into a standardized method for DM projects in the healthcare?

The MSP-DM was modeled by making use of Meta-Algorithmic-Modeling (Spruit & Jagesar, 2016; Spruit & Lytras, 2018). This is an engineering discipline that outlines the activities and concepts with their connected relations between each other. In chapter 6, the results of the MSP-DM of this research project are shown. Therein, the method is structured similar to the original CRISP-DM, wherein the extracted method fragments are designated to one of the following phases: domain understanding, data understanding, data preparation, modeling, evaluation, and deployment. Herein, a visualization and description of the activities and concepts is provided that explain the method in more detail.

8.2 Conclusion of the main research question

The sub-questions guided this research in answering the main research question.

How to develop a standard-based and enhanced data mining process method for researchers within the medicine domain to better guide them in the process of data mining projects considering the domain's specific challenges and unique characteristics?

This thesis project was centered in developing a standardized data mining process method for medical and IT researchers within the medicine domain, by considering various unique characteristics and challenges within this field that can generate insight and suitable (predictive) models to be deployed in real-life scenarios. However, before starting this endeavor it was required to find a suitable base structure or framework that contains the necessary foundations that would enable tailoring a method for its specific requirements. This structure was found in the CRISP-DM process method that presents a comprehensive foundation that can be used for the development of a standardized domain-specific method. Subsequently, it was essential to discover domain-specific method fragments that can be assembled with the original method and be modified. Those fragments were found in the literature, case study, and from domain expert interviews that revealed data mining activities and deliverables within the medicine domain. The construction of the MSP-DM is expressed with the use of a Meta-Algorithmic-Modeling. Finally, the MSP-DM was evaluated through expert interviews. In conclusion, the insight gained from answering the sub-questions facilitated the development of a domain-specific standardized process method in guiding data mining projects.

8.3 Discussion

The intention of this research is to provide some standardization to data mining activities in the healthcare with the added method fragments and modifications. However, more research lays ahead from the perspective of improving the proposed method, along with the VDS initiative. Hence, in this section the limitations of this study are discussed along with the reflection of the whole process of the research with its contribution to science.

8.3.1 Limitations

There are several limitations encountered during the course of the research. First, the research of theoretical frameworks was restricted only to three data mining process methods, which opened doors in missing possible better methods that could have been used. For example, the Three-phases method (3PM) was excluded from the research wherein aspects such as data retrieval, data mining, and result implementations were examined (Vleugel, Spruit, & Daal, 2010). Therein, the aspects of deployment could have been used because as earlier indicated in chapter 5 of the case study, not all phases were performed such as the data collection, partly the evaluation phase, and the deployment phase. Nevertheless, the CRISP-DM is one of the most cited, used, and extensively documented methods available and makes a great candidate as a base method. However, methods such as the 3PM could have been useful in supporting or amplifying the missing activities that were not conducted in the research.

Hence, the second limitation was that not all research activities were performed as previously mentioned in the case study and two challenges were not researched. This could mean that some of the relevant method fragments were missed. However, the data was collected by one of the stakeholders involved with the case study, who provided the dataset and explained his experience how this was performed. Regarding the evaluation and deployment phase, this was complemented with expert interviews that were conducted in identifying the missing method fragments. Nevertheless, the components that were not researched or conducted during this study due to the time constraints, still can be researched in discovering potentially unexploited method fragments that can improve this proposed method further. In addition, the focus of this thesis was mainly on the generic tasks instead of focusing much on the specialized tasks within a generic task. For this reason, the proposed method is not made for a specific setting within the healthcare but in a more generalizable fashion, it is applicable as a base method which can be specialized according to the specific needs of a project in the healthcare.

Moreover, the case study was initially meant to evaluate the proposed method as a proof-of-concept, although due to some circumstances explained in chapter 5 this changed. Thus, the case study was mainly used in extracting new method fragments instead of validating it. This resulted that the validation of the method was limited with only expert interviews. This could have been strengthened if a comparative study was conducted with other methods or comparing the original CRISP-DM with

this method during another DM project or if the initial case study was used to test this proposed method. However, due to the time constraint this was not possible.

8.3.2 Reflection

During the course of this thesis, field notes were written by hand on lined paper or noted on my mobile phone. These notes consisted of personal reminders of specific things or activities during my projects, as well as insights gained from the domain experts. This was relevant in understanding the culture and working environment. Likewise, this documented experience or insight helped in conducting my research, for example; the notes made during meetings helped in further researching the feedback provided on technical aspects or on researching other relevant issues. In addition, some meetings that I considered to be important were recorded. However, looking back on transcribing the recordings or explaining the notes, these were postponed to a later time which was not wise. This turned out to be a time consuming task at the final stage of the thesis.

Moreover, another advantage was working together with others on the case study projects. Especially, when time and resources were limited, good teamwork was the key to success. I was lucky having helpful colleagues with different ideas and own points of view that enabled me in broadening my understanding of the medicine domain, as well as in DM. Hence, the weekly and scheduled meetings helped in keeping my deadlines, as well as attaining a good understanding of the provided projects.

However, the case study did not initially go according to plan as mentioned before. This happened because much time was spent on the first VDS1 project which was originally considered to be a pilot or testing project to experience DM in the medicine and turned later to be more than that. Additionally, there were influential aspects that limited my options to move around because there was a deadline provided concerning the VDS1 project which caused me to focus more on the practical side of finishing the project instead of developing a method beforehand. Another issue was that in the initial phase of the thesis in the first two months, the subject and goals of this research were still unclear during the VDS1 project. This made it even more difficult once these were cleared up, because the focus was now on finishing the VDS1 project before developing a method. Later, the VDS2 was intended to be used as a proof-of-concept, but due to some data quality issues which were encountered later on along with time constraints, some activities and phases were not performed. This led to that the case study was not able to be used as an evaluation method but instead was used for constructing the proposed method.

Nevertheless, multiple method fragments were extracted from the research, case study, and expert interviews that were incorporated into a more standardized method. Hence, this proposed method could be considered to be an extension of the CRISP-DM that takes some current challenges into account, as well as being initially tailored for the healthcare but may comprise processes relevant to other domains. These extracted method fragments can also be taken as generic activities or concepts, which might be applicable in other industries such as the research domain in general. This can entail that the research space could be broadened to cover other cases outside the domain of healthcare.

References

- Altman, R. B., & Ashley, E. A. (2015). Using “Big Data” to Dissect Clinical Heterogeneity. *Circulation*, *131*(3), 232–233. <https://doi.org/10.1161/CIRCULATIONAHA.114.014106>
- 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>
- Backes, M., Grimm, N., & Kate, A. (2016). Data Lineage in Malicious Environments. *IEEE Transactions on Dependable and Secure Computing*, *13*(2), 178–191. <https://doi.org/10.1109/TDSC.2015.2399296>
- Behrns, K. E. (2015). Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System: Krumholz HM (Yale Univ School of Medicine, in New Haven, CT) *Health Aff* *33*:1163-1170, 2014§. *Yearbook of Surgery*, *2015*, 13–14. <https://doi.org/10.1016/j.ySUR.2014.10.001>
- Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, *77*(2), 81–97. <https://doi.org/10.1016/j.ijmedinf.2006.11.006>
- Bhoj Raj Sharma, Kaur, D., & Mnanju. (2013). A Review on Data Mining: Its Challenges, Issues and Applications. *International Journal of Current Engineering and Technology*, *3*(2), 1. Retrieved from <http://inpressco.com/category/ijcet>
- Boef, A. G. C., Dekkers, O. M., & Le Cessie, S. (2015). Mendelian randomization studies: A review of the approaches used and the quality of reporting. *International Journal of Epidemiology*, *44*(2), 496–511. <https://doi.org/10.1093/ije/dyv071>
- Bošnjak, Z., Grljević, O., & Bošnjak, S. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises’ data. *Proceedings - 2009 5th International Symposium on Applied Computational Intelligence and Informatics, SACI 2009*, *xx*(1), 509–514. <https://doi.org/10.1109/SACI.2009.5136302>
- Brazhnik, O., & Jones, J. F. (2007). Anatomy of data integration. *Journal of Biomedical Informatics*, *40*(3), 252–269. <https://doi.org/10.1016/j.jbi.2006.09.001>
- Brennan, P. F., & Bakken, S. (2015). Nursing Needs Big Data and Big Data Needs Nursing. *Journal of Nursing Scholarship*, *47*(5), 477–484. <https://doi.org/10.1111/jnu.12159>
- 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](https://doi.org/10.1016/0950-5849(95)01059-9)
- Catley, C., Smith, K., Mcgregor, C., & Tracy, M. (2009). Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. *System*, 0–4. <https://doi.org/10.1109/CBMS.2009.5255394>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Rudiger, W. (2000). Crisp-Dm 1.0: Step-by-step data mining guide. *CRISP-DM Consortium*, 76. <https://doi.org/10.1109/ICETET.2008.239>
- Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, *26*(1–2), 1–24. [https://doi.org/10.1016/S0933-3657\(02\)00049-0](https://doi.org/10.1016/S0933-3657(02)00049-0)
- DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical Education*, *40*(4), 314–321. <https://doi.org/10.1111/j.1365-2929.2006.02418.x>

- Dinov, I. D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *GigaScience*, 5(1), 1–15. <https://doi.org/10.1186/s13742-016-0117-6>
- Eapen, A. G. (2004). Application of Data mining in Medical Applications, 1–117.
- Esfandiari, N., Babavalian, M. R., Moghadam, A. M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, 41(9), 4434–4463. <https://doi.org/10.1016/j.eswa.2014.01.011>
- European Commission. (2018). 2018 reform of EU data protection rules. Retrieved July 12, 2018, from https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37. <https://doi.org/10.1609/aimag.v17i3.1230>
- General Data Protection Regulation (GDPR). (2018a). Art. 4 GDPR – Definitions. Retrieved July 12, 2018, from <https://gdpr-info.eu/art-4-gdpr/>
- General Data Protection Regulation (GDPR). (2018b). Art. 9 GDPR – Processing of special categories of personal data. Retrieved July 12, 2018, from <https://gdpr-info.eu/art-9-gdpr/>
- General Data Protection Regulation (GDPR). (2018c). Recital 53 - Processing of sensitive data in health and social sector. Retrieved July 12, 2018, from <https://gdpr-info.eu/recitals/no-53/>
- Gray, E. A., & Thorpe, J. H. (2015). Comparative effectiveness research and big data: Balancing potential with legal and ethical considerations. *Journal of Comparative Effectiveness Research*, 4(1), 61–74. <https://doi.org/10.2217/ce.14.51>
- Hansen, M. M., Miron-Shatz, T., Lau, A. Y. S., & Paton, C. (2014). Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives. *IMIA Yearbook*, 9(1), 21–26. <https://doi.org/10.15265/IY-2014-0004>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *Design Science in IS Research MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Hingorani, A. D., Van Der Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G. M., Steyerberg, E. W., ... Timmis, A. (2013). Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ (Online)*, 346(February), 1–9. <https://doi.org/10.1136/bmj.e5793>
- Hosseinkhah, F., Ashktorab, H., Veen, R., & Owrang O, M. M. (2009). Challenges in Data Mining on Medical Databases. *Database Technologies: Concepts, Methodologies, Tools, and Applications*, 1393–1404. <https://doi.org/10.4018/978-1-60566-058-5.ch083>
- Iavindrasana, J., Cohen, G., Depeursinge, A., Müller, H., Meyer, R., & Geissbuhler, A. (2009). Clinical data mining: a review. *Yearbook of Medical Informatics*, (December 2015), 121–33. <https://doi.org/me09010121> [pii]
- Jalali, S., & Wohlin, C. (2012). Systematic Literature Studies: Database Searches vs. Backward Snowballing. *ESEM'12: Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 29–38. <https://doi.org/10.1145/2372251.2372257>
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York (NY): Springer Publishing Company. New York: Springer-Verlag. <https://doi.org/10.1016/j.peva.2007.06.006>
- Jamshed, S. (2014). Qualitative research method-interviewing and observation. *Journal of Basic and Clinical Pharmacy*, 5(4), 87. <https://doi.org/10.4103/0976-0105.141942>
- Joyner, M. J., & Paneth, N. (2015). Seven Questions for Personalized Medicine. *Jama*, 55905, 2015–

2016. <https://doi.org/10.1001/jama.2015.7725>.
- Ketchersid, T. (2014). Big data in nephrology: Friend or foe? *Blood Purification*, *36*(3–4), 160–164. <https://doi.org/10.1159/000356751>
- Kleinberg J, Ludwig J, Mullainathan S, O. Z. (2014). Prediction policy problems. *Am Econ Rev*, *105*(5), 491–495. <https://doi.org/10.1038/nmeth.2839>
- Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *J Healthc Inf Manag*, *19*(2), 64–72. <https://doi.org/10.4314/ijonas.v5i1.49926>
- Koyuncugil, A. S., & OZgulbas, N. (2010). Donor research and matching system based on data mining in organ transplantation. *Journal of Medical Systems*, *34*(3), 251–259. <https://doi.org/10.1007/s10916-008-9236-7>
- Laborde-Castérot, H., Agrinier, N., & Thilly, N. (2015). Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: A systematic review. *Journal of Clinical Epidemiology*, *68*(10), 1232–1240. <https://doi.org/10.1016/j.jclinepi.2015.04.003>
- Lavecchia, A. (2015). Machine-learning approaches in drug discovery: Methods and applications. *Drug Discovery Today*, *20*(3), 318–331. <https://doi.org/10.1016/j.drudis.2014.10.012>
- Lee, C. H., & Yoon, H.-J. (2017). Medical big data: promise and challenges. *Kidney Research and Clinical Practice*, *36*(1), 3–11. <https://doi.org/10.23876/j.krcp.2017.36.1.3>
- Li J, Zhang Y, T. Y. (2016). Medical Big Data Analysis in Hospital Information System. In J. M. L. and A. C. ebastian Ventura Soto (Ed.), *World ' s largest Science , Technology & Medicine Open Access book publisher c* (pp. 65–96). INTECH. <https://doi.org/10.5772/711>
- Li, S., Kang, L., & Zhao, X. M. (2014). A survey on evolutionary algorithm based hybrid intelligence in bioinformatics. *BioMed Research International*, *2014*. <https://doi.org/10.1155/2014/362738>
- Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., & Tarczy-Hornoch, P. (2007). Data integration and genomic medicine. *Journal of Biomedical Informatics*, *40*(1), 5–16. <https://doi.org/10.1016/j.jbi.2006.02.007>
- Marban O, Mariscal G, S. J. (2009). A Data Mining & Knowledge Discovery Process Model. *RFID Technology, Security Vulnerabilities, and Countermeasures*, 75–100. <https://doi.org/10.5772/711>
- Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2009). Toward data mining engineering: A software engineering approach. *Information Systems*, *34*(1), 87–107. <https://doi.org/10.1016/j.is.2008.04.003>
- Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., ... Green, E. D. (2014). The National Institutes of Health's big data to knowledge (BD2K) initiative: Capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, *21*(6), 957–958. <https://doi.org/10.1136/amiajnl-2014-002974>
- Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review*, *25*(2), 137–166. <https://doi.org/10.1017/S0269888910000032>
- Menger, V., Spruit, M., Hagoort, K., & Scheepers, F. (2016). Transitioning to a Data Driven Mental Health Practice: Collaborative Expert Sessions for Knowledge and Hypothesis Finding. *Computational and Mathematical Methods in Medicine*, *2016*. <https://doi.org/10.1155/2016/9089321>
- Meulendijk, M., Spruit, M., Drenth-Van Maanen, C., Numans, M., Brinkkemper, S., & Jansen, P. (2013). General practitioners' attitudes towards decision-supported prescribing: An analysis of

- the Dutch primary care sector. *Health Informatics Journal*, 19(4), 247–263.
<https://doi.org/10.1177/1460458212472333>
- Milovic, B. (2012). Prediction and decision making in Health Care using Data Mining. *International Journal of Public Health Science (IJPHS)*, 1(2), 69–76.
<https://doi.org/10.11591/ijphs.v1i2.1380>
- Moro, S., Laureano, R. M. S., & Cortez, P. (2011). Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology. *European Simulation and Modelling Conference*, (Figure 1), 117–121.
- Murdoch, T., & Detsky, A. (2013). The Inevitable Application of Big Data to Health Care. *JAMA Evidence*, 309(13), 1351–1352. <https://doi.org/10.1001/jama.2013.393>
- Muskat, M., Blackman, D., & Muskat, B. (2012). Mixed Methods :Combining Expert Interviews , Cross- Impact Analysis and Scenario Development. *Electronic Journal of Business Research Methods*, 10(1), 9–21. <https://doi.org/http://dx.doi.org/10.2139/ssm.2202179>
- NCI Dictionary of Cancer Terms. (2018). Definition of prospective cohort study - NCI Dictionary of Cancer Terms - National Cancer Institute. Retrieved August 9, 2018, from <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/prospective-cohort-study>
- Neff, G. (2013). Why Big Data Won't Cure Us. *Big Data*, 1(3), 117–123.
<https://doi.org/10.1089/big.2013.0029>
- Niakšu, O. (2015). CRISP Data Mining Methodology Extension for Medical Domain. *Baltic J. Modern Computing*, 3(2), 92–109.
- Niakšu, O., & Kurasova, O. (2012). Data mining applications in healthcare: Research vs practice. *CEUR Workshop Proceedings*, 924, 58–70.
- Obenshain, M. K. (2004). Application of data mining techniques to healthcare data. *Infection Control and Hospital Epidemiology*, 25(8), 690–5. <https://doi.org/10.1086/502460>
- Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*, 375(13), 1216–1219.
<https://doi.org/10.1056/NEJMp1606181>. Predicting
- Onwubolu, G. (2009). An Inductive Data Mining System Framework. *International Workshop on Inductive Modeling*, 108–113. Retrieved from https://www.researchgate.net/publication/228537220_An_Inductive_Data_Mining_System_Framework
- Oxford Dictionaries. (2018). Method | Definition of method in English by Oxford Dictionaries. Retrieved May 21, 2018, from <https://en.oxforddictionaries.com/definition/method>
- Pal, J. K. (2011). Usefulness and applications of data mining in extracting information from different perspectives. *Nisclair-Csir*, 58(March), 7–16. Retrieved from [http://nopr.niscair.res.in/bitstream/123456789/11552/1/ALIS_58\(1\)_7-16.pdf](http://nopr.niscair.res.in/bitstream/123456789/11552/1/ALIS_58(1)_7-16.pdf)
- Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., & Abu-Hanna, A. (2009). The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine*, 46(1), 5–17. <https://doi.org/10.1016/j.artmed.2008.07.017>
- Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for Type-2 diabetic patients. *Expert Systems with Applications*, 37(12), 8102–8108.
<https://doi.org/10.1016/j.eswa.2010.05.078>
- Piatetsky-Shapiro, G. (1990). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. *AI Magazine*, 11(4), 68. <https://doi.org/10.1609/aimag.v11i4.873>
- Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science

- projects. Retrieved January 31, 2018, from <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>
- Piatetsky, G. (2016). Industries/Fields where you applied Analytics, Data Mining, Data Science in 2015? Retrieved January 31, 2018, from <https://www.kdnuggets.com/2016/01/poll-analytics-data-mining-data-science-applied-2015.html>
- Piatetsky, G. (2017). Industries / Fields where you applied Analytics, Data Mining, Data Science in 2016? Retrieved January 31, 2018, from <https://www.kdnuggets.com/2016/12/poll-analytics-data-mining-data-science-applied-2016.html>
- 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>
- Rivo, E., De La Fuente, J., Rivo, Á., García-Fontán, E., Cañizares, M. Á., & Gil, P. (2012). Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. *Clinical and Translational Oncology*, 14(1), 73–79. <https://doi.org/10.1007/s12094-012-0764-8>
- Rogalewicz, M., & Sika, R. (2016). Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. *Management and Production Engineering Review*, 7(4), 97–108. <https://doi.org/10.1515/MPER-2016-0040>
- Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: Promise and challenges. *Nature Reviews Cardiology*, 13(6), 350–359. <https://doi.org/10.1038/nrcardio.2016.42>
- SAS Institute. (2018). Data Mining and SEMMA :: Data Mining Using SAS(R) Enterprise Miner(TM): A Case Study Approach, Third Edition. Retrieved August 6, 2018, from <http://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm#n0pejm83csbja4n1xueveo2uoujy.htm>
- Seifert, J. W. (2004). *Data Mining: An Overview. CRS Report for Congress*. Retrieved from <http://www.fas.org/irp/crs/RL31798.pdf>
- Sessler, D. I. (2014). Big Data—and its contributions to peri-operative medicine. *Anaesthesia*, 69(2), 100–105. <https://doi.org/10.1111/anae.12561>
- Shah, S. J., Katz, D. H., Selvaraj, S., Burke, M. A., Yancy, C. W., Gheorghide, M., ... Deo, R. C. (2015). Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*, 131(3), 269–279. <https://doi.org/10.1161/CIRCULATIONAHA.114.010637>
- Sharma, S., Osei-Bryson, K.-M., & Kasper, G. M. (2012). Evaluation of an integrated Knowledge Discovery and Data Mining process model. *Expert Systems with Applications*, 39(13), 11335–11348. <https://doi.org/10.1016/j.eswa.2012.02.044>
- Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4), 13–22.
- Sinha, A., Hripcsak, G., & Markatou, M. (2009). Large Datasets in Biomedicine: A Discussion of Salient Analytic Issues. *Journal of the American Medical Informatics Association*, 16(6), 759–767. <https://doi.org/10.1197/jamia.M2780>
- Slobogean, G. P., Giannoudis, P. V., Frihagen, F., Forte, M. L., Morshed, S., & Bhandari, M. (2015). Bigger data, bigger problems. *Journal of Orthopaedic Trauma*, 29(December), S43–S46. <https://doi.org/10.1097/BOT.0000000000000463>
- Snoep, M. ., Jansen, N. J. G., & Groenendaal, F. (2018). Deaths and end-of-life decisions differed between neonatal and paediatric intensive care units at the same children 's hospital, 270–275. <https://doi.org/10.1111/apa.14061>

- Speckauskiene, V., & Lukosevicius, A. (2009). a Data Mining Methodology With Preprocessing Steps. *Information Technology and Control*, 38(4), 319–324.
- 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(Ic3k), 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>
- Stel, V. S., Dekker, F. W., Zoccali, C., & Jager, K. J. (2013). Instrumental variable analysis. *Nephrology Dialysis Transplantation*, 28(7), 1694–1699. <https://doi.org/10.1093/ndt/gfs310>
- Straughan, P., & Seow, A. (2000). Attitude as barriers in breast screening: A prospective study among Singapore women. *Social Science and Medicine. Social Science & Medicine*, 51(11), 1695–703.
- Stühlinger, W., Hogl, O., Stoyan, H., & Müller, M. (2000). Intelligent data mining for medical quality management. *Fifth Workshop Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP2000)*, (Workshop Notes of the 14th European Conf. Artificial Intelligence). Retrieved from http://scholar.google.de/scholar?q=Intelligent+data+mining+for+medical+quality+management&btnG=&hl=de&as_sdt=0%2C5#0
- Sun, J., & Li, H. (2008). Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems*, 21(1), 1–5. <https://doi.org/10.1016/j.knosys.2006.11.003>
- Tai, V., Grey, A., & Bolland, M. J. (2014). Results of observational studies: Analysis of findings from the nurses' health study. *PLoS ONE*, 9(10). <https://doi.org/10.1371/journal.pone.0110403>
- Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2012). Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. <https://doi.org/10.1145/2487575.2487629>
- van de Weerd, I., & Brinkkemper, S. (2009). 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.ch003>
- Van Giessen, A., Moons, K. G. M., De Wit, G. A., Verschuren, W. M. M., Boer, J. M. A., & Koffijberg, H. (2015). Tailoring the implementation of new biomarkers based on their added predictive value in subgroups of individuals. *PLoS ONE*, 10(1), 1–14. <https://doi.org/10.1371/journal.pone.0114020>
- Verschuren, P., & Doorewaard, H. (2010). Designing a Research Project: Project Design. *Designing a Research Project*, 1–25. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Vleugel, A., Spruit, M., & Daal, A. Van. (2010). Historical Data Analysis through Data Mining From an Outsourcing Perspective: The Three-Phases Model. *International Journal of Business ...*, 1–21. <https://doi.org/10.4018/jbir.2010070104>
- Wasan, S. K., Bhatnagar, V., & Kaur, H. (2006). The impact of data mining techniques on medical diagnostics. *Data Science Journal*, 5(October), 119–126. <https://doi.org/10.2481/dsj.5.119>
- Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2), xiii–xxiii.
- Weerd, I. van de, Brinkkemper, S., Souer, J., & Versendaal, J. (2006). A situational implementation method for web-based content management system applications: method engineering and validation in practice. *Software Process: Improvement and Practice*, 11(5),

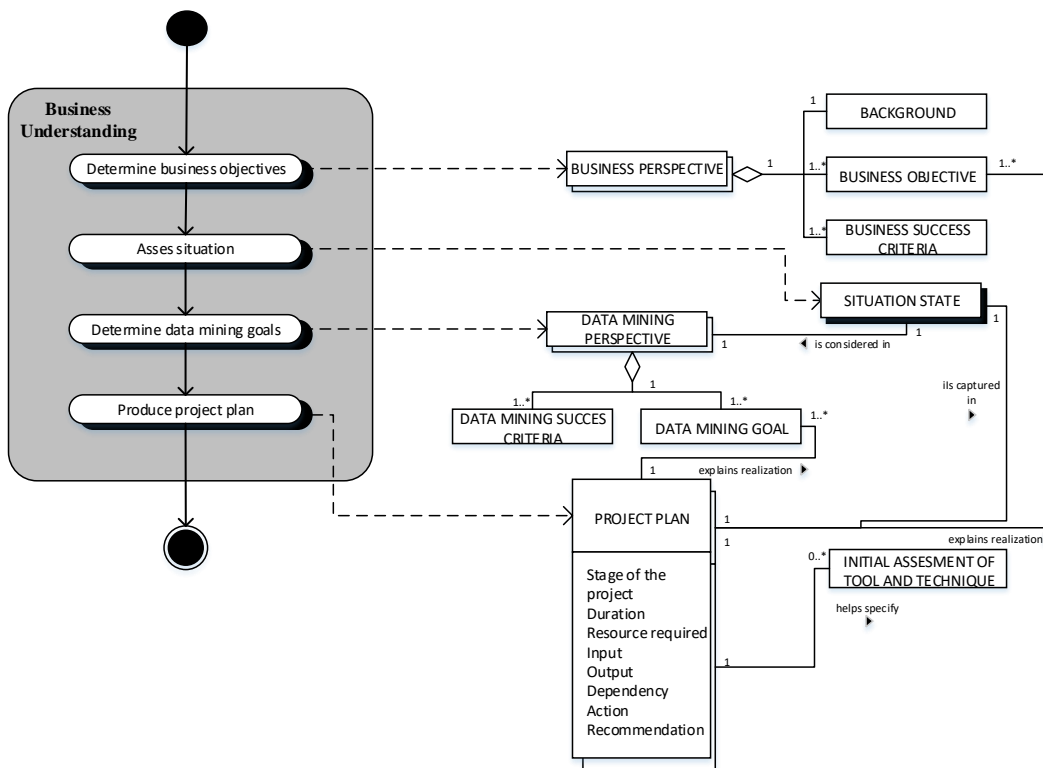
521–538. <https://doi.org/10.1002/spip>

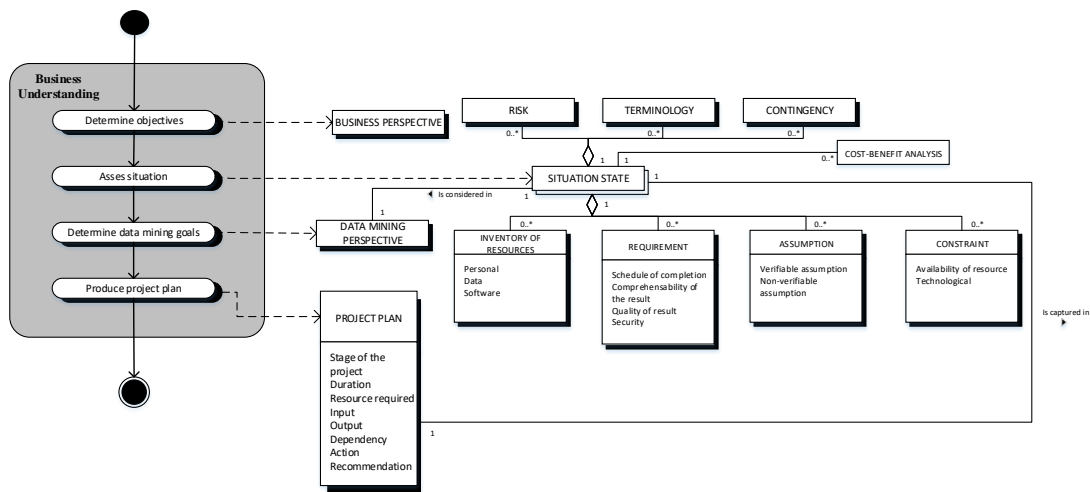
- Wirth, R., & Hipp, J. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), 29–39. <https://doi.org/10.1.1.198.5133>
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques. Complementary literature None*. <https://doi.org/10.1016/B978-0-12-374856-0.00001-8>
- Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, 36(4), 2431–2448. <https://doi.org/10.1007/s10916-011-9710-5>
- Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17. <https://doi.org/10.1080/08839510390219264>
- Zhang, Z. (2014). Big data and clinical research: focusing on the area of critical care medicine in mainland China. *Quantitative Imaging in Medicine and Surgery*, 4(5), 426–429. <https://doi.org/10.3978/j.issn.2223-4292.2014.09.03>
- Zhu, X. (2008). Semi-Supervised Learning Literature Survey Contents. *SciencesNew York*, 10(1530), 10. <https://doi.org/10.1.1.146.2352>

Appendices

A. Process-deliverable-diagram CRISP-DM

Business Understanding





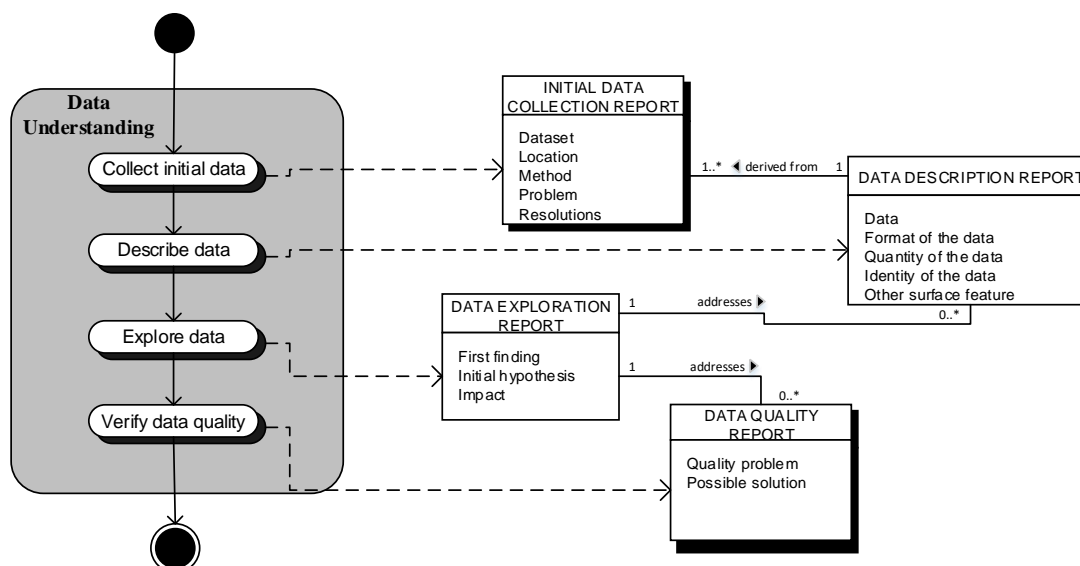
Activity	Sub-activity	Description
Business Understanding	Determine business objectives	It is important that the data analyst thoroughly understands what the customer really want to accomplish from a BUSINESS PERSPECTIVE. Herein, the BACKGROUND, BUSINESS OBJECTIVE and BUSINESS SUCCES CRITERIA are determined.
	Asses situation	This activity involves more detailed fact-findings of the INVENTORY OF RESOURCE, ASSUMPTION, CONSTRAINT, REQUIREMENT and other factors such as the RISK, CONTINGENCY, TERMINOLOGY and COST-BENEFIT ANALYSIS, that need to be considered in shaping the data analysis goal and project plan within a specific SITUATION STATE.
	Determine data mining goals	Likewise, within the BUSINESS PERSPECTIVE the DATA MINING GOAL and DATA MINING SUCCESS CRITERIA are stated in more technical terms within a DATA MINING PERSPECTIVE.
	Produce project plan	The intended plan for achieving the DATA MINING GOAL and BUSINESS GOAL are described in a PROJECT PLAN. In addition, an INITIAL ASSESSMENT OF TOOLS AND TECHNIQUES are performed to help specify the project.

Concept	Description
BUSINESS PERSPECTIVE	This entails a perspective from a customer viewpoint, which consist of the BACKGROUND information, BUSINESS OBJECTIVE and BUSINESS SUCCESS CRITERIA.

BACKGROUND	Consist of information about the organization's business situation at the beginning of the project (Chapman et al., 2000).
BUSINESS OBJECTIVE	A description about the customer's primary objectives from a business perspective (Chapman et al., 2000).
BUSINESS SUCCESS CRITERIA	A description of the criteria of successful or useful outcomes to the project from a business point of view (Chapman et al., 2000).
SITUATION STATE	This provides an explanation of the state of the project that should be considered in defining the data analysis goal and project plan. It consists of INVENTORY OF RESOURCE, REQUIREMENT, ASSUMPTION, CONSTRAINT, RISK, CONTINGENCY, TERMINOLOGY and COST - BENEFITS ANALYSIS.
INVENTORY OF RESOURCE	A list of available resources to the project including personnel, data, and software (Chapman et al., 2000).
REQUIREMENT	A list of all requirements of the projects that includes schedule of completion, comprehensibility and quality of results, security and legal issues (Chapman et al., 2000).
ASSUMPTION	A list of assumptions that is made by the project (Chapman et al., 2000). This may include verifiable assumptions and non-verifiable assumptions.
CONSTRAINT	A list of constraints on the project (Chapman et al., 2000). These may be on the availability of resources or technological constraints.
RISK	A list of the risks or events that may influence the project in a negative way such as delay or cause it to fail (Chapman et al., 2000).
CONTINGENCY	A list of corresponding contingency plans, on how to react if such risks or events take place (Chapman et al., 2000).
TERMINOLOGY	This is a compilation of glossary that is relevant to the project. This may consist of glossary that is relevant to business or data mining terminology (Chapman et al., 2000).
COST AND BENEFIT	This is a COST-BENEFIT ANALYSIS that compares the costs of the project with the possible benefits to the business if it succeeds (Chapman et al., 2000).
DATA MINING PERSPECTIVE	This entails a perspective from technical terms in data mining that consist of a DATA MINING GOAL and DATA MINING SUCCESS CRITERIA.
DATA MINING GOAL	A description of intended outputs of the projects in order to achieve the business objectives (Chapman et al., 2000).
DATA MINING SUCCESS CRITERIA	A description of the criteria of successful outcomes to the project in technical terms (Chapman et al., 2000).
PROJECT PLAN	A list of stages that need to be executed in the project. This includes the duration, resources required, inputs, outputs, dependencies, actions, and recommendations. In addition, within the project plan each phases are discussed in detail and what evaluation strategy will be used in the evaluation phase (Chapman et al., 2000).
INITIAL ASSESSMENT OF TOOLS AND TECHNIQUES	An initial assessment of the tools and techniques available should be performed that will be used during the different phases of the process. (Chapman et al., 2000).

Table 1: Concept table for the Business Understanding phase

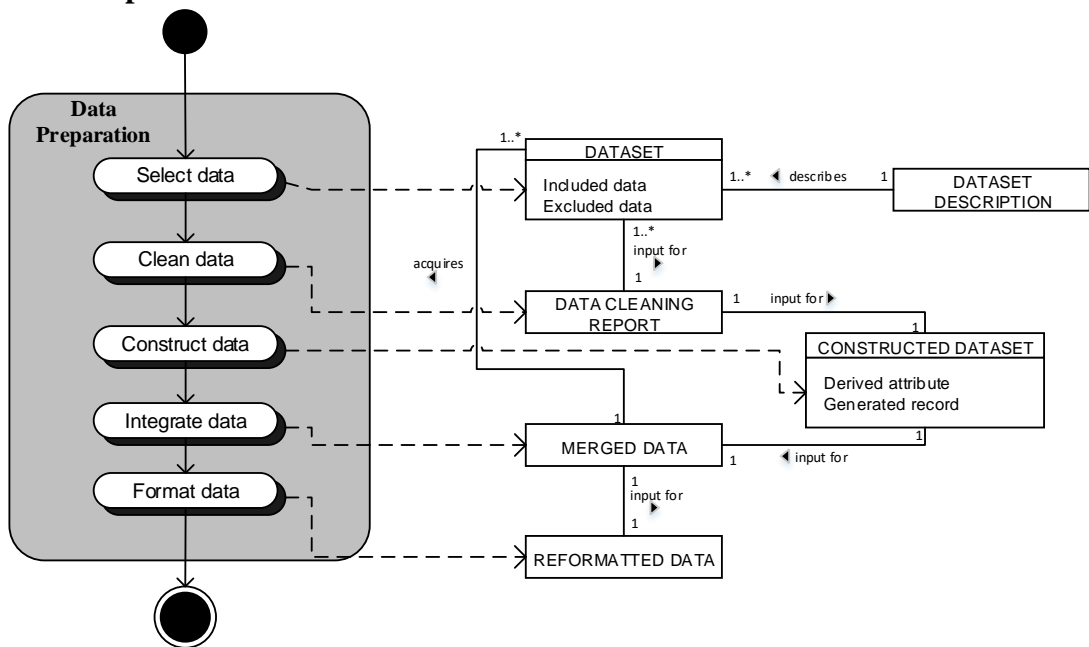
Data Understanding



Activity	Sub-activity	Description
Data Understanding	Collect initial data	The data needs to be acquired from the project resources, which is listed in the PROJECT PLAN in order to create an INITIAL DATA COLLECTION REPORT. This includes data loading, if necessary for data understanding.
	Describe data	The gross or surface properties of the acquired data needs to be examined and the results need to be reported in a DATA DESCRIPTION REPORT. The data is evaluated if it satisfies the relevant REQUIREMENT.
	Explore data	The data mining questions are addressed in the DATA EXPLORATION REPORT by using visualization, querying and reporting techniques. This process may address the DATA DESCRIPTION REPORT, DATA QUALITY REPORT and the DATA MINING GOAL by contributing or redefining them.
	Verify data quality	The quality of the data needs to be examined by addressing questions such as: if the data is complete or correct and if there are errors and how common they are. These kind of questions need to be addressed in the DATA QUALITY REPORT by verifying the data quality.

Concept	Description
INITIAL DATA COLLECTION REPORT	A description of the various data that is used for the project identifying whether some attributes are more important than others (Chapman et al., 2000). Besides, it contains a list of the datasets, locations, methods and problems.
DATA DESCRIPTION REPORT	Herein the acquired data needs to be described including the format of the data, the quantity of data, the identities of the fields and other surface features, which are found (Chapman et al., 2000).
DATA EXPLORATION REPORT	The results of the exploration are described that includes the first findings or initial hypothesis and the impact on the project (Chapman et al., 2000). This report can be illustrated by graphs and plots that indicate the data characteristics or other interesting initial findings.
DATA QUALITY REPORT	A list of the results of the data quality verification. In addition, a list of quality problems and their possible solutions are provided (Chapman et al., 2000).

Data Preparation

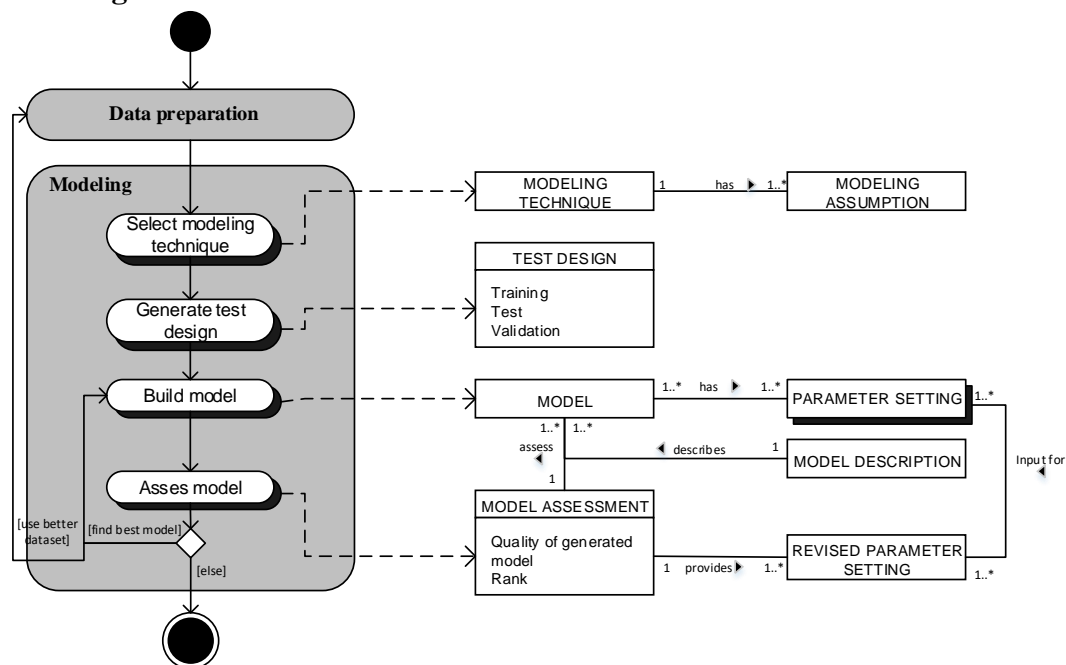


Activity	Sub-activity	Description
Data Preparation	Select data	The DATASET that will be used for analysis is selected. In addition, the DATA DESCRIPTION helps to specify the DATASET that will be used for modeling and other analysis activities.
	Clean data	The data needs to be cleaned in the DATA CLEANING REPORT and raised to the level that is required by the selected analysis techniques. This can be done by inserting

		suitable defaults or selecting clean subsets of the data and by using other techniques that help in preparing the data for the modeling phase. Moreover, the data quality problems need to be handled from the DATA QUALITY REPORT in which decisions and actions need to be taken.
	Construct data	This activity involves constructive data preparation operations such as creating entire new records in the GENERATED RECORD, or the production of DERIVED ATTRIBUTE or transformed values for existing attributes.
	Integrate data	In the MERGED DATA, information is combined from multiple records or tables of the DATASET in order to create new values or records.
	Format data	The REFORMATTED DATA is syntactic modified that does not change the meaning in order to be used by the MODELING TECHNIQUE in the modeling phase.

Concept	Description
DATASET	These are datasets that are produced during the data preparation phase that are being prepared for modeling or other major analysis work (Chapman et al., 2000). In addition, a list of included and excluded data is provided with its reasons for these decisions.
DATASET DESCRIPTION	A description of the datasets that will be used for the modeling phase (Chapman et al., 2000).
DATA CLEANING REPORT	A description of the decisions and actions taken that address the data quality problems from the DATA QUALITY REPORT (Chapman et al., 2000). The datasets need to be cleaned from irrelevant fields that create noise in the data that could have an effect on the results.
CONSTRUCTED DATASET	The CONSTRUCTED DATASET contains the derived attributes that are constructed from one or more existing attributes in the identical record. In addition, the creation of new generated records in the datasets are described (Chapman et al., 2000).
MERGED DATA	The datasets are merged with other relevant data that has similar information about a particular object (Chapman et al., 2000).
REFORMATTED DATA	The data is accustomed in accordance to the tools requirements that will be used in the modeling phase (Chapman et al., 2000).

Modeling

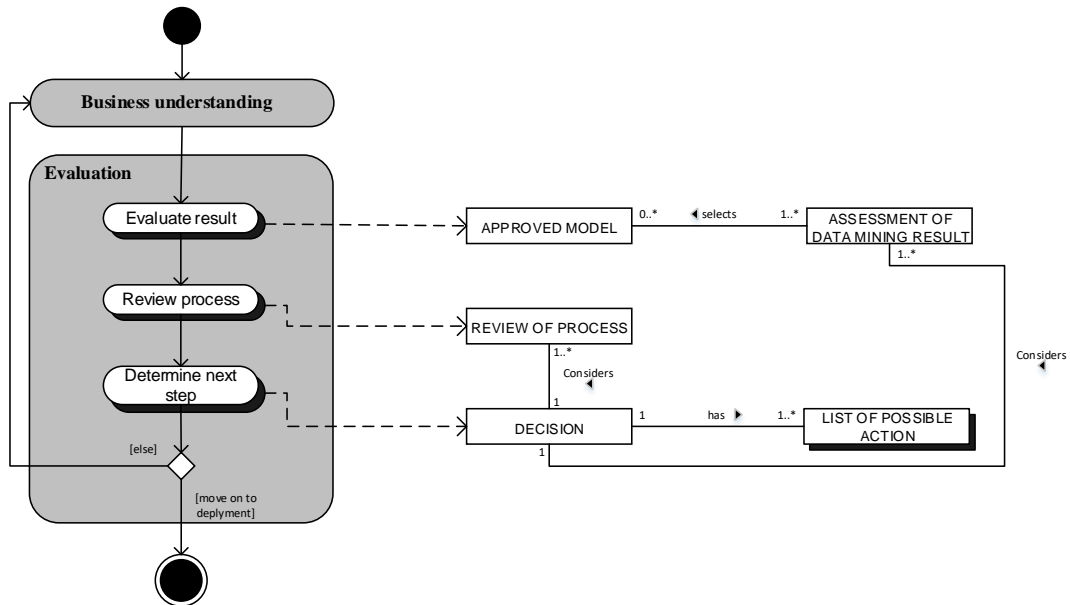


Activity	Sub-activity	Description
Modeling	Select modeling techniques	The first step in this phase is the selection of the actual MODELING TECHNIQUE, which describes how MODEL needs to be build. This technique makes a MODELING ASSUMPTION about the suitability of the data that is going to be used for a certain MODELING TECHNIQUE.
	Generate test design	Before the actual model is build, a procedure or mechanism is generated in order to TEST DESIGN the model in regards to quality and validity. Herein, the dataset can be divided in three components to test the model: training, testing and validating.
	Build model	The modeling tool is applied to the prepared dataset in order to produce a MODEL. It includes the PARAMETER SETTING that are selected for the modeling tool and the results of the MODEL are described in the MODEL DESCRIPTION.
	Asses model	In the MODEL ASSESSMENT, the models need to be ranked on how they perform. The BUSINESS OBJECTIVE and the BUSINESS SUCCESS CRITERIA are taken into account as well as the DATA MINING SUCCESS CRITERIA and the results of the TEST DESIGN during the assessment. Moreover, the parameters that are used can be revised in the

		RIVISED PARAMETER SETTING and iterated between the modeling building and assessment until the best model is found.
--	--	--

Concept	Description
MODELING TECHNIQUE	This is the selection of the actual modeling technique that has to be used for modeling.(Chapman et al., 2000).
MODELING ASSUMPTION	A list of assumptions that are made for the modeling technique about the data.(Chapman et al., 2000).
TEST DESIGN	This describes the intended plan for testing, evaluating and training the models (Chapman et al., 2000).
MODEL	These are the actual models that are produced by the MODELING TECHNIQUE (Chapman et al., 2000).
PARAMETER SETTING	A list of parameters and their chosen values that can be adjusted by the modeling tool (Chapman et al., 2000).
MODEL DESCRIPTION	Here the models are described and assessed with their expected robustness, accuracy, and possible shortcomings (Chapman et al., 2000).
MODEL ASSESSMENT	The outcomes are summarized including with a list of qualities of the generated models and their quality rank in relation to each other (Chapman et al., 2000).
REVISED PARAMETER SETTING	The PARAMETER SETTING is revised in accordance to the MODEL ASSESSMENT. Herein, the iteration with the model building and assessment is performed until the best model is found (Chapman et al., 2000).

Evaluation

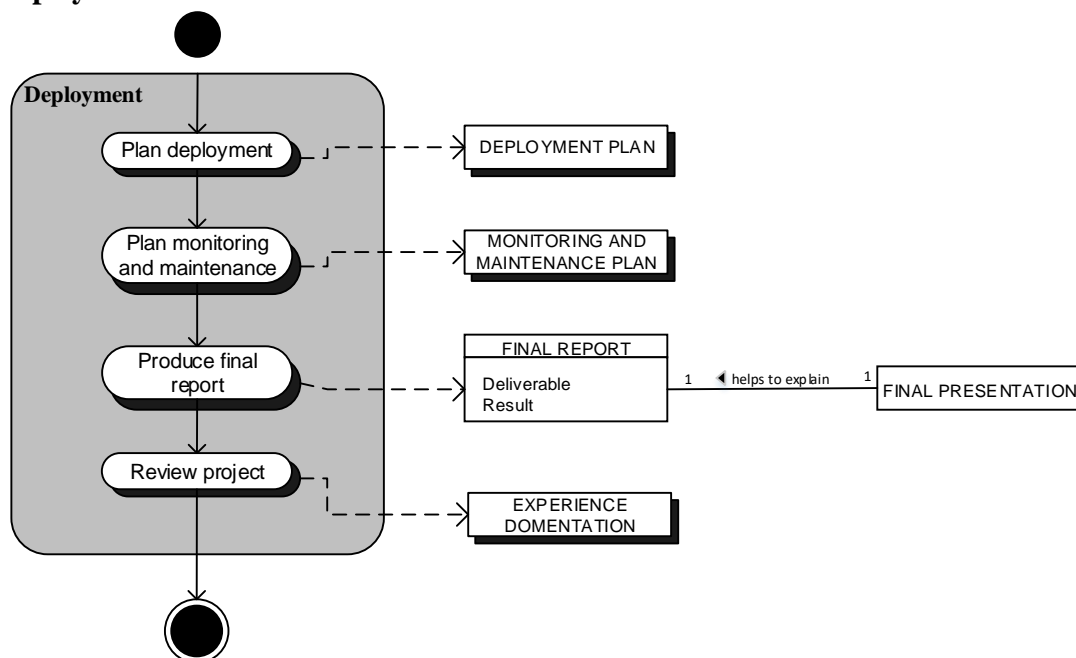


Activity	Sub-activity	Description
Evaluation	Evaluation results	The ASSESSMENT OF DATA

		MINING RESULT is evaluated to what extent a particular model meets the BUSINESS SUCCESS CRITERIA and BUSINESS OBJECTIVE and if there is a business reason, why this model might be deficient. After this assessment an APPROVED MODEL is selected that meets the needed requirements.
	Review results	A thorough REVIEW OF PROCESS is performed of the data mining engagement in order to check if an important task or factor has somehow been overlooked.
	Determine next step	This activity depends on the results of the ASSESSMENT OF THE RESULT and REVIEW OF PROCESS in order to make a DECISION on how to proceed. The team can decide to move on to the next phase or iterate between the phase again or quit and set up a new data mining project according to the LIST OF POSSIBLE ACTION

Concept	Description
APPROVED MODEL	The models that meets the BUSINESS SUCCESS CRITERIA are selected and approved (Chapman et al., 2000).
ASSESSMENT OF DATA MINING RESULT	The assessment results are summarized in terms of the BUSINESS SUCCESS CRITERIA and whether the project already meets the original BUSINESS OBJECTIVE (Chapman et al., 2000).
REVIEW OF PROCESS	The process review is summarized and the activities that have been missed and those that should be repeated are highlighted (Chapman et al., 2000).
DECISION	The decisions that are made are described along with the rational for them (Chapman et al., 2000).
LIST OF POSSIBLE ACCTION	A list of potential further actions with the reasons behind it is provided for each option (Chapman et al., 2000).

Deployment



Activity	Sub-activity	Description
Deployment	Plan deployment	The evaluation results are taken and the strategy for deployment is determined. This procedure is documented in the DEPLOYMENT PLAN and can be used for later deployment.
	Plan monitoring and maintenance	If the data mining results become a part of the day-to-day business and its environment a detailed MONITORING AND MAINTENANCE PLAN needs to be constructed. Therein, a maintenance strategy is developed in order to avoid incorrect usage of the data mining results and the specific type of the deployment are taken into account.
	Produce final report	A FINAL REPORT is written at the end of the project that may provide only a summary of the projects and its experience or it can be a FINAL PRESENTATION of the data mining result(s).
	Review project	A final assessment is performed in an EXPERIENCE DOCUMENTATION report in what went well or what went wrong during the project and what needs to be improved in the future.

Concept	Description
---------	-------------

DEPLOYMENT PLAN	The deployment strategy is defined including the necessary steps and how to perform them (Chapman et al., 2000).
MONITORING AND MAINTENANCE PLAN	The monitoring and maintenance strategy is summarized and how to perform the required steps (Chapman et al., 2000).
FINAL REPORT	A final written report of the data mining engagement in which all the threads are brought together (Chapman et al., 2000). This includes all the previous deliverables and their results.
FINAL PRESENTATION	This is a meeting in which the project is summarized and the results presented to the stakeholders (Chapman et al., 2000).
EXPERIENCE DOCUMENTATION	This is a summary of the important gained experience during the project (Chapman et al., 2000).

B. Experiment notebook – VDS1

This notebook contains the R code and procedure that was performed during the VDS1 project as indicated in chapter 5. Herein, the dataset was prepared and analyzed.

B.1 VDS1 code

```
## Data Preparation
```

1. Load base packages

```
```{r}
library(plyr)
library(dplyr)
library(lubridate)
library(stringr)
library(tidyr)
library(xts)
library(ggplot2)
library(tidyverse)
library(reshape)
library(nlme)
library(quantreg)
```
```

2. Load the dataset (servo_minute.csv)

```
```{r}
ds.df = read.csv(file.choose(), sep = ",", dec = ".")
```
```

3. Factor the categorical variables

```
```{r}
ds.df$icuid = factor(ds.df$icuid)
ds.df$period = factor(ds.df$period)
```
```

4. Splitting the data sets in two (before and after) without converting it to xts.

```
```{r}
```

```
ds.atbp.temp <- which(ds.df$period == "atbp")
ds.atbp <- ds.df [ds.atbp.temp,]
m(ds.atbp.temp)
```

```
ds.btps.temp <- which(ds.df$period == "btps")
ds.btps <- ds.df [ds.btps.temp,]
m(ds.btps.temp)
...
```

### 5. Remove outliers (outliers based on all columns except id(1), datetime(2), admissiondate(3), age(4), weight(5) & period(17))

```
```{r}
# atbp
mahal.atbp = mahalanobis(ds.atbp[, -c(1,2,3,4,5,17)], #
                        colMeans(ds.atbp[, -c(1,2,3,4,5,17)], na.rm = TRUE),
                        cov(ds.atbp[, -c(1,2,3,4,5,17)], use = "pairwise.complete.obs"))

cutoff.atbp = qchisq(1-.001, ncol(ds.atbp[, -c(1,2,3,4,5,17)])) #degree of freedom is 13
ds.atbp = ds.atbp[mahal.atbp < cutoff.atbp , ]

# btps
mahal.btps = mahalanobis(ds.btps[, -c(1,2,3,4,5,17)],
                        colMeans(ds.btps[, -c(1,2,3,4,5,17)], na.rm = TRUE),
                        cov(ds.btps[, -c(1,2,3,4,5,17)], use = "pairwise.complete.obs"))

cutoff.btps = qchisq(1-.001, ncol(ds.btps[, -c(1,2,3,4,5,17)])) #degree of freedom is 13
ds.btps = ds.btps[mahal.btps < cutoff.atbp , ]
m(cutoff.atbp, cutoff.btps, mahal.atbp, mahal.btps)
...

```

6. Make a complete dataset without the outliers (df.all)

```
```{r}
merge both files:
df.all <- rbind(ds.atbp, ds.btps)

add differences between set and measured TV
df.all$tv_dif = df.all$tv_set - df.all$tv_measured

remove NA's
df.all <- df.all[-which(is.na(df.all$tvweight)),]
...

```

### 7. Add amvweight2 (tv set x afreq)

```
```{r}
df.all[,19] <- 0
names(df.all)[19] <- "amvweight2"
df.all$amvweight2 <- df.all$afreq_measured * df.all$tv_setweight
ds.atbp[,18] <- 0
names(ds.atbp)[18] <- "amvweight2"
ds.atbp$amvweight2 <- ds.atbp$afreq_measured * ds.atbp$tv_setweight
ds.btps[,18] <- 0
names(ds.btps)[18] <- "amvweight2"
ds.btps$amvweight2 <- ds.btps$afreq_measured * ds.btps$tv_setweight
...

```

Data Analysis

1. Analysis: baseline table

```
```{r}
tv to weight, breathing frequency, amv to weight
for (i in c("tv_settoweight", "afreq_measured", "amvtoweight2", "ppeak")) {
 print(i)
 print(mean(df.all[,i], na.rm = TRUE))
 print(sd(df.all[,i], na.rm = TRUE))
 print(mean(ds.atbp[,i], na.rm = TRUE))
 print(sd(ds.atbp[,i], na.rm = TRUE))
 print(mean(ds.btps[,i], na.rm = TRUE))
 print(sd(ds.btps[,i], na.rm = TRUE))
}

a <- 0
b <- 0
age and weight
for (i in unique(df.all$icuid)) {
 a <- a + 1
 b <- c(b, unique(df.all[(which(df.all$icuid %in% i)), "age"]))
}
print ("age in all patients")
median(b) / 30
IQR(b) / 30

a <- 0
b <- 0
for (i in unique(ds.atbp$icuid)) {
 a <- a + 1
 b <- c(b, unique(df.all[(which(df.all$icuid %in% i)), "age"]))
}
print ("age in atpd patients")
median(b) / 30
IQR(b) / 30

a <- 0
b <- 0
for (i in unique(ds.btps$icuid)) {
 a <- a + 1
 b <- c(b, unique(df.all[(which(df.all$icuid %in% i)), "age"]))
}
print ("age in btps patients")
median(b) / 30
IQR(b) / 30

a <- 0
b <- 0
for (i in unique(df.all$icuid)) {
 a <- a + 1
 b <- c(b, unique(df.all[(which(df.all$icuid %in% i)), "admitweight"]))
}
print ("weight in all patients")
median(b)
IQR(b)

a <- 0
```

```

b <- 0
for (i in unique(ds.atbp$icuid)) {
 a <- a + 1
 b <- c(b, unique(df.all[(which(df.all$icuid %in% i)), "admitweight"]))
}
print ("weight in atpd patients")
median(b)
IQR(b)

a <- 0
b <- 0
for (i in unique(ds.btps$icuid)) {
 a <- a + 1
 b <- c(b, unique(df.all[(which(df.all$icuid %in% i)), "admitweight"]))
}
print ("weight in btps patients")
median(b)
IQR(b)

rm(a, b, i)
...

```

## 2. Analysis: different models (generalized linear / linear mixed effects)

```

```{r}
# Baseline intercept model
model1.etc2 = gls(etc2 ~ 1,
  data = df.all,
  method = "ML",
  na.action = "na.omit")
summary(model1.etc2)

# Random intercept model
model2.etc2 = lme(etc2 ~ 1,
  data = df.all,
  method = "ML",
  na.action = "na.omit",
  random = ~1|icuid)
summary(model2.etc2)

anova(model1.etc2, model2.etc2)
...

```

3: Add fixed effects to the model because model 2 with the random intercept is better

```

```{r}
Add fixed effects to the model
model3.etc2 = lme(etc2 ~ amvtoweight * period,
 control = lmeControl(opt = "optim"),
 data = df.all,
 method = "ML",
 na.action = "na.omit",
 random = ~1+ppeak|icuid)
summary(model3.etc2)

model4.etc2 = lme(etc2 ~ amvtoweight * period + ppeak,
 control = lmeControl(opt = "optim"),

```

```

 data = df.all,
 method = "ML",
 na.action = "na.omit",
 random = ~1+ppeak|icuid)
summary(model4.etco2)

model5.etco2 = lme(etco2 ~ amvtoweight * period + ppeak + admitweight,
 control = lmeControl(opt = "optim"),
 data = df.all,
 method = "ML",
 na.action = "na.omit",
 random = ~1+ppeak|icuid)
summary(model5.etco2)

model6.etco2 = lme(etco2 ~ amvtoweight2 * period + ppeak + admitweight + age,
 control = lmeControl(opt = "optim"),
 data = df.all,
 method = "ML",
 na.action = "na.omit",
 random = ~1+ppeak|icuid)
summary(model6.etco2)

model7.amv = lme(amvtoweight2 ~ period + ppeak + admitweight + age,
 control = lmeControl(opt = "optim"),
 data = df.all,
 method = "M",
 na.action = "na.omit",
 random = ~1+ppeak|icuid)
summary(model7.amv)

##compare all the models
anova(model1.etco2, model2.etco2, model3.etco2, model4.etco2, model5.etco2, model6.etco2)

write output to file
sink(choose.files())
print(summary(model6.etco2))
sink()
```

```

4: Build a visualization functions

```

```{r}
Multiple plot function
#
ggplot objects can be passed in ..., or to plotlist (as a list of ggplot objects)
- cols: Number of columns in layout
- layout: A matrix specifying the layout. If present, 'cols' is ignored.
#
If the layout is something like matrix(c(1,2,3,3), nrow=2, byrow=TRUE),
then plot 1 will go in the upperleft, 2 will go in the upperright, and
3 will go all the way across the bottom.
#
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
 library(grid)

 # Make a list from the ... arguments and plotlist
 plots <- c(list(...), plotlist)

```



```

numPlots = length(plots)

If layout is NULL, then use 'cols' to determine layout
if (is.null(layout)) {
 # Make the panel
 # ncol: Number of columns of plots
 # nrow: Number of rows needed, calculated from # of cols
 layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
 ncol = cols, nrow = ceiling(numPlots/cols))
}

if (numPlots==1) {
 print(plots[[1]])

} else {
 # Set up the page
 grid.newpage()
 pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

 # Make each plot, in the correct location
 for (i in 1:numPlots) {
 # Get the i,j matrix positions of the regions that contain this subplot
 matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

 print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
 layout.pos.col = matchidx$col))
 }
}
...

```

## 5: Visualize results

```

```{r}
# histograms
# all data
histo_all <- ggplot(df.all, aes(amvtoweight)) +
  geom_histogram(binwidth = 5) +
  ggtitle("Histogram all data") +
  xlab("Minute volume to weight")

# ATBP data
histo_atbp <- ggplot(ds.atbp, aes(amvtoweight)) +
  geom_histogram(binwidth = 5, fill = "#86c06a") +
  ggtitle("Histogram ATPD") +
  xlab("Minute volume to weight")

# BTPS data
histo_btps <- ggplot(ds.btps, aes(amvtoweight)) +
  geom_histogram(binwidth = 5, fill = "#f3997b") +
  ggtitle("Histogram BTPS") +
  xlab("Minute volume to weight")

multiplot (histo_all, histo_atbp, histo_btps, cols = 2)

# AMV to weight per ventilation mode
plot1 <- ggplot(data = df.all, aes (x = period, y = amvtoweight2)) +

```

```

geom_boxplot(fill = c("#86c06a", "#f3997b")) +
ylab("Minute volume to weight") +
xlab("Mode") +
scale_x_discrete(labels = c("ATPD", "BTPS"))

# Jitter dot plot etCO2 ~ AMV to weight
plot2 <- ggplot(data = df.all, aes(x = amvtoweight2, y = etco2, color = period)) +
  geom_jitter(aes()) +
  scale_color_manual(values=c("#86c06a", "#f3997b"), name = "Mode", labels = c("ATPD",
"BTPS")) +
  geom_quantile(data = ds.atbp, aes(amvtoweight2, etco2), formula = y ~ x, color = "#005321") +
  geom_quantile(data = ds.btps, aes(amvtoweight2, etco2), formula = y ~ x, color = "#e30613") +
  xlab("Minute volume to weight") +
  ylab("etCO2")

plot1
plot2
...

```

C. Experiment notebook – VDS2

This notebook contains the R code and procedure that was performed during the VDS2 project as indicated in chapter 5. Herein, the dataset was prepared and analyzed. Moreover, there were 3 analysis performed in this setting: on a smaller dataset of one patient, on a bigger dataset of one patient, and on the whole dataset with all patients. Hence, for convenience purposes the code of the bigger set is shown because the code is the same but differs only with the usage of a specific dataset.

C.1 VDS2 code

Data preparation

1. Load base packages

```

```{r}
library(tibble)
library(ggplot2)
library(neuralnet)
library(ggplot2)
library(dplyr)
library(plyr)
...

```

##### 2. Load the dataset

```

```{r}
df <- read.csv(file.choose(), header = TRUE, sep = ",", dec = ".", stringsAsFactors = FALSE)
...

```

3. Factor the categorical variable

```

```{r}
df$patientid <- as.factor(df$patientid)
...

```

##### 4. Delete NA's and irrelevant columns

```

```{r}
#Remove NA's or not relevant Columns
df2<- df[, -c(3:8, 20:27, 31:38)]
...

```

5. Check how many patients there are and finding the patients with the most data available

```
```{r}
#Show how many patients there are
length(unique(df2$patientid))

#Top 10
names(sort(summary(as.factor(df2$patientid)), decreasing=T)[1:10])
```
```

6. Finding out which patient has the most observations left after excluding the NA's: Below one example is given how to do it and the final results are below the code's

```
```{r}
Take from a patient with not much observations
pzero_b <- df2[which(df2$patientid == '21479'),]

#Remove NA's
pzero_b <- na.omit(pzero_b) #deleting the na's is needed for normalization in later phase

summary(pzero_b)
```
```

Results of patients in which the NA's are removed:

```
21479 = 68506
24688 = 46233
24421 = 49967
6203 = 28722
923 = 3980 obs
197 = 1958 obs
26984 = 13403
772 = 5997
```

For the implementation of neural networks on the data set, the patient with the most observations will be used. In this case this will be patient 21479 with 68506 observations. Additionally, for trying it on a much smaller data set in order to see how the neural network performs, the patient 9 will be used with 2200 observations.

7. Take the data from one patient with not much and one with less observations

```
```{r}
Take from a patient with not much observations
pzero<- df2[which(df2$patientid == '9'),]

Additional column needed to be removed because of no data input and the column of the patientID:
pzero<- pzero[, -c(2,6)]
pzero<- na.omit(pzero)
summary(pzero)

Delete the patientID column of the big data variant of patient 21479
pzero_b<- pzero_b[, -2]
```
```

8. Perform general check:

```
```{r}
str(pzero)
summary(pzero)
```
```

```
str(pzero_b)
summary(pzero_b)
```

```

### 9. Convert time to time series

```
```{r}
pzero[,1] <- as.POSIXct(pzero[,1])

pzero_b[,1] <- as.POSIXct(pzero_b[,1])
```

```

### 10. Plot the data

```
```{r}
ggplot(pzero, aes(x = 1:nrow(pzero), y = `mon_etco2`)) + geom_line()
hist(pzero$mon_etco2)

ggplot(pzero_b, aes(x = 1:nrow(pzero_b), y = `mon_etco2`)) + geom_line()
hist(pzero_b$mon_etco2)
```

```

### 11. Plot a more narrow one of the data

```
```{r}
ggplot(pzero, aes(x = time, y = `mon_etco2`)) + geom_line()
ggplot(pzero_b, aes(x = time, y = `mon_etco2`)) + geom_line()
```

```

### 12. Save new dataset (already done)

```
```{r}
#Save cleaned file
write.csv(pzero,file='pzero.csv', row.names=FALSE)
write.csv(pzero_b,file='pzero_b.csv', row.names=FALSE)
```

```

## ## Data Analysis

### 1. Delete the timestamp variable

```
```{r}
pzero_b <- pzero_b[, -1]
```

```

2. Set rownames that counts from 1 till the amount of rows. This is done to double check in the dataset for the next step of adding the hourly data for convenience.

```
```{r}
rownames(pzero_b) <- seq(length=nrow(pzero_b))
```

```

### 3. Add the hourly data

```
```{r}
# Create variable t_60 that indicate the etco2 the next hour
pzero_b$t.etco2_60 <- lead(pzero_b$mon_etco2, 60)
```

```

```
Create variable t_120 that indicate the etco2 in two hours
pzero_b$t.etco2_120 <- lead(pzero_b$mon_etco2, 120)
```

```
pzero_b <- na.omit(pzero_b) #deleting the na's which is needed for normalization
```

```

4. Create a training and testing dataset

```
```{r}
```

```

#Create a training and testing set
index_bd <- sample(1:nrow(pzero_b),round(0.80*nrow(pzero_b)))
train_bd <- pzero_b[index_bd,]
test_bd <- pzero_b[-index_bd,]
...

5. Normalize the dataset
```{r}
maxs_bd <- apply(pzero_b, 2, max)
mins_bd <- apply(pzero_b, 2, min)

scaled_bd <- as.data.frame(scale(pzero_b, center = mins_bd, scale = maxs_bd - mins_bd))

train_bd <- scaled_bd[index_bd,]
test_bd <- scaled_bd[-index_bd,]
...

6. Create the Neural Network predicting the next hour
```{r}
#create formula:
allVars_bd <- names(train_bd)
predictorVars_bd <- allVars_bd[!allVars_bd%in% c('t.etc2_60', 't.etc2_120')]
predictorVars_bd <- paste(predictorVars_bd, collapse = "+")
form_b <- as.formula(paste("t.etc2_60~", predictorVars_bd, collapse = "+"))

#create neural network
nn_bd<- neuralnet(formula = form_b, data = train_bd, hidden = c(14), linear.output = FALSE,
threshold=0.01, stepmax=1e6)

nn_bd$result.matrix
plot(nn_bd)
...

7. Creat the Neural Network predicting the next two hours
```{r}
#create formula:
form_b2 <- as.formula(paste("t.etc2_120~", predictorVars_bd, collapse = "+"))

#Efficient way:
nn_bd2<- neuralnet(formula = form_b2, data = train_bd, hidden = c(5,3), linear.output = FALSE,
threshold=0.05)

nn_bd2$result.matrix
plot(nn_bd2)
...

8. Compute results for bot NNs
```{r}
#nn1
nn.results_bd <- neuralnet::compute(nn_bd, test_bd[, 1:14])

#nn2
nn.results_bd2<- neuralnet::compute(nn_bd2, test_bd[, 1:14])
...

9. Scale back in order to make meaningful comparison or prediction
```{r}

```

```

#NN 1
pr.nn_bd <- nn.results_bd$net.result*(max(pzero_b$t.etco2_60)-
min(pzero_b$t.etco2_60))+min(pzero_b$t.etco2_60)

test.r_bd <- (test_bd$t.etco2_60)*(max(pzero_b$t.etco2_60)-
min(pzero_b$t.etco2_60))+min(pzero_b$t.etco2_60)

#NN 2
pr.nn_bd2 <- nn.results_bd2$net.result*(max(pzero_b$t.etco2_120)-
min(pzero_b$t.etco2_120))+min(pzero_b$t.etco2_120)

test.r_bd2 <- (test_b$t.etco2_120)*(max(pzero_b$t.etco2_120)-
min(pzero_b$t.etco2_120))+min(pzero_b$t.etco2_120)

```

10. Calculated the MSE

```

```{r}
NN error for hourly prediction
MSE.nn_b <- sum((test.r_bd - pr.nn_bd)^2)/nrow(test_bd)
MSE.nn_b

NN error for the next two hours prediction
MSE.nn_b <- sum((test.r_bd2 - pr.nn_bd2)^2)/nrow(test_bd)
MSE.nn_b

```

Results of `with` different setting of nodes:

```

MSE = 4.17285199 with 2, 1 NN
MSE = 3.854993038 with 3,2 NN
MSE = 3.482475715 with 5, 3 NN
MSE = 3.470389816 with 6, 4 NN

```

## 11. Plot the results

```

```{r}
plot(test.r_bd, pr.nn_bd, col='red', main='Real vs predicted NN', pch=18, cex=0.7)
abline(0,1,lwd=2)
legend('bottomright', legend='NN', pch=18, col='red', bty='n')

```

12. Measure accuracy

```

```{r}
comparison_bd <- data.frame(pr.nn_bd, test.r_bd)
deviation_bd <- ((test.r_bd - pr.nn_bd) / test.r_bd)
comparison_bd <- data.frame(pr.nn_bd, test.r_bd, deviation_bd)
accuracy_bd <- 1 - abs(mean(deviation_bd))
accuracy_bd

```

## **D. Expert interviews**

These interviews were conducted in Dutch and has been translated into English. Moreover, the interviews were semi-structured as explained in chapter two. In addition, an interview protocol was developed in guiding the interviews in a more structured manner, although some deviation from the questions occurred depending on the expert's experience and field of work as well as understanding of data mining projects.

### **D.1 Interview protocol form**

#### **Introduction:**

- Provide a brief introduction about yourself and explain the purpose of your research and this interview.

**Note\*** - request permission to record this interview for the purpose of transcribing it later on and explain that all information collect will be confidential and will not be shared outside the Utrecht University and will be only used for scientific research purposes.

#### **\*Begin recording\***

Ask the following questions related to the following topics:

#### **Background:**

- Can you tell me something about yourself and what your profession is? Your experience with Data mining / Data science in healthcare?
- Have you previously worked on projects through data analysis and developing models like the VDS-project?
- How involved were you in this project and how did you prepare?
- To which extent did you acquire knowledge relate to IT such as programming?
- How is your statistics knowledge?
- What is your opinion towards using big data to retract information and to which extent do you think hospitals will gain from this?
- Are you familiar with data mining processes? If yes, which?
- Can you share your experience from the moment you started your data mining project?
- During the data analysis, what were the stumbling blocks that you experienced besides the technical aspects?
- What have you done to overcome these stumbling blocks?
- What were your most important steps and activities from starting a data mining project till ending by the models?

- Was there certain expertise that was missing or that was helpful during your start of a data mining project? And is there still some sort of expertise still needed to be able to do better data mining projects?

In my research I have encountered a number of obstacles with regard to data mining in healthcare such as: Data integration, data quality, validation & analytical problems, Causal inference in observational data sets, Legal Issues, and User-friendliness.

1. How can we overcome these obstacles?
2. Are there any other obstacles that are not been mentioned above but are important? Could you also explain how you deal with them?

### **Evaluation of the process method:**

In this part, the task and deliverable (the extracted method fragments) table is shown as illustrated in chapter 6 as well as the life cycle of the CRISP-DM in chapter 3.

### **Show the MSP-DM method fragments and the overall method and discuss them:**

Herein, provide your method fragments found in your research and case study and discuss them with the domain expert. Ask for comments and if there are missing elements present which were overlooked.

### **End**

Thank the interviewee for their time and for their input into this research.

\*End recording\*



**Interviewee A1:**

**1. Which data mining process methods are used here in the hospital?**

- **I have found that there could be two separate objectives in a data mining project in the healthcare; clinical and managerial objectives. Do you agree with this?**

Well firstly, a hospital is a company, we have clinical and business processes. The business process means that products need to be bought and archived. Information needs to be shared, communication and technology needs to be arranged. The clinical process contains medical decisions and patient care. These two generally overlap yet we separate them in our mind.

**2. For which goal are these methods used?**

The data is currently used for business intelligence; this means that data is used to make a monthly or quarterly report. We also have a central dashboard, but we don't use them for high quality data science or data analytics. Clinically speaking we use a lot of data, also with current data. We use a standard scientific process, it starts with defining the problem and posing a question made out of certain components such as the dependent and independent variables, the determinant outcome and the domain.

**3. To which extent does the hospital comply with these methods and are these generally applicable?**

That's a difficult question, unfortunately there are no straight lines to follow, just a lot of legislation concerning certain aspects. It's very poorly protocolled.

**4. Can you give me a description of these methods and its activities and steps?**

The first step is the question or hypothesis, then we make a research plan and record this so it's not possible to manipulate the plan to the outcome. The point of this is that it's reproducible and goal oriented.

**5. What are the most important steps and activities herein?**

It's important to make a plan beforehand, to indicate the way of research and to work goal oriented. The goal needs to be transparent in the research.

**6. What is seen as optional and when is it relevant?**

Currently the plan is not required, yet more and more organizations are asking for it. Personally, I believe that a plan is important to have. Randomize in the plan is also optional, also reporting between times is not done often.

**Show the MSP-DM method fragments and the overall method and discuss them:**

**7. I have found that there could be some legal constraints during a data mining project and I have added it in the concepts of constraints, what do you think about this?**

There could be legal constraints although there is not a big issue here because frequently the data can be acquired with not that much problem. (Did not show much concern or emphasis)

**1. I have found that it is important to anonymize the data when working with others that do not have permission in accessing it, what do you think about this?**

This is true, like in your case. You received the data from A3 anonymized because you did not have access to personal data from patients.

**8. Do you think steps or activities are lacking? If yes, which?**

Business understanding:

**1. I have found that scientific literature can benefit the projects, what is your opinion about this?**

Yes, there is much written related to predictive modeling or other data mining topics. There can be matters that are related to a given data mining project. However, there is also much discussion what is the correct way to do things. So, you may find relevant papers but you will still need to do some research by yourself.

Produce project plan:

I haven't talked about this but when you have a research plan, it's possible to get funding for it. It depends on what the costs are i.e. hiring someone or time cost like with a student. The order usually is, first to make a plan and then you can ask for a funding.

Data understanding:

**2. During my case study I have found that discussing the origin of the data with a domain expert to be very beneficial in understanding the data, what do you think about this and how this fits in this phase as an activity?**

When it comes to data understanding, we rarely do it, we usually just do a pilot or a sample size. We do this to see if there is an effect or to see the feasibility.

The step verifying data quality is done during the process, the steps describe and explore data we do during the analysis. Data understanding in general is for export in a pilot or sample size, it's not data driven. However, I can understand that for newcomers in this domain can be useful and following this structure.

Data preparation:

**3. I have noticed that the order of preparing the dataset for modeling is a bit different, in which you first select, format, integrate, clean, and then construct the data. Is this correct?**

In the data preparation phase, is mostly collecting data and finding the right data. There is not a particular structure to follow. This mainly depends what kind of order you prefer.

Modeling:

**4. Do you find it useful in involving (clinical) practitioners during the model assessment?**

It depends on your own expertise of the domain. I can imagine after working for a long time in a certain domain setting that you be able to understand the most of the domain related data. However, checking your findings or decisions with a more experienced expert will not be a harm. I think this will apply more for newcomers in the medicine domain than experienced researchers.

Evaluation:

**5. Why is it important to involve both the health care professionals and other experts in reviewing the results?**

When evaluating the models, there should be other (clinical) experts involved in this process because of their insight that can benefit in finding mistakes or even improving the model because of the suggestions they can make.

**6. Do you think that after evaluating the results that there are other possible actions such as moving back to the preparation or modeling phase?**

Yes, there could be other possible action because during the assessment of the results you would probably get some feedback which can influence your progression. So, those suggestions can be possible.

**9. Why do think it's important to have more standardized DM process methods?**

I think it's to find the right tool for the right process, not to make the mistake to try to find the truth with a method that allows a lot of flexibility because it tries to avoid bias and casualties. I think there could be more gain in more machine learning in healthcare than simply assume it can replace a doctor without proof. I think that's injustice toward the patient, but more importantly machine learning could be a better way to link multiple fields. I believe it's a good way to achieve common goals, but neither one is going to take over the other.

- 1. I have found that before evaluating the whole results, to use a pilot to test the model on a smaller scale or environment to measure its effectiveness, what do you think about this?**

As mentioned before, we do skip some steps and go right to pilot or sample size to see if there is an effect or feasibility. So, it can be placed before evaluating the results.

## **Interviewee A2**

- 1. Can you tell me about your function within the hospital?**

I'm an anesthesiologist, a product developer in the WKZ.

- 2. Have you previously worked on projects through data analysis and developing models like the VDS-project?**

Yes, I worked on a project that monitored data from infusion pumps and how to quiet them in pediatric oncology.

- 3. How involved were you in this project and how did you prepare?**

I mainly prepared through collaborating and remain informed in terms of content. I was the supervisor and part of the initiative and the lay out, not the statistician or the data analyst.

- 4. To which extent did you acquire knowledge relate to IT such as programming?**

Mainly by understanding the terms of the technical properties of IT. Also the data layers and the extraction but not programming.

- 5. How is your statistics knowledge?**

I only had it during my study, after that I didn't. I'm mathematically sufficient but no in-depth knowledge.

- 6. What is your opinion towards using big data to retract information and to which extent do you think hospitals will gain from this?**

I think big data is the future, in an appropriate manner. The doctor won't be obsolete, it will have an effect on the performance quality and consistency of data. There will be a gain when data will lead to action, currently 70 - 90% of alarms don't lead to action. By simplifying and concretizing information we can accomplish that.

- 7. Are you familiar with data mining processes? If yes, which?**

Yes, the CRISP-DM but not specialized in it. I have heard of other processes but not in terms of content.

- 8. Can you share your experience from the moment you started your data mining project?**

In the project with the pumps we saw a lot of data from the alarms. They were linked to certain actions i.e. when a nurse would be working on it. There were also delays from the alarms, from a third party like the parents or child. These would suppress the alarm which would cause a delay in signal and data.

A learning moment was that not everyone is looking for the same data.

**Show the MSP-DM method fragments and the overall method and discuss them:**

**9. During the first phase the domain goals and the situation of the project are established. How did it go for the VDS project?**

The end goal was to track the ventilation of the patient and improve it in a sensible manner.

**10. What was necessary to make a start with the VDS project, such as request for approval, data collection, draw up goals and a plan of action?**

We didn't need to request anything. It depends on the needs of the University of Utrecht, the student and the hospital. The hospital doesn't focus, the student just needs guidance and support.

**11. How were the goals composed for the project and data mining?**

An outline was drawn and written. The goals were set during the project not beforehand because of the data restrictions.

**12. I have found that there could be two separate objectives in a data mining project in the healthcare; clinical and managerial objectives. Do you agree with this?**

Yes, there could be different objectives for data mining projects such as those you mentioned but I am not that familiar with them. Nevertheless, I can imagine that there could be multiple various objectives in a data mining project besides the clinical aspect.

**13. I have found that there could be some legal constraints during a data mining project and I have added it in the concepts of constraints, what do you think about this?**

There are not much legal issues or constraints that needs to be taken into account. I did not really encounter them.

**13.1. I have found that scientific literature can benefit the projects, what is your opinion about this?**

Consulting scientific literature which needs to be done always when conducting a research by yourself. So, including this in getting a better understanding of the project can be helpful indeed.

**14. What do you find important in evaluating a model like the VDS project or ATPD vs BTPS project.**

It's important to me that it's physiological correct, otherwise I wouldn't trust it. By simulating it, it's possible to control whether the data is correct. This is only possible when you involve the practitioners when evaluating the model in order to find the physiological correctness or in convincing the practitioners of the model that would ease the adoption and trust of the model. For instance; putting all the information in the model besides the last 2 days of information and see whether it would produce the same outcome.

**14.1. Do you think that after evaluating the results that there are other possible actions such as moving back to the preparation or modeling phase?**

Yes, during the evaluation of the results the actions can be discussed with others and this may entail the actions you suggested.

**14.2. I have found that before evaluating the whole results, to use a pilot to test the model on a smaller scale or environment to measure its effectiveness, what do you think about this?**

I don't know what the standard procedure is during this phase but this seems logical.

**15. In the deployment phase, a few activities are mentioned. Are there activities lacking? If yes, which?**

Implementation research, which has an effect for the patient. ICT infrastructure to be able to transfer, with evidence otherwise it's just a hypothesis.

**16. What were obstacles you faced during the VDS project?**

Various things i.e. where do you start, which data is available and which is useful. How to get useful information and of course the vision of the project.

**Interviewee A3**

**1. Can you tell me something about your position within the hospital?**

I'm a pediatrician at the Wilhelmina's Children's Hospital and currently doing my fellowship. My goal is to be a pediatrician intensivist.

**2. Since when did you start with data analysis or do you have any experience with making models?**

Before I did my specialization during my clinical research I needed to conduct data analysis, wherein basic analysis was performed with models such as regressions.

**3. To what extent did was prior knowledge required in doing such analysis?**

At the beginning of my academic life, I started with bioinformatics. Herein, I was taught in scientific methods as well as statistics and research in which data is used for the analysis. In addition, during my PhD research I needed to collect data beforehand and do analysis on it.

**4. To what extent did you gain knowledge in IT related to matters such as programming?**

Early in my teenage years and before moving starting my medical career I did a lots of programming as a hobby. For example, making websites, analysis in databases, php, java, mysquel (all sql language), little C# and now R.

**5. How is your statistical knowledge?**

Before starting with the case study project, it was above average. And now it's more advanced because more complicated models are being developed.

**6. Did had some experience in data mining before the VDS project?**

Yes, but more on a distance level. Together with the ICT I participated in the development of dashboards, as well as doing analysis on the schedules of doctors.

**7. What is your opinion about the use of big data to extract knowledge / insight and to what extent do you think hospitals will benefit from this?**

There is a lot of potential of using data in a smart way. Big data can help doctors in decision making that will eventually benefit patients.

**8. When you started with data analysis, did you follow a certain process method to get started in a structured way?**

Not particularly. However, I did follow a structure that I regularly use for research which can be called the clinical scientific research method. Herein, I research question or hypothesis is defined first. Then the research is measured or assessed. Then, the data is collected and analyzed. After, data mining techniques can be used. In the meanwhile, writing down the steps how it is researched and results and then implementing the outcome.

**9. Are you familiar with data mining process methods? If yes which one?**

Not really, just heard about the CRISP-DM but never used it.

**10. Did you apply one of these methods yourself during the VDS project? If so, what did you think of this?**

No.

**11. Can you share your experience from the moment you started a data mining project?**

Many things are the same with doing research in general. However, the collection of data in hospitals is quite challenging because the hospital is not set to this kind of analysis.

**12. During the data analysis, what were the stumbling blocks that you experienced besides the technical aspects?**

The implementation can be difficult of making a good model and extracting good results. Accessing data can be a challenge as well. Moreover, safely analyzing data can also be a problem in which good data management is required. This applies working with privacy sensitive data in which it is important that it does not get lost neither that others can access it without the necessary permission.

### **13. What have you done to overcome these stumbling blocks?**

The issues of collecting data was overcome by identifying the stakeholders and confronting them in order to be able to access the data. In order to overcome some difficulties in implementation of analysis of models, it was mainly done by consulting with experts that can explain or help in a specific situation and also making goals more concrete and tangible.

### **14. What were your most important steps and activities from starting a data mining project till ending by the models?**

- Asking the question: why am I doing this?
- Getting authorization for accessing data
- Having pilot cases or training cases to get some experience and acquiring knowledge in preparing and analyzing the data.
- Having regular meeting in which the projects are discussed with other experts
- Collaboration with others in a team.

### **15. Was there certain expertise that was missing or that was helpful during your start of a data mining project? And is there still some sort of expertise still needed to be able to do better data mining projects?**

Having a basic or sufficient knowledge of statistic was helpful. The technical part of data preparation was missing at the beginning which was important. Moreover, it is important to be political aware within the hospital and knowing the structure as well knowing the people where to go if needed.

### **16. Can you specifically share your experience with the preparation of data and where you need to pay attention?**

It is quite important to learn the technical part of preparing data which can take much time during the process of preparing the data and learning how to do it from a technical aspect.

### **17. How will you from now on look at when beginning with a data mining project?**

It is very important to define the problem or research question and understanding how to answer it. Another part is to start early thinking of collecting data and identifying the right people where to ask data from.

### **Show the MSP-DM method fragments and the overall method and discuss them:**

Business understanding:

**17.1. I have found in the literature that there could be two separate objectives in a data mining project in the healthcare; clinical and managerial objectives. Do you agree with this?**



- As you know, I have just started in doing some data mining projects and I am still figuring out how things should be done. However, I can imagine that there could be different kind of objectives but currently I am just familiar with the VDS project which has a clinical purpose.
- Assess situation: identifying stakeholders, understanding the culture of the hospital is really important in order to get things done. There have been some obstacles in attaining the required data for the project, wherein several request were done to the responsible for retrieving the data but it was very difficult in attaining it until I found the right person who provided me the necessary means to retrieve and get the data. There is a lot of politics within a hospital and it quite important to get familiar with it in order to get things done.
- Produce project plan: can be optional
- Legal constraints: As you know that during the case study there were some privacy issues in sharing the data with you because you did not have the authorization in overseeing patient data and therefore I had to anonymize the data for you in order to make it possible for you to access it and to use it as well.

**17.2. I have found that scientific literature can benefit the projects, what is your opinion about this?**

There are multiple ways of finding certain information or answers that are applicable in your project. I think using scientific literature can be a part of this.

Data Understanding:

**17.3. During my case study I have found that discussing the origin of the data with you to be very beneficial in understanding the data, what do you think about this and how this fits in this phase as an activity?**

I think this helped you in better understanding the data, as well as the project overall. For people that are not familiar within the medicine domain I can see that this can be necessary.

- Explore data: it is important during this activity to visual when exploring as well as trying to identifying if causality can be found or indicated for further research.
- Verify data quality: verify the data and its quality with an expert and doctor in order to get some understanding or context.

Data preparation:

- Format data: it is important when working with others that don't have access to patient's information to see what the privacy issues are or ethical issues in order to depersonalize the data for further use with others.

**17.4. I have noticed that the order of preparing the dataset for modeling is a bit different, in which you first select, format, integrate, clean, and then construct the data. Is this correct?**

As you know I am still learning how data mining should be done and currently I am learning from mistakes and seeing how others do it as well as following courses online. So, I am not sure what order is the best to do, although I have followed the same order you described until now and it works fine with me.

Modeling:

**17.5. Do you find it useful in involving (clinical) practitioners during the model assessment?**

Yes, as you know I am a clinical practitioner and do have a good understanding of the medicine domain. However, I as you have seen I still need some help from others in understanding what I am exactly doing and if it's right. So, consulting with other experts during this process is import if you do not have that much experience or domain knowledge.

**17.6. Do you think that after evaluating the results that there are other possible actions such as moving back to the preparation or modeling phase?**

I think that those are good possibilities, although I am not sure about it because I do not have that much experience in other data mining projects.

**17.7. I have found that before evaluating the whole results, to use a pilot to test the model on a smaller scale or environment to measure its effectiveness, what do you think about this?**

This should be a good idea and probably wise. However, I am not sure if that's the standard procedure here.

Deployment:

- Produce final report: During the presentation it is important to show to the stakeholders that the quality will be improved, it is trustworthy information, it will be more efficient if implemented, don't show or explain it complicated and make it user friendly.
- Other remarks: if the success rate of the implementation is not high then it should be returned back to the first phase. Hence there should be are link between deployment and business understanding.

**Interviewee A4**

- 1. Can you tell me something about yourself and what your profession is? Your experience with Data mining / Data science in healthcare?**

I did artificial intelligence here at the UU and currently I am doing my phd in informatics / data science and work at the department of psychiatry of the UMC.

**2. What is your opinion about the use of big data to extract knowledge / insight and to what extent do you think hospitals will benefit from this?**

There is much that can be discovered. There are many unanswered questions in which big data can be used in finding the answers.

**3. Are you familiar with data mining process methods? If yes which one?**

Yes, I know the innovation funnel (it's a bit as scrum) and the CRISP-DM.

**4. Show the MSP-DM method fragments and the overall method and discuss them:**

In my research I have encountered a number of obstacles with regard to data mining in healthcare such as: Data integration, data quality, validation & analytical problems, Causal inference in observational data sets, Legal Issues, and User-friendliness.

How can we overcome these obstacles?

Data integration:

- at psychiatry department of the UMC, we have our own service that works with pipelines that enables easy access of data. This solves a bit the problem of data integration. However, before this it was quite difficult in collecting the data. The data manager was the one where you request the data from or doing it manually at the location.

Data quality:

- this can be indeed an issue which happens frequently at the healthcare. The main thing is that understanding what data you have on which analysis can be done and thinking out the box. Otherwise, not usable or bad quality data will be thrown out.

Validation & analytical problems:

- We see often that registration is not well filled in which makes it a bit difficult because the registration for many instances is not designed for research purposes and due to the workload mistakes can happen in the data. It is then a decision what data is useful or how can it be used. This part is not that different outside the medicine. The analysis needs to be well performed. It is necessary to be update with the validation methods and its applicability. Hence, it is important to have the people that understand such validation method or analysis because there can be methods that indicate that the results are positive and if you don't know that adjustment is needed in such method you will make a mistake in thinking that the results are positive.

Causal inference in observational data sets:

- Good question, but in my work at the hospital we do not really encounter such dilemmas of causality and focus more on the data that is present that we can understand and use it. This a bit out of our scope in our team. If such instances would occur we would rather focus first on the information what is

available. After that we could look how we can expand our research. However, consulting domain experts or the literature that is available in that field in finding some causal influences.

Legal issues:

- we have form called privacy risk assessment that is filled for the every project. However, because there can be multiple projects in a short run, we have created in our department a general privacy risk assessment that applies to other cases as well in which we use that as our form. Usually, this needs to be done for each project separately although practically this is not achievable. Therein, it explains how you handle the data and this needs to be approved.

**1. I have found that there could be some legal constraints during a data mining project and I have added it in the concepts of constraints, what do you think about this?**

Personally, we don't have that much problem with privacy issues since most of the time our data acquired is already anonymized and we don't use much poor raw data on which patient information is available. However, I can imagine that perhaps for other projects there could be some legal constraints.

**2. I have found that it is important to anonymize the data when working with others that do not have permission in accessing it, what do you think about this?**

Yes, if you do not have permission in accessing the data you can only retrieve it if its anonymized. However, as mentioned earlier we do not encounter such issues since such data is most of the time already anonymized.

User-friendliness:

- we always try to explain the findings to those who are involved as simple as possible on how it works or why it works. You try to give some insight of the findings to the practitioners but in reality it is quite difficult that they will understand it what kind of statistics or modeling is behind a certain analysis because it is quite complicated. It is quite important that there needs to be some prior knowledge or current knowledge that will help in making certain decision and if this exist that will bring some trust in your decision. Therefore, if trying to explain certain results than this should be kept in mind.

**5. Are there any other obstacles that are not been mentioned above but are important? Could you also explain how you deal with them? No.**

**1. Show the MSP-DM method fragments and the overall method and discuss them:**

Business understanding: Remarks

- Domain understanding: perhaps change the name of the phase to domain understanding which fits the healthcare better because it also has a clinical perspective and provides a broader view within analyzing a domain.
- Clinical and managerial objectives: there could be clinical as well as managerial objectives within a data mining project in the healthcare.
- Privacy risk assessment: is a lot of work but it is important and should be more highlighted here perhaps in the assess situation.

**2. I have found that scientific literature can benefit the projects, what is your opinion about this?**

Yes, I agree that scientific literature is missing here which can be consulted during a project.

Data Understanding: Remarks

- Assess data lineage and or assess registration: knowing the process how data is stored and collected will help you in getting a better understanding of the data or finding interesting insights that can help in the analysis. This can be in a form of retrieving the data lineage of a particular dataset.

Data preparation: Remarks

- The order is first selecting then formatting, integrating, cleaning and then constructing.

Modeling: Remarks

- Model assessment: In the model assessment, the clinical experts should be involved in evaluating the selected parameters or doing it together. This can be beneficial because those clinical experts have a better understanding of these parameters and its relations and therefore can assist in selecting or revising the parameters that will be used for the model.

Evaluation: Remarks

- Assessment of data mining results: assessing with various experts is needed be it with clinical or technical experts or other data scientists if the model is suited for a particular case.

**3. Do you think that after evaluating the results that there are other possible actions such as moving back to the preparation or modeling phase?**

Yes, those actions can be taken if needed.

**4. I have found that before evaluating the whole results, to use a pilot to test the model on a smaller scale or environment to measure its effectiveness, what do you think about this?**

This is probably correct. Testing the model on a certain setting seems logical. In our cases, we do not have really implementation assignments or projects but are more exploratory driven. So, we do not test our models in certain environments but use only the data that has been provided or retrieved.

## Interviewee A5

### 1. Can you tell me something about yourself and what your profession is? Your experience with Data mining / Data science in healthcare?

I come from a bioinformatics background and now I am working as a data scientist here at the UMC. Moreover, I am also a IT developer. I have been in multiple projects related to data science in the hospital and also in the development of the data flow infrastructure in order to enable better accessing of data for research.

#### 1. What is your opinion about the use of big data to extract knowledge / insight and to what extent do you think hospitals will benefit from this?

It is immensely important, because there is a lot of data being collected and with the technology that we have today the possibilities great. The benefit can be for hospitals in enabling better decision making which will at the end benefit patient care.

#### 2. Are you familiar with data mining process methods? If yes which one?

I know the CRISP-DM a bit as well as the innovation funnel. But never really used neither of them but follow kind of my own method.

#### 1. Show the MSP-DM method fragments and the overall method and discuss them:

In my research I have encountered a number of obstacles with regard to data mining in healthcare such as: Data integration, data quality, validation & analytical problems, Causal inference in observational data sets, Legal Issues, and User-friendliness.

How can we overcome these obstacles?

#### Data integration:

- if there is not a good data flow infrastructure present within a hospital then there is nothing much do be done but to create one. This cannot be done alone but in cooperation with the management in order to make it a priority. Otherwise, the traditional way needs to be done which is going to the data managers and requesting certain data or locating the departments and collecting it manually which can be really time consuming. However, the UMC is currently working on a better infrastructure that will make this process easier. Moreover, if there is inconsistency or instability within the data first you need to ask yourself why this is the case and what you can do with it.

#### Data quality:

- it's not that different from other industries on how to deal with such problems. Here, you need to think what are the consequences are and what can you use.

#### Validation & analytical problems:

- Validation is a bit more important in the healthcare because there decisions being made about the state of people and having the correct information or analysis is quite important.

Causal inference in observational data sets:

- Experiments can be conducted in trying to retrieve causal inference or with statistical means.

Legal issues:

- the patients Id should not be traced but there is not much limitation with it because usually the dataset can be easily be requested in which they are anonymized. However, this also depends if you know your way around because the data manager for me is nearby and we are in good relation.

**2. I have found that there could be some legal constraints during a data mining project and I have added it in the concepts of constraints, what do you think about this?**

Yes, there could be legal constraints although as earlier mentioned anonymized data can be easily requested here which makes it easier to avoid such constraints.

**3. I have found that it is important to anonymize the data when working with others that do not have permission in accessing it, what do you think about this?**

Here we really do not encounter it, although yes if you have no authorization in accessing directly patient data then indeed it needs to be anonymized first upon receiving it.

User-friendliness:

- When there is output from data mining which needs implementation then it is important to be able to explain why certain predictions are made and why they are correct because clinical practitioners are heavy argument based and need good evidence.

**3. Are there any other obstacles that are not been mentioned above but are important? Could you also explain how you deal with them? No.**

**1. Show the MSP-DM method fragments and the overall method and discuss them:**

Business understanding: Remarks

- Assess situation: I miss here the scientific literature when assessing the situation, because it is highly possible that for certain cases it can be scientifically written or done before. Hence, consulting such literature can be beneficial when conducting a data mining project when assessing the situation.
- Produce project plan: The privacy risk assessment needs to be added here which is needed for starting a project. It is required for each project that uses patient data to fill in this assessment before getting approval in using certain data.

- Objectives: there could be various types of projects in respect to data mining activities such as performing an analysis on administrative matters within the hospital, as well as clinical research.

#### Data Understanding: Remarks

- Assess data lineage: the activity of assessing the data lineage needs to be added in order to get to know the data and its origin as well how the registrations were done for better insight.

#### Data preparation: Remarks

- The order is a bit different here compared to the CRISP-DM. Here, we start with select data, format data, integrate data, clean data and then construct data.

#### Modeling: Remarks

- Model assessment: involve experts in checking what the important factors are or variables within a model. This is important because they have greater insight within the domain which can help in identifying the right parameters.

#### Evaluation:

**2. I have found that before evaluating the whole results, to use a pilot to test the model on a smaller scale or environment to measure its effectiveness, what do you think about this?**

- Before evaluating the model, there should be a setting in which the model is tested. And as you mentioned this can be similar to a pilot setting in a particular environment. This is recommended because this will provide you some initial understanding in regards to the effectiveness or usefulness of the model.
- Assessment of data mining results: Assessing the data mining results should be done by involving experts as well clinical practitioners in evaluating the model in order to receive feedback that can help in improving the model or finding faults.
- Produce final report: this activity fits better in this phase than in the deployment phase because during the deployment in which the results are already implemented it is not necessary to make a final report. However, doing it before it is more appropriate after reviewing the process.

**3. Do you think that after evaluating the results that there are other possible actions such as moving back to the preparation or modeling phase?**

- Yes, there could be more actions after evaluating the model. And I agree that you can move back to the modeling phase or to the data preparation phase if other models needs to be constructed after the feedback or adjusted or the model can be accepted.

#### Deployment:

- Review end users: at the end of this phase reviewing with the end users is missing. It is important to get their feedback after the results are implemented. This feedback can be asked every month or half year or week. This depends of course how often the results of the data mining projects are used by the end users. This could mean that the feedback received can mean



that more insight is gained about the initial problem of the data mining project which could mean that from the deployment to the business understanding an action can be performed in which changes are necessary in this phase.