

**A UNIFIED STRUCTURED PROCESS MODEL FOR
HEALTH ANALYTICS**

SUPUNMALI AHANGAMA

(B. Sc. IT (Hons), University of Moratuwa, Sri Lanka)

**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF INFORMATION SYSTEMS
NATIONAL UNIVERSITY OF SINGAPORE**

2015

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Supunmali

Supunmali Ahangama

22/01/2015

ACKNOWLEDGEMENT

First and foremost, I wish to express my sincere gratitude to my supervisor, Prof. Danny Chiang Choon Poo for his guidance, valuable advices, constructive criticism and continuous support given in completion of my doctoral research study and I admire his patience, motivation and enthusiasm shown in guiding me throughout my graduate study program at the National University of Singapore.

I am grateful to all the members of the staff and my colleagues and my friends in the Information Systems Laboratory and the School of Computing at the National University of Singapore for encouraging and assisting me in my research and helping me in various other ways.

I thankfully acknowledge the National University of Singapore for granting me the research scholarship and giving me the opportunity to pursue my graduate studies in this prestigious National University of Singapore.

Last but not least, I would like to thank my parents and my brother for their love, inspirations and continuous support extended during this intensive learning period and in every step of my life.

CONTENTS

	Page
DECLARATION	i
ACKNOWLEDGEMENT.....	ii
CONTENTS	iii
SUMMARY	vi
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF ABBREVIATIONS	x
CHAPTER 1. INTRODUCTION.....	1
1.1 Problem Definition.....	4
1.2 Objectives and Significance of the Process Model Development	6
1.3 Uniqueness of Medical Data Mining	9
1.4 Definition of Health Analytics Process Model	11
1.5 Organization of the Thesis	16
CHAPTER 2. LITERATURE REVIEW.....	17
2.1 Software Engineering Frameworks.....	17
2.2 Data Mining Frameworks	19
2.3 Unified Modelling Language with Data Mining.....	26
2.4 Health Analytics Frameworks.....	28
2.5 Supporting Dimensions.....	33
2.6 Application of Related Work in the Proposed Model.....	37
2.7 Summary	39
CHAPTER 3. METHODOLOGY.....	40
3.1 Introduction.....	40
3.2 Design Science Research Approach	41
3.3 Research Process.....	44
3.4 Summary	49
CHAPTER 4. SURVEY STUDY.....	50
4.1 Introduction.....	50
4.2 Conceptual Background.....	51
4.3 Research Model and Hypotheses	54
4.4 Research Methodology	59
4.5 Data Analysis and Results	62
4.6 Discussion.....	65

4.7	Application of the Survey Results in Process Model Development	67
4.8	Summary	69
CHAPTER 5. DEVELOPMENT AND EVALUATION OF THE PROCESS MODEL.70		
5.1.	Introduction.....	70
5.2.	Research Setting.....	73
5.3.	Initial Model Development-Evaluation by Applying the Model in an External Project	74
5.3.1.	Case Description	75
5.3.2.	Application of the Process Model in <i>Hospital X</i> Project	75
5.3.3.	Revisions to the Model	86
5.3.4.	Revised Model	91
5.4.	Final Model Development-Evaluation in a Hospital while Working as an Intern...92	
5.4.1.	Case Description	93
5.4.2.	Project Variations in the Case Organisation	95
5.4.3.	Application of the Process Model in General	96
5.4.4.	Application of the Process Model in Complex and Ambiguous Projects (Project A) in <i>Hospital Y</i>	104
5.4.5.	Application of the Process Model in Simple and Clear Projects (Project B) in <i>Hospital Y</i>	107
5.4.6.	Revisions to the Model	108
5.5.	Evaluation Outcome.....	111
5.6.	Summary	112
CHAPTER 6. PROCESS MODEL FOR HEALTH ANALYTICS.....113		
6.1.	Introduction.....	113
6.2.	Overall Structure of USAM	114
6.3.	Process Management of the USAM.....	115
6.3.1.	Step 1: Project Initiation	117
6.3.2.	Step 2: Domain Understanding	119
6.3.3.	Step 3: Data Understanding	127
6.3.4.	Step 4: Conceptualization	131
6.3.5.	Step 5: Data Preparation	134
6.3.6.	Step 6: Data Modelling	135
6.3.7.	Step 7: Validation	136
6.3.8.	Step 8: Presentation of the Data Model.....	137
6.3.9.	Variations to the Methodology	140
6.4.	Project Management	142

6.5.	Communication Management	148
6.6.	Knowledge Management	149
6.6.1.	Technical Documentation Approach.....	151
6.6.2.	Extending UML Diagrams.....	158
6.7.	Discussion	160
6.8.	Summary	162
CHAPTER 7. DISCUSSION AND CONCLUSION.....		164
7.1.	Implications.....	166
7.2.	Limitations	168
7.3.	Future Work.....	170
7.4.	Conclusion	172
BIBLIOGRAPHY		174
APPENDIX A.....		183
APPENDIX B.....		185
APPENDIX C.....		189
APPENDIX D.....		190

SUMMARY

Health Analytics (HA) is the use of statistical, predictive, quantitative and various other models on healthcare data in informed healthcare decision making. The progress in HA has been curtailed due to issues such as user resistance, essential dependence on the skills and experience of a data analyst and approaching HA in an ad hoc manner. These problems could be addressed through a well-designed analytic process model tested specifically in healthcare context. Such a process model will facilitate the performance of all the relevant projects as a structured process, with clearly defined objectives, proper project planning and with systematically documented prior knowledge, data, methodologies and results. Numerous examples and possible best approaches could be drawn from data mining and software engineering projects. Most of the existing methodological approaches of data mining such as CRISP-DM, SEMMA etc. are not been popularly utilized by users.

Thus, a unified structured analytics model is proposed in this research which could be easily adopted even by analysts with limited skills. The model was developed by synergising prior knowledge from literature and predetermined requirements of the users in healthcare context. The ultimate objective was to assist the novice data analysts to develop a strong sense of the nature of the target HA task as well as to provide them with a clear effective strategy to perform the analytic process. The proposed process model is developed based on four dimensions, namely, (1) process management, (2) project management, (3) knowledge management and (4) communication management where, the latter three dimensions are considered as supporting dimensions for process management. With the elements of the input/output and tasks of each stage in the process model, visual diagrams using UML are proposed from domain understanding to deployment of the HA project.

Available published literature on behavioural and software engineering research was examined to conceptualize the problem. Initially, a survey was carried out to determine the factors affecting novice user's intentions to use a methodology for analytics. The core of the project is the construction of the process model (as a method). It is presented as the design artefact of this Design Science Research (DSR) based study. Finally, the application of the model is validated using the action case research approach while working as an intern in a large hospital in Singapore. The development of the process model for HA and proposing a methodology for constructing and evaluating the process model can be considered as the major contributions of this study.

LIST OF TABLES

Table 1: Correlations between constructs and the dependent variable	62
Table 2: Terminology related to radiation oncology.....	78
Table 3: Factors influencing fraction duration of radiation treatment	80
Table 4: Re-categorization of activities at the Radiology Department, <i>Hospital X</i>	82
Table 5: Mean fraction duration over radiation treatment intent and activity	83
Table 6: Comparison of models	86
Table 7: Satisfying the design criteria by the model design in <i>Hospital X</i>	87
Table 8: Model improvement satisfying the limitations in the design criteria observed in <i>Hospital Y</i>	109
Table 9: Variations in USAM for Project B	141
Table 10: Summary details of application of DSR guidelines	165
Table D. 1: Modified dataset.....	207

LIST OF FIGURES

Figure 1: Use of data mining methodologies (in %) (KdNuggets.com 2014)	21
Figure 2: CRISP-DM model (Chapman et al. 2000).....	23
Figure 3: Data mining knowledge discovery process (Cios and Moore 2002)	30
Figure 4: The design science research method applied to HA process model development	45
Figure 5: Development -evaluation process	49
Figure 6: Research model for the survey study	54
Figure 7: Results of hypothesis testing.....	64
Figure 8. Revised CRISP-DM model	92
Figure 9: Project types classification	96
Figure 10: High level process model of the USAM	115
Figure 11: Methodological steps of the USAM.....	116
Figure 12: Activities in the domain understanding stage.....	121
Figure 13: Activities in the data understanding stage	128
Figure 14: Activities in the conceptualization stage.....	132
Figure 15: Activities in the data preparation stage	134
Figure 16: Activities in the data modelling stage.....	136
Figure 17: Activities in the validation stage	137
Figure 18: Activities in the presentation of the model stage.....	138
Figure 19: Overview of the UML diagram used	158
Figure 20: UML profile extension.....	160
Figure 21: UML profile extension for actor	160
Figure D. 1: Business use-case diagram	199
Figure D. 2: Business goal diagram	201
Figure D. 3: Analytics use-case diagram.....	203
Figure D. 4: Analytic goal diagram.....	205
Figure D. 5: Data diagram.....	206
Figure D. 6: Data preparation activity diagram	208
Figure D. 7: Technique diagram for identify patient treatment profile	209
Figure D. 8: Technique diagram for identify factors influencing treatment duration	210
Figure D. 9: Algorithm diagram – OLS regression	210
Figure D. 10: Analytic model diagram for identifying factors influencing treatment duration.....	211
Figure D. 11: Technique diagram – Predict treatment duration using regression model	211
Figure D. 12: Technique diagram – Predict treatment duration using decision tress	212
Figure D. 13: Algorithm diagram – GEE regression.....	212
Figure D. 14: Analytic model diagram – GEE regression.....	213
Figure D. 15: Validation diagram	214

LIST OF ABBREVIATIONS

BI – Business Intelligence
CRISP-DM - Cross Industry Standard Process for Data Mining
DOI - Diffusion of Innovation
HA – Health Analytics
HIMSS - Healthcare Information and Management Systems Society
HIPAA - Health Insurance Portability and Accountability Act
ICT – Information and Communication Technology
IS – Information Systems
MST - Media Synchronicity Theory
PHI - Protected Health Information
SEMMA - Sample, Explore, Modify, Model, Assess
TAM – Technology Acceptance Model
UML – Unified Modelling Language
USAM – Unified Structured Analytic Model

CHAPTER 1. INTRODUCTION

Analytics has gained a great deal of importance in Information Systems (IS) research with the tremendous advancements in social networking, mobile technology, remote sensing technology and electronic health records. The progress has been hindered by issues such as provider resistance, availability of data in heterogeneous sources and unstructured and ad hoc approach to analytics (Marban and Segovia 2013; Yang and Wu 2006). Moreover, the accuracy of results depends entirely on the skills of the data scientist working on the data. These limitations could be attributed to undefined project objectives, non-availability of user-accepted methodologies and also to lack of systematic documentation. Most of these issues could be solved in a systematic way by adopting suitable methodologies leading to timely, cost effective and pragmatic solutions (Mohan and Ahlemann 2011).

Numerous examples and possible best approaches could be drawn from data mining and software engineering projects (Marban et al. 2009b). Several authors have proposed methodologies and documentation approaches for such projects. For example, CRISP-DM (Chapman et al. 2000), SEMMA (SAS 2008), DM-UML (Marban and Segovia 2013) and other specific approaches for each data mining technique (Luján-Mora et al. 2006; Prat et al. 2006; Zubcoff and Trujillo 2006) have been proposed. However, these approaches had not been diffused into the general population of analysts. Leading reason highlighted by organisational theorists is that any new methodology usage is resisted by individuals as it does not meet their needs (Mohan and Ahlemann 2011). According to these authors, this is due to the failure of methodology developers to consider the individual attitudes towards methodology

use. Thus, it is important to recognise what characteristics in a method drive the individual users for its deployment. Moreover, the developed methodologies should be tested in a given context. As these criteria are applicable for Health Analytics (HA) as well, a study was carried out to develop a unified methodology for analytics with prior recognition of user requirements and testing it in HA context.

According to the definition of Healthcare Information and Management Systems Society (HIMSS), Health Analytics is “the systematic use of data and related business insights developed through applied analytical disciplines (e.g. statistical, contextual, quantitative, predictive, cognitive, and other [including emerging] models) to drive fact-based decision making for planning, management, measurement and learning in the healthcare industry. Analytics may be descriptive, predictive or prescriptive” (Cortada et al. 2012). This definition (of HIMSS) is adopted from the definition of analytics given by IBM. HA can be further described as a “way of transforming data into action through analysis and insights in the context of healthcare decision making and problem solving” (Raghupathi and Raghupathi 2013).

HA applications can be defined as “collection of decision support technologies for the healthcare provider aimed at enabling knowledge workers such as physicians, nurses and health officials, health policy makers and pharmacists to gain insight knowledge and make better and faster health decisions” (Raghupathi and Raghupathi 2013). HA applications allow a healthcare system to be more efficient (improved outcomes, improved coordination, reduced time and cost, and better value) by providing constant or better quality care. However, most of the current health IT systems are deployed in clinics merely to assist physicians to diagnose and treat patients rapidly, without taking the need to integrate and aggregate data for analysis and reporting into account. This indicates that, there is a necessity to use HA in

healthcare industry to enable personalized healthcare, to predict health behaviour and to initiate clinical improvements by discovering new insights hidden in healthcare data (Chen et al. 2012). However, due to the special status of medicine, certain tests may not be performed, certain questions may not be asked or certain conclusions may not be made (Cios and Moore 2002). Thus, decisions should always be supported with valid justifiable explanations (Bellazzi and Zupan 2008) and certain significant relationships found may not be bio-medically valid.

HA is anticipated to be pervasive across clinical healthcare delivery, personal health management and public health promotion. Even though a paradigm shift from volume based to value based healthcare through HA could be expected within next few years (Horner and Basu 2012), it has been hindered by lower user acceptance of such methodologies. Usage of a unified methodology will improve the process and output of HA. Such a process model will facilitate performance of all the inclusive projects as a structured process by dividing a complex process of HA into plausible and coherent steps (Chan and Thong 2009; Fitzgerald 1996), with clearly defined objectives, proper project planning and with systematic acquisition and documentation of prior knowledge, data, methodologies and results (Bellazzi and Zupan 2008). In the health analytics process model proposed by Raghupathi and Raghupathi (2013) too, the specific methodologies and relevant documentation approaches for each stage have not been proposed.

The unified structured health analytics process model proposed in this thesis will facilitate the performance of analytics without much difficulty, independent of skills of the data scientist while providing a systematic documentation as a communication tool for various stakeholders in this sector. This process model will

avoid any duplication of the tasks and will enable traceability while assisting result oriented effective project management.

User requirements of a HA process were examined using Design Science Research (DSR) approach (Hevner et al. 2004) prior to the development of this process model. Based on these user requirements, the model was developed under four dimensions, namely, process management, project management, communication management and knowledge management. The latter three dimensions are the supporting dimensions for process management dimension. Applicability of the proposed process model in a real world scenario will be illustrated using an example. The following sections will provide problem definition, scope and aims of the proposed unified structured process model.

To avoid any confusion, the ‘data model’ will be used specifically to refer the output generated by applying various data analytic methodologies on data. The ‘analytic process model’ refers to the approach followed to develop the data model.

1.1 Problem Definition

While the two fields, data mining and software engineering grow parallel to each other, data mining is still behind software engineering as it focuses primarily on design methodologies while software engineering is focussing on programming environments, automated programming, software quality, human resource management etc. (Marban et al. 2009b). Most of the existing analytic projects are performed in an impromptu manner without addressing proper project, communication and knowledge management. As analytic projects progress and become too complex, the need arises for a standardized process model. Even though, there are several data mining methodologies available (e.g. CRISP-DM), only a handful of organisations are using such methodologies and in many cases those

methodologies are failing to meet the specific needs of the users. Moreover, these methodologies have ignored to consider the organisational and other corresponding activities not directly related to data modelling (Marban et al. 2009a). It is important to note that data analytics and data mining have been used interchangeably in the literature. While data analytics deals with the complete process consisting of insights generation from data and communication of them to recommend actions (Cortada et al. 2012), data mining is a particular tool used solely for determining the relationships in data.

According to a critical study carried out by Fitzgerald (1996), the lower user acceptance of some methodologies could be due to several reasons, namely, (1) individuals simply ignoring the newly introduced methods; (2) existing methodologies treating the analytic process as an orderly rational process even when it is not; and (3) assuming process models to be universally acceptable though they should be modified to suit the application context in real world scenarios. Neglecting such critical factors in developing a process model may lead to its rejection by data analysts. Thus, in developing a model, it is important to consider such factors influencing its acceptance among practitioners and to evaluate its applicability in a given specific context.

Similarly, the authors Bellazzi and Zupan (2008) have indicated the importance of having process models specific to a particular problem domain. In this thesis study, the development of the model would be for HA context. As HA market which was worthy of \$3.7 billion in 2012 and is targeted to reach a worth of \$10.8 billion by 2017 growing at a rate of 23.7% (Osborne 2012), HA is decided to be used as the context to focus on in this thesis. Moreover, it is decided to select healthcare

domain as it will be a part of patient care and it will be the most rewarding of all to analyse effectively (Cios and Moore 2002).

In summing up, it could be stated that the omission of paying attention to the user needs in developing the process models, in built rigidity of the processes, failure to design the projects based on individual user requirements (all existing models are one fit all projects), exclusion of support elements such as project management, communication management, knowledge management (present models are focussing mainly on model development), and failure to modify to suit a specific application context are the weaknesses of existing analytic processes. Experienced data analysts may develop their own personal approaches to mitigate these shortcomings. However, the lack of an applicable methodology to the data analytic process will put novice users in a difficult situation of having to face a steep learning curve. This thesis work has been carried out to develop a unified structured process model addressing these important issues based on the following research questions.

Research Questions:

(1): What methodological steps are needed to be followed by a novice user in health analytics?

(2): How supporting dimensions (project management, communication management and knowledge management) are utilized in a HA project based on user requirements?

1.2 Objectives and Significance of the Process Model Development

The main aim of this thesis was to propose a new Unified Structured Analytic Model (USAM) to perform HA projects in a standardized way so that, individual data analysts will be able to carry out better quality HA projects with control and with less time and effort (Mohan and Ahlemann 2011). As specific protocols available and

followed in medicine, having a standardised process model for data analytics will guide healthcare data analysts through the analytics process where some steps could have been neglected if performed in an ad-hoc manner (Bellazzi and Zupan 2008).

Healthcare being a dynamic and patient centric field, the user requirements to be considered in developing the analytic model can vary with time and new developments. In such cases, depending on a rigid process model for data analysis required for efficient improved management may become counter-productive. The process model to be used should have the flexibility to alter based on changing requirements of healthcare institutes. The proposed methodology is developed to meet such needs of the healthcare personnel (and HA data analysts) thus enabling its acceptance by them for effective use. Also an evaluation of the developed model for user acceptance too is important to identify the unforeseen shortcomings which may lead to user resistance and to make necessary improvements to the model.

This study was carried out with the following objectives formulated upon a critical review of the available literature.

- to identify the best practices in software engineering and data mining process models and determine their applicability to HA
- to determine the factors affecting the intention to use a process model by novice users in HA
- to propose a HA process model as a complete process model
- to determine the subsequent variations to the HA process model based on user requirements
- to evaluate the applicability of the model through an action case base approach in a hospital in Singapore

Responding to the research questions and objectives mentioned, the process model was developed using a Design Science Research approach. It was used for the explanation of the problem and the related theoretical principles for the proposed model and to develop the new artefact. Design Science Research approach was selected as the methodology to address two concerns in IS research (Arnott 2006). Determining the role played by IT artefact in IS research (Orlikowski and Iacono 2001) and determining the reasons for the lower professional applicability of many IS related studies (Benbasat and Zmud 1999) were the two specific concerns. While the method developed was used as the unit of analysis, research outcomes were evaluated in an organisation context. The process model developed can be considered as the design artefact type 'method'. A participatory research approach (action case based approach) was used as the evaluation strategy (Baskerville and Pries-Heje 2014) to incorporate social and technical needs of the users.

The findings of this study will have a significant impact on both theoretical discourse and the practical discourse of HA. First, this unified structured analytics model may be used as a standardized process and as a reference model to provide a better understanding of the flow of the HA process. This will offer a clearer comparison of existing and future models. While this process model allows an uncomplicated performance of HA without having to depend on the skills of the data scientist, it provides a systematic documentation as a communication and knowledge management tool for various stakeholders in this sector too. Secondly, the model highlights the determinant factors affecting the user acceptance of novel methodologies. HA methodological attributes that lead to acceptance among novice users were determined through several behavioural theories. The method used to

develop and evaluate the process model using Design Science Research approach too can be considered as a contribution from the study.

As the scope of the study, we decided to develop and evaluate the process model focussing on the healthcare context. It is important to note that the model developed will be a generalised model applicable to any analytics context though it had been evaluated specifically in HA context. The main target users of the study will be novice users who are new to a HA project carried out in a healthcare institute. However, even an experienced user can use relevant components of this model or even the analysts who are not involved with a healthcare institute or a major project can use it as a reference.

1.3 Uniqueness of Medical Data Mining

In a study on uniqueness of medical data, Cios and Moore (2002) have identified (1) heterogeneity of medical data (Kwiatkowska et al. 2007), (2) ethical, legal and social issues, (3) statistical philosophy to address heterogeneity of data and social issues and (4) special status of medicine as the four key factors that differentiate it from other data.

First, medical data is voluminous and is collected from various sources (images, patient interviews, physicians' notes, and biomedical data) (Bellazzi and Zupan 2008). Though the standard HL7 (v3.0, RIM): international health informatics interoperability standards provides a framework for retrieval, integration, dissemination and sharing of electronic health information; processing of numerous data types and integrating them into a single repository is a major concern (Esfandiary et al. 2014). Medical data is complex and difficult to analyse compared to financial data that is well organized and could be easily used by automated analysis systems (Bellazzi and Zupan 2008). Specially, case notes with physician's interpretation of

clinical data are unstructured, ambiguous, not standardized and are using different grammatical constructions varying from one physician to another.

Moreover, no canonical form (a standard form of representation for data) exists even for most simple concepts in medicine (Cios and Moore 2002). Thus, the tabulation and indexing of equivalent concepts together becomes tedious. ICD-X (latest version is ICD-10): international classification of diseases, NANDA-II: Standardized nursing language and classification of diagnoses, SNOMED CT: systematically organized clinical terminology, and MEDCIN: proprietary medical vocabulary allow a consistent form of expression of diagnosis.

Many other complex ideas like logical quantifiers (e.g. for every, for some), conditionals (if there is... else...) and logic operations (e.g. logical-and, logical-or and logical-not) are yet to be standardized into a consistent form. Another difficulty associated with heterogeneity of medical data is the inability to be characterized mathematically like many other types of data where formulas, models can be effectively applied in determining the relationships.

Second, with medical data, there are complications on (1) data ownership as data is scattered in different health establishments distributed in multiple geographical locations, (2) privacy and security as it could infringe patient confidentiality and damage patient-doctor relationship (it is essential to conceal individual identifiers when sharing and only authorized persons are allowed to access them) (Li and Qin 2013; Yoo et al. 2012) and (3) rigid administrative guidelines (e.g. IRB-Institutional Review Board, privacy rules in HIPAA of USA) (Chen et al. 2012). Such administrative policies are normally not required for non-medical data mining.

Third, it is important to consider the statistical philosophy related to medical data. For example, there may be common or rare occurrences of certain medical

events. They need to be clarified by a domain expert. The data is collected (or not collected) to use for the patient care and not as a source of data for research. Thus, the data collection will be narrowly focused and may be incomplete and imprecise (Eggebraaten et al. 2007). Furthermore, most of the datasets are small in number of data points (instances) but they will be having thousands of data attributes compared to other standard datasets (Bellazzi and Zupan 2008). Thus, it is important to find means to handle these attributes (e.g. dimensionality reduction).

Similarly, there will be incomplete, missing, inconsistent or redundant values. For example, during a patient's visit to a doctor, certain tests may not be performed as the patient is weak. As such, the data set could be incomplete. As another example, there could be mutually exclusive categories (e.g. male patient having positive pregnancy test results) mentioned for a certain data point. Moreover, it is important to consider the comprehensibility of the data models generated (Schmidt et al. 2008) (e.g. decision trees vs. artificial neural networks). In HA it is essential that decisions should always be supported with valid justifiable explanations (Bellazzi and Zupan 2008) as these applications are deployed in a safety critical context.

Fourth, due to the special status of medicine, certain tests may not be performed, certain questions may not be asked or certain conclusions may not be made (Cios and Moore 2002). The outcome of the healthcare will lead to a life-or-death situation.

1.4 Definition of Health Analytics Process Model

The Unified Structured Analytic Model (USAM) is developed to carry out HA projects in multidisciplinary fields following a methodical procedure for knowledge discovery. This process model includes a set of processing steps that should be followed by HA practitioners and researchers involved with healthcare projects. A

methodology can be described as an instance of a process model (*'what to do'*) with sets of inputs, outputs, tasks and specifications on *'how to perform'* a certain activity (Mariscal et al. 2010). Conversely, in the literature the terms, 'process model' and 'methodology' have often been used interchangeably (Marban et al. 2009a). Thus to avoid confusion, in this thesis both terms will be used loosely while the former will be specifically used when the broad view of the proposed process model is considered and the term methodology will be used in referring to the exact steps and tasks in the process.

The developed model will follow an iterative and incremental life cycle based on agile approach. This non-trivial process will be documented with standard notations for repeated usage and to provide support for novice users to ease the learning curve of these projects. This process model will allow flexibility to alter steps (based on organizational objectives, project requirements and project limitations allowing for creativity) rather than restricting them to a rigid structure where even the unwanted phases or components have to be followed as per the process model.

The main dimensions of the process model are:

- Process management – Considers overall structure of the HA process and the activities managing different phases of HA. A comprehensive, generalized and structured process model will allow smooth adoption of the model in a specific context. It will focus on the technical component of the data analytics process (where a data model will be developed as output from data gathered).
- Project management – Considers the management of resources and task coordination of a HA project.

- Communication management – Considers different types of stakeholders in a HA process and their requirements when collaborating and communicating with other stakeholders.
- Knowledge management - Considers the support available for knowledge capture, retention and transfer. Knowledge management will be required from initiation till the end of the project considering the amount of information and knowledge generated in the process.

To address the uniqueness of medical data mining compared to standard data mining (Cios and Moore 2002), the USAM process model is developed considering the following factors.

- Heterogeneity of medical data –
 - As a solution for the non-standard representation of medical data, standards such as SNOMED, ICD 10 are introduced to represent medical data (Bellazzi and Zupan 2008). Specially, electronic medical records are represented using these standards. Even the medical data extracted from numerous other sources like case notes and images needs to be codified using these standards. Thus, the ambiguity existing in medical data sources could be avoided.
 - Considering the difference in training, knowledge and approaches existing among the medical professionals and computer scientists, heterogeneity of data sources is a barrier to work across these professions (Schmidt et al. 2008). It is necessary to gain domain knowledge for data understanding and model results understanding and clarification by continuous collaboration and consultation with domain experts. Experts from both domains (physicians and data analysts) need to inspect the data set and

clarify the content. Thus, the communication management and knowledge management are important.

- Visual representation is also a worthy approach to reduce the knowledge divides between the two professions. Visual representation of user requirements to output results makes it easy to comprehend (Schmidt et al. 2008).
- Ethical, legal and social issues – De-identification and anonymization of patient data is done using the HIPAA standards. Usually the access to the dataset will be authorised only for a specific time period based on the data analyst's requests. Gaining prior approval from the relevant internal review board for access to the data before commencement of a major project is very important.
- Statistical philosophy
 - There will be a high volume of attributes in a medical dataset even though numbers of instances are less (Bellazzi and Zupan 2008). Thus, it is important to consider the feature selection strategies. All the attributes should be maintained in the dataset and only a group of attributes will be filtered, depending on the type of analytics to be performed. Then the model should be conceptualized using the selected attributes as the use of all the attributes in data analytics is not advisable and not possible (Bellazzi and Zupan 2008).
 - Quality of medical data is inferior compared to other datasets. There will be many missing and incomplete data. Thus, data pre-processing is an important component in handling medical data.
 - There may be certain redundant, insignificant and inconsistent data instances and attributes (Cios and Moore 2002). To avoid such data

objects, it is important to get expert advice before the removal or correction of them.

- The selection of the data modelling technique should depend on the comprehensibility and the ability to explain. ‘Black box’ like methods (e.g. neural networks) is not transparent to data analysts and the users.
- Also, rather than using accuracy to evaluate the models, specificity and sensitivity are important measures to be used in medical context (Cios and Moore 2002).
- Considering the special status of medicine, it is important to use a standardised approach to perform analytics. As the decisions made through data modelling could lead to life or death situations, the data analysts cannot afford to miss certain components in the analytic process leading to incorrect results. Thus, it is essential to consider the three supporting dimensions (project management, communication management and knowledge management) in data modelling in the healthcare domain.

In developing a process model, it is essential to consider how it will be accepted by the data analysts. A higher acceptance of the developed model by intended users (novice data analysts) is to be assured by:

- Developing the process model with due consideration to characteristics determining methodology use
- Using an agile based evolutionary approach
- Using an action case approach to evaluate the model

Thus, the proposed process model will be able to meet the needs of the individual novice users and will guide them in initiating their work.

1.5 Organization of the Thesis

In Chapter 1 and Chapter 2 of this thesis, the problem, the objectives and the background of this study have been described. Chapter 3 describes how the artefact (process model) was developed using DSR approach. The Chapter 4 describes the survey study carried out to identify the factors influencing the usage intention of a process model. Then Chapter 5 describes the process model development and the evaluation approach using action case approach. The HA process model is described in Chapter 6 and finally, Chapter 7 is used to present the discussion of results and the conclusion.

CHAPTER 2. LITERATURE REVIEW

This chapter presents a literature review pertinent to studies on data mining process models with particular reference to documentation using Unified Modelling Language (UML) along with the previous studies that have attempted to develop joint models by combining data mining process models with software engineering processes. Moreover, the limited number of studies related to health analytics (HA) process models available in published literature are evaluated to highlight the importance of a structured unified process model for HA projects. There is a dearth of literature, specifically on studies relevant to HA process models and methodologies of HA. To date, most of the HA studies have been performed as impromptu projects where analysis steps vary according to the expertise of the data scientist. Thus, it is necessary to study the literature related to data mining process models and joint models (data mining process models and software engineering processes) to avoid complexities that have arisen due to unsuitable methodologies.

In addition, theoretical background related to project management, communication management, and knowledge management are discussed in this chapter. They are considered as supporting dimensions of the HA process.

2.1 Software Engineering Frameworks

During the early years of software development, the main focus was on programming languages and algorithms. Programmers implicitly designed the programs (in their mind rather than documenting the design) and developed them according to their personal style (Marban et al. 2009b). With time, software programs became much more complex. However, the lack of a standard approach led to many issues like 88% of the software to be substantially modified, 30% to be not completed and 68% of

software overrunning delivery schedules (Jibitesh Mishra and Mohanty 2011). These issues in software development and delivery led to 'software crisis' in 1968 (Naur and Randell 1969). Many of these shortcomings are due to the failure to use a standardized procedure and faults in the methodology. Thus, to improve the efficiency, to reduce the maintenance expenses and to meet the user expectations, a requirement aroused to propose formal models, methods and methodologies for software development (Kozar 1989; Mohan and Ahlemann 2011). System development methodology can be defined as "a documented collection of policies, processes and procedures to improve software development process" (Chan and Thong 2009). Thereby, software development led to a new discipline called software engineering and was developed by adopting techniques used in engineering.

While waterfall model, iterative model and spiral model are the most common software development life cycle models; Unified Process (UP) and agile process too are very popular in software development industry. Due to advancements in technologies and changes in user demands, these methodologies are evolving continuously (Chan and Thong 2009). Unified process is a software engineering process used to transfer user requirements to a software system. It can be considered as a generic process model that could be used in very large scale application developments. Unified Modelling Language (UML) is an integral part of the unified process and it uses UML to prepare the outline of a software system. Iterative and incremental growth and use-case driven nature can be taken into account as two key aspects of the unified process (Jacobson et al. 1999). UML use-cases are used in the software engineering projects to capture functional requirements and based on them developers design and develop the system and review the systems. Thus, the unified process is known to be a use-case driven process. Here, the projects are broken into

mini projects and iterate through the mini projects. The project grows incrementally with iterations to reach the final end product. Considering the uniqueness of unified process, we believe that we could adopt these two aspects into HA projects as well. Thus, HA projects could be iterative and incremental while being a use-case driven process.

In dealing with dynamic business environments, agile methodology is claimed to be more suitable compared to traditional approaches in software engineering (Paetsch et al. 2003). Agile software development manifesto (Beck et al. 2001) too provides interesting principles which can be adopted in HA projects. Agile methodologies can deal with changing requirements even late in the project (volatile requirements) (Chan and Thong 2009; Dybå and Dingsøy 2008) by allowing business people and developer to work together (Dybå and Dingsøy 2008), building projects around motivated individuals and by reviewing in regular intervals reflecting on how to improve. Thus, there is a significant distinction between traditional software engineering methodologies and the agile methodology. These factors could be taken into account when developing a process model for HA too.

2.2 Data Mining Frameworks

Data mining has been considered by many as an ‘art’ (creative process) and data analysts followed their own styles when carrying out data mining projects (Westphal and Blaxton 1998). In comparing the history of data mining against software engineering, Marban et al. (2009b), have shown the parallelism between the two and have indicated the importance of having methodologies for data mining as in software engineering. Otherwise, data mining too could have faced similar issues such as ‘software crisis’ in software engineering.

According to a survey carried out to determine the 10 most challenging problems in data mining, non-availability of a unifying theory for data mining (it is the top priority problem in the list of 10 problems) and issues related to data mining process have been identified as two of them (Yang and Wu 2006). The former refers to the lack of a theoretical framework that unifies different data mining tasks (classification, clustering, association, etc.) and data mining approaches (databases, statistics, machine learning, etc.) as various techniques created for individual projects (e.g. for classification or clustering problems). The latter identifies issues such as automating different data mining process operations and building a methodology into data mining system. As a result, methodology related issues are created where the success of the data mining project depends on the skills and the knowledge of the team member analysing the data but giving no prospect for repetition of successful practices in future assignments (Wirth and Hipp 2000). Numerous process models are being proposed, to avoid these complexities and to facilitate a standardized approach in performing data mining studies.

CRISP-DM (Cross Industry Standard Process for Data Mining) (Chapman et al. 2000) and SEMMA (Sample, Explore, Modify, Model, Assess) by SAS (Matignon 2007; SAS 2008) are the two most popular data mining process models among data analysts (Figure 1). As such, CRISP-DM and SEMMA were considered in the review of literature. CRISP-DM and SEMMA are derived from KDD (Knowledge Discovery in Databases) process (Mariscal et al. 2010) and data analysts tend to use KDD in addition to their own personal methods (Fayyad et al. 1996a). It is important to note that CRISP-DM and SEMMA are derived from KDD process (Mariscal et al. 2010) and as such the KDD process is not discussed further on in this thesis.

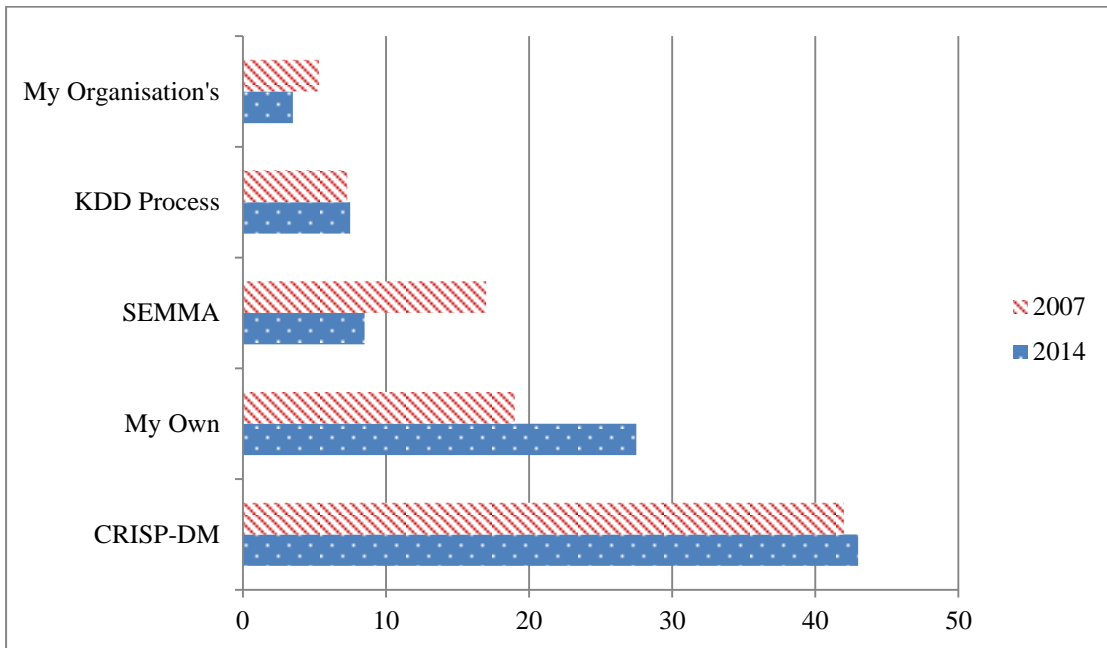


Figure 1: Use of data mining methodologies (in %) (KdNuggets.com 2014)

Compared to CRISP-DM, SEMMA had failed to provide an adequate attention to rigorous requirements of a complete data mining process. SEMMA focuses only on the technical portion of the project (statistical, modelling and data manipulation sections in a data mining process) rather than on the complete process. This inadequate representation of the complete process (e.g. absence of analysis, design and implementation sections), could be recognized as a common problem in most of the process models available in data mining (Marban et al. 2009a). SEMMA does not consider data mining as a central element within a system and as such it does not include roles of the organization and the stakeholders in a project. Moreover, its design approaches correlate strongly with the SAS Enterprise Miner Software package (SAS 2008) and it is reflected as a proprietary methodology. In contrast to SEMMA process model, CRISP-DM provides a comprehensive description and a representation of the complete data mining process.

As a result of limitations existing in other models (including SEMMA model), CRISP-DM is implied as the de-facto standard in data mining (Mariscal et al. 2010) for several reasons: (1) it is a standardized step by step approach to data mining (Chapman et al. 2000; Wirth and Hipp 2000), (2) it is based on pre-CRISP-DM models and has incorporated some of their substantial features (Wirth and Hipp 2000), (3) it is used as the foundation for many forthcoming models (Mariscal et al. 2010), (4) it is the most frequently used model in data mining projects (Bellazzi and Zupan 2008; Marban et al. 2009a; Mariscal et al. 2010) and (5) it is vendor independent (Wirth and Hipp 2000).

CRISP-DM shown in Figure 2 (adapted from (Chapman et al. 2000)) is composed of 6 stages, namely, (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation and (6) deployment. In this model, it is possible to move through the different stages successively or return to the precedent stage if any error is encountered in the current stage. Thus, CRISP-DM is known as a waterfall life cycle model with feedbacks (Cios and Kurgan 2005).

As indicated in Figure 1, the usage of CRISP-DM has not shown a wider spread from 2007 (42%) to 2014 (43%) in spite of its benefits and this may be due to rapidly increasing usage of their own methods (19% in 2007 to 28% in 2014) by data analysts (KdNuggets.com 2014) due to the limitations in existing methods and availability of tool specific methods (Mariscal et al. 2010).

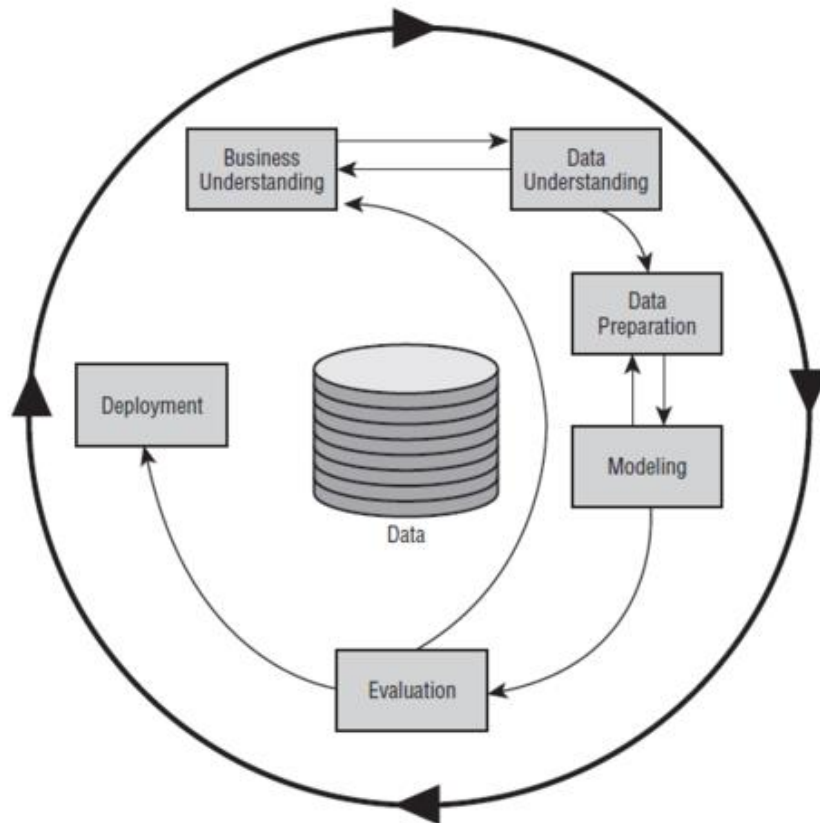


Figure 2: CRISP-DM model (Chapman et al. 2000)

There are several disquiets in CRISP-DM model when compared to a software engineering process model or when real world scenarios are considered in carrying out data mining projects. First, it is a model with a rigid structure (techniques mentioned may be applied because they are included in the tools even though they may not be required). Thus, the models developed may not be in accordance with the organization’s objectives and may not represent the actual problem.

Second, CRISP-DM does not support new data collection during later stages of the process (e.g. data processing and modelling) as it assumes that the required data are identified at the initial phases and continues to be valid till the end of the project. However, in actual scenarios when the project progresses (with a better understanding of the project), new data requirements may arise and sometimes the way data is represented or formatted may need to be modified (Jacobson et al. 1999).

Third, it lacks project management processes and an integral process to ensure the project completeness and quality. Concerning the uncertainty involved in data mining projects, proper project planning is important to meet the needs.

Fourth, CRISP-DM (even SEMMA) assumes that sufficient knowledge of the requirement is already available (Britos et al. 2008). However, in actual settings the clients use a different terminology compared to data analysts making it hard to translate the requirements. Also, most of the time, the requirements will be correctly identified only at the end of the process. The available tools do not support it.

Finally, CRISP-DM selects the data mining techniques based on the data collected. However, the selection of the technique should not depend only on data (Chapman et al. 2000) but should consider the organization goals in addition to the conceptualization of the problem.

Therefore, contemplating some of the disquiets in CRISP-DM model, there is a necessity to develop an iterative life cycle model or an extreme programming based model to avoid the rigid structure in the waterfall model. To avoid some limitations in CRISP-DM, Marban et al. (2009b) proposed a joint data mining engineering process model by comparing and contrasting data mining process models against software engineering process models. They integrated software engineering process models, namely, IEEE 1074 (IEEE 1997) and ISO 12207 (ISO 1995) with the CRISP-DM model. Software engineering process models make project tasks repeatable, and easily and effectively manageable. Furthermore, they include a methodology indicating inputs, outputs, tasks and tools.

The joint data mining engineering model introduces 3 main components; namely, (1) project management processes, (2) integral processes to ensure the project completeness and quality and (3) organisational processes to ensure the effective

organisation. Project management process aims at managing resources through the project life cycle. Integral processes cover aspects such as user training, evaluation of the outputs generated and its process as well as documentation. Organisation processes focus on the whole organisation including their business goals rather than focusing only on the project; namely on its infrastructure available to carry out projects and training. Thus, the joint model ensures the completeness of the functions of a data mining project while guaranteeing the quality and the project management aspects too.

A major drawback of this proposed joint model is that it has failed to address the knowledge management and communication requirements for data mining projects. Even though documentation is identified as an important component in data mining process, they have failed to describe how it should be carried out and other aspects in knowledge management related to creation, storage and transfer of knowledge. Communication is an important component in a successful project (Goodwin 2011) though the authors have failed to incorporate it in their model. It allows coordinating various stakeholders and in delivering the product as per user expectations. Furthermore, the use of waterfall life cycle model is not plausible since it significantly intensifies the cost of modification of data models and serious errors will be discovered only at the later stages of the project. Notably, these methodologies had failed to achieve expected user acceptance levels even though they seem to provide many benefits to the users. Thus, it is essential to contemplate on using new techniques like unified process and agile software development.

Although medical data mining shares a great deal in common with HA, as both are striving to achieve better patient care, they are not identical. Analytics includes entire methodology of data analysis consisting of insights generation from

data and communication of them to recommend actions (Cortada et al. 2012). That is, determining and communicating important patterns in data using various visualization techniques (e.g. charts, tabulation). Data mining can be recognized as a specific tool capable of determining relationships in enormous quantities of data. Data mining is a subset of data analysis. Moreover, as data mining is purely data driven, it cannot be applied in prescriptive analytics where expertise of the physicians is required. Data mining can only be linked with predictive analytics (Watson 2013). Nevertheless, both data mining and HA have been used interchangeably, as reported by many authors. As such, to develop a complete analytic framework, other perspectives such as knowledge management (includes knowledge capture, retention and transfer), communication management (includes communication and coordination with the stakeholders) and project management (includes resource management throughout the project life cycles) too were considered in using the data mining process. These three will be elaborately discussed in subsequent sections with theoretical background as well as how they are applied in analytics processes.

2.3 Unified Modelling Language with Data Mining

Data mining process models have developed adopting software engineering processes to assure the completeness and quality of data mining projects and to support effective management of those projects (e.g. (Marban et al. 2009b)). However, these process models have overlooked how the documentary support for a project can be provided (Yang and Wu 2006).

It is common in software engineering, that each individual developer in a project uses his/her own personal documentary strategy (Marban and Segovia 2013). Thus, it is hard to manage all these documentations and as such systematic project documentation is required (Becker and Ghedini 2005; Fayyad et al. 1996b).

Systematic documentation is required for the repetition of projects (enables prospective projects to follow parallel steps as in the documented project) and for the management of software engineering or data mining project stages/steps. Since Unified Modelling Language (UML) (OMG 2011; Rumbaugh et al. 2004) is a popular modelling language in documenting software engineering projects, it can be applied in data mining as well.

UML is a general purpose graphic notation technique to model each and every stage of a software engineering process model visually. UML is considered as the de-facto standard for design, specification and modelling in software engineering projects (Koch et al. 2008; Podeswa 2005; Zubcoff and Trujillo 2007). Even though it had originally been deliberated for object oriented design documentation, it has been extended to be used in process oriented design documentation as well (Zubcoff and Trujillo 2007). It could be used for business modelling, object modelling, component modelling and data modelling (Ambler 2004; OMG 2011). Thus, UML graphical multiple blocks/diagrams can comprehensively represent a data mining project.

UML has been introduced to data mining projects as well. However, most of the UML extensions are limited to data model development (Luján-Mora et al. 2006; Prat et al. 2006). For example, a UML extension is proposed for clustering models (Zubcoff et al. 2007) and classification models (Zubcoff and Trujillo 2006) in data warehouses and for association rules (Rizzi 2004). Since these extensions overlooked the full process model except for the model development stage, Marban and Segovia (2013) proposed a UML focused approach to model the CRISP-DM based projects concealing the whole data mining process. As a result of the aforementioned issues in data mining, documentation based on UML will provide a standard approach for easy communication and understanding.

UML represents activities, actors, business processes, programming language statements, reusable software components as well as database schemes. The latest version UML 2.X, includes 14 diagrams, where they could be categorized as static (structural) view and dynamic (behaviour) view (Podeswa 2005). The static view focuses on things that should present in the system being modelled (e.g. class diagrams, component diagrams) while the dynamic view emphasizes on the functionality of the system (what happens in the system) and is used to illustrate the interactions and the changes to the internal status (e.g. use-case diagrams, sequence diagrams, activity diagrams).

UML can be extended to model processes with different needs, using (1) UML extension via profiles and (2) extension to Meta Object Facility (MOF) (Koch et al. 2008; Marban and Segovia 2013). The UML profile based extensions are useful when customizing the standard model elements for a specific purpose and domain (Koch et al. 2008). This includes stereotypes (meta-class), tagged values (meta-attributes) and constraints (Aldawud et al. 2003). These extensions can be flexible, mixable and mutable. However, it provides only the customization of existing meta-model rather than defining a new one as in meta object facility (Jacobson et al. 1999). The MOF based meta-models are stable (do not evolve) and formal (semantics are completely defined). Thus, the type of extension to be used can be decided upon depending on the project specifications. UML extensions could be introduced to HA projects as well.

2.4 Health Analytics Frameworks

Previous sections provided an important insight into the benefits of having a structured framework (with UML based documentary strategy) for management of software engineering and data mining projects. It is considered that the lack of a framework is a hindrance for the further development of the field (Dzeroski 2007).

Similarly, HA projects too require a process model (Nelson 2010). HA can borrow the best practices from data mining and software engineering process models to develop a unified and structured process model. Moreover, UML can be used to document each stage of a HA project and prevent any other difficulties in managing data mining projects in general.

Cios and Moore (2002) and Eggebraaten et al. (2007) proposed a data mining knowledge discovery (DMKD) process (Figure 3) for medical applications as an extension to CRISP-DM considering the uniqueness of medical data mining. It is proposed as a semi-automated process where user input (as knowledge on domain and data) is required to perform the complete DMKD process from problem specification to application of the results. It is a six step DMKD process model and the authors have shown its application in the medical domain (Kurgan et al. 2001). Furthermore, it is imperative to note that they have tried to use an iterative and incremental process with feedback loops. In their paper, the authors (Cios and Moore 2002) have focused mainly on introducing a consistent nomenclature using XML. Though they have mentioned about proposed extensions to CRISP-DM, no distinguishable extensions could be identified (Figure 2). Most importantly, even though they have mentioned about the uniqueness of medical data mining they have failed to incorporate any such specific components into their process model.

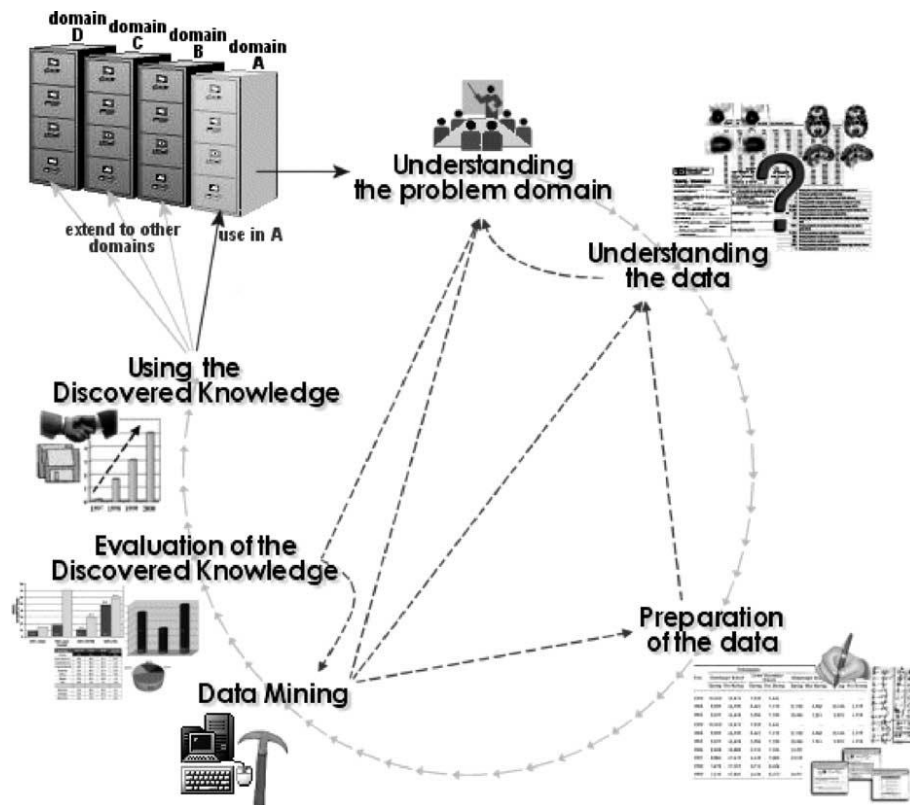


Figure 3: Data mining knowledge discovery process (Cios and Moore 2002)

In a recent review (Esfandiary et al. 2014) on medical data mining research, using 291 journal publications (in 81 journals) between 1999 and 2013, data mining, CRISP-DM has been adopted as the standard model for medical data mining. The authors have highlighted, the dearth of a standard in the overall knowledge extraction process (data collection to evaluation) as a weakness in medical data mining up to now (Bellazzi and Zupan 2008) and also that it is required to find a means to transfer the knowledge to medicine process as if not, the medical data mining will be of no use.

In another study Schmidt et al. (2008) applied CRISP-DM to a dataset related to a condition of asthma, and found that CRISP-DM methodology cannot be directly applied to the clinical data due to the limitations of scope in several areas of the CRISP-DM and the knowledge gap among professions (computer science and medical field). They suggested that consultation and collaboration at the data

understanding stage and visualisation are essential to merge the knowledge divide in the two domains.

Catley et al. (2009), emphasized the importance of extending the CRISP-DM model when modelling clinical systems integrated with data mining and temporal abstraction to deal with time series data using a case study carried out by them. They proposed a new CRISP-DM model named as CRISP-TDM considering temporal data mining (TDM) and identified several factors that need to be taken into consideration. First, for the business understanding phase, they highlighted the significance of the clinically relevant and population-based information. Thus, the goal is to get patient centric outcomes based on the clinical data and population based data. Second, for the data understanding phase, they recommend to reflect the temporal characteristics of data. Third, they proposed the inclusion of temporal abstraction details and integrated models (e.g. temporal abstraction with data mining) to the data modelling phase. Temporal abstraction can be applied on data to extract trends and temporal relationships and then those data can be analysed using data mining techniques.

Finally, for the deployment stage, the authors suggested including a methodology to describe system storage. To conduct a dynamic data mining study, it is vital to store raw data and temporal abstractions and then use them in the subsequent temporal data mining analysis. Even though CRISP-DM is extended, the authors have failed to handle the earlier mentioned issues in CRISP-DM and ignored the other supporting elements that are useful in creating a complete process (e.g. project management and knowledge management).

In addition, CRISP-DM has been used in many recent individual studies, including healthcare related data mining projects. For example, it has been used as the data mining methodology to study on data from 501 patients operated for lung cancer

with curative intention (Rivo et al. 2012). They used logistic regression to predict the post-operative death of lung cancer patients. Similarly, it has been used as the process to identify patterns in bed overflow and to formulate strategies to solve such problems in hospitals (Teow et al. 2012). In another study, CRISP-DM has been used as the process model to apply machine learning to predict the mortality of using allogeneic hematopoietic SCT in various hematologic malignant and non-malignant diseases (Shouval et al. 2014). Thus it could be noted that CRISP-DM has been utilized as a process model in diversified clinical settings and sub domains for prediction modelling (Bellazzi and Zupan 2008). However, further revisions are required to be made to the existing CRISP-DM model to improve the usability (not only in predictions), applicability and repeatability in healthcare.

As an emerging field, up to now only one HA based framework can be identified (Raghupathi and Raghupathi 2013) in the literature to the best of our knowledge. This could be identified merely as a HA methodology as it describes '*how to do things*' in a HA project. This methodology includes 4 stages, namely, (1) concept design (project description), (2) proposal (abstract, introduction and background), (3) methodology (hypothesis development, data collection, model development, etc.) and (4) presentation and evaluation. However, we can consider several related shortcomings in this HA framework. It lacks a proper socio-technical based process model and it considers only the documentations. Furthermore, the proposed documentary strategy lacks proper methodological steps (implicit) and there is no visible direct link between input and output from one stage to another. Thus, to perform healthcare projects as a structured process, a new process model is required to clearly define objectives, to systematically document prior knowledge, data, methods and results (Bellazzi and Zupan 2008).

Despite the abundance of information on HA studies in the literature, most of them have used their individual approaches rather than following a standardized approach, making it hard to manage and repeat successful project steps or identify mistakes in certain steps. Thus, it is hard to translate the findings to specific actionable steps. Furthermore, it is interesting to note that none of these studies have used a proper research methodology in developing the process model. Also, they have not paid any attention to the reasons for individual user resistance to such methodologies even though there are many benefits to gain from using the methodology. Therefore, it is important to develop a complete and structured process model to support the needs of the users in healthcare sector and specifically the novice users and evaluate the acceptance of the model by such users to improve it for higher user acceptance.

2.5 Supporting Dimensions

According to Nambisan (2003), IT plays four roles in new product development in Information Systems, namely, process management, project management, communication management and knowledge management. As mentioned earlier, these four dimensions were considered in the development of the model. Project management, communication management and knowledge management can be considered as supporting dimensions on HA process management. It was decided to utilize a similar line in the development of the process model for HA based on prior experience as all these components are important in developing a unified model. Even other software engineering methodologies and data mining methodologies have considered some of these dimensions even though they are not specified explicitly. For example, CRISP-DM model gives some indications on project management at the initiation of the project. Chan and Thong (2009), mention the importance of

knowledge management in developing a software methodology. Even though these four dimensions are identified as separate perspectives of the proposed process model (USAM), there are inter-relationships among them. For example, knowledge transfer is linked with communication (Chan and Thong 2009).

The first dimension is the process management and it deals with the overall structure of the HA process model directly dealing with the data modelling tasks. This includes input, outputs and activities in each phase of the process model. CRISP-DM model is used as the foundation of the process as it is considered to be the industry standard for data mining.

The second dimension is project management. This deals with the management and coordination of the activities performed in each stage of the process. In project management, it is important to consider about the initiation phase (involves creating and updating project infrastructure), project planning phase (plan evaluation, installation, integration, documentation, training, etc.) and monitoring and control phase (involves identification of potential problems, likelihood of their occurrence, their impact and steps to mitigate them) (Marban et al. 2009b).

In addition to those three phases in project management, in dealing with unknown project outcomes with ambiguity in project direction, there should be a methodology like agile (as used in software engineering) to guide the project in the right direction. In agile concept there are four main attributes that are considered, namely, evolutionary approach, story driven approach, continuous collaboration and testing agile projects (Collier 2011). This is especially useful in long-term projects where the problem is not specified at the beginning of the project. Agility is useful in responding to changes in a timely and an effective way (Highsmith 2009).

Gartner (Goodwin 2011) has reported that, around 70% to 80% of corporate business intelligence projects fail due to poor communication. This indicates the importance of communication for the success of an analytic project. To be in line with agile methodology utilized in the project management, it is important to maintain continuous and effective communication with the users and the other stakeholders. Unlike in traditional (sequential) data mining methods (where the communication with the users tend to be merely at the beginning of the project to get their requirements with very limited interaction like status update during data modelling stage), in agile approach, continuous collaboration will be promoted throughout the project (Collier 2011). Here, regular interaction will be maintained with personnel working on data modelling, direct or indirect users and beneficiaries, and sponsors setting the requirements.

This is especially, useful in healthcare context, due to the unfamiliarity of the analyst with health domain. Thus, it is important to maintain the frequent communication to guide the analyst in the right direction and to provide necessary feedback (e.g. significant but meaningless findings could be detected and should be dropped from further consideration). Furthermore, it leads to getting new ideas and directions to explore data. Media Synchronicity Theory (MST) (Dennis et al. 2008) is used to ask the right question, present the results and to maintain the right mode of communication between individuals working together to accomplish meaningful findings from HA projects (from rich media like face to face communication to documentation) in line with knowledge transfer.

According to Media Synchronicity Theory, two communication processes are conveyance and convergence. Dennis et al. (2008) defined conveyance as the *“transmission of a diversity of new information-as much new, relevant information as*

needed-to enable the receiver to create and revise a mental model of the situation". A variety of information is exchanged at this stage and extensive information processing is required. Dennis et al. (2008) have further described conveyance as *"the discussion of the pre-processed information about each individual's interpretation of the situation, not the raw information itself"*. At this stage a mutual understanding will be reached among the individuals and less information processing will be required. In accordance with Media Synchronicity Theory, while media of low synchronicity is used to carry out conveyance tasks, media of high synchronicity is used in carrying out convergence tasks.

A similar approach has been used in software development teams such as virtual team where Media Synchronicity Theory is used as the theoretical basis (Baker 2002; DeLuca and Valacich 2005; Niinimaki et al. 2009). In a case study (Edström 2009) on changing software development from ad hoc approach to agile, the authors have used Media Synchronicity Theory as well. As such we believe that it could be applied in our study as well to study the communication process in data analytic projects among stakeholders.

Knowledge management is the final dimension. It deals with information on knowledge outcomes- creation, retention and sharing. Organizational knowledge management framework proposed by Argote et al. (2003) was used as the basis for this study. Similarly, this had been used in agile software development methods as well (Chan and Thong 2009), thus, it was considered to use a similar theoretical basis for this study as well.

The knowledge, needs to be retained within the groups and transferred among the members (Chan and Thong 2009). Moreover, successful knowledge management depends on ability, motivation and opportunity (Argote et al. 2003) and it aims to

assist meeting knowledge needs of a team. As indicated by Lindvall and Rus (2002), knowledge needs are (1) gain knowledge about the domain, (2) gain knowledge about different tools and HA algorithms, (3) share knowledge about local policies and practices (e.g. Personal Data Protection Act – PDPA, data de-identification practices), (4) capture knowledge within data analysts and (5) transfer knowledge among the members. A proper documentation approach enables to achieve these needs in HA projects. As such, for knowledge and information management, documental steps are proposed.

In Argote et al. (2003)'s framework, authors have considered two dimensions, namely, (1) knowledge management outcomes and (2) properties of knowledge management context. While the former refers to knowledge outcomes, the latter refers to properties of units (individual, group), properties of relationship between units and properties of knowledge itself (tacit and explicit). A unit can be an individual, a team or an organisation. The knowledge outcomes depend on the characteristics of the unit (Argote et al. 2003). In this study, the consideration will be on an individual level. Individual knowledge sharing and seeking behaviour depends on physiological factors (Kankanhalli et al. 2011; Kankanhalli et al. 2005). Individual's knowledge management depends on ability, motives and opportunity to create, retain and transfer knowledge (Argote et al. 2003).

How these supporting dimensions are applied with consideration on the theoretical background will be discussed in the next section.

2.6 Application of Related Work in the Proposed Model

Considering the benefits highlighted through this chapter on CRISP-DM, it was decided to base the proposed model on using it. However, to safe guard from the pitfalls in CRISP-DM and data mining engineering model, other factors like project

management, communication management and knowledge management too were considered in the development of the new proposed model. Moreover, it is important to consider the variations of the projects based on changing user requirements (e.g. complexity and ambiguity of the requirements) as it is not possible to use one-fit-all model for all the data analytic problems.

Agile based approach has been successfully used in software engineering projects. Considering the evolutionary development process, continuous stakeholder collaboration and flexibility allowed through agile based approach, it was introduced into the proposed model as well. This will play a significant role in projects with complex and ambiguous requirements.

For communication management, application of Media Synchronicity Theory on two communication processes - conveyance and convergence- was incorporated in the HA process model in line with the variations of the project type. Communication requirements and means of coloration with the stakeholders will vary based on the project requirements and familiarity of users with the project.

For knowledge management, how individual's ability, motives and opportunity to create, retain and transfer knowledge was incorporated in the proposed model. Even though process models state the importance of documenting the steps performed, many data mining processes have omitted to direct how the project documentation should be carried out. Documentation plays an important role in knowledge management. Going through the programming code alone (even if having appropriate comments) is not practical. Even though less documentation is emphasized in agile based project management, this could lead to poor knowledge management in long term projects (Lagerberg et al. 2013). This is especially required if the project is complex and ambiguous. As they have found in their comparative

study (comparing traditional and agile approach), internal daily documentation is an important part of a project. As such, a documentation strategy is proposed here.

Knowledge transfer and retention can be effective if the members share a common language. Similar findings have been shown in the study performed by Weber and Camerer (2003). This could be achieved by documentation of the projects using a standard notation based approach where a short hand language is used for knowledge retention and transfer (Argote et al. 2003). A modelling language like UML could be used to represent information and the system structure. UML notations that have been successfully used in software engineering documentation were introduced into data analytics context considering the need for a documentation approach. UML was used for documentation of methodological steps of the proposed model developed due to its popularity and wide acceptance (Marban and Segovia 2013; Zubcoff and Trujillo 2006). By using a universal visual modelling language, the users and analysts can focus on the main objective, the HA process. It is important to note, that new notations are required to be introduced as existing notations will not allow representation of data preparation and data modelling tasks.

2.7 Summary

An evaluation of software engineering process models and data mining process models by reviewing published literature was presented in this chapter. Shortcomings of these models and the inability to apply these models in HA context were discussed while elaborating on the importance of introducing UML into HA for more clarity and objectivity. A conceptual framework integrating different recommendations given by some of the industry standards and findings of most cited studies on HA into a coherent whole process to confront issues in HA projects will be introduced in the following chapters.

CHAPTER 3. METHODOLOGY

This chapter describes the approach taken to design and evaluate the unified structured analytics model (USAM) for HA. Design science research approach and several behavioural research approaches were used in identifying the requirements and in evaluating the proposed model. A detailed description of the methodology used in this study is given below.

3.1 Introduction

The socio-technical approach in the field of Information Systems (IS), aims to integrate social and technological systems in implementing an ICT artefact (Lee 2001). As technologies are socially located, it is important to consider the features of any technological system and the social norms and rules of use (Sawyer and Jarrahi 2014). However, implementation of ICT artefacts taking both social and technological systems into consideration at the same time is rare (Eason 2008; Sawyer and Jarrahi 2014). Similarly, Enid Mumford, the most influential researcher to initiate socio-technical research within IS (Davenport 2008) had indicated that most of the IS research is limited to engineering approaches. Due to competitive business environments observed since 1990s, organisations had to adopt methods like lean production, outsourcing and business process reengineering (Carr 2008; Kling and Lamb 1999). These methods provide less emphasis on user needs compared to the socio-technical approach.

Considering the popularity gained over the past decade for design science as another approach to IS research, design science research (DSR) could be a good means for socio-technical researchers to follow (Sawyer and Jarrahi 2014). The methodology used to explain the problem and the related theoretical principles for the

proposed process model in this thesis is the DSR approach (Hevner et al. 2004; Pries-Heje and Baskerville 2008).

In IS discipline there are two paradigms, namely, behavioural science paradigm and design science paradigm. While former focuses on developing and testing theories used to explore or predict human and organizational behaviours (interactions among humans, technology and organizations) the latter focuses on creating innovations to solve problems (Hevner et al. 2004). It was decided that exploring and confirming the hypothesis research approach (in behavioural paradigm) is not suitable for this study as the main aim is on explicating the goals of the research artefact, followed by the development and evaluation of its utility (Gregor and Hevner 2013; Hevner et al. 2004). Moreover, DSR helps to overcome one of the major concerns in IS research, that is, artefact's low level of professional relevance (Arnott 2006). Thus, it was decided to use DSR in this study where the unit of analysis will be the method (HA process model) designed and evaluating it in an organizational context (in a real application scenario).

3.2 Design Science Research Approach

The design science research (DSR) approach in IS discipline is a problem solving paradigm, where new innovations are tried to be created to define ideas, practices and products to achieve effective and efficient analysis, design, implementation, management and use of IS (Hevner et al. 2004). According to the DSR knowledge contribution framework proposed by Gregor and Hevner (2013), in this study, it was attempted to extend the known solutions to new problems, which is known as 'exaptation' in DSR. This allows adoption of existing process models in data mining and software engineering to the HA context by making certain modifications along

the three supporting dimensions (project management, communication management and knowledge management).

In design science research approach the artefact is the most important outcome of the research and as such in the next section, the artefact developed through this research will be described.

Artefact

An artefact in IS design science research can be a construct (it is the language used to specify the problem and solution e.g. concept, symbol), a model (representations of the problem and possible solutions using constructs mathematical models, logical models and diagrammatical models), a method (processes to guide on how to solve a problem, e.g. textual descriptions, algorithms for best practices) or an instantiation that can be converted into a material existence (problem specific aggregates of constructs, models, methods in a working system) (Hevner et al. 2004; March and Smith 1995; Pries-Heje and Baskerville 2008; Winter 2008).

Based on Winter (2008)'s description on methods and models in design science research, this study aimed at developing a 'method' for Analytics. According to Winter (2008), if procedural aspects are considered in developing the artefact, it can be classified as a 'method'. This methodology (the final revised method) used process management, project management, knowledge management and communication management as focusing constructs (or as the dimensions of the proposed method). It conceptualized an eight step analytics process mainly grouped under two cycles: data cycle and modelling cycle and was developed as a generic method for analytics. The model was evaluated specifically focusing on healthcare context. The core of this study, the design artefact developed through this thesis will be presented in Chapter 6.

Unique points of the artefact:

- Inclusion of project management, communication management and knowledge management as supporting dimensions of the data analytic process (referred to as process management).
- Consideration of variation of supporting dimensions and the data analytic process based on the requirement type (based on complexity and ambiguity of the requirements in the healthcare sector) as all existing models are one-fit-all projects.
- Modification of the CRISP-DM model (considers the process management) and introduction of new components to the model.
 - Changed to an iterative loop structure with two main cycles as data cycle and modelling cycle. Thus, the limitations due to waterfall structure used in CRISP-DM could be avoided.
 - Introduction of two steps as data access and conceptualization
 - Addition of new sub-steps. e.g. to domain understanding step as determine stakeholder requirements and determine compliance needs (specially required in healthcare context) and their related tasks, to data understanding step as decoding of data and related tasks and to presentation step as post-implementation.
- Specific consideration to address issues relevant to the uniqueness of medicine. Codification of extracted data, free text and other media files, anonymization and de-identification of data, visual representation to bridge the knowledge gap, etc. (Section 1.3 and 1.4).

Part of the process involves business requirements identification and project/process management. As can be seen in other process models, for example, software engineering process models a significant part is based on generic project and process management components. An important factor is how these generic components could be applied in data analytics.

It is important to note that most of the IS design science research is focusing on models and specific instantiation development while there is a dearth of studies on methods (Winter 2008). According to the author, even the available method development studies are on construction and evaluation of algorithms, mathematical/statistical techniques rather than on developing methodologies. In contrast, this thesis study was focussing on procedural aspects in carrying out a HA project.

3.3 Research Process

The research method used in this thesis is illustrated in Figure 4. It is composed of five distinct steps; namely, identification of the problem, suggestion, development, evaluation and conclusion (Vaishnavi and Kuechler 2005). A Similar, research method has been followed by Arnott (2006) in designing a methodology as the artefact.

The method shown in Figure 4 can be linked to other methods and approaches available for DSR. For example, this is in line with the approach proposed by Peffers et al. (2007), with six steps namely, (1) problem identification, (2) description of objectives; (3) designing and developing the artefact; (4) demonstration; (5) evaluation; and (6) communication of results. First three phases in Peffers et al. (2007)'s method will be effectively covered by the first three phases in Figure 4, and demonstration and evaluation will be covered by the evaluation in Figure 4. March

and Smith (1995) state that “build” and “evaluate” are the two phases in DSR. They are represented by first three phases in Figure 4.

Research Process	Current Project
1. Problem awareness ↓	To understand “what to do” and “how to do” a HA project based on project requirements.
2. Suggestion ↓	Use software engineering and data mining methodologies and Media Synchronicity Theory, knowledge management framework (Argote et al. 2003) and agile approach for process management, project management, communication management and knowledge management.
3. Development ↓	Develop HA process model for novice users that use project management, communication management, knowledge management
4. Evaluation ↓	Use the HA process model in actual setting to evaluate its effectiveness
5. Conclusion	Reflect on the instantiation and determine amendments of the process model developed

Figure 4: The design science research method applied to HA process model development

Right hand side of Figure 4 illustrates how the DSR methodology is applied in this thesis.

The **first step** - problem awareness has already being addressed in Chapter 1 where the problems are being defined by research questions as (1) *What methodological steps are needed to be followed by a novice user in health analytics?* and (2) *How supporting dimensions (project management, communication management and knowledge management) are utilized in a HA project based on user requirements?*. Furthermore, a survey was carried out with the aim of understanding the novice user’s intention to use a methodology for analytics (implementation details and results are given in Chapter 4).

Moreover, several novice users who are in internships (M.Sc. students in a business intelligence program) in healthcare context were interviewed to understand how a process model approach could be used by them and they indicated that having a proper methodology will help them to understand how work can be commenced rather than doing their work in an ad hoc manner. In addition, I did attend some of the weekly capstone project meetings of those M.Sc. students with their supervisor as an observer to understand how they had approached the problem and progressed weekly. They indicated their preference for having methodologies with sufficient flexibility instead of methodologies with a rigid number of steps.

As analytics is not a straightforward problem it is important to employ an iterative approach in carrying out the project.

In the **second step** - suggestion - project management, communication management and knowledge management are proposed as focusing constructs while using software engineering and data mining methodologies and Media Synchronicity Theory, knowledge management framework (Argote et al. 2003) and agile approach as the conceptual background. The aim of this step was to determine the problem and search through the existing data mining approaches like CRISP-DM.

The **third step** – development - is the heart of the DSR process where the design artefact, the HA process model will be developed for the novice users. The instantiation of the artefact in this thesis is the development of the analytical data model using the method built.

For the **fourth step** – evaluation - researchers can use approaches from positivist to interpretive IS traditions (Arnott 2006). According to Hevner et al. (2004), to evaluate an artefact five classes of methods can be identified. The first class -evaluation, observational; comprises case studies and field studies. This thesis study

used a participatory case study (action case based approach) to evaluate the HA process model in a hospital. It was decided to do a case study as it captures more specific details than a survey and it allows identifying the nature and the key attributes of the development process (Arnott 2006).

For the **fifth step**- conclusion (or reflection) - an attempt was made to determine refinements to the HA process model. The success of the study, refinements, contributions as well as the limitations of the research will be described subsequently in Chapter 7.

Action Case Approach

As per the socio-technical approach, user participation in IS development tasks is essential (Sawyer and Jarrahi 2014). In line with that, we decided to use the participatory based approach to improve the USAM model. For the method development and evaluation, an action case methodology (or participatory case study (Arnott 2006)), which integrates action research with interpretive case study approach (Baskerville and Pries-Heje 2014) was used.

In action research, there is a close cooperation between practitioners and researchers to introduce changes and evaluate them. Here, the researcher was a member of the team to understand the problem and she worked with the practitioners to come up with a solution. This was an iterative process and used interviews with the practitioners to determine the utility of the model. Thus this study can be identified as a hybrid of action research and case study.

A similar strategy has been used by Arnott (2006), where the design artefact was a decision support system development method using cognitive bias as a focusing construct (uses 'method' as the design artefact). The author has presented a model of the system development method with major cycles as initiation, analysis and delivery.

In another design science study (using a similar approach) by Tjørnehøj et al. (2014) a distributed global project management model is developed by facilitating informal processes in project management. Moreover, to study the diffusion of best practices in project management procedures in an organisation, an action case based approach has been used as a design science study (Baskerville and Pries-Heje 2014). Thus, through action case approach, it was expected to test the feasibility of using the development method in an organisation context and to test the effectiveness of it in use.

As illustrated in Figure 5, there were two development-evaluation cycles in this development -evaluation process. In cycles 1 and 2 in Figure 5, the attention was on design, development and evaluation of the artefact. Then the model was adjusted based on the findings. This will be a fluctuating design process between searching for theoretical input and looking for new possibilities that could be incorporated in the proposed model. As cases, two hospitals were used in cycle 1 and cycle 2 in Figure 5, which will be elaborately described subsequently in Chapter 5. Due to the regulations existing in the healthcare sector, it is extremely hard to gain access to perform a case study. Thus, the selection of the case was opportunistic (Pettigrew 1990).

The development-evaluation process related two studies (Figure 5) were carried out in two prominent healthcare institutes in Singapore. The first study (Cycle 1 in Figure 5) was carried out on machine utilization in one of the case organisation's Radiology Department. The final evaluation iteration was carried out at in a Health Analytic Department of another hospital using an action case based approach. The details of the evaluation and case organisations are given in Chapter 5.

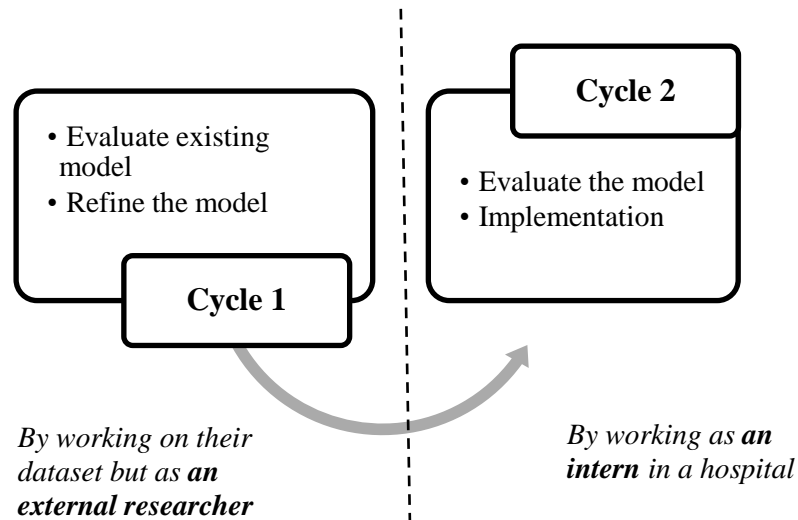


Figure 5: Development -evaluation process

3.4 Summary

In this chapter, it is elaborated on how the process model was developed using the design science research approach. The artefact of this project is a process model for analytics. Research process comprised of problem awareness, suggestion, artefact development, evaluation and conclusion was used as the design science research method to develop the process model. The core of this research is the development of the process model. The study was carried out at an individual level and targeting novice users to HA. The artefact, the HA process model development-evaluation approach will be discussed in Chapter 5 and the HA process itself will be described in Chapter 6.

CHAPTER 4. SURVEY STUDY

This chapter presents the survey study of the research process carried out to understand the novice user's intention to use a methodology for analytics. The relevant conceptual background as well as the conceptual model (and hypothesis), the data analysis and the discussion of the results are described in this chapter.

4.1 Introduction

The aim of the survey was to identify what methodological attributes novice users look at in a process model. The research question was, "what attributes of a methodology affect the novice analyst's decision to use that methodology". Thus, through this survey study it will be possible to understand the methodological attributes that will persuade an individual towards using the analytic process model and incorporation of them will lead to the development of a methodology that is deemed suitable for its users.

Since it is considered that the initial decision to adopt a particular methodology will be made at individual user level, this study was performed at individual level rather than at organizational level. Furthermore, as most of these analytic projects are usually carried out by one or two individuals in the organization (with interactions with many stakeholders); the decisions will be made at individual level rather than at organizational level based on their personal preferences.

In this study, the focus was on the perception of the aspects of the methodology instead of looking at the primary methodological attributes. It was considered that their perception of the artefact will depend on how they perceive these primary attributes (Mohan and Ahlemann 2011) and the individual perception about an innovation's potential effect on his/her work will have an impact on the intention

to use (Hardgrave et al. 2003). Potential individual novice users will adopt the methodology based on their perception (Moore and Benbasat 1991) of how its attributes fulfil their requirements.

4.2 Conceptual Background

Even though there are several data mining methodologies, there is a dearth of empirical studies related to adoption of such methodologies. The available studies are confined into case studies carried out in organization context on adoption of business intelligence (e.g. (Catley et al. 2009)). Thus, it was necessary to examine the literature related to software engineering methodology adoption. Several authors have carried out empirical studies on the adoption of a software engineering methodology by individual users in an organization. Most of these works too are carried out as case studies (Dybå and Dingsøy 2008).

Recently, researchers have started to look at methodologies as innovations, just because they are new to the potential users (Mohan and Ahlemann 2011). Most of the authors have carried out these user acceptance studies as technology innovations rather than considering them as new processes (Chan and Thong 2009; Mohan and Ahlemann 2011). Similarly, the Diffusion of Innovation (DOI) (Rogers 2010) Theory with Technology Acceptance Model (TAM) (Davis et al. 1989) was used as the theoretical foundation of this study.

Roger's DOI was selected due to the following reasons. First, based on DOI, the innovation's adoption rate is most extensively determined by its characteristics. Second, DOI is applied at individual level. Third, previous studies related to software engineering methods, have used DOI in studying the methodological characteristics (Hardgrave et al. 2003). Even though, it was acknowledged earlier that these theories are used merely to study the acceptance and diffusion of products, several researchers

(Chan and Thong 2009; Mohan and Ahlemann 2011) have used DOI and TAM to examine technical characteristics of the methods (Fichman and Kemerer 2012; Riemenschneider et al. 2002). In a similar sense, Raghavan and Chand (1989) suggested that DOI is suitable for methodological acceptance studies (Hardgrave et al. 2003). In previous methodological studies, DOI characteristics had given mixed results relevant to the significance of their influence on adoption (Hardgrave et al. 2003; Riemenschneider et al. 2002).

Similarly, TAM also provides a suitable theoretical foundation on intention to use an innovation based on ease of use and usefulness (Davis et al. 1989) as used in software engineering methodology related studies. Riemenschneider et al. (2002) used TAM, TAM2, Theory of Planned Behaviour (TPB), Perceived Characteristics of Innovating (PCI), Model of Personal Computer Utilization (MPCU) to examine the acceptance of software engineering processes and found the relationship between perceived usefulness, voluntariness, compatibility and subjective norm to be significant with intention to use the software engineering process. Hardgrave et al. (2003) reported similar findings using DOI and TAM. Thus, DOI and TAM will provide the necessary theoretical basis to study the Research Question.

Several authors have considered the exploration of personal traits and organizational characteristics. In an empirical study carried out among potential software developers, Mohan and Ahlemann (2011), have tried to examine the psychological needs of the users (through motivation theories) in addition to the technical aspects of the method. Similarly, some prior research had focused on the individual developer's experience (Hardgrave et al. 2003). Moreover, in a conceptual framework proposed, Chan and Thong (2009) have studied the acceptance of agile methodology from a knowledge management perspective. However, the experience

and other personal characteristics were not considered in this study, as our target group of novice users' level of understanding and experience may be limited and all of them will be new to the projects.

On the other hand, some authors have examined the effect of organizational characteristics on the acceptance of software engineering processes and shown organizational culture (Iivari and Iivari 2011), management support, training and external support influencing the acceptance of those processes (Roberts et al. 1998). In our study, organizational characteristics were not considered as undergraduate students who do not have prior work experience were used for the study as novice data analysts.

Johnson et al. (1999) identified a list of beliefs underlying intention formation to use object oriented development and it includes several usefulness elements like process usefulness and communication usefulness. As indicated by Nambisan (2003), IT is involved in four extents in new product development (NPD) in IS; namely, process management, project management, communication management and knowledge management. Latter three are considered as supporting dimensions for process management. Thus, the perceived usefulness of each of these three dimensions can be considered as separate usefulness elements.

Based on the literature review, it is observed that TAM and DOI provide well established constructs to study the characteristics of a process model acceptance and adoption (Chan and Thong 2009). By synthesizing the literature from innovation diffusion and intention formation related to methodology acceptance, our model attempts to capture the technological factors influencing the adoption of the analytic process model.

4.3 Research Model and Hypotheses

The proposed research model developed based on the conceptual background outlined above is presented in Figure 6. We identified seven antecedents, namely, (1) ease of use, (2) relative advantage, (3) compatibility, (4) result demonstrability, (5) trialability, (6) project management usefulness, and (7) knowledge management usefulness. These are the perceived characteristics of a process model. The dependent variable is the intention to use a process model.

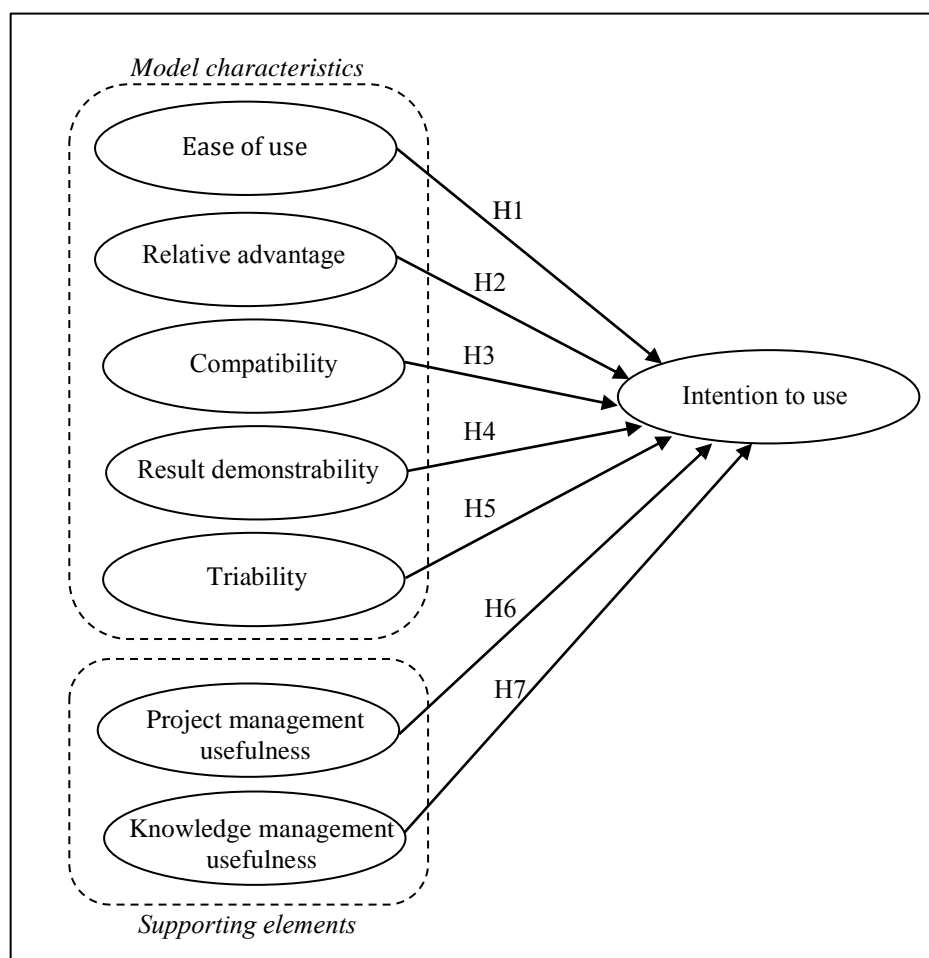


Figure 6: Research model for the survey study

The former five constructs represent the perceived methodological attributes. According to Rogers (2010), perceived characteristics of innovations are relative advantage, compatibility, complexity (replaced as ease of use), trialability and

observability (replaced as result demonstrability). The variations to the characteristics were made based on the prior literature and according to the context studied. The justifications for the replacement for each construct are given in subsequent sections. Process management (analytical data model development process) is represented by the five model characteristics. The latter two constructs represent the usefulness of supporting elements to the main model development process.

Ease of use

Ease of use refers to ‘the degree to which a person believes that using a particular system would be free of effort’ (Davis 1989). Ease of use has been used to address complexity construct in technology adoption literature (Mohan and Ahlemann 2011). As such, instead of using complexity, ease of use is considered (Moore and Benbasat 1991). The decision to use a methodology will depend on whether it is perceived to be easy to understand and use.

Therefore, if the users find a process model is free of mental and physical effort and it is easy to learn, they are likely to use it.

HYPOTHESIS 1 (H1): Ease of use has a positive effect on the intention to use a process model

Relative advantage

Relative advantage refers to ‘the degree to which an innovation is perceived as being better than its precursor’ (Moore and Benbasat 1991). This is the value of process models like CRISP-DM over using an ad-hoc approach. Excellence of a methodology can be measured through improvement of its acceptance rate as well as through improvement of efficiency and productivity (Hardgrave et al. 2003; Mohan and Ahlemann 2011) or meeting intended purpose (Moore and Benbasat 1991).

Similarly, perceived usefulness in TAM demonstrates conceptual equivalence to the relative advantage (Moore and Benbasat 1991). The expectation of developing a structured process is to improve the application of the analytics techniques to the processed data based on the user requirements and coming up with better results while having a low learning curve which would not have been possible by using an ad hoc approach.

Therefore, if the novice users find that using a process model for analytics will be useful for their work there is a prospect of successful deployment of it.

HYPOTHESIS 2 (H2): Relative advantage has a positive effect on the intention to use a process model

Compatibility

Compatibility refers to ‘the degree to which an innovation has been consistent with existing values, needs, and past experiences of potential adopters’ (Moore and Benbasat 1991). If an individual is used to certain habits, there may be resistance from users towards a new process. In analytics, if the users are used to their own personal styles of carrying out analytics projects which have been developed based on their experiences, they may find it hard to change their practices. Even for novice users, if there is a certain style learnt earlier, they may find it hard to deviate from it as it is the initial practice that had been engraved in them.

Therefore, if the methodology is compatible with past experiences and learning of the users, then they will use a new process model.

HYPOTHESIS 3 (H3): Compatibility has a positive effect on the intention to use a process model

Result demonstrability

Result demonstrability refers to ‘the degree to which the results of using an innovation are observable by others’ (Mohan and Ahlemann 2011; Moore and Benbasat 1991). Thus, as indicated by Moore and Benbasat (1991), if it is perceived that the methodology provides observable results which can be communicated then it is considered that the results are demonstrable. Poor communication of usage benefits and quantification of results in an analytic method will not depict the results as highlighted in any other methodological domain (Mohan and Ahlemann 2011). Particularly, as novice users, they will be more concerned about the quantification of results.

Therefore, if the results are demonstrable the novice users will intend on using a process model.

HYPOTHESIS 4 (H4): *Result demonstrability has a positive effect on the intention to use a process model*

Triability

Triability refers to ‘the degree to which an innovation may be experimented with before adoption’ (Moore and Benbasat 1991). Ability of the users to test the method before making the final decision will allow them to make an informed decision about the method. This allows users to understand the un-communicated benefits of the method (Mohan and Ahlemann 2011).

Therefore, if the novice users can try out a process model before adopting, there will be a positive influence on the prospect of using it.

HYPOTHESIS 5 (H5): *Triability has a positive effect on the intention to use a process model*

Usefulness

Perceived usefulness is ‘the degree to which an individual expects that following a methodology will improve job performance’ (Hardgrave et al. 2003). Even in HA projects, project management, communication management and knowledge management are playing a key role. Since no (or minimum) attention has been given to communication management in existing process models, it was not considered in this study even though the result demonstrability focuses on some attributes of communication management. As perceived usefulness of the process is evaluated through relative advantage from DOI (Moore and Benbasat 1991), the process management was not considered here. Thus, only the influence of usefulness of project management and knowledge management on usage intention of the process model was considered here.

Considering the risk involved in analytic projects, having project management elements in the process model is useful (Marban et al. 2009b). Project management is to establish reasonable plans for performing and managing the project (Weber et al. 1991) and it includes estimating the work to be performed (milestones), identifying necessary resources and creating schedules. In considering the uncertainty involved in analytic outputs, project management is useful in scheduling the resources and keeping the project on track.

Therefore, novice users will find project management useful to plan out and perform their tasks.

HYPOTHESIS 6 (H6): Usefulness of project management has a positive effect on the intention to use a process model

Knowledge management is an important part in a process model. Chan and Thong (2009) used knowledge management as a strategic perspective to be

considered in implementation of agile methodologies in software engineering. Similarly, in analytic process models too, achieving positive knowledge management outcomes (create, retain and transfer of knowledge) is crucial for learning and in replicating the best practices (Argote et al. 2003). Success of an analytic project depends on how knowledge is retained within the project teams and how it is transferred to team members.

Therefore, having a suitable strategy for knowledge management will be useful for novice users in coping with and adopting the organisational context in less time thus increasing their intent to use a process model.

HYPOTHESIS 7 (H7): Usefulness of knowledge management has a positive effect on the intention to use a process model

4.4 Research Methodology

Since this model was developed targeting novice users dealing with data analytics, a survey was carried out among senior undergraduate students studying a module related to HA and a module related to business intelligence at a local university having around 30,000 students. Even though, both modules are related to analytics, one module deals with analytics in general and the other module is designed specifically for HA. It was assumed that the differences between the two modules increase the generalizing ability of the results (Kim et al. 2012).

Also, as a requirement for the module, they are assigned to read research papers related to analytics every week. Thus, those students were considered to have sufficient understanding of analytics and as they are new to analytic context we considered them as novice users. The survey was carried out at the end of the semester (during the last lecture), with the assumption that students would have gained a satisfactory knowledge of their subject through lectures, assignments and

reading material (research papers). The basic aim of this survey was to identify the factors affecting usage intention of novice users.

It is important to note that we did not use experienced data analysts as they have already used their own personal styles in performing data analytic tasks and they will be biased in their judgments based on their experience and skills developed in the past. Furthermore, since the aim is to develop a process model for novice users, we decided to consider users without prior experience in working in the industry.

The survey was based on the CRISP-DM as it is considered to be a de-facto standard and even if the students have not known specifically the name of CRISP-DM as so, they have learnt similar steps during the course of their module. For example, domain understanding, data understanding, data processing, data modelling, evaluation and presentation are the main steps that they had learnt during the course even though it is not explicitly defined as CRISP-DM.

Operationalization of Constructs

To develop the survey instrument, existing validated scales were used. To measure, the *intention to use a process model*, scales were adapted from Venkatesh et al. (2003) considering the research context of analytics. Items for compatibility and usefulness were adapted from Hardgrave et al. (2003). Items from previous literature were adapted to measure the other perceived characteristics of a process model (Moore and Benbasat 1991).

Seven-point Likert scale ranging from 1 (strongly-disagree) to 7 (strongly-agree) was used in the questionnaire for all the constructs except for usage intention. Usage intention was measured using a scale ranging from 1 (no) to 3 (yes). The survey items (questions used to measure each construct) are given in APPENDIX A. In addition, the gender was used as a control in the model analysis. To ensure the

appropriateness of the questions, the questionnaire was reviewed by three IS researchers prior to the actual survey. Then a separate pilot study was conducted among 20 3rd and 4th year undergraduate students to improve the validity and reliability of the instrument.

Data collection

As survey participants we used undergraduate students studying analytics in two courses. The questionnaire was given as paper based surveys to students. It was decided to not to use online surveys as the students may not be receptive to them and may not be enthusiastic in providing responses to the survey. Even though, online surveys are flexible and one can create and distribute surveys (via emails, social networks) and collect and organize data very swiftly, we decided to use the paper based surveys to ensure participation of all the selected students in the survey. However, the participation in the survey was totally on a voluntary basis. The questionnaire was distributed during the break of the lesson on the last day of the module at the end of the semester with prior permission from the respective lecturers. A brief verbal explanation on what is an analytic methodology and about the survey was given in addition to the explanations on CRISP-DM given in the front page of the questionnaire. As illustrated in Chapter 2 (Figure 2), CRISP-DM is self-explanatory (Swanstrom 2013) and students have learnt these steps in studying their course contents.

A total of 114 completed and valid responses were collected. As a general rule, the minimum sample size should be at least 10 times of the number of constructs (Hair et al. 2006; Kankanhalli et al. 2011). As there were only seven constructs, it was decided that the sample size of 114 is adequate. The correlations of the sample are

given in Table 1. The descriptive statistics indicates that students are between the ages of 20-28 years (mean 23.75 years and standard deviation of 1.75).

Table 1: Correlations between constructs and the dependent variable

	I	RA	C	EU	RD	T	KM	PM	CR
I	0.82								0.86
RA	0.21	0.83							0.92
C	0.22	0.36	0.88						0.91
EU	0.14	0.28	0.58	0.85					0.89
RD	0.42	0.36	0.42	0.41	0.77				0.81
T	0.08	0.20	0.18	0.22	0.13	0.79			0.76
KM	0.29	0.25	0.12	-0.11	0.17	-0.01	0.77		0.84
PM	0.24	0.36	0.31	0.10	0.18	0.14	0.53	0.79	0.85

Notes. Leading diagonal shows the squared root of AVE of each construct, I=intention, RA=relative advantage, C=compatibility, EU=ease of use, RD=result demonstrability, T=triability, KM=knowledge management, PM=project management, CR=composite reliability

Negative values in Table 1 indicate the negative correlation. So for example, KM and EU indicate a negative correlation. However, that is not a concern in this study. The correlation should be less than 0.8 (Gujarati 2003; Gujarati and Porter 2009), and according to our results there is no indication of potential for multi-collinearity. Also, square root of AVE for each construct should be greater than its correlation with other constructs (Kim et al. 2012).

4.5 Data Analysis and Results

The data analysis was performed using the partial least squares (PLS) technique with *SmartPLS*. PLS was selected as it enables to analyse measurement model (relationship between items and constructs) and structural model (relationship among constructs) (Kankanhalli et al. 2004) with multi items constructs and not restrictive on the sample as in covariance based structural equation modelling (SEM) (Kim et al. 2012). Since PLS is primarily intended to be used in early stages of theory

development (Kankanhalli et al. 2004) and as this is one of the first attempts to do a causal predictive analysis on the behavioural intention to use a process model for analytics, PLS was considered to be suitable for this study. Testing the validity of the measurement instrument and subsequently the hypothesis testing were carried out.

Instrument validation

The convergent validity and discriminant validity of the constructs were assessed to demonstrate the construct validity. Convergent validity indicating the extent to which two or more items measure the same construct is examined using (1) standardised path loadings of items, (2) composite reliability (CR), and (3) average variance extracted (AVE), (Kim et al. 2012). The standardised path loadings are significant (at t-value > 1.96) with a threshold of 0.7. It is considered appropriate to have at least 0.7 for CR and 0.5 for AVE (Kim et al. 2012). Thus, based on the results it could be noted that the construct's convergent validity was acceptable. The squared root of AVE of each construct and the CR are shown in Table 1.

The discriminant validity indicates the degree to which items that measure different constructs differ (Kankanhalli et al. 2011). This is satisfied by having a square root of the average variance extracted for each construct greater than its correlation with other constructs (Kim et al. 2012). This is shown in Table 1. Based on the results discriminant validity is supported.

Hypotheses Testing

After establishing the instrument validity, PLS was used for hypotheses testing. Gender was used as the control variable as it is expected that the males may be more willing to take advantage of available opportunities (Arch and Cummins

1989) and prefer a structured process. Age is not considered as a control variable as all the users are from the same age category.

Path coefficients and significant results are indicated in Figure 7. Perceived relative advantage, result demonstrability, triability and usefulness of knowledge management indicate a significant effect on the intention to use the process model for analytics. However, the direction of relationship between triability and intention to use is negative (path coefficient = -0.047), and as such the hypothesis H5 is not supported. All the other significant relationships indicate a positive influence and as such H2, H4 and H7 are supported.

The explanatory power (R^2) is 0.31 and it is above the threshold of 0.10 as specified by Falk and Miller (1992).

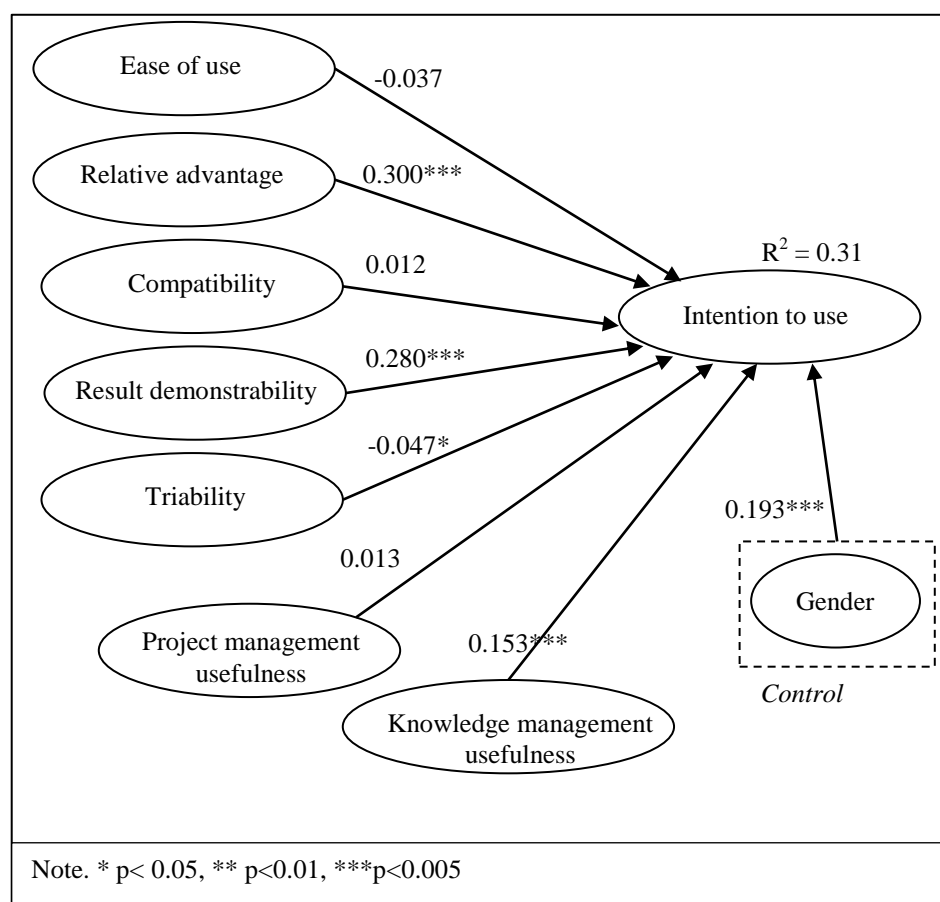


Figure 7: Results of hypothesis testing

4.6 Discussion

Several important relationships were found from this study. First, characteristics such as relative advantage and results demonstrability are shown to be important attributes in a process. Novice users may also like to get a relative advantage over others by using a process model. They will see that using a process model will enable them to start the project satisfactorily rather than going in ad hoc directions. Similar results could be observed in considering the previous studies related to methodology adoption too. Consistently, relative advantage is the only attribute that is significant in those studies while other attributes are insignificant (Mohan and Ahlemann 2011). Even through the study carried out by Riemenschneider et al. (2002) using five theoretical models, perceived usefulness (referred as relative advantage in DOI) was the only construct found to be significant in all models.

Novice users will like to see if the progress of their performance is shown or demonstrated giving them an opportunity of showing their progress even to their seniors. Specially, this will be a motivator and will allow getting further assistance from the senior analysts.

Second, it is noted that novice users value the knowledge management components in a process model. Thus, having documentation will be useful in managing (creation and transfer of) knowledge. In their study, Chan and Thong (2009) also indicate the usefulness of knowledge management in software engineering methodology usage. However, it is important to explore how knowledge management is used in successful HA teams.

Third, it is interesting to note that triability is showing a negative relationship. It is a negative relationship of low significance. Nevertheless, individuals might not try out a new innovation if they perceive risks in doing so or if there is no continued

accessibility (Agarwal and Prasad 1997). Accessibility should be provided through proper information management (access to specific information on usage, e.g. user manual). Furthermore, though it is hard to explore a process without actually using it in a real context, novice users may be reluctant to test a new method by trying it out. The negative relationship shown may be indicating that reluctance and it should be explored further with knowledge management.

Fourth, compatibility and ease of use are not proving significant relationships. Hardgrave et al. (2003) found the relationship between compatibility and software engineering methodology usage to be significant but weak. CRISP-DM like process models are introduced independent of the data, analytic tools or analytic algorithms that are being used. As such compatibility may not be a relevant issue. However, if a practice is more compatible with the type of projects that are been carried out and if they are compatible with existing work practices, the users will be more willing to use a process model (Hardgrave et al. 2003). As such, when developing new methodologies it is important to look into components that are having a greater alignment to actual settings and project types to be included in the process model.

It is interesting to note that ease of use (complexity in DOI) was not significant among all five models used by Riemenschneider et al. (2002). Hardgrave et al. (2003), also found similar results in their study. This is a variation from the technology acceptance studies (Chan and Thong 2009). Rather than considering the ease of use, a higher focus should be given to providing comprehensive and complete specification of the phases and tasks to be followed in the full HA process.

Finally, the relationship with usefulness of project management is not significant. For novice users, project management may not be useful in carrying out their university projects. However, as project management is important in real settings

(Marban et al. 2009b), it is essential to explore how project management can be incorporated in the model in a useful manner to the novice users starting projects in real organisational settings.

There are several limitations encountered in this study and suggestions for future research. First, additional antecedents and interaction effects could have been considered. For example, personal characteristics and individual needs could have been considered as factors that can affect the motivation to use the process model. Mohan and Ahlemann (2011) conceptualize that, acceptance of a methodology will depend on the individual needs and it will motivate them to use the methods. They have considered individual needs as moderators. Second, a large sample size could have been used to further test the robustness of the results and the study could be further extended to other user groups, such as new recruits in an analytics organisation.

4.7 Application of the Survey Results in Process Model

Development

According to the findings from the survey study relative advantage, result demonstrability, and usefulness of knowledge management indicate significant positive relationship with intention to use. In the development of the analytic process model, we considered these three constructs in the following manner.

- Relative advantage (usefulness of the process) – If an innovation is better than the existing approaches, the process model will be having a relative advantage (Moore and Benbasat 1991). Identification of the problems found in previous process models (specifically CRISP-DM as it is the process model considered in this survey) and attempting to solve those issues are essential.

The problems identified are: rigid structure (use a waterfall structure), identification of user requirements and data needs at the beginning, non-selection of the technique based on data collected and limited project planning (only looking at resource management). Addressing the problems identified will allow achievement of effectiveness of the job performance and quality of the work performed.

In addition, it is important to consider the uniqueness of the process model. By incorporating unique components we will be able to achieve higher relative advantage in HA domain compared to other generic process models.

- Result demonstrability – It is important to have higher transparency and better communication of results achieved by using a methodology (Mohan and Ahlemann 2011). Novice users can be more exposed and confident to a methodology if they can see their seniors and peers using such a methodology (Mohan and Ahlemann 2011).
- Usefulness of knowledge management – knowledge management components can be introduced to improve creation and transfer of knowledge about the process used and output generated. Considering the complexity of the health analytic requirements, complexity of the data (Cios 2000) and importance of the decisions in healthcare delivery, management of the knowledge is important.

Thus, based on the identifications made through this survey, USAM model development was initiated with due consideration on addressing issues in previous models and achieving better result demonstrability and knowledge management throughout the iterative model development-evaluation process.

4.8 Summary

Determination of methodological attributes affecting the intention to use the model by novice analysts through the survey was elaborated in this chapter. The conceptual model was developed based on DOI and TAM. Based on the analysis performed on the survey data, it was found that relative advantage (usefulness) and result demonstrability of the analytical model development process and the usefulness of knowledge management are significant on usage intention of a process model for analytics. Findings from this survey were considered in the model development.

CHAPTER 5. DEVELOPMENT AND EVALUATION OF THE PROCESS MODEL

The importance of the development and evaluation of the artefact (USAM) is emphasized in the Design Science Research (DSR) approach (March and Smith 1995) as they are essential for the successful adoption (Stockdale and Standing 2006). Details on the process model development and evaluation are discussed in this chapter.

5.1. Introduction

The development and evaluation of the USAM are carried out as a cyclic process (develop -> evaluate) with two iterations as described in Chapter 3 (Methodology) and in Figure 5. Initially the problems of the latest process models were identified and based on them the design criteria to modify the process model were decided on. The model was developed (or modified) to satisfy the design criteria. As per the action case research approach, there will be an evaluation of the process model after its development. Based on the initial requirement identification by the survey on the usage intention of CRISP-DM (Chapter 4), the process model development was initiated as explained in the following sections.

In Information Systems (IS) research both ex-ante (prior to artefact construction) and ex-post perspectives post (after construction of an artefact) are used in evaluations (Johannesson and Perjons 2014; Pries-Heje et al. 2008). Former refers, to evaluation of the candidate systems and deciding whether to develop an artefact and which features should be adopted. Thus, in DSR approach as explained by Pries-Heje et al. (2008), it is “*theoretically evaluating a design without actually*

implementing the material system or technology". The ex-post perspective can be described as evaluation conducted after implementation of the artefact (Johannesson and Perjons 2014). However, the above authors have considered the evaluation of an operating technology (e.g. tool), and have identified two stages as designing the artefact and then the construction of the artefact. Conversely, when developing a process (not an operating technology) is considered, there will not be an independent stage as construction of the artefact. The design and construction of the artefact are not two independent stages and can be simply called as 'build artefact' as per the categorization of March and Smith (1995). Thus, as highlighted by Pries-Heje et al. (2008) as a means of evaluation of a method, we consider the design/construction as the anchor point (rather than considering the construction as the anchor point as done by other authors).

We decided to use ex-ante evaluation as it allows assessing the prototype quickly without access to users and organizations and it is a useful strategy to get feedback for further improvement (formative evaluation). However, since it assesses a preliminary prototype or design (Johannesson and Perjons 2014) it was decided to incorporate some ex-post evaluation strategies at the end of the implementation of the artefact to get further feedback. This is achieved by applying the process model in an organization and by carrying out interviews among practitioners (in naturalistic settings).

The first model development-evaluation was carried out by the researcher while working as an external analyst in a hospital. This facilitated the evaluation of the process management dimension and documentation approach. As a participant based study by observing the actual work setting was not possible as an outsider to the organisation, supporting dimensions were not considered at this stage. This evaluation

can be considered as ex-ante. Considering the difficulties in obtaining access to the organisation, the process model was evaluated as a preliminary prototype (Johannesson and Perjons 2014).

At the initial step of the process model design (or development), the refining of the data model development process was considered and as such, evaluation of the process using an external project was considered to be adequate. However, when the socio-technical factors in an organisation setting are to be considered it is important to evaluate them in an actual organisation setting.

The final model development-evaluation was carried out while the researcher was working as an internal employee of the organisation. This is considered as an ex-post evaluation (Arnott 2006; Pries-Heje et al. 2008). For ex-post evaluation of intangible benefits, interpretive evaluation approach could be used (Hevner et al. 2004; Stockdale and Standing 2006). Moreover, it allows attaining deeper understanding of the context, which is not feasible to measure through quantitative measures. As mentioned by Arnott (2006), “the assessment of success is a difficult problem for design research studies because it is impossible to determine if an alternative invention would have been more successful or have led to a different outcome, after the research intervention”, the evaluation criteria for our model was based on the perception of its success by the members of the HA department in a hospital. It could be observed that Arnott (2006) and Baskerville and Pries-Heje (2014) have used a similar strategy to evaluate a decision support system methodology and a design case to diffuse best practices among various groups respectively (as the evaluation process an action case research method was used and evaluation criteria was the perceived success of the method built).

5.2. Research Setting

First, the problems in existing models were identified through the literature (Chapter 2) and the survey (Chapter 4). The initial process model developed for HA context was based on CRISP-DM (as it is the mostly used data mining process model as illustrated in Figure 1) and it was modified to address the issues identified through the literature review on existing process models for data mining. The results of the survey on the intention to use a process model in Chapter 4 were considered throughout the thesis study to achieve the relative advantage, result demonstrability and knowledge management to users.

The two main development-evaluations were carried out in two healthcare institutions. For reference the first institute will be referred to as *Hospital X* and the second institute will be referred to as *Hospital Y*. *Hospital X* is one of the major hospitals in Singapore. The Business Intelligence (BI) maturity level (Eckerson W. 2004) of *Hospital X* can be considered as at level 2 – ‘*Tactical*’ as there were limited users and limited focus on application in organisation needs. The BI maturity level is an indication of the nature of requirements of an organization and the type of projects they are involved with. The radiology department of the *Hospital X* provided the access to machine data (around 28294 records) to be used as a test data set to validate the proposed model comparing with a standard model and the model used by the *Hospital X*.

The final development-evaluation iteration was carried out at a Health Analytic Department of *Hospital Y*, which has more than 500 beds. They had a HA team of five members working on different projects with several interns assisting in those projects at a given time. I worked as an intern for 4 months along with two other interns (graduate students studying for masters) at *Hospital Y* getting involved with

their activities and occasionally participating as an observer. The BI maturity level (Eckerson W. 2004) of *Hospital Y*'s HA Department can be considered as at 'Focused' level (level 3) as there is a successful focus on the specific institute needs and funding available as grants on a project basis. Moreover, the management is interested in HA and that interest is created and enhanced among other employees too through internal workshops and presentations.

The main benefit of being an employee (intern) of the organisation was gaining access to senior staff and opportunities to attend meetings and project discussions as such involvements are not permitted for total outsiders. The involvement with the *Hospital Y* paved the way to understand how user requirements and necessary project management, communication management and knowledge management practices vary according to project types which were used in the development of the process model.

5.3. Initial Model Development-Evaluation by Applying the Model in an External Project

First, the literature was reviewed to understand the existing approaches to determine how the artefact should be implemented. Several problems were identified in existing process models (e.g. CRISP-DM). To explore the validity of the issues mentioned in literature and to get a clear understanding on those issues, in this thesis study, CRISP-DM was used in an external HA project as an initial step. As such this is referred to as the initial development-evaluation cycle of USAM.

This development-evaluation cycle was carried out while working as an external analyst in *Hospital X*. I was involved in the project as an outsider, and used their dataset to exercise the process model developed. By going through the proposed

process model using an actual dataset, I was able to identify certain shortcomings and the model was refined accordingly. In this project, three employees were used as informants: a physician, a radiologist and a data analyst assisted us in solving the HA problem. They provided the domain knowledge and direct experience on handling HA projects.

5.3.1. Case Description

Patients of *Hospital X* are provided with radiation oncology services and they use very expensive and complex technologies like linear accelerators. Attempts should be made for optimum utilization of such limited essential resources to provide the maximum possible service to patients. The productivity (treatment workload per day) of linear accelerators can be increased by pre-determining the actual demand for them according to various factors relevant to individual patients like treatment complexity, treatment technique, etc. This study was involved with developing a model to predict the duration needed for each radiotherapy treatment.

An elaborate explanation of the problem and expectations of the project were provided along with descriptions on the tasks that are to be carried out in access gain, requirement gathering to modelling and validation.

5.3.2. Application of the Process Model in *Hospital X* Project

Here, the main consideration was on the evaluation of the process management component of the model development process in performing an external project. The application details of the process model are given below. It is important to note that the application of documental steps is illustrated using clinical data obtained from the *Hospital X*.

Data access

In this study we considered the access to a hospital data source (*Hospital X*) as the starting point (as an external researcher). The access to data sources was obtained by initial collaborations with the *Hospital X* and considering their interests in incorporating HA and discussions on how we can assist them. Discussions were held with key physicians and members of the HA Department, mainly through a gatekeeper (member in HA Department) who helped in identifying the requirements and refer informants having the required domain knowledge to conduct the study. An initial data access document was created describing data sources that are available. In this study we were able to access machine data of radiotherapy equipment (linear accelerators) available in *Hospital X* from January 1, 2013 to August 30, 2013. The data samples included patient treatment types, treatment techniques and patient information with more than 28294 records of 1758 patients' radiotherapy treatments carried out in 2013.

Step 1: Domain Understanding

In the domain understanding stage, it is important to understand the specific requirements of the hospital as well as the problem domain. Non-technical articles (e.g. Wikipedia articles, *Hospital X's* web site) were initially used to study the domain. This provided us with background knowledge of the organization and their expectations. Having a clear understanding of the services they provide and the daily operations they carry out will be of value in this kind of collaborative work. This information and necessary clarifications were obtained from the physicians, radiologists and the gatekeeper through set appointments. For example, the radiologist explained the complete process that they carry out from taking a patient to a radiotherapy room, types of predictions they make, how they make schedules and

how the treatment is carried out, etc. The discussions with physicians allowed us to get basic domain knowledge on the focused disease and how it is diagnosed and treated. In addition, treatments specified in related medical articles too were read to get the domain knowledge. Thus, the stakeholders will be the radiologist, data analysts and the patients.

The organization objectives and requirements were identified as follows:

- Set treatment time – This is performed by the radiologist based on the doctor's prescription. The time taken for each treatment is decided by radiologist based on the severity and location of the tumour.
- Schedule patients – Then the patients are scheduled by the radiologist. Time will be taken for equipment setup and treat patients. There will be several rooms housing the necessary radiotherapy equipment. Based on the treatment time assigned for each patient, the patients will be assigned to each room in a particular order. This will be known as the waiting list and accordingly each patient will be given an appointment to arrive for the treatment.
- Treat patient – patient will be treated on the assigned time and machine utilization details will be documented. This includes machine set up time, treatment time, treatment techniques used, etc.
- Develop KPI (Key performance indicator) – based on the details gathered, time allocation for patients will be refined to achieve maximum productivity in radio therapy equipment usage.

Then the business goals were identified. They were mainly related to improving the productivity and meeting organisation KPIs. Productivity can be improved by treating more patients and reducing delays (patient waiting times and machine idle times). To treat more patients it is important to identify the number of patients that can be

allocated per room correctly. Then to reduce delays it is important to determine the treatment time that will be taken based on the complexity of the patient's tumour.

For our study, it was important to identify the HA goals as per the CRISP-DM model. They are:

- Determine patient treatment duration per patient based on the treatment complexity
- Determine the number of patients per room
- Determine KPIs (key performance indicators)

Also, at this stage it is important to determine the terminology used. Some of the important terms relevant to the study are specified in Table 2 considering the importance of specific terminology used.

Table 2: Terminology related to radiation oncology

Term	Description
Radiotherapy	Medical use of radiation to control or kill malignant cells
Adjuvant therapy	To prevent reoccurrence of tumours after surgery
Curative therapy	To prevent reoccurrence of tumours after surgery
Palliative therapy	To local disease control and symptomatic relief (not possible to cure).
Fraction duration	Time from patients entry into the room until the patient left the room
Non-operational time	Time the device is not treating patient
Dose	Amount of radiation used in the therapy. This is fractionated over a time period.
Treatment fraction	Single treatment dose where the total dose is fractionated over a time period

A project plan was made specifying the project scope, resources required, schedule and the communication plan. The communication was carried out mainly

through emails and informal meetings held at *Hospital X* or at School of Computing when it was necessary.

Step 2: Data Understanding

In this data set, there was protected health information (PHI) data that need to be de-identified. Specific details are mentioned in the data de-identification report. For example, this includes data such as patient ID, NRIC (National Registration Identity Card) number, admission date, appointment time, etc. NRIC number was removed from the data set. The patient ID number was replaced with a new code as such that it will be possible to recognize each individual as some patients may have visited more than once to undergo treatment. However, the admission time and treatment duration data were kept as that information was necessary for the study. The appointment time was removed from the dataset.

The dataset included two tables: patient specific data like patient Id, NIC, name, whether inpatient, appointment time, whether on subsidiary or private (payment type) and machine utilization data like treatment room, number of fields, treatment start/end time, activity, treatment intention, etc. Data indicated that certain patients had undergone treatments more than once.

Main variables in the dataset are explained in the Table 3. This includes explanation on the categories of certain factors (e.g. activity, treatment intention). Furthermore, it is important to specify the mean, standard deviation of continuous values and count for categorical variables.

Data quality was assessed to check whether it is complete and correct. The missing values in the dataset were represented as “NULL” or kept blank. Moreover, there can be repetitions of the data records. Such information is specified in the data quality report.

Table 3: Factors influencing fraction duration of radiation treatment

Factors	Name in data set	Type
Treatment start time	tx_start	Datetime
Treatment end time	tx_end	Datetime
new case	new_case	Yes, No
Inpatient	Inpatient	Yes, No
No. of fields	no_of_fields	Numeric
Treatment intent	tx_intent	Curative, palliative
Activity	tx_activity	See Table 4 for the list of activities
Beam type	beam_type	Electron, Photon, Mixed
Whether wedges used	wdg_appl_yesNno	Yes, No
No. of wedges	no_wedges	Numeric
Whether bolus used	bolus_yesNno	Yes, No
No. of wedges	no_bolus	Numeric

Conceptualization

Based on the literature review, we were able to determine several research questions for the current scenario. However, in this study we focused only on (1) what are the factors influencing the prediction of treatment duration and (2) how to measure the fraction duration? Here, fraction duration will be the dependent variable and the variables like new patient, number of fields, number of wedges, treatment intent and activity were independent variable.

Step 3: Data Preparation

After initial data preparation (e.g. missing data, outliers, etc.) at the data understanding stage data was modified based on the HA goals.

The two tables with patient specific data and machine utilization data are integrated as treatment data. Thus, there was a duplication of patient specific data when associating with their specific treatments (as one patient can undergo more than one treatment fraction).

Several actions were carried out to clean the dataset. The tasks carried out are:

- There were 127 records that were duplicates with same machine utilization and patient data but with different schedule set id and appointment time. We kept only the final recording as an error had occurred due to change in appointment times when saving the timestamps in the system.
- There were 39 rows where the activity, treatment intent and number of fields were with NULL. Thus, we removed those records from the dataset.
- Using “tx_start” and “tx_end”, the treatment duration (“tx_duration”) was calculated. A new column was included in the dataset as “tx_duration”.
- Activities were re-categorized as IGRT, IMRT, VMET, Others and BTE based on the technology. Mapping of activities to technology is shown in Table 4.
- Certain columns had “wedges_count” as blank, as wedges were not used for certain treatments. If “wdg_appl_yesNno” is No, then the “wedges_count” was filled with 0.
- Similarly, for “bolus_count” was filled with 0 if the “bolus_yes/no” is No.
- After consulting with personnel from *Hospital X*, we identified that “no_of_fields” and “NoPF” column both represent the same value. Also, it was confirmed that “no_of_fields” column is more accurate (as there were some discrepancies in values) and such we removed “NoPF” from the dataset.
- Correlation of the independent variables were considered and found “MLC_fields” and “no_of_fields” are correlated more than 0.90 and as such we removed “MLC_fields” from the dataset.
- At the end, further 53 rows had more than 4 columns blank or null. Thus they were removed. Exact records were noted down to replicate the tasks to be carried out in future.

Table 4: Re-categorization of activities at the Radiology Department, *Hospital X*

Technology	Activity
IGRT	IGPROSTATE, IGRTH&N
IMRT	IMH&N, IMNPC, IMPROSTATE, IMRTOthers, IMTHORAX
OTHERS	SBRTLUNG, SBRT0th, SBRTPELVIS, T BODY, T ELECTRON
VMET	VMBLADDER, VMPROSTATE
BTE	C1 -H&N, C1-ABD, C1-BREAST, C1-CRANIUM, C1ELECTRON, C1-EXTREME, C1-MULTI S, C1-PELVIS, C1-SPINE, C1-THORAX, C2-H&N, C2-BREAST, C3-BREAST, S-H&N, S-BREAST, S-CRANIUM, S-EXTREME, S-MULTI S, S-PELVIS, S-SPINE, S-HORAX, S-ABD, S-ELECTRON

Finally there were 28051 records from 1756 patients. Many patients had undergone many fractions. For example, S000001 had undergone 20 and S000002 had undergone 5 treatment fractions during the period considered.

Step 4: Data Modelling

There were several sub tasks performed at the data modelling stage. They are given below.

A. Identify patient treatment profile

Patient treatment profile is identified to get a concise description of the characteristics of the data related to treatment activity and intent depending on whether it is a new case (first fraction). This is only a sampled descriptive statistics on treatment profile. Visualization could be used to understand the variations in each segment.

Descriptive statistics are given in Table 5. Mean values obtained indicates that, the patient undergoing first fraction takes more average time compared to other

fractions. Also it could be seen that certain technologies are not used for certain treatments.

Table 5: Mean fraction duration over radiation treatment intent and activity

Mean of fraction duration	Treatment Activity					
Treatment Intent	IGRT	IMRT	VMET	Other	BTE	Total
Curative	18.9	18.7	13.1	48.1	12.1	16.1
New case = No	18.8	18.5	13.1	46.6	11.7	15.8
New case = Yes	21.8	30.4	13	56.1	22.8	27.1
Curative (adjuvant)	18.4	17.9	12.9	40.5	11.8	13.3
New case = No	18.3	17.6	12.9	39.1	11.2	12.7
New case = Yes	20.6	26.6	14.3	46.5	22.8	23.5
Curative (primary)					10.6	10.6
New case = No					10.2	10.2
New case = Yes					14.6	14.6
Induction-Primary					14.2	14.2
New case = No					13.6	13.6
New case = Yes					23	23
Other, NOS		20.3			17.4	18.9
New case = No		20.3			16.8	18.7
New case = Yes					22.6	22.7
Palliative	28	19.6		49	12.1	12.8
New case = No	28	19.4		47.9	10.9	11.7
New case = Yes		27.7		56.3	20.2	20.6
Primary-Neoadjuvant		18.7			12.2	12.8
New case = No		18.5			11.7	12.4
New case = Yes		30			20.9	21.1
Total	18.5	18.3	12.9	45.2	11.9	13.8

B. Identify factors influencing treatment duration

This was the second goal where most influential variables on treatment (fraction) duration are identified. Usually, in medical data there are a large number of variables in a dataset. Thus, to avoid using all the variables it is important to perform feature selection. For example, R^2 , gini index, principal component value could be used as the selection criteria to select variables.

In this study with *Hospital X* project, ordinary least square (OLS) was used as the HA technique to determine the effect of each factor on fraction duration (attribute evaluation). Partial R^2 value (fraction of total variation accounted for by a variable) was used to determine the influence of each factor on the dependent variable (fraction duration). Selection of this technique is reliant on the HA goal as well as on data (machine utilization data). Here, variables with R^2 value greater than 0.01 were included in the prediction model. We used 13 attributes for the analysis. Parameters selected (using OLS regression) were `number_of_fields`, `activity`, `newcase`, `bolus_count`, `beam_type` and `inpatient`.

C. Predict treatment duration

This was the third HA goal where the treatment fraction duration will be predicated. By using regression we planned to identify the relationship of each variable to the dependent variable through standardized beta coefficients. Later on the decision trees was applied on the same dataset. Here, the dependent variable had to be transformed into a categorical variable. As such in the USAM process model, we had to iterate back in the loop to data preparation step and had to create a new data model. A new data model was created with a version number (e.g. V 1.2) and was stored. Then we had to re-achieve the HA goal of identifying the most influencing factors on treatment duration. Here, we used the technique information gain to identify the order of the factors. There were further iterations using other different techniques. For example, we used neural networks as a HA technique.

Generalized estimation equation (GEE) regression model was used as the algorithm for the prediction technique. Here GEE regression was used instead of OLS as the dataset included unequal repetitions of individual patients having different

fractions of treatments (Delaney et al. 1997a). In OLS it is assumed that there is independence among observations. However, as GEE adjust highly correlated observations, it could accommodate the dependence among them.

The GEE regression model is given below:

$$\begin{aligned}
 \text{Fraction duration} = & 8.940 \\
 & + 4.023 * \text{no_of_fields} \\
 & + 25.203 * \text{activity=other} \\
 & + 4.189 * \text{newcase=yes} \\
 & + 0.388 * \text{bolus_count} \\
 & - 1.549 * \text{beam_type=1} \\
 & - 5.709 * \text{beam_type=3} \\
 & + 0.979 * \text{inpatient=yes}
 \end{aligned}$$

Step 5: Validation

In data analytics it is essential to have at least two datasets for training and testing. The model was developed using the training dataset. The test dataset was collected from *Hospital X* (from September 1, 2013 till December 31, 2013). Other than that we used 10 fold cross validation. The validation dataset was obtained from treatment records from 2014.

Mean squared error (MSE) and root relative squared error were computed based on the test dataset. MSE was calculated as the average of squared error

$$\left(= \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \right). \text{ Root relative squared error was calculated as } \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\bar{y} - y_i)^2},$$

where \hat{y}_i is predicted value for item i , y_i is actual value for item i and \bar{y} is the mean.

As indicated below (Table 6), the results of the 3 models were compared. BTE original model is based on the original model developed in Australia (Delaney et al. 1997a) and the *Hospital X's* BTE model is an adoption of the BTE model by *Hospital X* with some variations to the original model. GEE model is our model. The testing was performed on the same test dataset.

Table 6: Comparison of models

Model	Mean squared error (MSE)	Root relative squared error
BTE original	97.40915	1.398253
<i>Hospital X</i> BTE	63.58833	0.912774
GEE Model	62.86595	0.902405

Even though there was no direct relationship in achieving higher accurate outcome by using the process model (as it aims to provide a guide to users to carryout data analytic projects easily), the GEE data model generated by applying the process model gave better results compared to other data models. However, the developed process will allow streamlining the necessary activities to be carried out in initiating new projects along with maintaining documentation as shown above.

Step 6: Deployment

The process model developed was given as a report to the Hospital X, so they can use it in their machine utilization predications and in setting up of organisation KPI (key performance indicators).

The limitations identified in the application of the model and steps taken accordingly to modify the process model are explained in the following section.

5.3.3. Revisions to the Model

In working on the project in *Hospital X*, it was found that most of the tasks had to be revisited with new ideas and new approaches to be looked into as the initial approach did not work nor gave expected results or the process model could not be straightforwardly applied into HA context. The problems identified by applying CRISP-DM in *Hospital X* project and through literature review were used as additional design criteria to modify the model. The revisions made to the process model with the necessary details are given in the Table 7.

Table 7: Satisfying the design criteria by the model design in *Hospital X*

Problem 1	The process is linear, however actually it is evolutionary
Design criteria 1	Support evolutionary design
Assumption	Not possible to define requirements upfront
<p>This is achieved by</p> <ol style="list-style-type: none"> 1. Having minimally sufficient upfront design, so that, the team can evolve the design when project progresses 2. Modelling small increments and demonstrate the findings to stakeholders. 3. Refactoring without having an undesirable influence on things done in previous iterations (without breaking previously developed models). 4. Configuration management 	
Problem 2	Need to collaborate with the users till the end of the project, Not only at the beginning of the project
Design criteria 2	Support establishment of a collaborative process
Assumption	To maintain collaboration among stakeholders there should be mutual understanding and a commitment to work together.
<p>The design criterion is satisfied through clear guidelines on communication modes, frequency and content to discuss. It is important to have a high degree of communication to avoid conflicts. Also documentation is important.</p>	
Problem 3	No consideration on de-identification of data
Design criteria 3	Protect patient data
Assumption	Privacy of the patients is protected through de-identification and richness in the data is available after that to perform the analytics.
<p>This is achieved by</p> <ol style="list-style-type: none"> 1. De-identification and anonymization of patient data using the HIPAA standards 2. Controlling access to the data. Thus, only a limited number of personnel have access to the dataset and having an authorisation process to gain access to them 3. Gaining internal review board approval before commencing a project 	
Problem 4	No conceptualization of what is to be studied
Design criteria 4	Conceptualization of the problem
Assumption	To use data modelling algorithms the constructs should be determined.
<p>A new phase is introduced into the CRISP-DM model after domain and data</p>	

understanding to conceptualize the model.	
Problem 5	It is not possible to commence the project from domain understanding stage without getting access to a healthcare institute or a dataset
Design criteria 5	Inclusion of data access stage at the beginning of the project
Assumption	It is difficult to gain access to healthcare projects due to data protection regulations.
A new phase is introduced into the CRISP-DM model as the step 1 – Data Access. The healthcare projects (specially external projects) should be opportunistic.	
Problem 6	No version control of multiple files (scripts, data, and documents)
Design criteria 6	Configuration management
Assumption	Multiple versions of data, models and documents are generated
This is achieved by version control to manage versions and changes made in data, models and documents. 1. Organize the files into directories 2. Maintain a version control repository with tagging and branching	
Problem 7	No visual documentation approach
Design criteria 7	Visual documentation approach
Assumption	-
This is achieved by introducing a visual documentation approach to be used along with the textual documentation. AS UML has been successfully used in software engineering and had been extended to various data mining techniques, it is introduced in HA process model too.	

The existing models with a rigid structure make it hard to carryout data analytic projects especially when dealing with complex and bigger projects. As identified in the above project, various data modelling techniques were used on the dataset and based on the technique selected the dataset was required to be modified. Sometimes, it was necessary to ask for new data types and for clarification on how certain values should be considered. Waterfall approach used in other available models was not suitable with ambiguity of the requirements. Thus, it was noted that

agile approach is more appropriate for HA projects. In the application of agile methodology, there are several factors that need to be considered.

First, it uses an incremental, iterative and evolutionary approach. A preliminary design plan (not all content of the analysis) will be made to initiate the project and to support interaction with stakeholders (to get feedback). Thus, the conceptual model built at the beginning will evolve to a physical analytic model with necessary flexibility to allow changes during the project. Furthermore, success of an advanced analytic project depends on the ability to improve the outcomes through an iterative feedback loop.

Second, unlike in waterfall method and other sequential methods where interaction among stakeholders occurs only at the requirement gathering stage (at initial stages of the project) and limited interaction on the project later; agile method supports continuous collaboration between the analyst, sponsors and users of the system. The collaboration can be maintained through establishment of proper communication channels between stakeholders.

Other than the introduction of agile approach, other challenges such as access to data, protection of patient data, conceptualization of the model and configuration management are shown in Table 7 on the basis of satisfying the design criteria mentioned in the previous section.

Gaining access to dataset is also an important factor. Especially when dealing in healthcare context, it is important to get access to the dataset first as it could be a bottleneck to commence the project.

In healthcare context for the protection of the patient privacy there is a need to de-identify the data. There are certain challenges in de-identification of the data while maintaining sufficient richness to be used in the analytic process. It is important to

note that data is collected for the purpose of treating the patient rather than to be used in secondary use (Cios and Moore 2002).

While carrying out the project, the importance of having a conceptualization stage where the research questions will be determined (may not be the final research question that will be ultimately handled by the project) was identified. Without, a preliminary idea of the direction of the study, it is difficult to commence the data preparation and data modelling. To perform the data processing and data modelling, it is important to determine the research question that is to be solved and what attributes to be used. Thus, even though it was not in CRISP-DM, we had to conceptualize the problem for this project. Conceptualization may not be the development of a set of hypotheses (it is useful in statistics but not in data mining) (Schmidt et al. 2008). For data mining projects, it is important to distinguish the relevant attributes for the model rather than using all the attributes. Using all the attributes in a dataset is not a viable option (kitchen sink approach) and there should be necessary justification for using a variable in the data model. This step is mainly performed through expert advice.

Multiple versions of data, models and documents were generated while performing the HA project as an external analyst. This becomes more complicated when dealing with a team of analysts working on the same data in an organisational environment. Thus, while it is important to maintain an original copy of data, model etc. in a known location the other versions created with changes made should be properly managed. There should be a proper version control (Marban et al. 2009b) enabling easy access to earlier versions and also to avoid any mix-up of versions.

It is a good practice to store all the files (e.g. scripts, data, documents, lookup tables, etc.) in a central version control repository by organising into different directories (Marban et al. 2009b). This is especially useful in managing the data files

as well as the script files (e.g. R code). Also, tagging (to label a group of files) could be used at each iteration to mark significant variations and branching (create several paths from the main project) for modifications made to a particular model based on the different requirements (Collier 2011).

Furthermore it is identified that it is hard to use textual documentations to represent the association between user requirements and organisation goals as well as HA project goals. Having a diagrammatic representation would be easy to comprehend the details by both the medical practitioners and data analysts. Also, having a means to represent association of requirements, goals, techniques and tools used etc. too will be useful.

The consideration of the patient data protection (uniqueness of medical data mining) and addressing issues related to CRISP-DM (e.g. linear process model, no conceptualization of the problem) will allow achieving relative advantage. Furthermore, version control and establishment of user collaboration will facilitate knowledge management.

5.3.4. Revised Model

The revised process model for HA is given below (Figure 8). The new steps data access and conceptualization are included in the revised model. Furthermore, while maintaining the same connections as in CRISP-DM, a new connection is included between data validation and data preparation. This is to represent the changes made to the dataset when the data model developed has not given expected results after the validation. Initially the model iterates to domain understanding stage after validation. However, after validation of the model, if expected results are not obtained there is a possibility of moving back to refine the model after modifying the dataset.

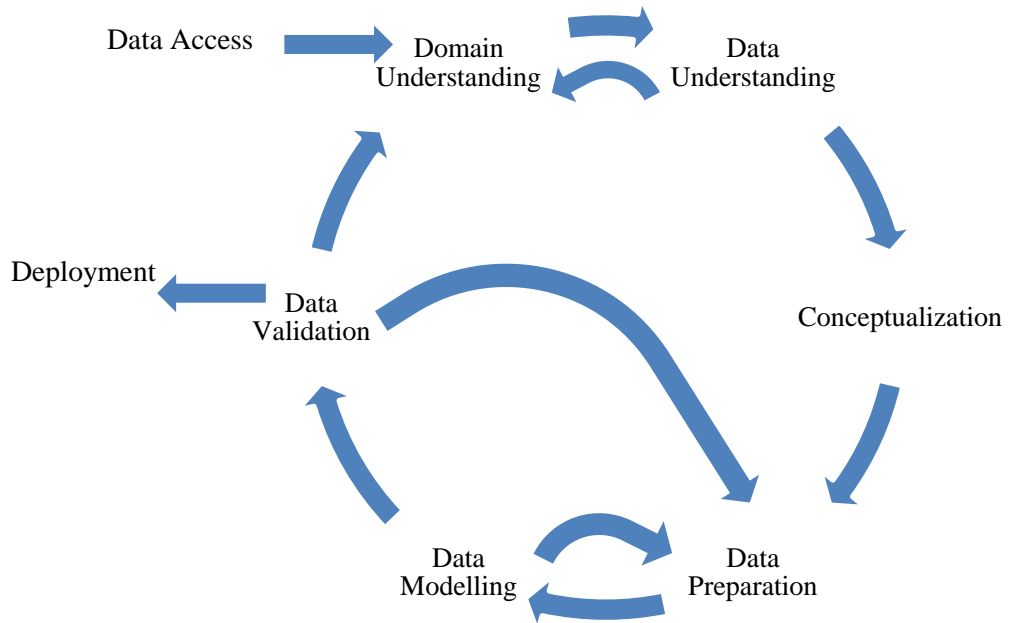


Figure 8. Revised CRISP-DM model

In addition, emphasis is given to continuous user collaboration and evolutionary model development as discussed in the previous section above. The model was further evaluated and refined by working in an internal project of a hospital.

5.4. Final Model Development-Evaluation in a Hospital while

Working as an Intern

Action case research approach was used for development-evaluation of the model as mentioned above as an internal employee of the *Hospital Y*. This was very important as it was carried out in a real organisational setting. In this stage, the aims were to get an actual understanding of the work carried out in a Hospital's HA department and to evaluate the applicability of the process model. The advantage of this cycle is that in working as an intern in a HA department of a hospital, I was able to gain access to staff members, other stakeholders and the organization processes (Arnott 2006). The

staff members in the HA department were aware that the case was utilized for a research study in developing a HA process model.

5.4.1. Case Description

Context

There are many projects running simultaneously in the HA department of *Hospital Y*. Most of the HA projects are focussed on operational activities of the hospital to support their daily activities (e.g. forecasting patient flow to Accident and Emergency department, patient discharges, etc.). The HA department had undertaken several clinical based projects (e.g. risk stratification of patients) with necessary bio-medical validations too. In addition, programmes related to population health too are carried out there.

The staff-members are guided by a team head widely recognised in the industry with indispensable experience in HA as well as in management. The head provides the necessary guidance and resources to the team to carry out projects effectively. These relatively young staff-members demonstrate familiarity from versatile backgrounds and are keen to learn about numerous health domains while participating in multiple HA projects at a given time. Whenever the staff members become unsure of the direction to proceed, they consult other senior members who had worked in similar projects and they depend on the Head of HA department too for guidance.

The data processing and modelling in projects are mainly carried out using R statistical language (*RStudio*), excel and SAS tools including *JMP*. They presently use *Qlikview* software as the dashboard for viewing data.

Before commencement of the internship, I was asked to refer the eBook; “*Forecasting: principles and practice*” which gave explanations on forecasting time

series data using R. At the beginning I was given selected patient movement data and some previously used codes related to prediction of accident and emergency department data after necessary de-identification to be used in forecasting as part of my involvement as an intern.

Approach

As an intern, a meta-diary was kept by me to record the daily activities performed and observed in the HA department to reflect the operational activities. Also the interviews with staff members of the HA department participating in various types of HA projects were transcribed for data analysis. The mirroring technique proposed by Myers and Newman (2007) was used to draw the interviewee's opinions and understandings in their own language. It was used as it is one of the most commonly used approaches to extract information from qualitative interviews (Lyytinen et al. 2009; Newman and Zhao 2008). First, the staff members were requested to explain their daily work activities to be followed by their experiences, practices and later their view of developing a process model.

The interviews were carried out as formal interviews mostly lasting for about 30 minutes each and with a few exceeding more than one hour. The interviews were open ended while maintaining the freedom and control using non-leading and non-passive questions. In addition, more information was gathered through daily informal discussions held during work and lunch breaks. Later on, the model was presented to the members of HA department for feedback and model was further revised based on their comments. *Hospital Y* has vetted the specifications only for factual accuracy. However, I was free to express my own observations, opinions and postulations.

The data from each narrative (personal and interviews) was organised and coded based on the three supporting dimensions determined previously. Later, the

narratives were compared against the literature to synthesize and amend the model. For data collection and data analysis, the SPS framework of Pan and Tan (2011) was followed considering its detailed instructions on carrying out a case study.

5.4.2. Project Variations in the Case Organisation

The projects handled in that HA department can be separated according to difficulty (simple or complex) and clarity (clear or ambiguous) of the requirements.

Figure 9 illustrates the variation of projects based on project management, communication management and knowledge management in a grid. Other than that, the projects could vary based on the profile of the project and the urgency. For example, projects requested by the CEO (high profile project), are considered as of high priority prompting regular meetings. If the project is urgent (mostly a simple task that can be done in a short time), then it would be generally sequential (request → response) or with a limited number of iterations.

The case organisation (*Hospital Y*) has useful and efficient practices to carry out their HA projects. Novice users commencing HA projects can use these fundamentally useful practices. These practices and challenges are explored from three perspectives; namely, project management, communication and knowledge management. Any specific details related to process management (model development oriented) were not specified due to concerns of the case organisation.

		Difficulty	
		Simple	Complex
Clarity	Clear	PM – No or less iterative with fewer revisions CM – Less frequent and less use of rich media KM – Less documentation	PM – Iterative with revisions CM – Frequent with less use of rich media KM – More documentation on the analytic process and the model
	Ambiguous	PM – Iterative with revisions CM – High frequent and use rich media KM – More documentation on requirements	PM – Iterative with many revisions CM – High frequent and use rich media KM – More documentation requirements, analytic process and model
(Note: PM = project management, CM = communication management, KM = knowledge management)			

Figure 9: Project types classification

5.4.3. Application of the Process Model in General

The following practices can be considered to achieve successful HA project outcomes.

Project management

Self-organization among team members could be observed in the HA department of *Hospital Y*. First, the data analysts take initiative to perform the project and whenever there is an issue with the shared understanding, they communicate with the client through emails rather than waiting for the project manager or department head. Second, with how many new people they engaged as collaborators or partners is considered as a performance indicator of the data analysts. Third, the management

depends on the data analysts to detect signs of trouble in their projects and inform promptly to take necessary actions. Thus, the project manager can be considered as a facilitator. That is, rather than managing tasks, the team head will be focussed on removing the barriers (avoid disruptions by providing what is required and buffering external pressure) to do the project and management of the team.

As in other projects, top management support is important for the smooth functioning of the project. Other than that different levels of the organisation hierarchy are involved in the project. Thus, the stakeholder coordination is important. Project team is composed of planners, doers and consumers (Collier 2011). At the beginning of the project, it is important to identify the roles of each member. It is important to note that individual members will play multiple roles and teams require personnel with necessary skills and expertise. While planners are mainly the senior management, project sponsors who act as facilitators and project champions may not be directly involved in the analytic process. Doers (data analyst, ground staff of the requirement providing department) are involved with performing the data modelling and work in the project daily. Consumers will use (directly or indirectly) the outputs generated by the doers.

Furthermore, case data indicates that simultaneous project handling by each individual analyst is useful in dealing with unforeseen interruptions in projects. In other contexts focussing on one project at a time is encouraged to avoid confusions. In HA context (based on case organisation), this improves productivity of the team. Task switching using alternating-runs procedure could be applied (Rogers and Monsell 1995). However, it is important to schedule projects in a way that the deadlines of concurrent projects do not fall in the same period. Moreover, prioritization of these

multiple projects is important. It could depend on the urgency as well as on the manageability.

Deciding on a time frame to terminate a project is another issue faced in carrying out the project. There are so many possibilities that should be looked into for continuous improvement of the results making it an unending process. However, with proper project planning and due consideration to the main focus of the group working on the HA project, it is important to decide on the appropriate time and conditions under which the project can be terminated. An analyst would be interested in improving the accuracy of the results. In addition, users should be confident that the model meets their requirements. User acceptance is facilitated through user collaboration and as such at the end of the project, the sponsors are able to articulate their requirements (may have changed and more refined) and gain a better understanding of HA. This can be represented through alignment of expectations of the analyst and the users (Collier 2011).

Due to the size of the HA team and the scope of the project, most of the projects are performed with one analyst. There are some projects involving 2 or 3 analysts. The individuals in the HA team establish relationships and develop a shared understanding. As such the requirements providers who worked in previous HA projects will usually work with the same analyst. This enables analysts to specialize in a particular domain area in HA, making it easy to understand the problem and the model as they have prior understanding on the domain, data, and user expectations. Moreover, when there is shared understanding the need for face-to-face discussions will be low. Sometimes analysts are purposefully rotated to increase the breadth of domain knowledge and this will enable newly recruited analysts to find where they are more comfortable with their expertise.

Communication management

Sometimes there may be miscommunication between the person who is articulating the requirement and the analysts. It is noted that there are regular face-to-face discussions with the stakeholders to understand the problem from different user perspectives and to review the process (conveyance of project information to convergence). This allows the participants to observe the facial expressions and body language of others to confirm whether they understand the message or whether further clarification is required. It is noted that the stakeholder discussions are more structured and more focussed as an agenda is prepared prior to the meeting. Furthermore, face-to-face meetings create a strong social presence allowing them to collaborate effectively and easily with a sense of togetherness. In addition, meetings create soft deadlines making it easier to plan. However, the meetings may sometimes extend the project completion date. When dealing with busy senior management and clinicians, finding a common time will be hard leading to project delays. As too many face-to-face meetings can cause project delays, sometimes it is advisable to use other asynchronous media for transmission of information.

According to MST's communication capabilities, this provides an immediate feedback (answers for the questions will be received immediately). Symbol variety is higher as the gestures and voice tones are cues to realize the reception to the message. However, the parallelism will be lower as it needs full attention of the participant.

It is observed in the *Hospital Y* that presentations are used to pass information to users. Analysts' make use of presentations to demonstrate the current results. Presentations are useful to indicate the progress of the work and to get the necessary feedback at the same time as presentations allow immediate feedback by the participants though it may take some time to process the information conveyed.

Visual cues used in the slides and the tone of the presenter can be used to highlight important issues (symbol variety). Furthermore, it is observed that presentation slides (more formal and structured) are used as a substitute for a document repository too.

Emails are used to communicate and pass information between the team members and the clients. As per literature, emails (less rich media) are used for tasks with low complexity and high certainty. Similarly, in *Hospital Y*, most of the messages relevant to various tasks are communicated through emails as there are only a few requirements to make someone understand the problem. Emails are useful for task assignment and status reporting. Moreover, the meeting minutes are emailed to pass the details of the contents discussed and agreements made.

During the formal interviews in the *Hospital Y*, the employees indicated that emails are used as a formal mode of communication and used as a document repository (folders with proper labelling). It is observed that some of the emails sent (queries or results) are not responded immediately by the receiving party (sometimes taking even days) due to very busy schedules of the partners (as for them HA is secondary when compared to providing patient care). Thus, the promptness of the feedback is low in emails. Emails save time as one can perform other projects or tasks while waiting for a reply. Moreover, the ability to rehearse is high as the sender can rethink and rephrase the message before sending. In addition, mode of communication could vary based on the individual preference as some clients may prefer face-to-face communication over emails and vice versa.

Knowledge management

It is observed in *Hospital Y*, that most of the projects are not performed isolated in one department. For example, if the project is carried out in the pharmacy, the analyst may need to understand the operations of other departments like specialist

operation clinics, accident and emergency units etc. too. As the data analysts require data from other departments for data modelling, they should be aware of operations and kind of data available in those departments too. Thus, occasional rotation of members provides an opportunity for the analysts to learn about the other domains (departments) and kind of data they gather and their workflows. Moreover, HA department staff meetings provide a chance for the analysts to learn from others. Other than that documentation is a good practice as other analysts can use those as reference material.

In model development, the knowledge pertains to (1) process used (e.g. domain, different standards to extract and process data, coding standards, tools to use for different situations) and (2) the output (the analytic model and the interpretations). Under knowledge generated from output, it is important to have knowledge on what to do with the model (Chan and Thong 2009). It is essential to know the limitations and under what real conditions they are applicable. By observing steps followed by another experienced data analyst (especially as a novice user working under a senior analyst) their best practices can be adopted.

Moreover, the organisational standards and approaches used to extract data can be available as reports or as presentation slides. Documentation made at different stages of the project such as at the requirement gathering, project goal identification etc. is useful for knowledge transfer. Thus, it allows to take full advantage of the HA process model by learning, capturing and reusing experience (Lindvall and Rus 2002). The quality of documentation depends on the amount of effort an individual is prepared to put, deciding on what can be shared and selecting appropriate dealings to be documented. The level of comprehensiveness of documents depends on the complexity and the clarity of the projects.

Furthermore, knowledge about the process of modelling can be gathered by going through the code used. For example, R script will be written for the whole process of data extraction, processing and modelling with necessary comments where required. This provides a guideline to the novice user on type of processing performed on data, shortest approach (codes) to process the data, and how to model the data. The knowledge about the output that is the resultant data model (e.g. forecast model) will be presented through presentation slides and could be shared with users by explaining the model and rules generated from the model.

As found in analysing the data accessed from the *Hospital Y* (case organisation), the ability of an individual is improved through training by going through presentation slides and codes on a previous similar project and polishing up knowledge by going through text books on analytics. Furthermore, this depends on the past experience and thus, it can be associated with new projects. Social rewards (e.g. improve recognition, for reciprocity) are important in motivating individuals in sharing their knowledge. The opportunity could be provided to share knowledge by reduction of distance between individuals through informal networks. HA department team having lunch together and discussing about projects during lunch allows informal interaction and sharing of knowledge. Team colocation positively impacts on project success (Ambler 2009). As noted in the HA department of *Hospital Y*, team members sitting next to and facing each other (through physical setting and the seating arrangement) provides the opportunities to share and collaborate (ultimate colocation). Also, if the experts are in close proximity they can learn through observing how others approach a particular problem. Moreover, encouraging working in pairs purposefully can create opportunities as it allows retention and transfer of

knowledge and reduces the risk of depending merely on one analyst. However, this may not be possible with smaller teams.

It was observed in *Hospital Y*, that the organisation wide knowledge is created, retained and shared through formal practices like sharing experiences by attending conferences, book review presentations, senior management presentations on 'vision alignment', inviting external speakers, admin meetings on quality improvement and sharing good ideas and encouraging others to give feedback to them. Furthermore, this allows shared understanding among employees.

With the agile methodology, since much emphasis is given to collaboration, it is important to consider how relationships between individuals and groups exist. The connection will depend on the intensity of communication, frequency of communication and the social similarity. A strong connection between parties can be achieved through frequent discussions and direct relationships. This type of harmony was observed in *Hospital Y*, where the users too are considered as a part of the project and as such they will have a sense of the ownership of the project. It leads to knowledge transfer among stakeholders.

In a study carried out, Weber and Camerer (2003) have indicated that it is hard to transfer knowledge to unfamiliar partners as they focus on different aspects and even longer explanations will not work. Similar observations could be made in the case analysed too. This could be avoided by having meetings with stakeholders frequently showing the progress made allowing continuous transfer of knowledge from users to analysts on domain, data and requirements instead of communicating with them only at the end of the project. Continuous interaction and communication between members allow to set a form of transactive memory systems (Wegner 1987)

unknowingly. Thus, whenever there is a problem in understanding the data or the process, they can directly refer to the person with particular expertise.

Individual skills

Other than the three dimensions considered at organisation level, individual factors are also important in carrying out HA projects successfully. The skills of the analysts can be separated as hard skills and the soft skills. Technical skills are considered as hard skills. As most of the projects are not too technically complex (e.g. developing computationally intensive data science approaches) and since *Hospital Y* is using established practices and approaches to analytics, it is more straightforward to learn and teach technical skills (e.g. by searching online blogs or forums, reading research papers or reading a reference book) compared to soft skills. Some of the soft skills important to HA are understanding requirements, grasping views of others and separating positive and negative ideas, problem solving skills, presentation skills, negotiating skills, etc. In contrast to software engineering, the data analyst should give due consideration to soft skills too in addition to hard skills. As the results need to be presented to the senior management, it is important to craft the ideas.

Since HA work is highly domain specific, domain understanding is vital to succeed in HA projects. Especially analysts are not familiar to the health domain and since the projects for different departments are performed in the centralised HA department, the analysts should have a willingness to learn the unfamiliar contexts.

5.4.4. Application of the Process Model in Complex and Ambiguous Projects

(Project A) in *Hospital Y*

Project A's requirements are complex and ambiguous. For example, a requirement in developing a productivity matrix for hospital staff is complex and ambiguous. It is

ambiguous, as the expectations are not clearly defined (measure productivity of whom, on what basis/perspective, used for what, etc.). Also, it is complex as one needs to consider different levels and do background study, as it is not purely data driven. As such, the project could take a longer period to complete.

Project management

Particularly for these types of projects, project management is essential. It becomes more important as the final outcome of the project is not known beforehand and project becomes more complex and new directions are identified as the data is explored in deep; as such it is very hard to manage the project on time. The views expressed by the data analysts indicated the appropriateness of using the agile approach as highlighted in the previous section. Besides, application of agile concepts in *Hospital Y's* practices was noticeable even though informants did not mention it specifically. Considering the uncertainty and changes made to the problem statements in the long term projects as the project progresses, envision-explore (rather than request-response) cycle in project management can be observed. This is in accordance to the APM framework proposed by Highsmith (2009). Envision is understanding what is to be done and how it is to be done (Collier 2011). Explore stage focuses on starting the HA with a simple iteration, reviewing with users and exploring possibilities of expanding the project. This is an iterative planning process with review of project scope. Thus, collaboration with stakeholders is an important aspect.

Communication management

It is important to have both conveyance (to make correct conclusions about the problem) and convergence (to move forward in the project) in dealing with the

projects. As the stakeholders are from different backgrounds using different terminology and there is ambiguity in the requirements richer media formats are appropriate for both processes. Face-to-face communication from requirement elicitation till the presentation of the results is important (in all stages). The communication will be less regular with the senior management compared to the middle managers and the junior staff (ground level staff). Senior management as planners will be more involved at the beginning of the project and there will be less involvement during the analytic process. Middle managers will be more involved throughout the project and will aid in getting the data and other resources. Ground level staff will be actively involved in providing the necessary aid in understanding the domain and data. Also, before communicating with senior managers and middle managers, it is important to email the questions and meet them to discuss the issues.

Knowledge management

With the complexity of the project, the knowledge management should be done from requirement gathering stage till the completion of the project. The created knowledge should be stored for reuse in future projects. This allows to complete, complex and ambiguous projects with similar requirements or background in less time by learning from the steps that are followed in previous successful projects. The requirements, users (and their expectations), situational assessments (risks, feasibility), policies and regulations to adhere to and approval process, types of data used (with reasons for their usage), data processing steps and data modelling approaches can be noted. Moreover, it is important to include the interpretations of the results generated after data modelling as well as how the results are deployed or used by the users.

5.4.5. Application of the Process Model in Simple and Clear Projects (Project B) in *Hospital Y*

Simple projects with clearly defined requirements are considered here. As an example, descriptive analytic project or forecasting patient flow to accident and emergency units can be considered.

Project management

Most of these projects can be performed with less iteration. That is, with fewer revisions. It could be observed that some projects are performed in a sequential manner where a response is given as a request. Usually, sequential projects (no iterations) are very short duration projects lasting few hours to 2 or 3 days. Thus, scheduling of milestones and meetings are not structured as in other complex projects.

Communication management

Communication would be available in all stages of the project. The selection of the communication process depends on the stage of the analytic process other than the type of the work. In *project B*, for conveyance of the domain knowledge and the requirements at the beginning, project team members have a face-to-face communication. During these meetings, analysts and project sponsors will come to a shared understanding of the requirements. Since there is low uncertainty and complexity, the convergence can be performed using less rich media like emails. The frequency of the meetings will be less compared to *project A*. Furthermore, during the data understanding, data processing and modelling phases, most of the communication will be maintained through emails with occasional meetings. At the end, results may be emailed to users. If more clarification is required then there will be a face-to-face meeting.

Knowledge management

In *project B*, there will be less generation of knowledge compared to *project A*. Since the project is simple and clear, another analyst will easily understand the requirement for future projects too. However, as noted in *Hospital Y*, most of the content is documented as email, or in programming code and presentation slides. In documentation it is important to maintain the versions of data used as well as the actions performed to generate the results in a shared folder even though it is not detailed as in *project B*. For example, requester names, analyst's names, results and interpretations are important.

5.4.6. Revisions to the Model

Based on the suggestions made by the data analysts in *Hospital Y* and based on the observations made there, a set of new design criteria were identified (Table 8). Basic assumptions considered and the actions taken to satisfy the design criteria are indicated.

One of the major issues identified in existing models is that the variations to the model are not considered based on project type. Initially, at the beginning of the process model development (at problem awareness stage in Design Science Research), it was considered that projects can vary as descriptive analytic projects (simple projects) and advanced analytic projects (predictive and prescriptive analytic projects). However, during the action case study at *Hospital Y*, it was identified that the projects actually do not vary as such. Instead, they vary based on the requirements as given in Figure 9.

Similarly, the existing models do not consider the importance of communication and organisational level knowledge management in data analytic projects. However, during the experience gained through working as an intern in

Hospital Y, the importance of communication and organisational level knowledge management was identified especially in complex projects they handle. Particularly, with the limited domain knowledge those two components were found to be very important in a HA process. Based on the observations made in the case organisation project management, communication management and knowledge management are included into the HA process model as supporting dimensions. The data model development technical process will be supported by these three dimensions.

It is important to note that the model was fine-tuned responding to uniqueness of medicine (Cios and Moore 2002), I personally experienced while working in the hospital. A higher emphasis is given to deal with heterogeneous data, privacy and social issues, statistical philosophy and special status of medicine. In the previous model development-evaluation cycle, the model was refined to include components to ensure privacy of patient data. To address uniqueness of medical data the additional factors mentioned in Table 8 are considered.

Table 8: Model improvement satisfying the limitations in the design criteria observed in *Hospital Y*

Problem 1	One model fit all projects
Design criteria 1	Project variations are available based on requirements
Assumption	One fit all model is not suitable as the usage of the process model varies based on the project type
This was achieved by identifying that HA activities vary according to the complexity and clarity of the project requirements. Thus, variations are identified for simple and clear projects Vs. complex and ambiguous projects.	
Problem 2	No consideration on communication
Design Criteria 2	Communication of information between stakeholders
Assumption	Communication is required at all stages of the project. The selection of the communication process depends on the type of the work.

<p>This was achieved by</p> <ol style="list-style-type: none"> 1. Using media of low synchronicity for conveyance activities and media of high synchronicity for convergence activities. In low uncertainty and complexity, the convergence can be performed using less rich media like emails. 2. Having regular team meetings and overall project stakeholder meetings 3. Establishing guidelines on communication form, schedules and content (set agenda). 	
Problem 3	No consideration on organisation level knowledge management
Design criteria 3	Support knowledge management
Assumption	Knowledge about the process will be created, retained and transferred.
<p>This was achieved by</p> <ol style="list-style-type: none"> 1. Improving the ability (through observation, going through previous work and training) 2. Improving motivation (through social rewards like recognition, appreciation, organisation culture) 3. Improving opportunity (through team colocation, informal meetings, teamwork) 	
Problem 4	Bottlenecks in projects due to delays in response from stakeholders
Design Criteria 4	Handle concurrent projects
Assumption	There are external interruptions (e.g. clarifications, access rights)
<p>This was achieved by alternating between projects rather than waiting for the feedbacks. Through continuous alternating-run procedure, team members can keep in touch with tasks to be performed in each project.</p>	
Problem 5	Limited consideration to uniqueness of healthcare
Design Criteria 5	Include components to address uniqueness of healthcare
Assumption	-
<p>This was achieved by</p> <ul style="list-style-type: none"> • Heterogeneity of medical data – code data using a standard codification system (e.g. ICD 10) to avoid complexity in data, reduce the knowledge gap between medical professionals and data analysts by close collaboration and consultation, visual representation of user requirements and project goals in manner easy to comprehend by professionals from both domains. • Ethical, legal and social issues – De-identification and anonymization of patient data when accessing data, gaining internal review board approval for a project and controlling access to data 	

- Statistical philosophy – dimension reduction through feature selection at the data preparation stage, data pre-processing to handle missing, incomplete and inconsistent data based on advice from medical experts, selection of data models with high transparency and comprehensibility and visualization of the results.

Moreover, with the busy schedule of the other stakeholders and their differences in priorities there were continuous delays in projects in *Hospital Y* due to delays in their responses and feedback. Specially, finding a feasible common time to have a discussion is very limited. Also, there are significant delays in gaining access rights to data because of the requirement to protect patient data. The required portions of data are extracted and given to the data analysts by the IT infrastructure handlers (different groups handles the IT system from the data analysts). Thus, these observations made the requirement of working on concurrent projects simultaneously to achieve maximum productivity of data analysts.

The final refined model is given in Chapter 6 with full description of the process model with the supporting dimensions project management, communication management and knowledge management.

5.5. Evaluation Outcome

It is impossible to assess the success after a research intervention in design research studies to conclude whether an alternative intervention could have been more successful or could have directed to a different result (Arnott 2006). In a prior study (Finlay and Forghani 1998), it is stated that success of an intervention depends on repeat use and user satisfaction.

The main indication of the success of the unified model developed by our study is indicated by the acceptance of the process model by the staff members and senior management of the *Hospital Y* and their request for a report for future reference

by the new analysts. Furthermore, continuous involvement of the team head of the HA department during the refining of our model through observations made in the *Hospital Y* indicated his interest in the project and its repeat use. A similar approach had been used by Arnott (2006), to indicate the success of a new project by considering the opinion of the managing director.

5.6. Summary

The model development-evaluation carried out using an action case research method is explained in this chapter. Initially, an ex-ante evaluation was carried out as an external data analyst. In working as an intern in a Hospital with a satisfactory HA department, I was able to gain insight into the actual HA processes carried out by a centralised team. The action case study showed that the process model is feasible and effective. The use of the process model provides the data analysts a clear strategy to improve the data modelling process. In the action case study, the process of improving the model involved a group of experienced data analysts working in HA projects. The final model was developed based on the suggestions made by them and the problems experienced in that hospital. The success of the project was argued based on the opinion of the HA team and the senior management of the *Hospital Y*.

CHAPTER 6. PROCESS MODEL FOR HEALTH ANALYTICS

This chapter describes the Unified Structured Analytic Model for Health Analytics. A detailed interpretation is given for each stage of the process model. Subsequently the documentation steps are discussed.

6.1. Introduction

With the increased use of HA and the recognition of its significance to the healthcare sector, numerous new studies have been conducted and published using healthcare data by relevant professionals and researchers. However, they have not provided a proper consolidated structure representing the complete process of HA nor considered the variations to the process depending on the project requirements. Thus, it is important to develop a well-defined process facilitating necessary adjustments to accommodate requirement changes in HA. Such a new process model for HA using Designed Science Research (DSR) approach by adopting significant and related components from software engineering processes and data mining processes is proposed in this thesis.

This unified structured process model is developed specifically targeting novice users carrying out HA projects. The term ‘structured’ in the Unified Structured Analytic Model (USAM) refers to the arrangement of steps in a highly organized and in a definitive pattern. That is, the proposed process model will be a well-organized methodology with distinctly defined steps intending to improve the completeness, ease of use, consistency and relative advantage. Here, ‘unified’ stands for consolidated or full representation of an entity.

Through this chapter, the USAM is explained along four dimensions. To maintain the research rigor (in DSR), several theories were considered in designing

the process model as specified in previous chapters. The steps necessary for documentation of the complete HA process using UML too are provided in this chapter.

6.2. Overall Structure of USAM

As depicted in Figure 10, there will be associations between project management, communication management and knowledge management. These supporting dimensions assisting the process management are explained in Chapter 2 with appropriate theoretical support. A modified CRISP-DM model is used to manage the process addressing the limitations of various analytic process models found in reviewing the literature. The process model is presented as a methodology along with the supporting dimensions in each of the steps. Therefore, practitioners can follow the process model and its steps without having to worry about the segregation into individual dimension. The changes made to each step based on ex-ante and post-ante evaluations carried out before and after the initial designing of the model are mentioned in the modified model.

Application of these methodological steps depends on HA project type. HA projects are classified based on the difficulty and the clarity of project requirements (*project A*– complex and ambiguous and *project B*– simple with clear requirement). For projects of type A, where the problem is complex and the project requirements are not clear the agile approach will be more suitable. As projects of type B have clearly defined requirements and a simple problem, a sequential process (or less number of iterations) could be utilised with less interaction with the stakeholders. In this study, only complex and ambiguous projects (*Project A*) and simple and clear projects (*Project B*) will be considered.

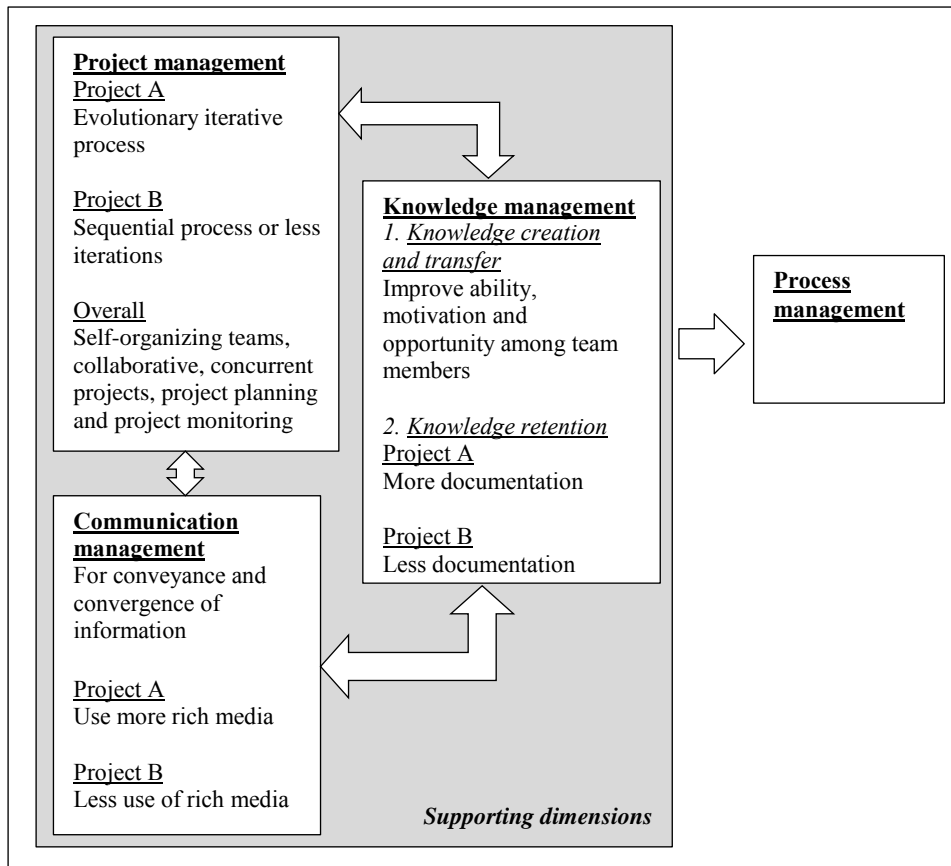


Figure 10: High level process model of the USAM

Process management and its supporting dimensions for USAM are described in the following sections.

6.3. Process Management of the USAM

Process management includes the data modelling component of a process model. It could be considered as the core of the process model, where the technical oriented component of data analytics is considered. The HA process management component consists of eight steps and it is an iterative-incremental life cycle model. As shown in Figure 11, the process iterates in a cycle (data, model cycle) until there is high confidence on the validity of the data prepared and the model built. At the initiation of the project the problem to be solved needs to be identified.

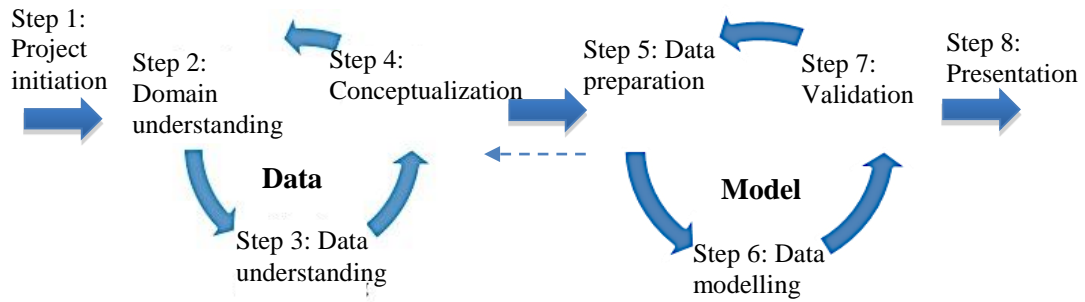


Figure 11: Methodological steps of the USAM

Then the process enters the data cycle. The data cycle starts with attaining relevant domain knowledge. To understand data, user needs to explore the dataset and extract only the relevant data to facilitate the preliminary stage of theorizing. The research questions and relevant hypothesis will be developed based on the collected data, prior literature and understanding of the domain. A model will be conceptualized going through the loop until there is high confidence on the quality and usefulness of the data collected related to the problem defined.

Similarly, the data model will be fine-tuned until the model is validated in the loop in the model cycle. The first step in the model cycle is data preparation based on the conceptualization of the problem. Then an appropriate analytic model will be selected and the data model is built. The emergent data model is validated with a new set of data to ensure that it has reached expected accuracy levels. Following the model cycle, the process enters the step where the results and tasks performed are documented and subsequently the completion of the project.

This is not a rigid one way cycle as moving back and forth between steps is always possible. Thus, this is considered as an iterative process. Moreover, there is a feedback loop from one cycle to another cycle to correct, if there is any error in the current step (or cycle) or if expected results are not achieved. As a whole, the complete process is a life cycle model where, the HA does not end once the solution

is presented. New projects can be triggered by the lessons learnt during the HA process and based on the results obtained (and possible research areas and questions) (Chapman et al. 2000). Thus, such new projects will be more focused on the requirements as are planned based on the experiences from prior projects.

It is important to note that *projects A and B* can use the iterative loop methodology proposed above with some variations. Considering the clear definition of the problem and the simplicity of the project tasks in *project B*, it would be straightforward to use a sequential process with fewer phases. The next section will deal with outlining each and every step of the HA methodology related to data modelling oriented component.

6.3.1. Step 1: Project Initiation

Project can be initiated in two ways depending on whether the project is commenced after a problem is identified within the organisation (then an internal team or outsourced external party will do the HA project) or without identifying a problem (usually done by a researcher). Problem is identified based on the organisation requirements (reduce cost and time, improve productivity, etc.) or based on some interesting approaches that had been found (or used) in other similar organisations (to replicate).

Access to data sources will be a major concern for novice users (external researcher to the data provider) if they are working on data obtained from external sources. The notion of accessing data will vary with individual researchers working on analytics. The leading school of thought with regards to initiation of a study is that the researcher should begin by formulating the research question (Eisenhardt 1989). It is believed that a substantial understanding can be acquired by going through the literature and knowledge gaps and research questions can be identified based on them.

However, obtaining access to the data required for the study becomes a deciding or a limiting factor. The strict regulations in healthcare sector may make it hard to get access to the data required (Herzlinger 2006) and in HA, getting access to medical databases is extremely difficult. Moreover, it is considered that there will be a high likelihood of gaining access to data by being open to all possible data sources and adopting an analytical approach on those data gathered. This is known as planned opportunism (Pettigrew 1990) and it refers to how a person reacts to chance events or how we can use our competences to seize opportunities. Thus, there will be chances to access data (uncertain and at the beginning cannot plan on what to access) and an intentional choice has to be made of what to study (research question).

Normally in research, the access to data is granted as goodwill, and as such it is always good to offer something beneficial to these data providers in return (e.g. allow them to use the findings in their clinical setting). For example, they could be provided with a computer system or a dashboard, so that they could use the application in their day-to-day activities.

Even if they are working within an organisation, there will be significant delays in obtaining access to data due to maintenance of the security of information of the patients. In organisation context, the required data sources will be identified based on the user requirements (e.g. clinicians). Then the project manager will seek necessary approval to access the set of data required (complete access to the whole dataset will not be provided even for an internal analyst).

At the data access stage, a report will be created on initial data access. This will include a list of data sources accessible, their location, size of the datasets (number of records and the time period covered by the records) and type of data it contains (e.g. format-text, images, electronic medical records, type-clinical, personal

data). Moreover, it is important to record the procedure adopted in obtaining the permission to access the data sources and the problems encountered during the data access process. The information given in the initial data access report will be useful for replication in future projects.

6.3.2. Step 2: Domain Understanding

Domain understanding is quite important after gaining access to the data. As portrayed in the statement made by Albert Einstein '*If I had only one hour to save the world, I would spend fifty-five minutes defining the problem and only five minutes finding the solution*', it is important to understand the domain, requirements and the problem to be solved before performing HA modelling. This is useful specifically for complex and ambiguous projects. A major portion of time of the project should be allocated for clear definition of the project objectives. Especially, due to the unique nature of the healthcare domain (Cios and Moore 2002), a greater effort is required to understand the domain requirements and necessary objectives relevant to the field.

Perusing, general medical literature such as electronic articles, Wikipedia are quite useful in collecting background information about the specific domain handling. Furthermore, communicating with clinicians in the relevant specific areas is an effective and time saving source of information, as well as a way of getting probable doubts clarified. Here, the domain could be a particular disease group or a particular health unit (e.g. ICU, wards, radiotherapy units) or a particular activity (e.g. quality of service, scheduling, resource planning, and waiting time management). Thus, if the access to diabetic patient data is obtained (like snapshots of glucose readings, HbA1C readings, insulin dosages, patient demographic data, and calorie intake) then it is important to have background knowledge on diabetics. However, HA projects in real life do not function in isolation limited to one domain. Based on the findings from

interpretive study, many departments and their activities can be overlapped with others. Even the diseases will overlap with other diseases and treatment procedures. Thus, while getting in-depth knowledge, it should spread over different areas as well. This information allows the data analyst to recognize more sensitive distinct issues and work out potential research questions related to the domain under study.

Even though more detailed understanding is specifically required at the data cycle, having such an understanding is important at the initiation of the study to guide the project in the right direction. Therefore, it is useful to initiate the project with domain understanding. Britos et al. (2008) proposed a requirement elicitation process with a documentation template, as most of these proposed processes and methodologies for data mining had neglected the requirement specification aspects of projects including systematic documentation of requirements and a technique to extract necessary knowledge. In developing our model, several important components available in the model proposed by Britos et al. (2008) have been taken into account (such as cross referencing and common lexis). Rather than increasing the types of documents, a particular document could be used at different stages of the process.

A project documentation template to be used to document the domain understanding step is given in APPENDIX B. This was adopted based on the business requirement document template proposed by Podeswa (2005). The proposed template provides the actions necessary to be carried out in performing an analytics project. It will be used throughout the project with necessary amendments made in each stage. This includes a RACI (responsible, accountable, consulted and informed) chart specifying roles performed by each team member. Team members can identify the persons responsible to be consulted before making changes to the document. This chart will be a useful audit trail and will make the document more transparent. In

addition, such a project document will be of use to new personnel joining the project to get acquainted with the actions carried out and also to take note of mistakes made in carrying out the project. This will help to reduce the learning time for new recruits and will facilitate rechecking the essential steps for re-assurance by personnel already working in the project.

It is vital to discuss each of the activity (in domain understanding step) one by one to understand different tasks, input and outputs. The main activities in this step are illustrated in Figure 12. New activities added to the CRISP-DM are indicated as ‘*’. Even in other stages, similar symbols were used to indicate the new items added.

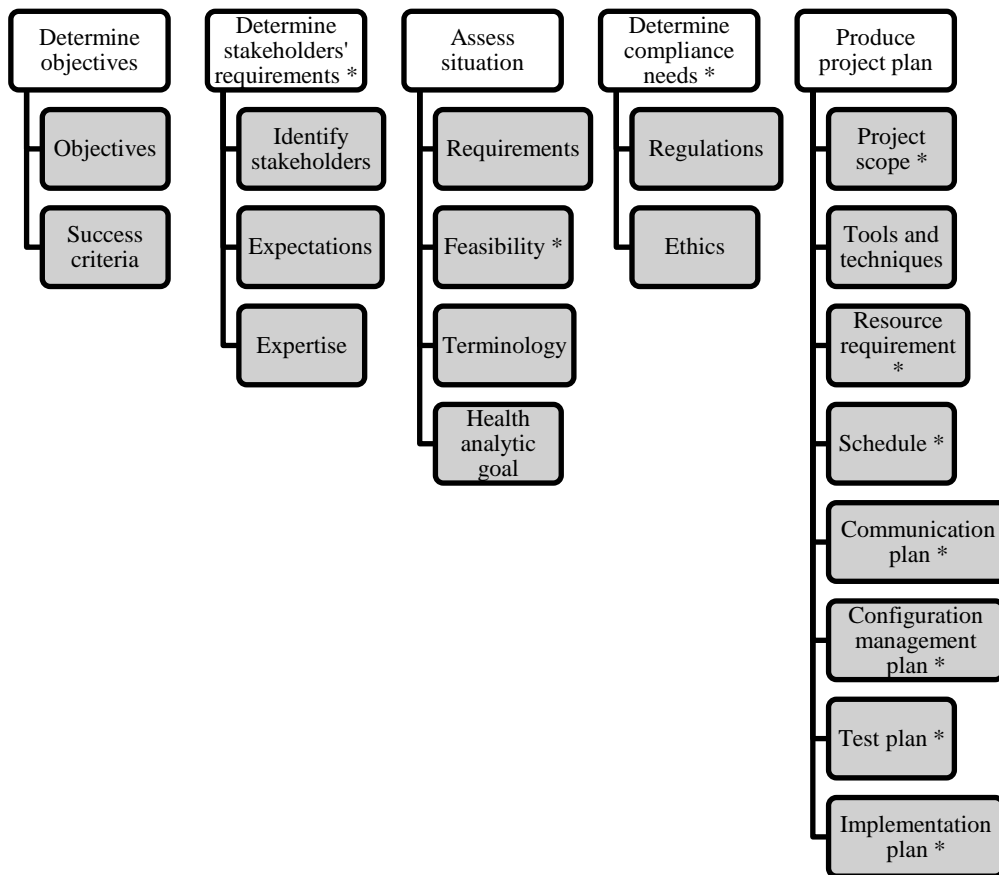


Figure 12: Activities in the domain understanding stage
(N.B. * - New items added)

A. Determine objectives

It is important to understand the objectives of the client (requirement provider). This could be a clinical institute like a hospital, polyclinic, nursing home or a laboratory providing clinical based data related to patients, administration, resource usage, workflow, etc. If the objectives of the clients are not determined at the beginning, the models may not be of any utility for them. Also, there is a danger of creating a “ripple effect” where one change in the objectives could lead to overall change in the modelling and deliverables (Podeswa 2005). When determining objectives of the client, the analyst can use the ACES approach to determine the type of objectives. They can use the ACES approach to determine the goals of the provider (Peterson et al. 2013) as stated below.

1. Achieve (things that the organization desires to accomplish in future)
2. Conserve (things that the organization desires to sustain)
3. Eliminate (things that the organization desires to get rid of)
4. Steer clear (things that the organization desires to avoid)

They could be operations (on day to day operation; e.g. order tracking), development (on acquiring new skills and expertise), innovations (new ways to perform), problem solving (to handle issues faced) or profit objectives based. It is important to note that even if the objectives are known, they may not have a clear idea on the project requirements.

It is important to define the primary objective and then the related business questions. For example, primary objective could be to reduce bed overflow (assign a bed in a different specialty) in hospital clinics during peak hours when there is no available bed matching the required ward (Teow et al. 2012). Then the possible related questions are “*whether the condition of a certain patient will be degraded if*

s/he is asked to wait in the Accident and Emergency Department?”, “Will it be possible to project discharges in the required specialty or in a close by unit?”, “When should the patient be assigned a bed in another specialty if beds are not available in the required specialty?”.

The success criteria indicating the effective performance of the project should be identified at this stage. Users will accept the findings if they are satisfied with the results and if they find them meaningful. Findings should be based on the objectives and should be specific and measurable. For example, possible success criteria could be reducing patient readmission within 30 days after surgery by 20% and reducing patient time to see a doctor to 10 minutes, etc. Here, it is important to identify the person who will be responsible in assessing the success criteria.

B. Determine stakeholders

Roles played by users and their expectations and expertise are important criteria to be assessed. A project team (Collier 2011) comprises of

- Planners (senior management and project sponsors who act as facilitators and project champions)
- Doers (data analyst and ground or junior staff appointed by the management to work on the project directly)
- Consumers (use the outputs generated by the doers)

Particularly, in an internally performed project, it is essential to determine the role played by each member in each of the three types (planners, doers or consumers) at the beginning of the project. A particular member could play several roles. In an external project, the clients of the project could be the users as well as owners of the project. As users of the project, they will know the exact requirements (expectations)

and domain knowledge of the system (expertise), thus, they will provide the relevant know how to understand the requirements and to validate the results.

User requirements vary from user to user and also based on the role they play. While a physician may consider the reliability of the results and ease of use of the data model, the project owner may have more requirements like scalability, ability to adapt to other scenarios and cost saving ability.

C. Assessment of situation

Situation assessment is carried out to understand the requirements, constraints, risks before making HA goals and project plans. The situation will be assessed based on the organization objectives and the specific requirements for the project already determined. In addition, it is essential to determine whether there is an existing solution to handle the defined problem. If there is a current solution, it is important to review it to understand the advantages and disadvantages of that and also any possible relevant issues.

Changes to the requirements are welcomed even at the last stages of the project and users can always revisit and modify the requirements and model can be fine-tuned based on the changes necessary.

We consider the feasibility of the project under four groups, namely, operational, technical, schedule and economic feasibility. It is important to determine the prerequisites of the project and determine whether the approval is already obtained to use HA in the organization, whether it is accepted by the users and if not how it is needed to be prompted in the organization. This is known as operational feasibility. In technical feasibility, the analyst will mainly consider the availability of technological capability. Schedule feasibility will be concerned of whether the project

can be successfully completed within a certain time frame and economic feasibility will focus on cost benefit analysis. The project should be economically viable.

It is useful to identify the already acquired data sources and their type (e.g. whether case notes, machine extracted data, online data or reports) at the initial stage. If the access is not obtained yet, it is good to identify the required data sources and commence the data understanding by gaining access to them. It is necessary to check whether all the data required are available after going through the available data sources and the HA goals. Setting selection criteria (get advice from a domain expert) to determine the irrelevant data and identifying them too is important. One needs to determine any additional information required and for how long the data should be available (e.g. records of last 2 years, last 5 years, and last 10 years).

The analyst needs to determine the type of knowledge sources required to commence the project. They may need data from other departments as usually HA projects are not carried out in isolation. The required data may vary based on the scope. For example, when forecasting patient admission to a ward, it is important to decide what point is considered as the actual admission point. It could be physical admission time or ordered admission time by a doctor (usually there will be a waiting time till a bed is freed). Moreover, other secondary sources to acquire knowledge like, written documents, online articles and videos too need to be considered. Thus, at this stage it is essential to consider whether the relevant domain knowledge is accessible to commence the project and should try to acquire them if not available.

Furthermore, medical jargons make it too complicated for the analyst (Britos et al. 2008). For a researcher it is important to have an idea on the related medical terms in the data sources for a better analysis. Since HA is dealing with a domain having a different terminology, it is important to prepare a glossary of specific terms

and vocabulary related to the domain of the healthcare project. This should include the healthcare nomenclature as well. For example, it is always better to link with healthcare standards like SNOMED, ICD10. Then a glossary should be made incorporating HA related terminology. This would be useful in presentation of the outcomes of the project.

Finally, the intended outcomes (goals) of the HA project have to be stated and it should match with the business objective. The HA goals should represent business goals in technical terms. For example, we can reflect on categorization of diabetic patients before providing treatment, forecast the time taken to perform radiotherapy on a cancer patient, segmentation of patients based on the adherence to physician instructions/guidelines, etc. Determination of whether to use an advanced analytic approach or a descriptive analytics will be made after conceptualization.

D. Determine compliance needs

With the development of national wide electronic medical record systems there is a growing concern on the privacy of the patient data. In healthcare projects, it is essential to ensure the safety of patient records and sever the patient identity while improving the quality of care (Li and Qin 2013). There are certain regulations to adhere to, for ensure the privacy and to protect patient data. USA's HIPAA (Health Insurance Portability and Accountability Act of 1996) provides the right to maintain confidentiality of individually identifiable health data. It describes policies, procedures and guidelines to preserve the privacy and security of health data (Narayanan and Shmatikov 2010). Moreover, it describes regulations for the use and distribution of health data. According to HIPAA, identifiable health data should be removed before the health data is released to a third party. Other than that, there are

other regulations like FISMA (Federal Information Security Management Act of 2002) and IT governance based on ISO/IEC 27002: 2005.

Personal Data Protection Act of 2012 (PDPA) has been passed to promote transparency and to maintain privacy and security of personal data in Singapore. There is provision on protection, collection, disclosure, transfer, access to and care of personal data related to healthcare institutions. In addition, PDPA applies to all personal collected data, used and disclosed in Singapore. Thus, healthcare providers and data collectors are obliged to put in place reasonable security arrangements (Yeo and Gaw 2013).

There are institutional review boards (IRB) or ethical review boards to protect human subjects in biomedical and behavioural research. They will approve, monitor and review those research works and may suggest amendments prior to approval. IRB approval is necessary to be obtained before accessing patient data in healthcare institutes.

Upon development of the project plan with a clear understanding of the domain, its objectives, stakeholders, regulatory obligations, the next step will be the data understanding stage. This is done in the data cycle taking user requirements and HA goals of the project into consideration.

6.3.3. Step 3: Data Understanding

This phase as illustrated in Figure 13, begins with the data collection and carries out certain tasks to get familiarized with the dataset. This involves determining interesting subsets of the data or insights from data and data quality issues. Moreover, since we are dealing with data requiring compliance with data protection regulations, it is important to de-identify the dataset.

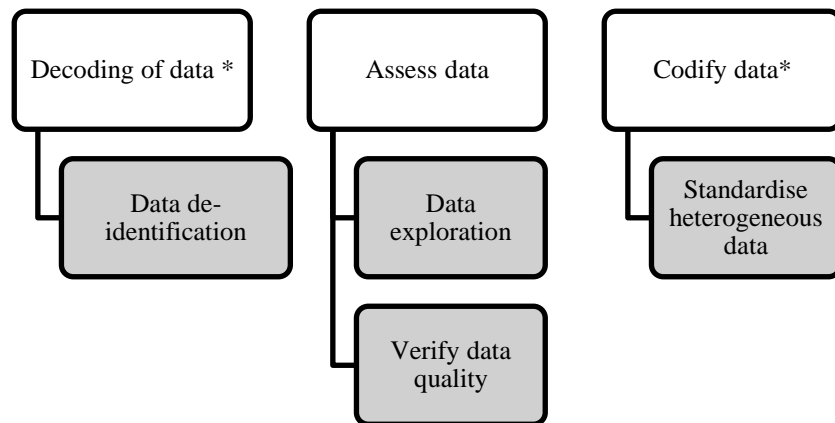


Figure 13: Activities in the data understanding stage
*(N.B. * - New items added)*

A. Decoding of data

This involves anonymization and de-identification of data. Anonymization could be defined as permanent removal of identity of the data contributor from a data set to avoid any future re-identification (Li and Qin 2013). De-identification could be defined as removal of the identity of the data contributor in a study but there will be identifiable information (de-code) that could be used to re-link with the actual contributor by a trusted party (Li and Qin 2013; Narayanan and Shmatikov 2010).

HIPAA specifies 18 Protected Health Information (PHI) attribute types that could explicitly or likely lead to identification. PHI attributes (includes individual and relatives, employers, or household members of the individual) specified in HIPAA are given in APPENDIX C.

Even though there are a plenty of studies on data privacy, there is a dearth of studies relevant to ensuring the privacy of PHI. As medical data are usually semi-structured or un-structured (e.g. case notes, pathology report, etc.) and as identifiable information are embedded inside the content of these reports, it is complicated to apply anonymization techniques in dealing with them. Creation of de-identified datasets and creation of limited datasets are options available in de-identification of

PHI data. In the former option all the 18 identifiable attributes are removed or replaced with a constant value after detecting them automatically (Meystre et al. 2010). However, this could be over-protective and valuable information useful for HA could get removed from the dataset. In the latter option PHI are partially removed. Limited datasets may have addresses other than street name, post office boxes and 5-digit zip code, all elements of dates of admission and discharge dates and unique codes or identifiers not listed as direct identifiers.

The other research friendly option is to use generated variables to replace the PHI content. Even though this removes PHI from the dataset, still these generated variables could be used to re-identify the records. However, these variables used to replace the PHI attributes should not be a derivation of the original value and only a trusted party should be able to link the data records to the original. For example, data contributor identification number could be used to replace the name and the social security number and the dates could be shifted using the study start date.

Thus, this report should contain information such as strategy used to de-identify the data, attributes replaced and removed from the dataset and the approach used to replace the data. Other than that, the report should maintain a record of personnel authorized to handle de-identification of data and to maintain the code to re-identify the data.

B. Assess data

Since the access to dataset is already granted and additional required data sources are identified after reviewing the dataset previously, this stage involves assessment of the data sets. In addition, there may be external systems such as national health registry, drug database, etc. interacting with the HA project.

It would be useful to tabulate the basic properties of the datasets, including the volume, attribute types and values and relationships. Specifically in healthcare data, there are many data columns (numerous timestamps representing different activities on patient, medical conditions and readings, etc.) and as such, it is important to align them in a list of columns rather than scrolling through the dataset. When analysing the quantity of the data, it is vital to report whether it contains free text and it is necessary to determine the tools that will be used to extract those relevant data. For flat files it is necessary to report the type of delimiters (e.g. comma separated, tab separated) used.

It is important to perform basic descriptive statistical analyses (e.g. min-max, mean, standard deviation, mode, distribution and skewness). The correlations among these attributes too need to be reported. In dealing with time series data, the trend and seasonality are needed to be determined at this phase. Even at this stage it is important to recheck the relevance of the attributes and whether new data are required. If required, it is possible to move to the previous cycle. Under relationships it is necessary to specify the table and their relationships and also the amount of overlap of key attributes between tables. As there are legacy systems, the same data could be tracked by different systems with some changes. As such it is important to clarify the exact point (or the definition of the data) at which these overlapping data is determined. Consumers of the project team can confirm the relevance of the data and suggest suitable datasets. For a certain type of work one dataset may be used while for another situation another overlapping dataset may be used based on the user requirements and HA goals.

Data quality will be assessed based on completeness and correctness. Data quality will be rectified by listing all the data quality issues and actions carried out. For example, if there are missing values, it is necessary to identify the relevant

attributes, how common they are and how they are represented in the dataset (e.g. whether it is kept blank, dot or -1). Furthermore, it is necessary to report whether same attribute is represented in different names or same attribute value is given with different names. Also, it is necessary to check for outliers in the dataset (important to determine whether they are noise) and whether some values for attributes are unrealistic (e.g. data records related to pregnancy had mentioned for some gender as male, age as negative value).

C. Codify data

Due to heterogeneity of medical data it is important to codify data into a common standard. Different systems may be using different standards to codify the data (SNOMED CT, ICD). Even some electronic medical records may not be up to date and could be using older versions (e.g. ICD 9 instead of ICD 10). Thus, it is difficult to integrate different data sources. Furthermore, a huge amount of available data is unstructured and it would be useful if the extracted data is converted into a common standard. For example, data extracted from case notes or x-ray images can be retrieved, integrated and shared using HL7 standard. Thus, it would allow a standard form of data representation avoiding any ambiguity in data interpretation.

6.3.4. Step 4: Conceptualization

The third step under data cycle is conceptualization (Figure 14). This is a totally new stage that will be introduced into USAM compared to other existing models (e.g. CRISP-DM). Conceptualization refers to abstract representation of some selected areas in the real world using entities and concepts to illustrate some interesting relationships.

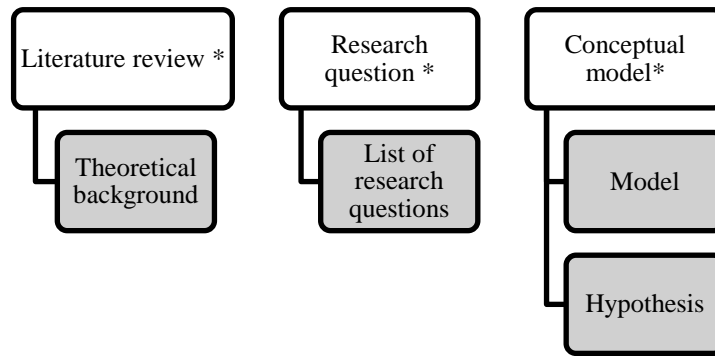


Figure 14: Activities in the conceptualization stage
*(N.B. * - New items added)*

A. Literature review

Here, it is important to review the literature related to the domain and the HA goals to determine related concepts of the problem domain. Thus, the user will identify relevant theory and past work carried out related to the area and will report them in the report on theoretical background. In this report user will specify an evaluation of those concepts as positive, negative and comments (neutral). Positive evaluations will be used in the conceptualization of the model and negative evaluations will be considered as gaps and will be used to modify existing concepts. Moreover, possible links among various concepts will be determined and it will be reported with how they can be merged together as a one concept.

B. Research question

The next step will be the formulation of the research question (RQ). There could be more than one RQ. It is a question that the study will answer and on which the study will be focussed (Easterbrook et al. 2008; Meltzoff 1998). In a knowledge focused RQ, there are three categories. First, exploratory RQs will be using qualitative methods to study unknown (less known) phenomena to get a better understanding. For example, this could have RQs like *Does X exist?*, *How does X differs from Y?* and *What are the properties of X?*. Second, base-rate RQs are set to

find out the common pattern of occurrence (how and when) of the problem under study. For example, possible RQs are *how frequently does X occur?*, *how does X normally work?* and *what is the process by which X happens?*.

Third, relationship RQs are set to study how the problem under study is related to other concepts or phenomena. Possible examples are, *Are X and Y related?*, *Does occurrence of X correlates with Y?*, *what causes X?*, *Does X cause more Y than Z does?* and *Does X cause more Y under one condition than others?*. Thus, it is important to determine the category of the RQ and justify the reasons for selection of the category before specifying the research question. Research opportunities and possible research questions in health are suggested by some of the authors (e.g. (Fichman et al. 2011; Hesse et al. 2010; Romanow et al. 2012))

C. Conceptual model

Finally, the conceptual model is set based on the literature review and the research questions. Here, under model it is important to mention the theories finally used to develop the model, description of variables in the model (dependent variables and independent variables) and interaction effects. Kitchen sinking is not a good practice and it is better to use meaningful variables that are justifiable based on the experience and the literature.

If it is a statistical problem, a hypothesis should be given for each and every RQ. Here, the problem statement is divided into several hypotheses (Raghupathi and Raghupathi 2013) and it will be a guidance for the HA process. In machine learning problems, the identification of the independent variables and the dependent variable (only if it is a classification or a prediction problem) will be sufficient (Schmidt et al. 2008). Moreover, relationships found through descriptive analytics could be explored using predictive analytics.

After conceptualizing the problem, the next stage will be the data preparation stage. Final dataset for modelling will be prepared at this stage based on the conceptualization of the problem under study. The next section will describe activities that will be performed and documented at the data preparation stage.

6.3.5. Step 5: Data Preparation

The data preparation step involves all the activities carried out to prepare the final dataset to be used in the modelling. Until the data preparation stage is finalized, the project will iterate through the three stages in the data cycle (data understanding, conceptualization and data preparation). At this stage, it is important to use version control on the dataset and one should be able to revert back to a specific prior version in the data set if a certain mistake has been made on a certain level of data preparation. Thus, it would save time and avoid the necessity to start from the beginning. Figure 15, illustrates the outputs in the data preparation stage. This includes data selection, cleaning, construction, integration and formatting tasks. Each component shown below is extracted from CRISP-DM model (Chapman et al. 2000).

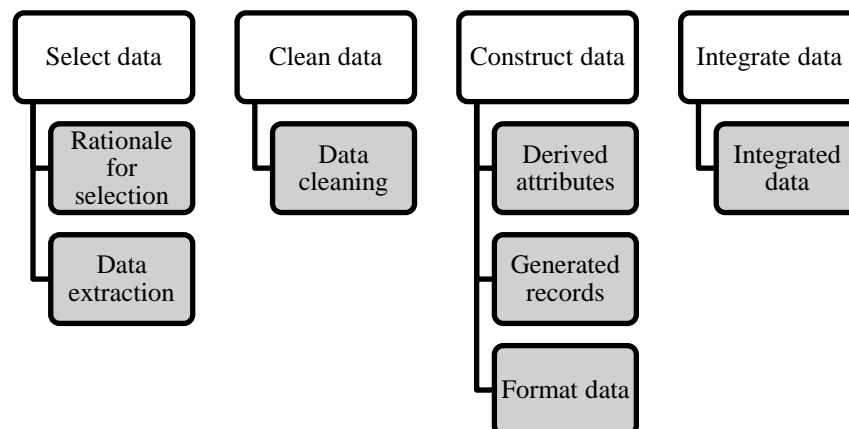


Figure 15: Activities in the data preparation stage

Since we are dealing with heterogeneous data sources, it is important to mention the details of the sources and the modes of data extraction. Also the programs used to extract those data and the parameter settings need to be documented. As such,

it would be easy to regenerate data later if there is a slight change in the dataset and if the data sources are changed still it could be used as reference. The dataset description report is removed from this phase compared to CRISP-DM and will be created at the data understanding stage with the amendments.

When formatting data, it is important to determine whether the dataset is correctly balanced. As this depends on the technique used, it should be mentioned in the report (with the dataset balancing technique used) along with the techniques decided in the previous stage.

After fulfilling the tasks in data preparation, one can proceed on to data modelling. Steps in data modelling will be explained in the next section.

6.3.6. Step 6: Data Modelling

Modelling step includes application of the selected modelling techniques where relevant algorithms and parameters are altered to get the optimal results. As the output illustrated in Figure 16, is extracted from CRISP-DM exactly in the same way as in data preparation, it will not be explained again in this section (Chapman et al. 2000). Compared to CRISP-DM model, we removed the generate test design from this stage. The planning of the test design will be carried out at the beginning and any changes to the design will be performed as amendments rather than creating a new document type.

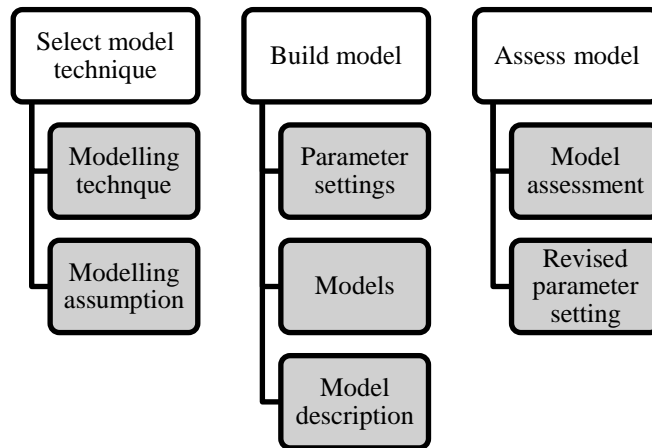


Figure 16: Activities in the data modelling stage

6.3.7. Step 7: Validation

The finalized data model needs to be evaluated in the validation step (Figure 17). Furthermore, at this step all the actions carried out to build the model will be reviewed, to detect any additional requirements or issues that had not been dealt with. For model evaluation there are four possible approaches, namely, holdout, k-fold cross validation, leave one out and bootstrap. The hold out evaluation strategy is suitable if there is a separate testing set. If not, one could use k-fold cross validations for larger samples and leave out and bootstrap if the sample size is small.

The selection of the evaluation strategy could be based on accuracy, speed and flexibility. However, since this is in healthcare model evaluation, its accuracy should be very high. The tasks and outputs are same as in CRISP-DM and as such, they will not be restated in this section (Refer (Chapman et al. 2000)). We believe that compared to other fields, in the medical field it is specifically important to interpret the results as the models will not be valid without any clear interpretation of the results.

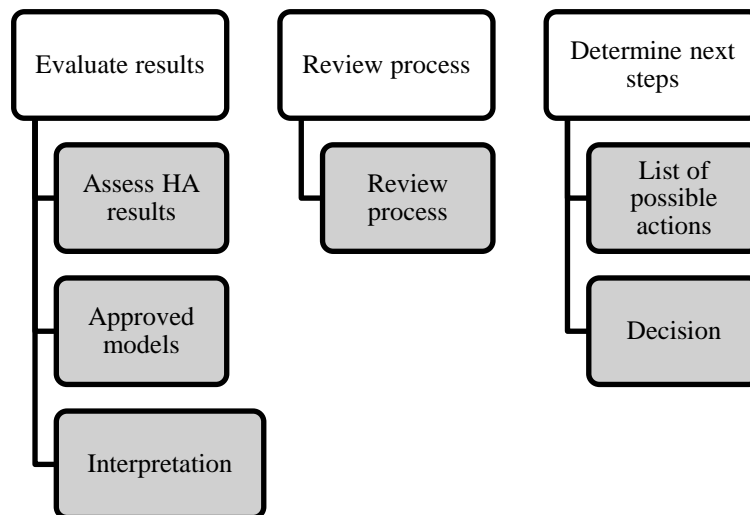


Figure 17: Activities in the validation stage

It is important to determine the limitations of the model and the conditions under which it works. The model built will be thoroughly evaluated before it is introduced to the client environment. The models are evaluated based on accuracy (ability to predict correctly the previously unseen data- e.g. reduction in false negatives), speed (computation cost of building and using the model), robustness (ability to make correct decisions even if there is noise and missing data), scalability (ability to use with a large amount of data), interpretability (level of insight provided by the model) and simplicity (easy to build the model and use it) (Stefanowski 2010). A decision matrix could be created based on these factors with a weight assigned based on the relative importance of each factor.

The next section will deal with the procedure to be adopted in presenting the data model to the client and the project implementation plan.

6.3.8. Step 8: Presentation of the Data Model

After creation of the model, it is necessary to organize the results and present it in a way that the customer can understand and use it effectively as the client has to understand the actions to be carried out in implementation of the project in the client environment. Furthermore, it is necessary to consider the storage of built models and

their interpretations for future reference. Therefore, presentation step could be considered as the final step in one increment. This will be a beginning for the next incremental loop created based on the feedback. In this stage, a deployment plan and monitoring and maintenance plan created in the step 1 will be adjusted based on the new requirements (to avoid creation of many new reports). The output for the presentation step is illustrated in Figure 18.

A. Present results

The model will be linked to an existing system in a live environment or will be embedded in a new system developed. Moreover, in the interpretive study, it is found that some of the outcomes are given to users by creating dashboards, representing the findings visually. The software system may link to another data modelling server and it will get only the output result to display on the system interface. In hospitals, as IT work is handled by a different group from data analysts, they can be involved with busy schedules. As such it is better to have a plan on delivery dates and infrastructure requirements.

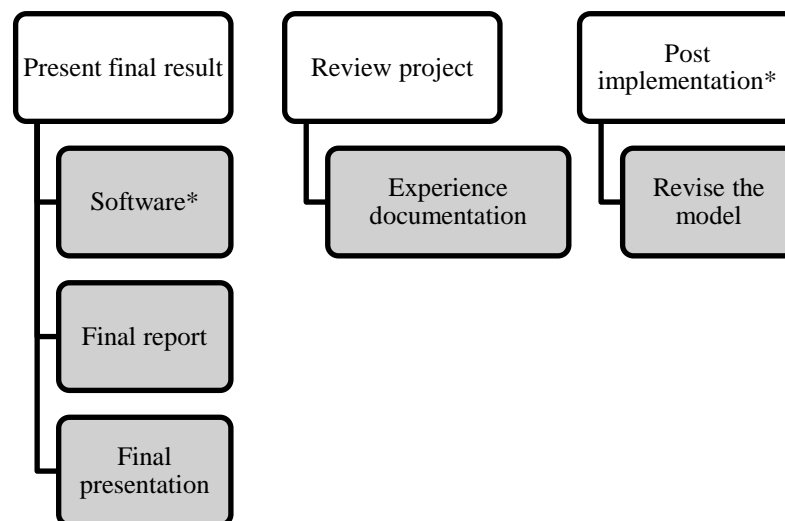


Figure 18: Activities in the presentation of the model stage
(N.B. * - New items added)

Communication of the results to the prospective users/clients can be considered as the most important component. The knowledge extracted and utilized in model development will be communicated through reports and presentations. The planners and consumers will be informed with the findings in such a way that they could use them in their future planning and strategies. In the final report, a summary of the deployable results will be presented with the costs incurred, deviations from the original plan and suggestions for future work in addition to a final presentation summarizing the whole project.

B. Experience document

The experience document will describe the expertise gained while carrying out the project (knowledge management). This document summarizes inappropriate approaches taken, any problematic issues faced, and necessary instructions to handle such issues. It is good to include viewpoints of each stakeholder at the end of the project including the feedback from the client and the users. Finally, the comments should be generalized for future reference (Bellazzi and Zupan 2008).

C. Post-implementation

Carrying out post-implementation follow-up too is required to recognize the problems and fine tune the system based on the feedback. As the environmental variables are changing, the model needs to be tweaked to accommodate new factors. For example, if a model to predict the monthly patient admission for 5 years had been created previously, it needs to be revised if a new ward is constructed in the hospital to accommodate more patients or if there is an expectation of having a new hospital in the same region during the five year period considered for forecasting. These changes should also be documented and new revisions can commence from there.

Conversely, most of the HA projects are not deployed after completing the model construction (Bellazzi and Zupan 2008). That is, many of those projects have not proceeded beyond the model building and validation. The deployment would have been hindered by the already existing decision support tools. Though these tools contain data models dedicated to different clinicians, they are devoid of support for interfacing new models. Thus, it is important to create possibilities of bridging new models developed with existing decision support systems by making these models compatible with existing tools (de Rooij et al. 2005). A possible solution is the use of XML based prediction models (e.g. PMML) (Bohlouli et al. 2013; Grossman et al. 2002). However, with proper planning and collaboration with the users and the system developers (IS developers), these models can be incorporated into operational use.

6.3.9. Variations to the Methodology

The proposed USAM process described in this thesis is a generalized process that could be used in dealing with different projects. Though it is directly applicable in a complex and ambiguous project, some of the components and outputs specified are not required for other projects (Table 9). Variations in all the steps excluding step 1 and 4 are given in Table 9 for the *project B* (project with simple and clear project requirements) where, even the detailed documentations are not required. The main variations and reasons for variations between *project A* and *project B* were given above with the explanation of the generic methodology.

Table 9: Variations in USAM for Project B

Step 2 Domain understanding	Step 3 Data Understanding	Step 5 Data Preparation	Step 6 Data Modelling	Step 7 Validation	Step 8 Presentation
1. Assess situation – requirements, feasibility, terminology, HA goals	1. Decode data – data de-identification	1. Select data – data extraction	1. Select modelling technique – find modelling technique and assumption	1. Evaluate results – assess results, write interpretations (not in detailed)	1. Present results – presentation (slides or email results)
2. Determine compliance needs – regulations and ethics	2. Assess data – data exploration, verify data quality	2. Construct data – derive attributes, generate records, format data	2. Build model – record parameter settings and model, model description (not in detailed)	2. Determine next steps – decision (less revisions)	2. Post implementation – revise model
3. Produce project plan – project scope, tools & techniques, resource requirement, schedule, communication plan, test plan, implementation plan		3. Integrate data	3. Assess model		

Gaining access to the data source and identification of the problem are important factors even in *Project B*. Since the problem is simple, it is not necessary to determine stakeholder expectations and objectives. Also, since there will be very few iterations, there is a high requirement to do project planning. The basic development work (model building) will be similar compared to other projects with lesser feedback loops. In conceptualization, it is not required to go through literature and definition of the basic model (attributes to be used in the model) would be sufficient as the problem

is simple and clear, and it is not required to understand how it is been done in previous studies as it can be performed directly.

6.4. Project Management

The elements in project management can be identified in three directions as project structure, project planning and project control. Project structure considers the composition of project process and members. Project planning involves activities performed before commencement of the project to manage resources. Project control aims at identifying possible problems and planning steps to mitigate them.

Project structure

Project structure involves the overall arrangement of the project process and the team. Based on the project type, it is worthy to determine how the agile project management components are incorporated in the model.

The elements of the model are as follows:

- Evolutionary Iterative process for projects with complex and ambiguous requirements. Since the requirements are not clear, planners and doers will have to go through an evolutionary approach to understand the exact requirements and data models to be built.
- Sequential process for projects with simple and clear requirements. There is no requirement to go through an iterative process to understand the requirements and thus can perform directly in a single iteration or a fewer number of iterations.
- Self-organizing teams - Self-organizing team members have the necessary autonomy to carry out their tasks rather than being led by an outsider (from the project team). Project manager will work as an enabler by supporting removal

of barriers to the project. The data analysts will have the necessary autonomy and the skill set to perform the project following the organisational guidelines with self-discipline. The team members have to review and revalidate their objectives and assumptions periodically through communication with planners and users.

- Collaborative work - Team composition should allow a collaborative environment where planners, doers and users can work together to identify valuable insights from the data. For advanced analytic tasks, it is important to get people with data, domain and modelling knowledge. However, it is not possible to find people having all three skills together. Therefore, creating teams to perform those tasks as collaborators (planner, doers and consumers) is a possible option. Moreover, in considering the knowledge divide between data analysts (doers) and medical professionals (planners and users) collaboration between them is required to reduce the knowledge gap.
- Concurrent Projects – Accessing data and feedback (e.g. clarifications on a data type or a finding made from data) on time is a bottleneck in healthcare institutions with their data protection regulations and prioritisation of planner's daily activities (higher priority given to administering to patient). Wasting of data analyst's time in such cases can be avoided by an individual analyst handling concurrent projects at the same time. This is achieved by alternating between projects rather than waiting for the feedback. Through continuous alternating-run procedure, team members can keep in touch with tasks to be performed in each project.

Project planning

Depending on the project requirements it is important to classify projects based on the difficulty and clarity of the problem. Project plan can vary according to this classification. If the project is simple and clear, the requirement for creating a detailed breakdown of milestones and project plan is not required and will not bring value to the project and knowledge outcomes.

Project planning includes the following tasks:

- **Scope creep** - The ‘scope creep’ is an effective strategy to handle complex projects. The scope of the project is incrementally expanded. This is possible with the iterative and incremental approach in USAM. Requirements represented through use-cases, allow prioritising the requirements and starting with the preliminary design. Requirements can be prioritised based on value (e.g. high value and high risk stories are completed first, then high value low risk use-cases and so on) and capability (categorise use-cases and work on a category at a time) (Collier 2011) to evolve the project incrementally. In addition, it is vital to declare parts (requirements, features of the project) that are out of scope of the project.
- **Techniques and tools** - It is important to determine the techniques and tools to be used in the project. Identification of these should be done at the initial phase as data collection and conceptualization of the problem will depend on the HA techniques to be used. These techniques should be determined based on the HA goals and the tools should be determined accordingly. At this stage it is good to prioritize the techniques to be used for each task/goal. Selection of the tool will depend on the (1) price of the tool (based on the software, hardware and maintenance cost), (2) performance of the tool (based on the data capacity,

speed, compatibility, platforms, data formats and management, and software architecture), (3) functionality (based on the type of data, HA technique, model exporting, model customization and model validation), (4) usability (based on the graphical user interface, intuitive learning, easy access, reporting and visualization features, error-proof design, navigation and predefined functions), and (5) support (based on the documentation, training availability, and services and resources) (Rohanizadeh and Moghadam 2009).

- Software, hardware, data and personnel requirements - Software, hardware, data and personnel requirements should be identified. Based on the tools identified in the previous steps, it is important to identify whether they are available in the organization. Specially, it is important to identify whether other supporting software tools for data capture (e.g. text mining tool) and data preparation (e.g. data transformation, synthesizing tools) are available. Another important resource to consider is the availability of the hardware facilities. Thus, it is important to determine the basic hardware in the organization and whether they are available for the project.

Timely availability of the required personnel is a necessity to carry out the project successfully. Thus, the relevant skills set for the project need to be identified and checked for their availability. It is important to check the availability of domain experts (usually junior staff representing consumer) for continued assistance throughout the project and at the end of the project for validation of results. Furthermore, soft skills too are more or equally important as hard skills.

As purchasing additional physical resources and recruiting additional skilled personnel may become necessary in carrying out the project, working

out the estimates for such additional requirements has to be done at the planning stage itself. In organisational context, additional personnel can be obtained by having interns and giving some parts as university projects.

- Scheduling - Usually, it is considered that 50-70% of the project time is allocated for data preparation, 20-30% for data understanding stage and 10-20% for data modelling and evaluation (Sattler and Schallehn 2001). Therefore, in scheduling, it is important to estimate the time scale for each phase. When estimating total project time, critical steps of the project and major iterations are required to be determined. Gantt chart and project network diagram could be used to indicate the sequence of tasks and their dependencies. Usually meetings create soft deadlines. If the projects are not urgent the deadlines can be fluid.

Since there could be delays in the project due to interruptions like delays in getting permissions to access new data, delays in setting meetings as planned (finding common time is hard) and delays in getting feedback or responses to clarifications, the project could extend and reduce the productivity of the HA team. Having concurrent projects is a possible solution. However, it is important to make sure that the projects are scheduled in a way that there is no overlap in deadlines.

- Test plan - This should be made at the inception itself for effective management of resources (including time and effort) even though it is always possible to make amendments to the components during the project implementation. In test plan, it is vital to determine the methods of obtaining test dataset and validation dataset as sometimes, these datasets will be a

prospective dataset or may be collecting from a different organization. This will be linked with configuration management of different versions of datasets and models.

- Implementation plan - This elaborates on the methods of product implementation in live environments and how the forecasts should be shown to the users (placement of results in the screen). Sometimes, it may be required to develop a software system (or amend an existing system) that includes the data model identified through the system and hence, it is important to plan how the project is going to be delivered to the client.

Project control

The difficulty of predicting the outcomes from data modelling can lead to many problems and it is important to identify such probable problems and the necessary solutions.

Project control includes the following tasks:

- Risk levels - It is important to analyse the risk levels (severity vs. likelihood). Likelihood could be defined as certain, likely, possible, unlikely and rare and severity could be defined as catastrophic, critical, marginal and negligible. This could vary as technological risk (e.g. incompatible format with the new version of the software), skill risk (e.g. withdrawal of a key player in the project) and requirement risk (e.g. change of requirements or incorrect capturing of requirements). For example, the resignation of a key data analyst would be a critical risk for a long-term project (especially if there is no proper knowledge management).

- Contingency plan - A contingency plan should be made to handle the risks. Risk acceptance (do nothing), transfer (pass the risk to another entity), mitigation (do something to lessen the harm) and avoidance (do something to avoid the risk) strategies could be used in such instances. This is useful specifically for long-term projects.

6.5. Communication Management

Communication mechanism will be effective only if relevant information is communicated with an understanding of the objectives of the stakeholders. This should facilitate timely and reliable project information dissemination to the stakeholders. There should be continuous collaboration with the project team.

Due to the overlap between knowledge transfer and communication, knowledge management focuses on the overall department perspective (not only at project level). The main focus will be on the knowledge creation and retention. Communication includes information and knowledge transfer and the facilitating media.

Communication management includes the following tasks:

- Target audience - Message to be delivered varies according to the target audience. For example, it is not correct to send the project plan and status report to the customer and the project briefing report and the status report to the project review team.
- Content - The content of each type of report should be distinguished. For example, the status report should be composed of status summary, schedule, accomplishments, next steps and issues and the project briefing should include status, checklist of activities and specific issues arising.

- Communication method – It needs to be pre-determined as different stakeholders may prefer different modes such as emails, presentations or reports. Furthermore, it varies based on the communication process. Less rich media could be used for conveyance of information compared to convergence of information. This depends on project type as well. While complex projects require rich media for communication, simple projects can depend on simpler types of communication. Nevertheless, it is appropriate to have the first meeting face-to-face to understand the requirements and to get domain knowledge.
- Frequency - The frequency of communication needs to be decided as weekly, bi-weekly or monthly. Most of the time, while monthly meetings are sufficient with the planners (if not urgent), daily or weekly meetings can be organised for doers. Frequent meetings (weekly or bi-weekly) are necessary among HA department members to keep them informed of what others are working on, and to find solutions for various problems faced. The person responsible for and the person delivering the communication have to be pre-decided in the communication plan.
- Feedback – Feedback is an important component in communication. There should be means to acknowledge the receipt of the message and a follow up mechanism and also to communicate the feedback after implementing a solution.

6.6. Knowledge Management

Knowledge is generated on the project process and the output produced. It is essential to store and disseminate this knowledge for future reference through actions such as:

- Working in pairs - An analyst can specialise into specific domains by working on projects repeatedly from the same department (usually in healthcare domain, project planners of one department will continue to work with the same individual analyst). Thus, they would have developed a shared understanding with the consumers and it allows easy communication and understanding of the requirements. Usually only one data analyst take part in a project (depends on the size of the HA staff and type of the project).

From knowledge management perspective and to avoid risk of an analysts resigning, data analysing in pairs (like pair programming in software engineering) is a suitable option. It will allow better insight generation, learning from others and reduce the risk of knowledge being with only one individual. However, if the HA department is small, having pairs will not be practical. The other possible option is having occasional rotation of analysts between projects related to different domains. Then they can learn from analysts who had worked previously in similar domains.

- Configuration management - Managing of various versions of models, data and documents is important. Having proper version control of project elements will make it easy to back track to a previous version whenever required and also to avoid confusions (failing to identify the correct version) (Marban et al. 2009b). The changes made, name of the person modified, and the location of them should be properly managed using version control software. In dealing with multiple versions of datasets, it is important to maintain the initial dataset as well as the modified versions for each model.

It is important to maintain all the files related to the project (including data) in a central repository organised into directories. Version control could be

used to track changes made. Besides, tagging (labelling of important milestones) and branching (variations to the project from main project) could be used to mark significant variations and modifications made to a particular model.

- Documentation - Documentation is also an important component in knowledge transfer. The reports and diagrams made in each step should be properly documented. Less documentation will be used for simple and clear projects. In contrast, complex projects should have more documentation in detail.

Complex projects can be performed using the user-story-driven approach to capture and organize the requirements (Collier 2011). UML based use-case is a practical strategy to identify the user requirements and associate it with actors. Here, the collaboration with the planners and consumers is essential. Even if it is simple, a similar approach with fewer use-cases can be used. If the same project is repeated with different datasets, it can be performed using a data driven approach without an explicit knowledge on the process.

6.6.1. Technical Documentation Approach

In this section, a technical documentation necessary to carry out the complete process will be proposed using UML and will be shown with an example application scenario. The user manual for documentation elements is given in APPENDIX D. In the documentation, two types of UML diagrams can be identified, namely, business related diagrams (business use-case diagram, business use-case realization diagram, business goal diagram and business analysis diagram) and HA related diagrams (analytic use-case diagram, analytic goal model, technique diagram, algorithm model and analytic model diagram). These diagrams are derived from the UML

definition of models and model extensions for data mining (Marban and Segovia 2013) and are revised to support HA.

An example of the application of each of the diagram is given in APPENDIX D based on the case description in *Hospital X*.

1. Domain understanding

At this stage there are several UML diagrams that will be used to document tasks that have been carried out. Even though it may be hard to capture all the necessary details (including the requirements), UML techniques like use-cases will allow making of a ballpark estimate at the initiation stage (Podeswa 2005). Use-case was used to capture functional requirements. Combination of use-cases is known as use-case model (or diagram) and it will record the complete functionality of the project (Jacobson et al. 1999). Thus, the traditional functional specification approach will be replaced by UML based use-cases. The developed USAM will be defined as a use-case driven process model. That is, use-cases will be used to capture the business requirements as well as it will drive the data preparation, modelling, testing and deployment of the project (Jacobson et al. 1999). Other than being the initiating step, use-cases will support to maintain the integrity of the project. Therefore, this process will continue through a series of workflows based on the use-case diagram.

Diagrams like business use-case diagram, business goal diagram and business analysis diagram representing the business perspective of the problem domain are used here. Furthermore, health analytics use-case diagram and health analytics goal diagram were designed to represent the HA perspective. The description of each diagram is given below.

1. **Business use-case diagram** will include the scope of the business and will correspond to the business processes. The general elements in a business use-case diagram are demonstrated in APPENDIX D. In each business use-case it should include actors (an individual or a system), goal, precondition (certain things that should occur or available before use-case begins), post conditions (outcome of the use-case), main flow (a sequence of events to move from precondition to post condition), exceptions (events that is possible to go wrong) and alternative flow (variation to the main flow).

2. **Business goal diagram** is used to indicate the business requirements. It represents the relationship between business use-cases and the business goals. The HA project will be carried out to achieve the business goal and will be related to at least one business use-case. The elements in a business goal diagram are illustrated in APPENDIX D. For example, a business goal could be increasing the volume (number of patients cared per day, number of patients undergone surgery per day), improve productivity, improve brand image, build personal health management portal to support patients. Here, it is important to determine the primary goal and other secondary goals. Moreover, generalization could be used to represent the overall goal and the sub goals.

3. **Analytic use-case diagram** is developed based on the business use-cases to illustrate how the interpretation of the knowledge extracted by HA is provided to its users. APPENDIX D presents the analytic use-case diagram elements and notations. This will depend on the business goals and one or more business goals could be represented by a HA use-case diagram (has many HA use-cases). This is used to indicate the interaction between the users and the

interpretation of the knowledge extracted. Potential users could be planners, doers or consumers. For example, ranking the factors that influence the long term clinical status after undergoing a surgery, forecasting the success after 2 years of surgery, and creation of a patient profile can be considered as HA use-cases. A particular HA use-case may include another use-case or may extend to another use-case.

4. **Analytic goal diagram** is used to represent the HA project requirements in HA perspective. As examples of HA goals we can consider creating a descriptive model of the medication adherence behaviour of a patient, creating a predictive model to forecast whether a certain patient will adhere to physician medication advices after discharge from hospital or a prescriptive model to understand what happens in the long run if not adhered to medication guidelines (e.g. not completion of specific dosage cycle of a drug). Here, the generalization could be used to indicate specific and abstract goals. Elements and notations in an analytic goal diagram are presented in APPENDIX D.

2. Data understanding

1. **Data diagram** indicates data sources, data types and the relationships. This will indicate the data integrations, derivations and transformations to the data sources as well as PHI attribute containing data sets. Moreover, it is important to indicate the data format as the data is obtained from heterogeneous sources in different formats. The elements and notations of HA data model are represented in APPENDIX D.
2. **Data component diagram** is used to indicate the relationships among the elements like documents, files (e.g. image files, flat files, web pages),

glossaries, folders, etc. This is used to represent the components that will support the data modelling. This includes files used by the project as well as any other element related to the project. For example, this could be used to organize and indicate the relationship between various related artefacts corresponding to the meaning of the data sources.

3. Conceptualization

There will not be any special UML document for conceptualization and hypothesis development. The literature reviewed will be optionally organized using a component diagram. Thus, it will be using hierarchy structure representing the specialization of documents. However, it could be provided as a text document in the literature review report. For the theoretical model, a separate UML diagram will not be used as it will be already represented as a diagram indicating independent variables, dependent variable and their relationships.

4. Data preparation

After initial data preparation (e.g. missing data, outliers, etc.) at the data understanding stage, data will be modified based on the HA goals. At the data preparation stage, there will be a *modified data diagram*. This model will include data construction information, data integration information and data formatting information represented in USAM model under data preparation phase. Elements in the modified data diagram are shown in APPENDIX D.

Furthermore, it is important to represent the actions performed to process the data visually. For different techniques different data formatting strategies may have used. Also, without knowing the exact sequence (sequential, concurrent or branched) of the data processing performed, it is not possible to regenerate the same modified

dataset from the original dataset. Thus, we are introducing an activity diagram to be used to present the data preparation activity flow. This is a new diagram introduced at the data preparation stage. Specially, as mentioned under uniqueness of medical data, most of the data are incomplete, missing, redundant or inconsistent (Cios and Moore 2002). To handle those issues in data it is important to perform certain modifications to data to be used in data modelling. APPENDIX D illustrates the elements in the activity diagram and later on in APPENDIX D the application example of the data preparation activity diagram is given.

5. Data modelling

In data modelling stage there are three UML diagrams, namely, technique diagram, algorithm diagram and analytic model diagram. At this stage various HA techniques are selected with corresponding algorithms and parameters to determine the optimal result.

1. *Technique diagram* indicates the HA techniques that had been applied to achieve HA goals. This indicates the data sources used, inputs and HA technique used. The elements in technique diagram are shown in APPENDIX D. For example, neural networks, regression, decision trees, what if analysis are some of the techniques available. The tool used in this model will be presented based on the technique. However, at first it is important to identify the health data type and analytic technique type.
2. *Algorithm diagram* is used to indicate the algorithms used by the HA techniques to extract knowledge. The algorithm depends on the data and the technique used as well as sometimes on the tool used. For example, decision tree could be using ID3 (Iterative Dichotomiser 3) or C4.5 or CART, neural

networks could be using feed forward or back propagation and support vector machines (SVM) could be using different kernel functions (e.g. linear, polynomial, Gaussian , etc.). The elements of the algorithm diagram are shown in APPENDIX D.

3. *Analytic model diagram* indicates the HA models used and where they are stored. The elements in this diagram are revealed in APPENDIX D. Analytic model diagram is used to specify and store the data analytic models derived from the data (e.g. forecasting model).

6. Validation

To evaluate the results at this stage an analytic test diagram will be created to indicate the transfer of results to interpretations and to indicate the approved data models. The elements are shown in APPENDIX D.

7. Presentation

The *health analytic deployment diagram* is used to indicate how knowledge extracted (interpreted) are deployed in the live environment. This will illustrate how physical hardware is used to deploy the software application developed based on the knowledge extracted from the data. This includes the server, monitor, caching server, medical devices, sensors/telemetric devices, modem, etc. The elements are shown in APPENDIX D.

UML diagrams used in this chapter are given with their connections in Figure 19. The arrows indicate the order of movement of content from one model to the other. The dotted arrows indicate the indirect relationship. When there are several HA goals, there could be many HA technique models, HA algorithm models, HA-model

models and HA validation models. To get a complete picture of the diagrams used with their connections, it is necessary to include all of them in the diagram.

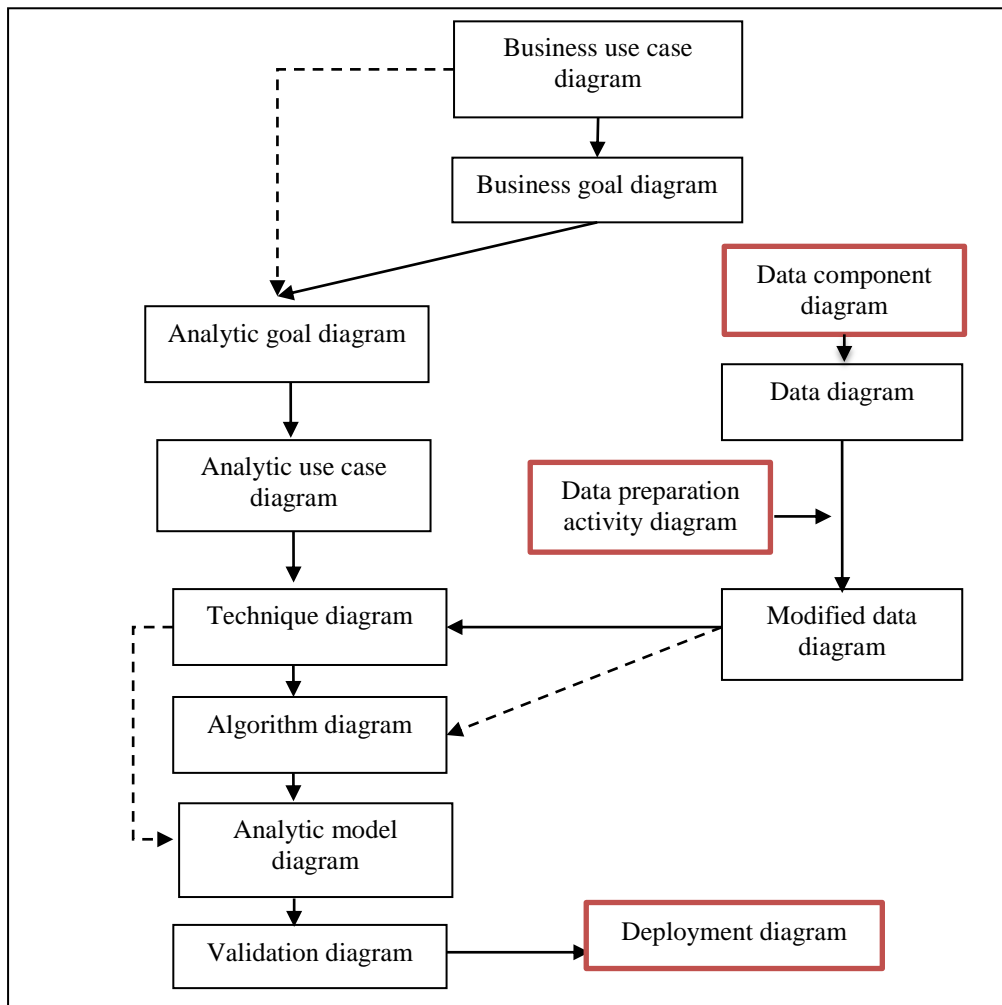


Figure 19: Overview of the UML diagram used

The unique components introduced to the UML diagrams compared to available UML diagrams for data mining (Marban and Segovia 2013) are the data preparation activity diagram, the data component diagram, and the deployment diagram.

6.6.2. Extending UML Diagrams

A modelling language like UML could be used to represent information and system structure. Considering the popularity and wide acceptance of UML in documenting systems, we propose to provide an extension to UML (Marban and Segovia 2013;

Zubcoff and Trujillo 2006). Even though there are documentation strategies based on UML proposed, they have failed to cover business and project requirements and to be a part of HA projects. In this study, we extended the UML by means of a profile to be used in each phase of the USAM. By using the extension mechanism, UML profiles customize the diagrams to a particular domain (for different use) (Zubcoff and Trujillo 2007). The extensions are specified through stereotypes, properties and restrictions (Zubcoff and Trujillo 2007) while respecting original semantics in UML (OMG 2011). Thus, in this study we extended the UML profile to facilitate the HA process proposed by us (USAM).

Two types of UML models can be identified here, namely, business related models (business use case model, business use case realization model, business goal model and business analysis model) and HA related models (HA use case model, HA goal model, HA technique model, HA algorithm model and HA models model). These models are adopted from the UML definition of models and model extensions for data mining (Marban and Segovia 2013) and are revised to support HA. We have proposed four new diagrams to represent heterogeneous data sources, activities performed in data preparation, and deployment.

The OCL (Object Constraint Language) 2.0 which is refined in UML 2.0 provides a means to express constraints in a model. As a query only language, it allows to present pre-conditions, post-conditions and invariants. An example application of UML profile extension is given below in Figure 20 and Figure 21.

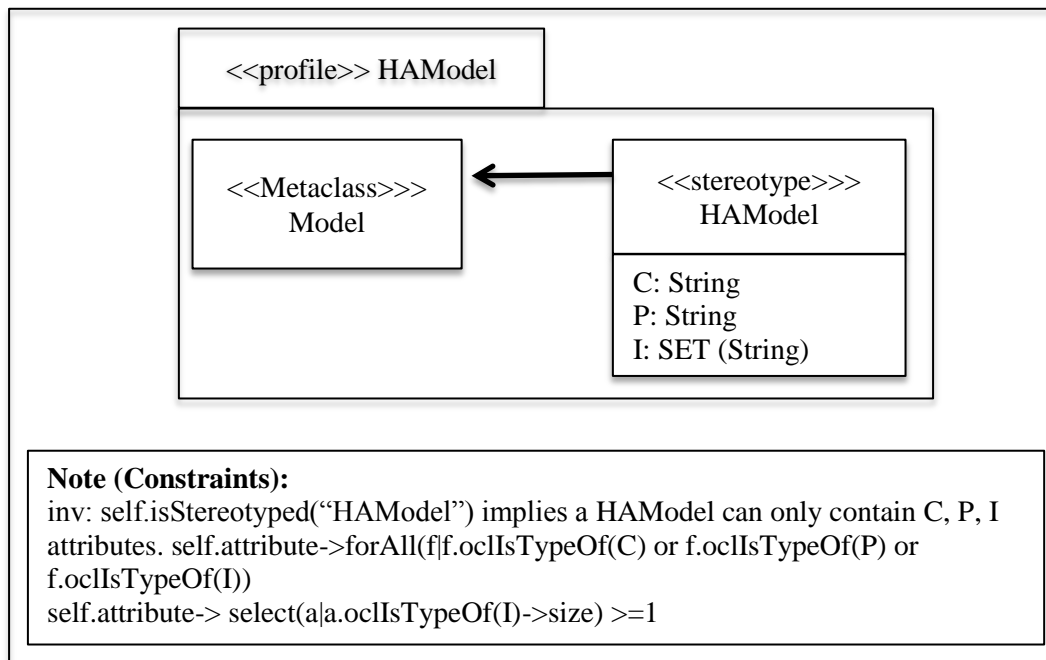


Figure 20: UML profile extension

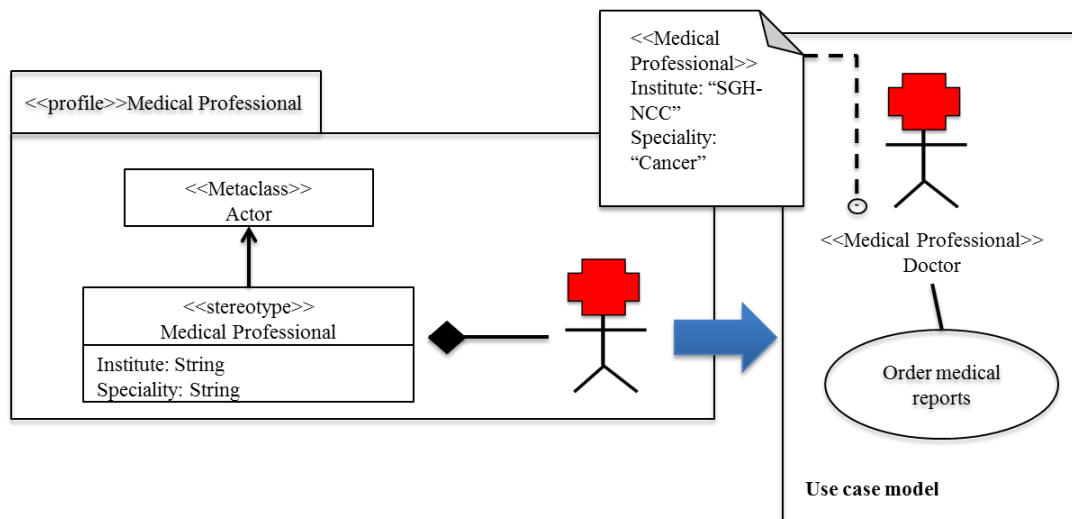


Figure 21: UML profile extension for actor

6.7. Discussion

In the evolving field of HA, there has been a necessity for a standard methodology with a set of best practices which are not too complicated, to deal with diversified and iterative processes in healthcare projects. We believe that this proposed model will (1) facilitate to articulate general guidelines to specific actionable steps (by a structured

process with detailed and repeatable actions), (2) hold true under real application scenarios and not merely under idealized conditions (by having practical techniques and by illustrating its application in real scenarios) and (3) have a gradual learning curve. As a means of achieving that, the USAM process model infused with project management, knowledge management and communication management is presented in this thesis.

The variations to project management, communication management and knowledge management are identified depending on the difficulty and clarity of the projects. The variations are described based on the two extremes as complex and ambiguous projects (*project A*) and clear and simple projects (*project B*). *Project A* will follow the 8 steps and necessary tasks associated. However, a simplified version can be used for the *project A*.

In the latter part of this chapter, it is explained how UML documentation could be used in a HA process. Importantly, a modelling language like UML could be used to represent information and system structure. There are several benefits. First, this would allow the users and analysts to direct their focus to the main objective, namely the HA process. Second, this is useful in reducing the textual documents and this assists as a communication tool to improve the customer understanding by having visual diagrams. Third, this provides an organization structure to represent the artifices in the project. The proposed UML diagrams take the interdependence between each other into account. Fourth, considering the acceptance and the popularity of UML diagrams, this will assist to reduce the project learning time.

Avoidance of considering specific components in HA algorithms is a limitation of this study. We have represented UML diagrams in a generalized manner to be used in a HA context. For example, in modelling we have not considered

specific details of an algorithm; like in association mining we could have considered the case stereotype, support and confidence constraints, etc. We believe that by considering the generic factors, the proposed use-case driven USAM will be able to be applied even in other models, which were not separately taken into consideration as all the algorithms specific to HA are impossible to be taken into account. Thus, if there are any specific UML extensions for a certain algorithm we suggest using them.

Identifying the exact business goals and the business cases at the planning stage itself is very useful. Thus, based on the business use-cases one can easily and clearly define the HA goals and the relevant HA use-cases. The identification of HA use-cases allow us to understand the complete process to be carried out in modelling the dataset and it will drive the subsequent phases in the USAM model too. In addition, as can be seen here, each UML diagram is interrelated. Thus, this allows cross referencing and maintaining a clear understanding of the process to the team members (especially if there are separate business analyst, data designer and a data analyst).

6.8. Summary

The final overall model developed based on the problems identified through literature review and the design science research method is explained using three supporting dimensions in process management. The designed process model is composed of eight steps starting from gaining access to the data and domain understanding to the presentation of results. This will allow implementing HA projects in a coherent manner. There will be variations to overarching process model based on the difficulty and clarity of the project requirements. These variations to the process model were discussed in this chapter. This will be a complete process which will be an iterative problem solving cycle (with data cycle and data model cycle). Finally, the

documentation approach using UML was presented as a visual representation of the process.

CHAPTER 7. DISCUSSION AND CONCLUSION

In this Thesis, a HA process model (Unified Structured Analytic Model - USAM) developed using the design science research (DSR) approach is presented. The process model was developed to carry out health analytic (HA) projects systematically by identifying project management, communication management and knowledge management aspects while dealing with data collection, sharing and analysis. The method was developed targeting novice data analysts.

The model was demonstrated as an iterative and incremental process. The inputs, outputs and tasks to be performed are clearly defined within eight steps in an iterative and incremental life cycle model. Moreover, a document template facilitating domain understanding, data preparation, data modelling, etc. was provided to capture necessities of each stage (using Unified modelling language - UML notations). Due to the acceptance and popularity of UML, its usage as the documentation strategy allows the analysts to direct their focus on main objectives rather than on different documenting approaches thus simplifying the representations compared to textual descriptions.

The model is developed based on current literature and extracting essential concepts from software engineering and data mining as well as prior work related to application of DSR in methodology development. The model development was carried out as an iterative process using DSR. The summary details of the application of DSR guidelines proposed by Hevner et al. (2004) are given in the Table 10. The approach we have used as per Vaishnavi and Kuechler (2005) was mapped to these seven guidelines.

Table 10: Summary details of application of DSR guidelines

Guideline	Application
1. Design as artefact	A ‘method’ to solve HA problems in a form of a textual description as the best practice. The uniqueness of the artefact was given in Chapter 3.
2. Problem relevance	The method was developed for novice users (practitioners) to understand “ <i>what to do</i> ” and “ <i>how to do</i> ” a HA project and to reduce their learning curve when commencing a project (Chapter 1).
3. Design evaluation	Method: An action case based approach (Chapter 5)
4. Research contribution	<ol style="list-style-type: none"> 1. The design artefact – method (that is the process model for HA) 2. The development process of the method and the evaluation approach
5. Research rigor	<ol style="list-style-type: none"> 1. Research methodology: Use of DSR approach proposed by Hevner et al. (2004) and Peffers et al. (2007). 2. Outset with the knowledge base (Literature study on data mining and software engineering methodologies and application of Theory – e.g. Diffusion of Innovations Theory, Media Synchronicity Theory, knowledge management framework proposed by Argote et al. (2003) and agile approach.) 3. Evaluation method: interpretive study (with a HA team based in a hospital).
6. Design as a search process	<p>Performance of an action case study in the centralized health analytic department of a hospital.</p> <p>Formal and informal interviews with the project manager and data analysts were carried out in addition to participating in their regular meetings as an observer.</p> <p>Systematic data analysis elucidates needs formulated as design criteria.</p>
7. Communication of research	<p>Technology oriented audience: details on how to be used within a HA project</p> <p>Management oriented audience: details on how to make it adopted in an organization (via Diffusion of Innovations Theory)</p>

Theory of Diffusion of Innovation (DOI) and Technology Acceptance Model (TAM) were used to identify what methodological attributes are looked for by novice users to HA. Through a survey, it was found that result demonstrability and relative

advantage are significant technical characteristics of a process model affecting its usage intention. Also, it was identified that perceived usefulness of knowledge management is a significant supporting element of the process model. The changes made to the model based on these findings were given in respective chapters.

The overall structure of the USAM was improved based on the data analysed, information gathered from the interviews and observations made in the two hospitals. An action case research methodology was used to evaluate the USAM as the focus of this study. Getting the opportunity to work as an internal employee of a HA department in a hospital, allowed me to gain better understanding of the organizational structure and actual project scenarios. Even though it was decided to consider analytic type (advanced analytic vs. descriptive analytics) to be used to identify the variations to the model initially, necessary modifications were identified from the interpretive study as the variations in projects in real working environments depend on clarity and the complexity of the project requirements.

The introduction of agile approach in the process model will allow data analysts to have greater control over their work while improving the quality of work and user effectiveness. For example, continuous user collaboration and evolutionary data modelling (as part of agile concept) enables understanding requirements and meeting user expectations at the end.

7.1. Implications

According to Hevner et al. (2004), a DSR study should provide contributions in the design artefact, design construction knowledge and design evaluation knowledge. Thus, the contributions of this research can be described as follows.

1. The design artefact – the USAM process model. The evaluated process model will assist the novice users at the commencement of a HA project as a guidance

to progress in the project. During the project implementation it will be a directive for them to understand how to conduct and present HA models and interpretations. The methodology built was developed based on the perception of the users and evaluated in a real scenario to be closer to their specific needs and they can easily apply it in carrying out their HA projects. Since there is no existing methodology evaluated specifically for HA context, it is considered that the developed artefact itself is an implication from this research.

2. Foundations - Theory of Diffusion of Innovation (DOI) with Technology Acceptance Model (TAM) was used to identify what methodological attributes novice users are looking for in a HA process model. Thus, we were able to extend the knowledge base by incorporating DOI and TAM with project management (framed using agile concept), knowledge management (framed using Organizational knowledge management framework proposed by Argote et al. (2003)) and communication management perspectives (framed using Media Synchronicity Theory - MST).
3. Methodologies – the use of the development and evaluation methods. An iterative process was used for the evaluation with the aim of further improvement of the methodology. As evaluation, an action case methodology was carried out at the end as a hybrid of action research and case study approach. This approach and the measures used in this study will be another theoretical implication from this study to Design Science Research related to methodology development as an artefact.

The practical contributions of the process model are:

1. Availability of formal processes will enable new comers in an analytic project to comfortably and easily establish and sustain in the organisation. Explicit

methodological steps developed will allow them to gain understanding on how to progress in their projects and to focus on the technical aspects of modelling.

2. Differentiation of projects based on clarity and difficulty of project requirements allows the analysts to apply the process model in different contexts. This allows focussing on the clarity and complexity of the requirements rather than considering the variations based on the analytic type (which should be decided at the end after conceptualization).
3. Consideration of overall process model with three supporting dimensions of project management, communication management and knowledge management along with process management, prompts the data analysts to pay special attention to the organisational context rather than following a mere data mining process. How organisational practices should be improved based on these three pillars was explained through the interpretive study in *Hospital Y*.
4. The HA process model was developed considering the uniqueness of medical domain with the aim of reducing the knowledge gap between the medical professionals and the data analysts.

It is important to identify the institutional environment in which the proposed model can be appropriate and feasible. The proposed model was tested in healthcare context where the requirements and working environments are unique.

7.2. Limitations

Failure to focus beyond prescriptive analytics can be considered as one limitation of this study. This may limit the generalization ability of the findings to other advanced analytic projects. Since decision on analytic algorithms will be made at the end phases

of the process, this will not make an impact on the overall phases of the model. Moreover, the ability of theoretical generalization of the findings is possible through a single case study as per the “analytical generalization” (Pan and Tan 2011; Yin 2014).

The post-ante evaluation could have been performed as a behavioural study to understand the adoption of the designed model. However, the action case research approach is another form of approach used in design science research to develop a model iteratively as part of an organisation (Arnott 2006; Pries-Heje et al. 2008) where the model will be designed and developed based in the real application environment rather than evaluating a set of hypothesis (Gregor and Hevner 2013).

The model is evaluated in a specific context. Carrying out of the final model evaluation in a hospital with 5 to 10 data analysts could be considered as a limitation. Based on our experience while working with other hospital contexts, it is the typical size in an analytic team in a hospital thus; we did not per see any problem with the team size. Importantly, the use of the USAM process model and the support dimensions will be more useful when the team size increases. When the number of team members grows, coordination of tasks can become difficult and as such a proper process model will be more useful in teams of larger size.

In the study context evaluated, we have worked with data analytic approaches related to descriptive analytics, predictive analytics and prescriptive analytics. In this model development, the main problems considered were related to resource allocation (e.g. productivity, waiting time, admission/discharge rates) and risk stratification (classification) in a hospital as we have used this with time series data at *Hospital Y*. However, the individual data modelling approaches were not considered in this thesis study. For example, for time series modelling there is a different set of modelling requirements to be fulfilled like checking for seasonality, additive or multiplicative

models, differentiation, etc. The specific individual modelling requirements were not considered, as it will limit the generalizability. Also, it is not practical to consider all the modelling algorithms individually, as there are many such and it will be a never-ending activity. Furthermore, there are individual process models developed targeting different algorithms. Time series data (Catley et al. 2009) and association rules (Rizzi 2004) are some examples. As such the specific details were ignored in this study.

7.3. Future Work

As future work, the model can be improved by considering the cognitive strategies of the data analyst. According to Arnott (2006), cognitive biases are defined as “cognitions or mental behaviours that prejudice decision quality in a significant number of decisions for a significant number of people”. When making a decision to use a method or a system, cognitive biases could play a role other than the rational choice of an individual. Thus, to avoid or reduce the biases, de-biasing could be used. Several de-biasing strategies are proposed by Bazerman and Moore (2012), Keren (1990) and other authors. These strategies could be incorporated in the methodology to make the practitioners make better outcomes (e.g. decision rules) by overcoming negativism of being biased (memory bias, situation bias, statistical bias, confidence biases, etc.) (Arnott 2006). In the current study, de-biasing was not considered as it is out of scope of the objectives of this study and it will be a different perspective to the current approach.

Second, industry wide workshops are planned to be carried out to educate the users about the model. Also presenting the model in international conferences will facilitate reaching higher user groups. After the workshops, as future work it is expected to carry out a survey on the user perception on performance improvement and user satisfaction on the use of the process model in analytics. Also we would like

to investigate the influence of the project type on the intention to use the process model.

Third, it is important to explore how SNOMED CT (Systemized Nomenclature of Medicine - Clinical Terms) could be incorporated into the process model. It is a standard to present and code medical data and can be considered as clinical terminology covering clinical specialties, disciplines and requirements. SNOMED CT facilitates consistency in data available in clinical data management systems. Electronic health records (EHR) and other users are using SNOMED CT to record and share clinical and related data (IHTSDO 2014). It is useful in developing high quality clinical content in EHRs while representing clinical phrases in a standardized way. Clinical information can be recorded in a hierarchical nature with relevant clinical concepts and additional details. To query these data SNOMED CT queries should be formed in a specific structure. However, they have not being exclusively used in data mining process (Bellazzi and Zupan 2008). As future work, UML extension mechanisms can be provided for SNOMED CT by profile extensions.

Fourth, a HA tool can be developed to support the workflow in USAM. Certain outputs from preceding steps will be used as the input in subsequent steps. Moreover, it is important that this tool is linked with a modelling language like R, as such the user does not need to flip over to different applications and it will provide the interconnectivity between requirements, data, modelling and the presentation of results. The features that should be incorporated in the tool are: (1) version control of data, data models and results; (2) documentation support; (3) collaboration and knowledge sharing tools and (4) project management tools.

Finally, we plan to explore the applicability of the process model in other contexts. Even though, the model was developed as a generalized analytic model,

presently, it was tested in HA context. As such, as future work the model will be validated in other contexts like financial analytics too. The variation of the project type will be applicable in other contexts as well. For example, there will be simple/complex and clear/ambiguous requirements for analytic teams even in institutions like banks. However, there are certain components introduced to the USAM model considering the uniqueness of medical domain. For example, heterogeneity of data will not be a major issue in financial sector as the data is mostly structured. Such specific steps can be avoided in non-healthcare contexts. The data protection and ethical and social issues will be valid even in dealing with personal bank accounts. Larger the team size of an analytic team better would be the planning and coordination of the tasks and achieving maximum expected benefits from a process model.

7.4. Conclusion

This study is a design science project developed to provide guidance to novice analysts working on health analytics. It is useful to have a standard method to perform HA, as otherwise certain activities that are necessary to be performed may be overlooked if the analytics are carried out in an ad hoc manner. Even though there are several methodologies that have been developed for data mining, they are not based on existing psychological research. As such in this research relevant behavioural research related to software engineering (system development) methodologies and decision support system development methodologies were examined to assist the conceptualization of the problem.

The study was carried out using DSR approach and necessary theoretical support (through Theory of Diffusion of Innovation, Technology Acceptance Model, agile concepts, organisational knowledge frameworks, and Media Synchronicity

Theory) was used in method (artefact) construction and in the evaluation. Findings from a survey study carried out at the beginning on a preliminary process model were used in refining the process model. For example, special focus was given to the result demonstrability and was achieved by using continuous user collaboration and a transparent data modelling process and documented (visual) communication.

In this study the unit of analysis is the method built and it was evaluated in an organizational context (in real application scenario). The process model was developed by infusion of four dimensions, namely, process management, project management, communication management and knowledge management. The evaluation was performed using the action case research approach. After going through several iteration loops the final model was developed to carry out HA projects. The success of the model was evaluated using the opinion of the senior management and the data analysts in the Health Analytic department of a hospital.

BIBLIOGRAPHY

- Agarwal, R., and Prasad, J. 1997. "The Role of Innovation Characteristics and Perceived Voluntariness in the Acceptance of Information Technologies," *Decision sciences* (28:3), pp. 557-582.
- Aldawud, O., Elrad, T., and Bader, A. 2003. "Uml Profile for Aspect-Oriented Software Development," *Proceedings of Third International Workshop on Aspect-Oriented Modeling*: Citeseer.
- Ambler, S. W. 2004. *The Object Primer: Agile Model-Driven Development with Uml 2.0*. Cambridge University Press.
- Arch, E. C., and Cummins, D. E. 1989. "Structured and Unstructured Exposure to Computers: Sex Differences in Attitude and Use among College Students," *Sex Roles* (20:5-6), pp. 245-254.
- Argote, L., McEvily, B., and Reagans, R. 2003. "Managing Knowledge in Organizations: An Integrative Framework and Review of Emerging Themes," *Management science* (49:4), pp. 571-582.
- Arnott, D. 2006. "Cognitive Biases and Decision Support Systems Development: A Design Science Approach," *Information systems journal* (16:1), pp. 55-78.
- Baker, G. 2002. "The Effects of Synchronous Collaborative Technologies on Decision Making: A Study of Virtual Teams," *Information Resources Management Journal (IRMJ)* (15:4), pp. 79-93.
- Baskerville, R., and Pries-Heje, J. 2014. "Diffusing Best Practices: A Design Science Study Using the Theory of Planned Behavior," in *Creating Value for All through It*. Springer, pp. 35-48.
- Bazerman, M., and Moore, D. A. 2012. *Judgment in Managerial Decision Making*, (8 ed.).
- Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., and Jeffries, R. 2001. "Manifesto for Agile Software Development." Agile Alliance.
- Becker, K., and Ghedini, C. 2005. "A Documentation Infrastructure for the Management of Data Mining Projects," *Information and Software Technology* (47:2), pp. 95-111.
- Bellazzi, R., and Zupan, B. 2008. "Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines," *International journal of medical informatics* (77:2), pp. 81-97.
- Benbasat, I., and Zmud, R. W. 1999. "Empirical Research in Information Systems: The Practice of Relevance," *MIS quarterly* (23:1), pp. 3-16.
- Bohlouli, M., Schulz, F., Angelis, L., Pahor, D., Brandic, I., Atlan, D., and Tate, R. 2013. "Towards an Integrated Platform for Big Data Analysis," in *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*. Springer, pp. 47-56.
- Britos, P., Dieste, O., and García-Martínez, R. 2008. "Requirements Elicitation in Data Mining for Business Intelligence Projects," in *Advances in Information Systems Research, Education and Practice*. Springer, pp. 139-150.
- Carr, N. G. 2008. *The Big Switch: Rewiring the World, from Edison to Google*. WW Norton & Company.
- Catley, C., Smith, K., McGregor, C., and Tracy, M. 2009. "Extending Crisp-Dm to Incorporate Temporal Data Mining of Multidimensional Medical Data Streams: A Neonatal Intensive Care Unit Case Study," *Computer-Based*

- Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on: IEEE*, pp. 1-5.
- Chan, F. K., and Thong, J. Y. 2009. "Acceptance of Agile Methodologies: A Critical Review and Conceptual Framework," *Decision Support Systems* (46:4), pp. 803-814.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. 2000. "Crisp-Dm 1.0 Step-by-Step Data Mining Guide." CRISP-DM Consortium.
- Chen, H., Chiang, R. H., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS quarterly* (36:4).
- Cios, K. J., and Kurgan, L. A. 2005. "Trends in Data Mining and Knowledge Discovery," in *Advanced Techniques in Knowledge Discovery and Data Mining*. Springer, pp. 1-26.
- Cios, K. J., and Moore, W. G. 2002. "Uniqueness of Medical Data Mining," *Artificial intelligence in medicine* (26:1), pp. 1-24.
- Collier, K. 2011. *Agile Analytics: A Value-Driven Approach to Business Intelligence and Data Warehousing*. Addison-Wesley.
- Cortada, J. W., Gordon, D., and Bill, L. 2012. "The Value of Analytics in Healthcare: From Insights to Outcomes," IBM Institute for Business Value.
- Davenport, E. 2008. "Social Informatics and Sociotechnical Research—a View from the UK," *Journal of information science* (34:4), pp. 519-530.
- Davis, F. D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS quarterly* (13:3), pp. 319-340.
- Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. 1989. "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," *Management science* (35:8), pp. 982-1003.
- de Rooij, S. E., Abu-Hanna, A., Levi, M., and de Jonge, E. 2005. "Factors That Predict Outcome of Intensive Care Treatment in Very Elderly Patients: A Review," *Critical Care* (9:4), p. R307.
- Delaney, G., Gebski, V., Lunn, A., Lunn, M., Rus, M., Manderson, C., and Langlands, A. 1997a. "An Assessment of the Basic Treatment Equivalent (Bte) Model as Measure of Radiotherapy Workload," *Clinical Oncology* (9:4), pp. 240-244.
- Delaney, G., Gebski, V., Lunn, A., Lunn, M., Rus, M., Manderson, C., and Langlands, A. 1997b. "Basic Treatment Equivalent (Bte): A New Measure of Linear Accelerator Workload," *Clinical Oncology* (9:4), pp. 234-239.
- DeLuca, D., and Valacich, J. S. 2005. "Outcomes from Conduct of Virtual Teams at Two Sites: Support for Media Synchronicity Theory," *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on: IEEE*, pp. 50b-50b.
- Dennis, A. R., Fuller, R. M., and Valacich, J. S. 2008. "Media, Tasks and Communication Processes: A Theory of Media Synchronicity," *MIS Quarterly* (32:3), pp. 575-600.
- Dybå, T., and Dingsøy, T. 2008. "Empirical Studies of Agile Software Development: A Systematic Review," *Information and Software Technology* (50:9), pp. 833-859.
- Dzeroski, S. 2007. "Towards a General Framework for Data Mining," in *Knowledge Discovery in Inductive Databases*, S. Dzeroski and J. Struyf (eds.). Springer-Verlag, pp. 259-300.

- Eason, K. 2008. "Sociotechnical Systems Theory in the 21st Century: Another Half-Filled Glass," *Sense in Social Science: A collection of essays in honour of Dr. Lisl Klein, Broughton*).
- Easterbrook, S., Singer, J., Storey, M.-A., and Damian, D. 2008. "Selecting Empirical Methods for Software Engineering Research," in *Guide to Advanced Empirical Software Engineering*. Springer, pp. 285-311.
- Eckerson W. 2004. "Be Prepared: Profile Your Data," *Business Intelligence Journal* (9:1).
- Edström, M. 2009. "A Case Study: Moving from Ad Hoc to Agile Software Development," *rapport nr.: Report/Department of Applied Information Technology 2009: 027*).
- Eggebraaten, T. J., Tenner, J. W., and Dubbels, J. C. 2007. "A Health-Care Data Model Based on the HI7 Reference Information Model," *IBM Systems Journal* (46:1), pp. 5-18.
- Eisenhardt, K. M. 1989. "Building Theories from Case Study Research," *Academy of management review* (14:4), pp. 532-550.
- Esfandiary, N., Babavalian, M. R., Moghadam, A.-M. E., and Tabar, V. K. 2014. "Knowledge Discovery in Medicine: Current Issue and Future Trend," *Expert Systems with Applications* (41:9), pp. 4434-4463.
- Falk, R. F., and Miller, N. B. 1992. *A Primer for Soft Modeling*. University of Akron Press.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996a. "From Data Mining to Knowledge Discovery in Databases," *AI magazine* (17:3), p. 37.
- Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. 1996b. "From Data Mining to Knowledge Discovery: An Overview," in *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds.). American Association for Artificial Intelligence, pp. 1-34.
- Fichman, R. G., and Kemerer, C. F. 2012. "Adoption of Software Engineering Process Innovations: The Case of Object-Orientation," *Sloan management review* (34:2).
- Fichman, R. G., Kohli, R., and Krishnan, R. 2011. "Editorial Overview-the Role of Information Systems in Healthcare: Current Research and Future Trends," *Information systems research* (22:3), pp. 419-428.
- Finlay, P. N., and Forghani, M. 1998. "A Classification of Success Factors for Decision Support Systems," *The Journal of Strategic Information Systems* (7:1), pp. 53-70.
- Fitzgerald, B. 1996. "Formalized Systems Development Methodologies: A Critical Perspective," *Information systems journal* (6:1), pp. 3-23.
- Goodwin, B. 2011. "Poor Communication to Blame for Business Intelligence Failure, Says Gartner." ComputerWeekly.com.
- Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS quarterly* (37:2), pp. 337-356.
- Grossman, R. L., Hornick, M. F., and Meyer, G. 2002. "Data Mining Standards Initiatives," *Communications of the ACM* (45:8), pp. 59-61.
- Gujarati, D. N. 2003. "Basic Econometrics. 4th." New York: McGraw-Hill.
- Gujarati, D. N., and Porter, C. 2009. "Basic Econometrics." McGraw-Hill International Edition, 5th Edd., Boston, page260-261, p. 338.
- Hair, J. F., Tatham, R. L., Anderson, R. E., and Black, W. 2006. *Multivariate Data Analysis*. Pearson Prentice Hall Upper Saddle River, NJ.

- Hardgrave, B. C., Davis, F. D., and Riemenschneider, C. K. 2003. "Investigating Determinants of Software Developers' Intentions to Follow Methodologies," *Journal of Management Information Systems* (20:1), pp. 123-152.
- Herzlinger, R. E. 2006. "Why Innovation in Health Care Is So Hard," *Harvard business review* (84:5), p. 58.
- Hesse, B. W., Hansen, D., Finholt, T., Munson, S., Kellogg, W., and Thomas, J. C. 2010. "Social Participation in Health 2.0," *Computer* (43:11), pp. 45-52.
- Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS quarterly* (28:1), pp. 75-105.
- Highsmith, J. 2009. *Agile Project Management: Creating Innovative Products*. Pearson Education.
- Horner, P., and Basu, A. 2012. "Analytics & the Future of Healthcare," in: *Analytics Magazine*. pp. 11-18.
- IEEE. 1997. "Standard for Software Life Cycle Processes," in: *IEEE Std 1074 -1997*. IEEE, Nueva York (EE.UU.).
- IHTSDO. 2014. "Snomed Ct Starter Guide." from www.snomed.org/starterguide.pdf
- Iivari, J., and Iivari, N. 2011. "The Relationship between Organizational Culture and the Deployment of Agile Methods," *Information and Software Technology* (53:5), pp. 509-520.
- ISO. 1995. "Iso/Iec 12207:1995, Software Life Cycle Processes." Geneva (Switzerland): International Organisation for Standardization.
- Jacobson, I., Booch, G., and Rumbaugh, J. 1999. *The Unified Software Development Process*. Addison-Wesley Reading.
- Jibitesh Mishra, and Mohanty, A. 2011. *Software Engineering*. Pearson Education India.
- Johannesson, P., and Perjons, E. 2014. *An Introduction to Design Science*. Springer.
- Johnson, R. A., Hardgrave, B. C., and Doke, E. R. 1999. "An Industry Analysis of Developer Beliefs About Object-Oriented Systems Development," *ACM SIGMIS Database* (30:1), pp. 47-64.
- Kankanhalli, A., Lee, O.-K. D., and Lim, K. H. 2011. "Knowledge Reuse through Electronic Repositories: A Study in the Context of Customer Service Support," *Information & Management* (48:2), pp. 106-113.
- Kankanhalli, A., Tan, B. C., and Wei, K.-K. 2005. "Contributing Knowledge to Electronic Knowledge Repositories: An Empirical Investigation," *MIS quarterly* (29:1), pp. 113-143.
- Kankanhalli, A., Tan, B. C., Wei, K.-K., and Holmes, M. C. 2004. "Cross-Cultural Differences and Information Systems Developer Values," *Decision Support Systems* (38:2), pp. 183-195.
- KdNuggets.com. 2014. "What Main Methodology Are You Using for Your Analytics, Data Mining, or Data Science Projects?" Retrieved 17 May 2015, from <http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html>
- Keren, G. 1990. "Cognitive Aids and Debiasing Methods: Can Cognitive Pills Cure Cognitive Ills?," *Advances in Psychology* (68), pp. 523-552.
- Kim, H.-W., Chan, H. C., and Kankanhalli, A. 2012. "What Motivates People to Purchase Digital Items on Virtual Community Websites? The Desire for Online Self-Presentation," *Information systems research* (23:4), pp. 1232-1245.

- Kling, R., and Lamb, R. 1999. "It and Organizational Change in Digital Economies: A Socio-Technical Approach," *ACM SIGCAS Computers and Society* (29:3), pp. 17-25.
- Koch, N., Knapp, A., Zhang, G., and Baumeister, H. 2008. "Uml-Based Web Engineering," in *Web Engineering: Modelling and Implementing Web Applications*. Springer, pp. 157-191.
- Kozar, K. A. 1989. "Adopting Systems Development Methods: An Exploratory Study," *Journal of Management Information Systems* (5:4), pp. 73-86.
- Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., and Goodenday, L. S. 2001. "Knowledge Discovery Approach to Automated Cardiac Spect Diagnosis," *Artificial intelligence in medicine* (23:2), pp. 149-169.
- Kwiatkowska, M., Atkins, M. S., Ayas, N. T., and Ryan, C. F. 2007. "Knowledge-Based Data Analysis: First Step toward the Creation of Clinical Prediction Rules Using a New Typicality Measure," *Information Technology in Biomedicine, IEEE Transactions on* (11:6), pp. 651-660.
- Lagerberg, L., Skude, T., Emanuelsson, P., Sandahl, K., and Stahl, D. 2013. "The Impact of Agile Principles and Practices on Large-Scale Software Development Projects: A Multiple-Case Study of Two Projects at Ericsson," *Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on: IEEE*, pp. 348-356.
- Lee, A. S. 2001. "Editorial Comments: Mis Quarterly's Editorial Policies and Practices," *MIS Quarterly* (25:1), pp. iii-vii.
- Li, X.-B., and Qin, J. 2013. "A Framework for Privacy-Preserving Medical Document Sharing," *Thirty Fourth International Conference on Information Systems*, Milan, Italy.
- Lindvall, M., and Rus, I. 2002. "Knowledge Management in Software Engineering," *IEEE software* (19:3), pp. 26-38.
- Luján-Mora, S., Trujillo, J., and Song, I.-Y. 2006. "A Uml Profile for Multidimensional Modeling in Data Warehouses," *Data & Knowledge Engineering* (59:3), pp. 725-769.
- Lyytinen, K., Newman, M., and Al-Muharfi, A.-R. A. 2009. "Institutionalizing Enterprise Resource Planning in the Saudi Steel Industry: A Punctuated Socio-Technical Analysis," *Journal of Information Technology* (24:4), pp. 286-304.
- Marban, O., Mariscal, G., and Segovia, J. 2009a. "A Data Mining & Knowledge Discovery Process Model," in: *Data Mining and Knowledge Discovery in Real Life Applications*, J. Ponce and A. Karahoca (eds.). InTech, p. 8.
- Marban, O., and Segovia, J. 2013. "Extending Uml for Modeling Data Mining Projects (DM-UML)," *Journal of Information Technology & Software Engineering* (3:121).
- Marban, O., Segovia, J., Menasalvas, E., and Fernández-Baizán, C. 2009b. "Toward Data Mining Engineering: A Software Engineering Approach," *Information systems* (34:1), pp. 87-107.
- March, S. T., and Smith, G. F. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4), pp. 251-266.
- Mariscal, G., Marbán, Ó., and Fernández, C. 2010. "A Survey of Data Mining and Knowledge Discovery Process Models and Methodologies," *The Knowledge Engineering Review* (25:02), pp. 137-166.
- Matignon, R. 2007. *Data Mining Using Sas Enterprise Miner*. John Wiley & Sons.
- Meltzoff, J. 1998. *Critical Thinking About Research: Psychology and Related Fields*. American Psychological Association.

- Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., and Samore, M. H. 2010. "Automatic De-Identification of Textual Documents in the Electronic Health Record: A Review of Recent Research," *BMC medical research methodology* (10:1), p. 70.
- Mohan, K., and Ahlemann, F. 2011. "What Methodology Attributes Are Critical for Potential Users? Understanding the Effect of Human Needs," *Advanced Information Systems Engineering*: Springer, pp. 314-328.
- Moore, G. C., and Benbasat, I. 1991. "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation," *Information systems research* (2:3), pp. 192-222.
- Myers, M. D., and Newman, M. 2007. "The Qualitative Interview in Is Research: Examining the Craft," *Information and organization* (17:1), pp. 2-26.
- Nambisan, S. 2003. "Information Systems as a Reference Discipline for New Product Development," *MIS quarterly* (27:1), pp. 1-18.
- Narayanan, A., and Shmatikov, V. 2010. "Myths and Fallacies of Personally Identifiable Information," *Communications of the ACM* (53:6), pp. 24-26.
- Naur, P., and Randell, B. 1969. *Software Engineering: Report of a Conference Sponsored by the Nato Science Committee, Garmisch, Germany, 7-11 Oct. 1968, Brussels, Scientific Affairs Division, Nato.*
- Nelson, G. S. 2010. "Business Intelligence 2.0: Are We There Yet," *SAS Global Forum*.
- Newman, M., and Zhao, Y. 2008. "The Process of Enterprise Resource Planning Implementation and Business Process Re-Engineering: Tales from Two Chinese Small and Medium-Sized Enterprises," *Information systems journal* (18:4), pp. 405-426.
- Niinimäki, T., Piri, A., and Lassenius, C. 2009. "Factors Affecting Audio and Text-Based Communication Media Choice in Global Software Development Projects," *Global Software Engineering, 2009. ICGSE 2009. Fourth IEEE International Conference on*: IEEE, pp. 153-162.
- OMG. 2011. "Omg Unified Modeling Language (OMG UML) Superstructure Specification Version 2.4. 1," document formal/2011-08-06. Technical report, OMG.
- Orlikowski, W. J., and Iacono, C. S. 2001. "Research Commentary: Desperately Seeking the "It" in It Research—a Call to Theorizing the It Artifact," *Information systems research* (12:2), pp. 121-134.
- Osborne, C. 2012. "Healthcare Analytics Market to Reach \$10.8 Billion by 2017?" *ZD Net, Innovation* Retrieved 17 May 2015, from <http://www.zdnet.com/article/healthcare-analytics-market-to-reach-108-billion-by-2017/>
- Paetsch, F., Eberlein, A., and Maurer, F. 2003. "Requirements Engineering and Agile Software Development," *2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*: IEEE Computer Society, pp. 308-308.
- Pan, S. L., and Tan, B. 2011. "Demystifying Case Research: A Structured–Pragmatic–Situational (Sps) Approach to Conducting Case Studies," *Information and organization* (21:3), pp. 161-176.
- Peffer, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 45-77.

- Peterson, S. D., Jaret, P. E., and Schenck, B. F. 2013. *Set Goals and Objectives in Your Business Plan*, (4 ed.). Dummies.
- Pettigrew, A. M. 1990. "Longitudinal Field Research on Change: Theory and Practice," *Organization science* (1:3), pp. 267-292.
- Podeswa, H. 2005. *Uml for the It Business Analyst: A Practical Guide to Object-Oriented Requirements Gathering*. Thomson Course Technology PTR Boston.
- Prat, N., Akoka, J., and Comyn-Wattiau, I. 2006. "A Uml-Based Data Warehouse Design Method," *Decision Support Systems* (42:3), pp. 1449-1473.
- Pries-Heje, J., and Baskerville, R. 2008. "The Design Theory Nexus," *MIS quarterly* (32:4), pp. 731-755.
- Pries-Heje, J., Baskerville, R., and Venable, J. R. 2008. "Strategies for Design Science Research Evaluation," *ECIS 2008 Proceedings. Paper 87*.
- Raghavan, S. A., and Chand, D. R. 1989. "Diffusing Software-Engineering Methods," *Software, IEEE* (6:4), pp. 81-90.
- Raghupathi, W., and Raghupathi, V. 2013. "An Overview of Health Analytics," *J Health Med Informat* (4:132), p. 2.
- Riemenschneider, C. K., Hardgrave, B. C., and Davis, F. D. 2002. "Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Theoretical Models," *Software Engineering, IEEE Transactions on* (28:12), pp. 1135-1145.
- Rivo, E., de la Fuente, J., Rivo, Á., García-Fontán, E., Cañizares, M.-Á., and Gil, P. 2012. "Cross-Industry Standard Process for Data Mining Is Applicable to the Lung Cancer Surgery Domain, Improving Decision Making as Well as Knowledge and Quality Management," *Clinical and Translational Oncology* (14:1), pp. 73-79.
- Rizzi, S. 2004. "Uml-Based Conceptual Modeling of Pattern-Bases," *PaRMA*.
- Roberts, T. L., Gibson, M. L., Fields, K. T., and Rainer Jr, R. K. 1998. "Factors That Impact Implementing a System Development Methodology," *Software Engineering, IEEE Transactions on* (24:8), pp. 640-649.
- Rogers, E. M. 2010. *Diffusion of Innovations*. Simon and Schuster.
- Rogers, R. D., and Monsell, S. 1995. "Costs of a Predictable Switch between Simple Cognitive Tasks," *Journal of experimental psychology: General* (124:2), p. 207.
- Rohanizadeh, S. S., and Moghadam, M. B. 2009. "A Proposed Data Mining Methodology and Its Application to Industrial Procedures," *J. Industr. Eng*:4), pp. 37-50.
- Romanow, D., Cho, S., and Straub, D. 2012. "Editor's Comments: Riding the Wave: Past Trends and Future Directions for Health It Research," *MIS quarterly* (36:3), pp. III-A18.
- Rumbaugh, J., Jacobson, I., and Booch, G. 2004. *Unified Modeling Language Reference Manual, The*. Pearson Higher Education.
- SAS. 2008. "Sas Enterprise Miner: Semma." Retrieved 27 February 2014, 2014, from <http://www.sas.com/technologies/analytics/datamining/miner/semma.html>
- Sattler, K.-U., and Schallehn, E. 2001. "A Data Preparation Framework Based on a Multidatabase Language," *Database Engineering and Applications, 2001 International Symposium on*: IEEE, pp. 219-228.
- Sawyer, S., and Jarrahi, M. 2014. "Sociotechnical Approaches to the Study of Information Systems," in *Computing Handbook, Third Edition: Information Systems and Information Technology*. pp. 5.1-5.27.

- Schmidt, S., Vuillermin, P., Jenner, B., Ren, Y., Li, G., and Chen, Y.-P. P. 2008. "Mining Medical Data: Bridging the Knowledge Divide," *eResearch Australasia 2008*.
- Shouval, R., Bondi, O., Mishan, H., Shimoni, A., Unger, R., and Nagler, A. 2014. "Application of Machine Learning Algorithms for Clinical Predictive Modeling: A Data-Mining Approach in Sct," *Bone marrow transplantation* (49:March), pp. 332-337.
- Stefanowski, J. 2010. "Data Mining - Evaluation of Classifiers." Retrieved 04/04/2014, from <http://www.cs.put.poznan.pl/jstefanowski/sed/DM-4-evaluatingclassifiersnew.pdf>
- Stockdale, R., and Standing, C. 2006. "An Interpretive Approach to Evaluating Information Systems: A Content, Context, Process Framework," *European Journal of Operational Research* (173:3), pp. 1090-1102.
- Swanstrom, R. 2013. "Data Mining Standard Processes." *Data Science 101* Retrieved 15 May 2015, from <http://101.datascience.community/2013/07/18/data-mining-standard-processes/>
- Teow, K. L., El-Darzi, E., Foo, C., Jin, X., and Sim, J. 2012. "Intelligent Analysis of Acute Bed Overflow in a Tertiary Hospital in Singapore," *Journal of medical systems* (36:3), pp. 1873-1882.
- Tjørnehøj, G., Balogh, M. B., Iversen, C., and Sørensen, S. 2014. "Designing Project Management for Global Software Development," in *Creating Value for All through It*. Springer, pp. 113-132.
- Vaishnavi, V., and Kuechler, B. 2005. "Design Research in Information Systems." Retrieved 15 August 2014, from <http://desrist.org/desrist/content/design-science-research-in-information-systems.pdf>
- Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS quarterly* (27:3), pp. 425-478.
- Watson, H. J. 2013. "All About Analytics," *International Journal of Business Intelligence Research (IJBIR)* (4:1), pp. 13-28.
- Weber, C. V., Paulk, M. C., Wise, C. J., and Withey, J. V. 1991. "Key Practices of the Capability Maturity Model," DTIC Document.
- Weber, R. A., and Camerer, C. F. 2003. "Cultural Conflict and Merger Failure: An Experimental Approach," *Management science* (49:4), pp. 400-415.
- Wegner, D. M. 1987. "Transactive Memory: A Contemporary Analysis of the Group Mind," in *Theories of Group Behavior*. Springer, pp. 185-208.
- Westphal, C., and Blaxton, T. 1998. *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*.
- Winter, R. 2008. "Design Science Research in Europe," *European Journal of Information Systems* (17:5), pp. 470-475.
- Wirth, R., and Hipp, J. 2000. "Crisp-Dm: Towards a Standard Process Model for Data Mining," in: *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*. Citeseer, pp. 29-39.
- Yang, Q., and Wu, X. 2006. "10 Challenging Problems in Data Mining Research," *International Journal of Information Technology & Decision Making* (5:04), pp. 597-604.
- Yeo, T., and Gaw, B. 2013. "Developments in Singapore's Healthcare Regulatory Landscape," in: *Asian Legal Business*. DREW & NAPIER, p. 53.

- Yin, R. K. 2014. *Case Study Research: Design and Methods*, (5 ed.). Sage publications.
- Yoo, I., Alafairet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., and Hua, L. 2012. "Data Mining in Healthcare and Biomedicine: A Survey of the Literature," *Journal of medical systems* (36:4), pp. 2431-2448.
- Zubcoff, J., Pardillo, J., and Trujillo, J. 2007. "Integrating Clustering Data Mining into the Multidimensional Modeling of Data Warehouses with Uml Profiles," in *Data Warehousing and Knowledge Discovery*. Springer, pp. 199-208.
- Zubcoff, J., and Trujillo, J. 2006. "Conceptual Modeling for Classification Mining in Data Warehouses," in *Data Warehousing and Knowledge Discovery*. Springer, pp. 566-575.
- Zubcoff, J., and Trujillo, J. 2007. "A Uml 2.0 Profile to Design Association Rule Mining Models in the Multidimensional Conceptual Modeling of Data Warehouses," *Data & Knowledge Engineering* (63:1), pp. 44-62.

APPENDIX A

Measurement Instrument

Construct	Item	Measurement	Sources
Intention to use process model (INT)	INT1	I intend to use an analytics method in the future.	Venkatesh et al. (2003)
	INT2	I predict I will use an analytics method in the future	
	INT3	I plan to use an analytics method in the future	
Ease of use (EOU)	EOU1	I believe that it is easy to get an analytics method to do what I want to do	Moore and Benbasat (1991)
	EOU2	Overall, I believe that an analytics method is easy to use	
	EOU3	Learning to operate an analytics method is easy for me	
Relative advantage (ADV)	ADV1	Using an analytics method enables me to accomplish tasks more quickly	Moore and Benbasat (1991)
	ADV2	Using an analytics method improves the quality of work I do	
	ADV3	Using an analytics method makes it easier to do my job	
	ADV4	Using an analytics method enhances my effectiveness on the job	
	ADV5	Using an analytics method gives me greater control over my work	
Compatibility (COM)	COM1	Analytics method is compatible with the way I develop systems	Hardgrave et al. (2003)
	COM2	Using an analytics method is compatible with all aspects of my work	
	COM3	Using an analytics method fits well with the way I work	
Result demonstrability (RDE)	RDE1	I would have no difficulty of telling others about the results of using an analytics method	Moore and Benbasat (1991)
	RDE2	I believe I could communicate to others the consequences of using an analytics method	
	RDE3	The results of using an analytics method are apparent to me	
	RDE4	I would have a difficulty explaining why using an analytics method may or may not be beneficial	
Triability (TRI)	TRI1	Before deciding whether to use any analytics method, I was able to properly try them out	Moore and Benbasat (1991)
	TRI2	I was permitted to use an analytics method on a trial basis long enough to see what I could do	
Project	PMA1	Using project management elements in	Hardgrave

management (PMA)		the model improves my job performance	et al. (2003)
	PMA2	Using project management elements in the model increases my productivity	
	PMA3	Using project management elements in the model enhances the quality of work	
	PMA4	Using project management elements in the model makes it easier to do my job	
	PMA5	The advantages of using project management elements in the model outweigh the disadvantages	
	PMA6	Project management elements in the model are useful in my job	
Knowledge management (KWM)	KWM1	Using knowledge management elements in the model improves my job performance	Hardgrave et al. (2003)
	KWM2	Using knowledge management elements in the model increases my productivity	
	KWM3	Using knowledge management elements in the model enhances the quality of work	
	KWM4	Using knowledge management elements in the model makes it easier to do my job	
	KWM5	The advantages of using knowledge management elements in the model outweigh the disadvantages	
	KWM6	knowledge management elements in the model are useful in my job	

APPENDIX B

General document template for domain understanding phase

Project document							
Project No.:							
Priority: <i>(If there are several HA projects going on)</i>							
Target date:							
Approved by:			Date:				
Prepared by:			Date:				
Version No.:							
Version Control:							
<ul style="list-style-type: none"> • Version History 							
Version No.	Phase changed	Date	Authorization	Author	Description		
<ul style="list-style-type: none"> • RACI chart <p>This specifies the roles (RACI- responsible, accountable, consulted and informed) played by team members and stakeholders in producing this project document.</p>							
Name	Position	*	R	A	S	C	I
Where;							
*	Authorize	Ultimate signing authority for any changes to the document					
R	Responsible	Responsible for creation of the document					
A	Accountable	Accountable for the accuracy of the document					
S	Supported	Supported in creating the document					
C	Consulted	Provided input when creating the document					
I	Informed	Must be informed of any changes					
Organization objectives							
<ul style="list-style-type: none"> • Background (Record the information about the organization and reasons for considering the project) <ul style="list-style-type: none"> ○ Develop organization chart (departments, specialties) ○ Determine key personnel in the company and their role ○ Determine the departments that will be affected by the HA project 							

- Objectives (Specifies business objectives addressed by the project)
 - Use a ACES approach to determine the goals of the provider
 - Define SMART objectives
- Success criteria
 - States the criteria for the outcome to be successful (The criterion should be related to the objectives, specific and measurable.)
 - Determine the person to assesses the criteria

Situation assessment

- Requirements (Specify the requirements of the project)
 - Problem to be addressed
 - Current solutions available to address the problem (including benefits and issues in the solution)
- Risk analysis
 - Technological risk (new technological issues that may impact the project)
 - Skill risk (unavailability of staff with required expertise for the project)
 - Requirement risk (risk of not correctly capturing the requirements)
 - Other risks
 - Risk matrix (likelihood * severity) for each risk

		Severity			
		Negligible	Marginal	Critical	Catastrophic
Likelihood	Certain				
	Likely				
	Possible				
	Unlikely				
	Rare				

- Contingency plan\ strategy to handle each risk based on the risk level (based on risk matrix)
- Feasibility
 - Operational feasibility (Prerequisites for the project - e.g. does the organization is using HA, Current status of the project - e.g. whether the project is already accepted, whether HA needs to promoted as a new technology to the organization)
 - Technical feasibility (Availability of necessary technology)
 - Schedule feasibility (Determine whether the project expectations can be fulfilled within the planned time frame)
 - Economic feasibility (Determine whether the economic benefits make it attractive to be implemented - Cost/benefit analysis to compare costs against the potential benefits of the project)
- Glossary of terminology (Relevant to the project)
 - Glossary of relevant healthcare related terminology (including healthcare standards – e.g. ICD10, SNOMED)
 - Glossary of HA related terminology
- HA goal (State the intended result of the HA project in technical terms)
 - Match business objective and HA objective

<table border="1" style="margin: auto;"> <tr> <th style="padding: 5px;">Business objective</th> <th style="padding: 5px;">HA objective</th> </tr> <tr> <td style="height: 20px;"></td> <td style="height: 20px;"></td> </tr> <tr> <td style="height: 20px;"></td> <td style="height: 20px;"></td> </tr> </table>	Business objective	HA objective							
Business objective	HA objective								
<ul style="list-style-type: none"> ○ HA problem type (e.g. descriptive, predictive, prescriptive, discovery and exploratory analytics) 									
<p>Business use-cases (End-to-end business processes affected by the project)</p> <ul style="list-style-type: none"> ● Business use-case diagrams (Specifies stakeholder involvement in each use-case, See 0) 									
<p>Stakeholders</p> <ul style="list-style-type: none"> ● Target group (State the profile of the target group to whom the results of the project will be presented) ● Role map (states the role, capability/expertise played by users and external systems) <table border="1" style="width: 100%; border-collapse: collapse; margin-top: 10px;"> <thead> <tr> <th style="padding: 5px;">User (position)/System</th> <th style="padding: 5px;">Role</th> <th style="padding: 5px;">Expertise</th> </tr> </thead> <tbody> <tr> <td style="height: 20px;"></td> <td style="height: 20px;"></td> <td style="height: 20px;"></td> </tr> <tr> <td style="height: 20px;"></td> <td style="height: 20px;"></td> <td style="height: 20px;"></td> </tr> </tbody> </table> <ul style="list-style-type: none"> ● User requirements and expectations (State the needs of each users) 	User (position)/System	Role	Expertise						
User (position)/System	Role	Expertise							
<p>Compliances</p> <ul style="list-style-type: none"> ● Regulations (security compliance and audit) <ul style="list-style-type: none"> ○ Personal Data Protection Act (PDPA) 2012 (Yeo and Gaw 2013) ○ IT governance based on ISO/IEC 27002:2005, FISMA, HIPAA compliant checklist ● Ethics <ul style="list-style-type: none"> ○ IRB approval 									
<p>Project plan</p> <ul style="list-style-type: none"> ● Project scope (State the in-scope and out-of-scope items) ● Tools and techniques (depends on HA goals) <ul style="list-style-type: none"> ○ HA technique for the task ○ HA tool for the technique ○ Prioritize the techniques to use ● Resource requirement (Determine accessibility, function and involvement in HA project) <ul style="list-style-type: none"> ○ Software requirement (e.g. software tools) ○ Hardware requirement (e.g. processing power, storage) ○ Data requirement (check if all the data necessary to work out HA goal available and check which data are unrelated, identify additional data required to achieve the HA goal and how to access them, consideration time period) ○ Personnel requirement (required skill set) ● Schedule (List the stages to be carried out with duration with their interdependencies) <ul style="list-style-type: none"> ○ Gantt chart (Illustrates the project schedule) ○ Project network diagram (Indicate the sequence of tasks and their dependencies) ● Communication plan (Provides consistent, timely and accurate information to the stakeholders and allows effective communication of deliverables to them.) 									

<ul style="list-style-type: none"> ○ Communication objectives (target audience and the message to deliver to them; e.g. customer – project plan and status report, review team – project briefing and status report) ○ Key content of the communication (e.g. project plan – Current future plans, project deliverables, issues ; status report – status summary, schedule, accomplishments, next step, issues; project briefing – status, checklist, issues) ○ Communication method (format and delivery mechanism – e.g. email, phone, formal presentation) and frequency (e.g. weekly, monthly) ○ Messenger (Describes who is responsible for the communication and who will present the content) ● Test plan (testing and validation of the data models to avoid biasness) <ul style="list-style-type: none"> ○ Test dataset ○ Validation dataset ● Implementation plan <ul style="list-style-type: none"> ○ Conversion (State existing data that must be converted) ○ Training (State who is responsible, how it is done) ○ Grant privilege to others to access the data models ○ Programs to promote the results ○ Post implementation follow up (determine whether there is a requirement to improve the outcomes) ● End user procedures (Write up of procedures to be carried out by the affected departments)
Other issues
Sign off

APPENDIX C

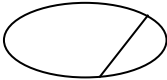
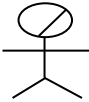
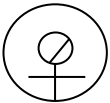
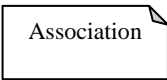

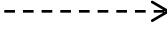
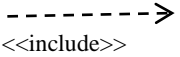
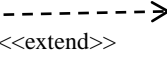
Protected Health Information (PHI) attribute types specified by HIPAA:

1. Names
2. Locations: All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geo-codes, except for the initial 3 digits of a zip code if the correspond area contains more than 20,000 people and the initial three digits of a zip code is changed to 000 if the correspond area contains 20,000 or fewer people.
3. Dates: All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicating such an age.
4. Telephone numbers
5. Fax numbers
6. E-mail addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code



APPENDIX D

User manual for UML based documentation




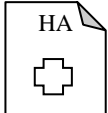

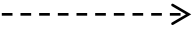
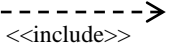
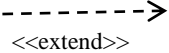
1. Elements in business use-case model

Element Name	Symbol	Description
Business use-case		Business use-case represents what should happen in the business when it is performed and it describes sequence of actions that generate a valuable result to an actor (Podeswa 2005).
Business actor		Business actor represents someone or something external to the business (e.g. in clinical setting it will be patient, supplier, external registry) who interacts with the system to attain desired goals.
Worker		Worker represents someone who is employed in the business (e.g. in a clinical setting it will be physician, nurse, radio therapist).
Association		Association represents the line that link an actor (business actor or worker) to a business use-case. This indicates that the actor interacts (by initiating or conducting) with the business within the use-case.
Business goal		Business goal represents the purpose of the project in business perspective.
Dependency		Dependency represents that some UML elements need or depends on other model elements for specification or implementation. This is shown as a dashed arrow line directing from the dependent at the tail to the contributor at the arrow head.
Include		Include relationship represents links to additional use-cases that depends on the result of the base use-case.
Extend		Extend relationship represents links to additional use-cases that are optional and which are not required to understand the main purpose of the use-case.

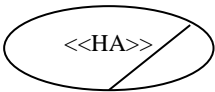

2. Elements in business goal model

Element Name	Symbol	Description
Business use-case		See elements in business use-case model
Business goal		See elements in business use-case model

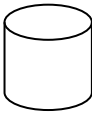
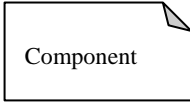
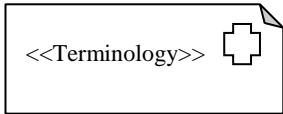
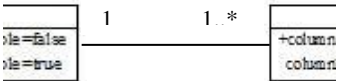
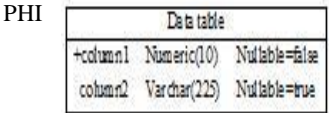
3. Elements in health analytic use-case model


Element Name	Symbol	Description
Health Analytic use-case		Health Analytic use-case represents the output from the HA perspective based on the expectation of the users (what they plan to do with the output). The output will be an interpretation of the results.
Health Analytic goal		Health Analytic goal represents the HA requirements that expected to be achieved by the HA use-case.
Health Analytic actor		Health Analytic actor represents the final user of the knowledge extracted from HA use-cases.
Health Analytic documentation		Health Analytic documentation represents a document composed with the results (list of individual or integrated output) from the HA use-case.
Health Analytic application		Health Analytic application represents a software application developed incorporating knowledge extracted in the HA use-case.
Dependency		See elements in business use-case model
Include		See elements in business use-case model
Extend		See elements in business use-case model

4. Elements in health analytic goal model

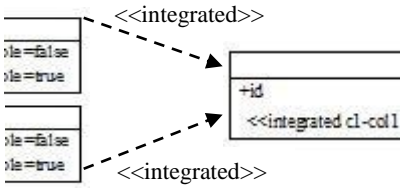
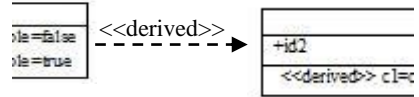
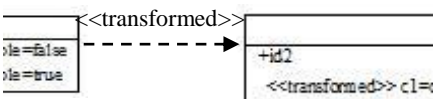
Element Name	Symbol	Description
Health Analytic use-case		See elements in health analytic use-case model
Health Analytic goal		See elements in health analytic use-case model

5. Elements in health analytic data model






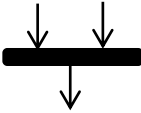
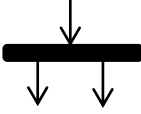
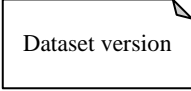
Element Name	Symbol	Description												
Data source	 <<vendor: type: version: location: user: password>>	Data source represents the datasets that are used for data modelling. It contains details of the vendor, type, location of the data source stored, user and password of the data source to be used when accessing it.												
Component		Component represents a physical aspect of elements that are used to describe the data sources and certain data records itself. This includes files, documents used for data modelling and other files relevant to the project.												
Health analytic terminology		HA terminology document represents a glossary of words mapping medical terms and HA terms. This could be two separate documents representing medical terminology and HA terminology.												
Data table	<table border="1" data-bbox="520 1503 963 1648"> <thead> <tr> <th colspan="3">Data table</th> </tr> </thead> <tbody> <tr> <td>+column1</td> <td>Numeric(10)</td> <td>Nullable=false</td> </tr> <tr> <td>column2</td> <td>Varchar(225)</td> <td>Nullable=true</td> </tr> <tr> <td>column3</td> <td>date</td> <td>Nullable=false</td> </tr> </tbody> </table>	Data table			+column1	Numeric(10)	Nullable=false	column2	Varchar(225)	Nullable=true	column3	date	Nullable=false	Data table represents the tables in the data sources. This indicates column name, data type, null, and primary key (by '+').
Data table														
+column1	Numeric(10)	Nullable=false												
column2	Varchar(225)	Nullable=true												
column3	date	Nullable=false												
Data relationship		Data relationship represents the relationship between data tables. This could be 0 to 1, 1 to 1, 0 to many, 1 to many.												
Protected health information		PHI represents the data that is decoded or removed for sensitivity of health data. This includes word 'PHI removed' or 'PHI replaced'. If PHI replaced then the strategy used to remove will be												

		mentioned in the data de-identification report.
Data standard		A note will be given to data table indicating the data standard used (e.g. SNOMED, ICD10).


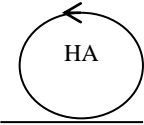
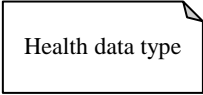
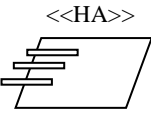
6. Elements in health analytic modified data model

Element Name	Symbol	Description																
Integration		Integration represents how tables are integrated (e.g. to avoid duplication). In data description report transformation details could be provided.																
Derived data		Derived data represents a new data column derived from the original (name of the table will be same). The derived column mentions the derivation formula with the tag 'derived'.																
Transformed data		Transformed data represents a change of format of data from the original (name of the table will be same). The transformed column mentions the tag 'transformed'.																
Modified data table	<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th colspan="4">Table derived</th> </tr> </thead> <tbody> <tr> <td>+c1</td> <td>Numeric(10)</td> <td>Null.=false</td> <td></td> </tr> <tr> <td>c2</td> <td>Varchar(50)</td> <td>Null.=true</td> <td></td> </tr> <tr> <td><<derived>> c3= formula</td> <td>Numeric(10, 2)</td> <td>Null.=false</td> <td></td> </tr> </tbody> </table>	Table derived				+c1	Numeric(10)	Null.=false		c2	Varchar(50)	Null.=true		<<derived>> c3= formula	Numeric(10, 2)	Null.=false		Modified data table represents columns that have been derived or transformed.
Table derived																		
+c1	Numeric(10)	Null.=false																
c2	Varchar(50)	Null.=true																
<<derived>> c3= formula	Numeric(10, 2)	Null.=false																
Generated data	<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th colspan="4">Table</th> </tr> </thead> <tbody> <tr> <td>+c1</td> <td>Numeric(10)</td> <td>Null.=false</td> <td></td> </tr> <tr> <td>c2</td> <td>Varchar(225)</td> <td>Null.=true</td> <td></td> </tr> <tr> <td><<generated>> c3</td> <td>Numeric(10)</td> <td>Null.=false</td> <td></td> </tr> </tbody> </table>	Table				+c1	Numeric(10)	Null.=false		c2	Varchar(225)	Null.=true		<<generated>> c3	Numeric(10)	Null.=false		Generated data represents new columns created in the data table. The generated column mentions the tag 'generated'.
Table																		
+c1	Numeric(10)	Null.=false																
c2	Varchar(225)	Null.=true																
<<generated>> c3	Numeric(10)	Null.=false																

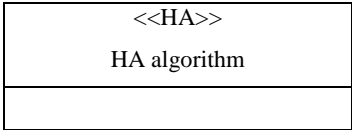
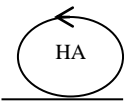
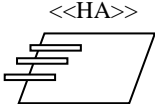
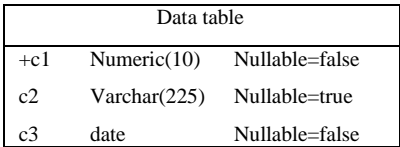
7. Elements of data preparation activity diagram

Element Name	Symbol	Description
Activity		Activity indicates the behaviour. It represents the data pre-processing tasks performed on the dataset for its modification.
Start		Start indicates the beginning of a process.
End		End indicates the end of a process. It represents the completion of all the flows in an activity.
Decision		Decision indicates the branching or merging of different flows.
Connector		Connector indicates the directional flow of the activities. End of an activity and start of an activity is connected by the arrowed line.
Join		Join indicates the merging of two concurrent activities and bringing them back to the single flow activity. Join is represented by the thick horizontal line. In HA context this is used when merging two datasets.
Fork		Fork indicates the brunching of a single activity flow to two or more concurrent activities, If the actions are performed on two datasets concurrently fork is used.
Dataset version		Dataset version indicates the version of the original and the processed dataset used in the data pre-process. There will be at least two dataset versions (one at the beginning and one at the end of the process).

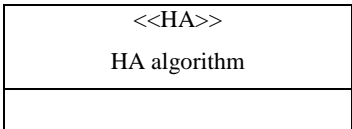
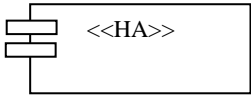
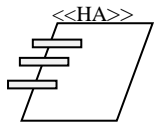
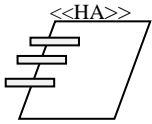
8. Elements in health analytic technique model

Element Name	Symbol	Description												
Health analytic technique		HA technique represents the technique used to model the data. If it is an ensemble technique then <<ensemble>> is indicated below the symbol.												
Health analytic technique type		HA technique type indicates the generalization of the HA technique. That it represents descriptive, predictive and prescriptive.												
Health data type		A note will be given to indicate the data type (e.g. clinical, public and personal data).												
Health analytic tool		HA tool represents any tool that is being used perform the HA technique. HA tool will depend on the HA technique.												
Data table	<table border="1" data-bbox="520 967 906 1111"> <thead> <tr> <th colspan="3">Data table</th> </tr> </thead> <tbody> <tr> <td>+c1</td> <td>Numeric(10)</td> <td>Nullable=false</td> </tr> <tr> <td>c2</td> <td>Varchar(225)</td> <td>Nullable=true</td> </tr> <tr> <td>c3</td> <td>date</td> <td>Nullable=false</td> </tr> </tbody> </table>	Data table			+c1	Numeric(10)	Nullable=false	c2	Varchar(225)	Nullable=true	c3	date	Nullable=false	See Elements in health analytic data model
Data table														
+c1	Numeric(10)	Nullable=false												
c2	Varchar(225)	Nullable=true												
c3	date	Nullable=false												

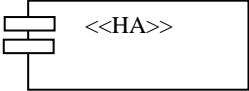
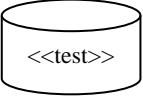
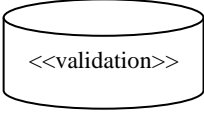

9. Elements in health analytic algorithm model

Element Name	Symbol	Description
Health analytic algorithm		HA algorithm represents the algorithm used based on the technique.
Health analytic technique		See Elements in health analytic technique model
Health analytic tool		See Elements in health analytic technique model
Data table		See Elements in health analytic data model

10. Elements in health analytic-model model

Element Name	Symbol	Description
Health analytic algorithm		See Elements in health analytic algorithm model
Health analytic model		HA model represents the output from modelling.
Health analytic tool workspace file		HA tool workspace file represents the workspace of the HA modelling in tool (tasks, inputs and outputs in the tool) and the saved location of it.
Health analytic tool model file		HA tool model file represents the model created by the tool and the saved location.

11. Elements in health analytic test model

Element Name	Symbol	Description
Health analytic model		See Elements in health analytic-model model
Test dataset	 test_dataset<<vendor: type: version: location: user: password>>	Test dataset represents the data set used to test the model. See Elements in health analytic data model
validation dataset	 validation_dataset<<vendor: type: version: location: user: password>>	Validation dataset represents the data set used to validate the model after testing. See Elements in health analytic data model.
Health analytic tool model result file		HA tool model result file represents the accuracy results of the created model after validating with a new dataset and the saved location of the output (knowledge extracted).

12. Elements in health analytic deployment diagram



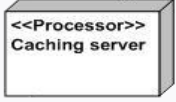



Element Name	Symbol	Element Name	Symbol
Server		User	
Caching sever		Modem	
Sensors		Medical devices	

Illustration of the application of the UML diagrams in USAM

Problem Description:

The UML diagrams are applied into the project performed for *Hospital X*. As a summary, to improve the productivity of machine (linear accelerators) utilization of the radiotherapy department of *Hospital X*, by predicting the duration needed for each radiotherapy treatment.

The UML diagrams relevant to this step are given below.

Step 2: Domain Understanding

Business use-case diagram

Figure D. 1, illustrates the requirements (use-cases) as well as elements involved outside the institute (actors). The business actors are patients (who will be undergoing the radiotherapy treatment), radiologist (who will operate the radiotherapy equipment and provide the treatment), data analyst (who will analyse what is happening in the operations and identify actions that need to be carried out to resolve any issues) and senior management of the institute (who will be interested in the performance/through put of the institute). Several business use-cases can be identified as core business cases and support business cases. The core business cases are dealing with the main task of the institute. The latter deals with other supporting activities related to the main task at hand.

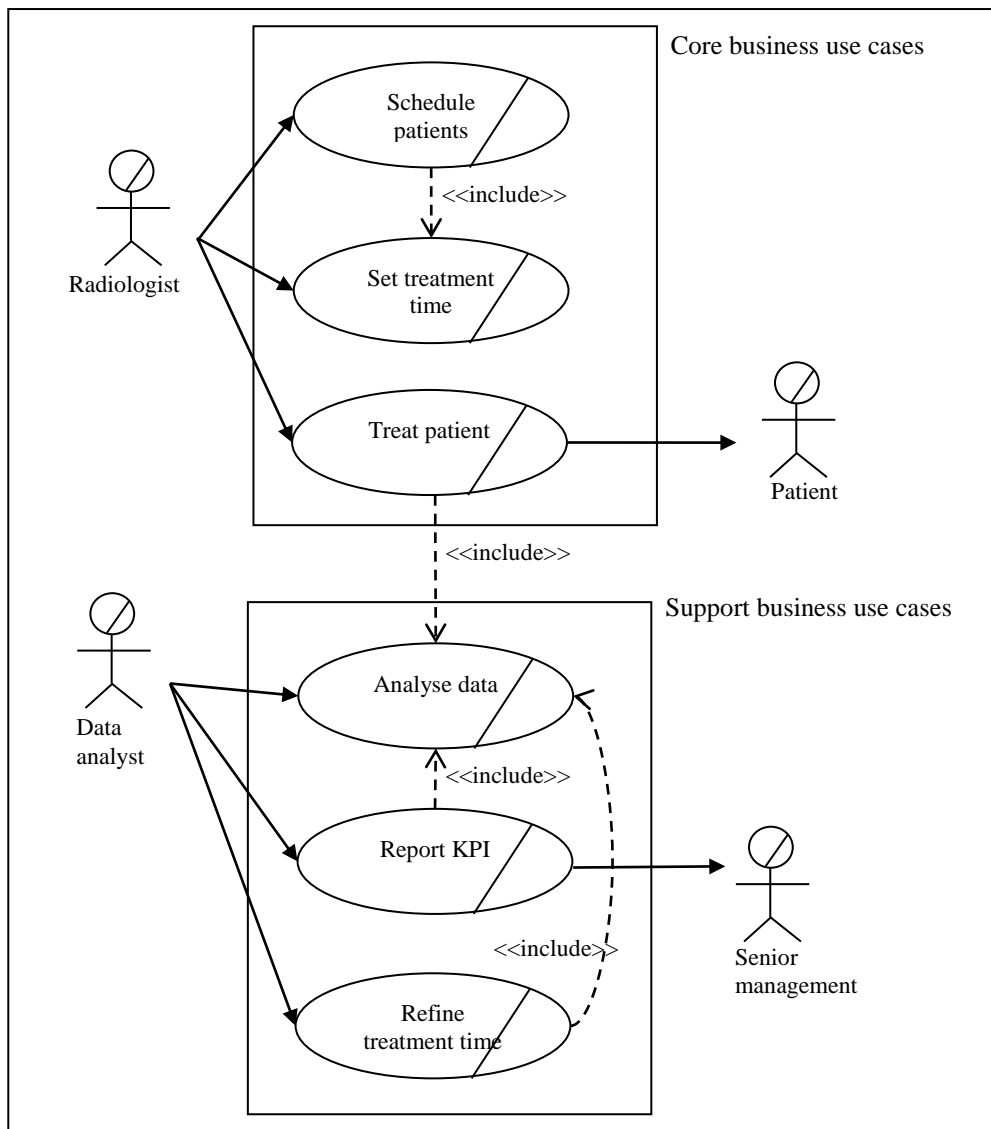


Figure D. 1: Business use-case diagram

Core business use-cases are:

- “Set treatment time”, this includes the action representing the setting time taken for each therapy. When a doctor prescribes a patient to undergo radiotherapy, the radiologist will decide the time taken for each radiotherapy treatment based on the complexity of the patient’s tumour (based on a matrix).
- “Schedule patients”, this represents the actions taken to schedule all the patients in a waiting list to relevant radiotherapy room. Thus, based on the assigned

treatment time for each patient, they will be given various appointment times to come for a treatment.

- “Treat patient”, use-case deals with the action related to carrying out the therapy. Here, the setup time and the treatment time will be clocked and other treatment technique related information will be noted down as well.

Support business use-cases:

- “Analyse data”, this includes the action carried out to analyse the stored treatment data to make right decisions in line with organization objectives.
- “Report KPI”, this includes calculating the key performance indicator (KPI) to measure the success of conducting radiotherapy treatment. This could be represented as number of fields treated per unit time (Delaney et al. 1997b) and will be done after the data analysis.
- “Refine treatment time matrix”, this represents the action of refining the assignment of treatment time based on the suggestions of the data analyst after carrying out the data analysis.

In Figure D. 1, it could be noted that there are dependencies (“*include*”) among use-cases. The “*include*” dependency between the assign patient use-case and the set treatment time indicates that once former is completed, the latter will also be executed as a result. The treat patient use-case includes the analysis of data. That is once base use-case treat patient is completed, analyse data will also be executed. Similarly, report performance KPI and refine treatment use-case always includes analyse data use-case.

Business goal diagram

As illustrated in Figure D. 2, business goals are linked with business use-cases defined previously in Figure D. 1. Thus, each use-case will be connected to at least one goal. The main goal as shown in Figure D. 2 is “improving quality” of the treatment services.

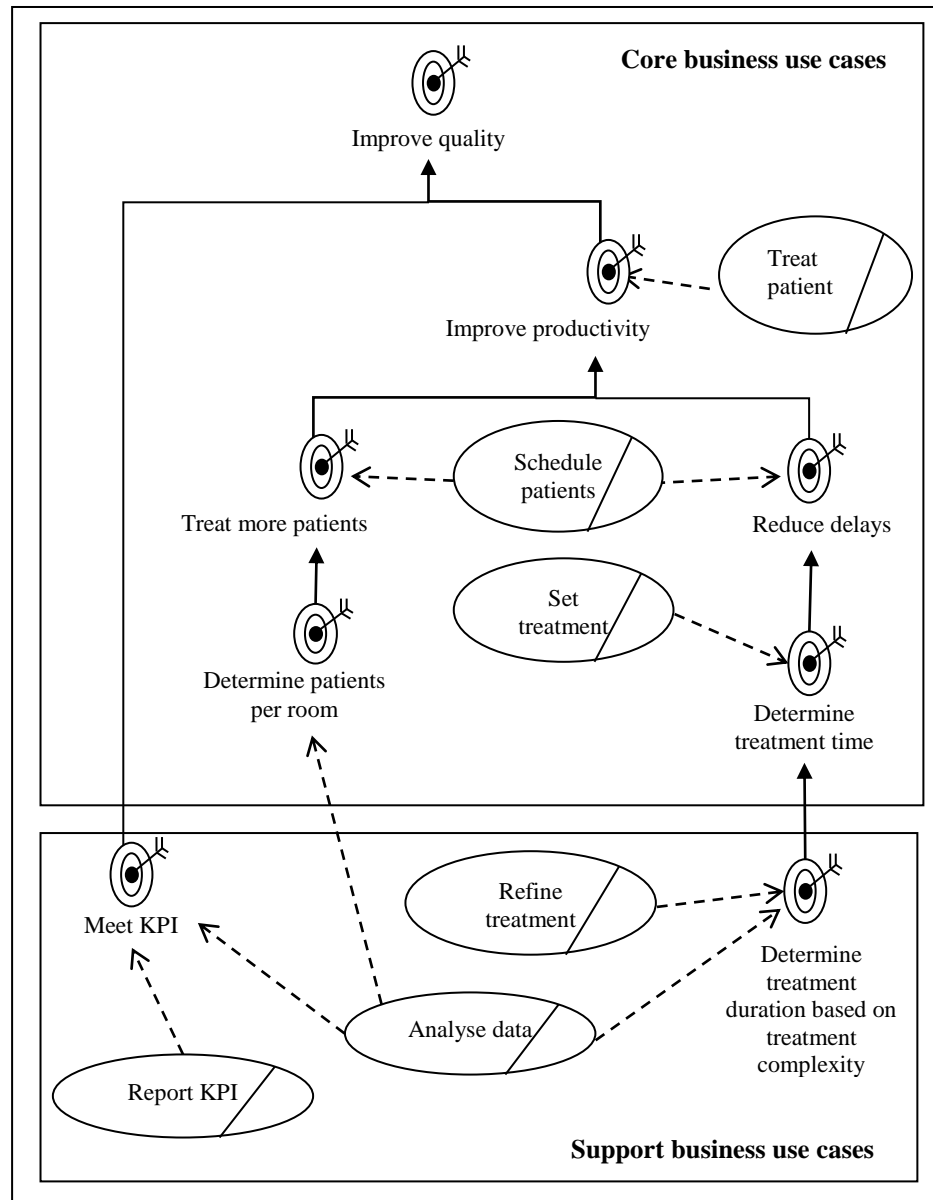


Figure D. 2: Business goal diagram

As illustrated, it is specialized into two goals as “improve productivity” of the radiotherapy equipment (that is increase number fields treated per unit time) and meet

success criteria or “meet KPI” (that is achieve KPIs set by the management). “Reduce delay” business goal is to indicate reduce waiting time as incorrect assignment of treatment time leads to delays in treatment of other patients. Therefore, generalizations of goals are indicated in the diagram below. The “determine treatment time” goal indicates when to start the treatment. This could be further specialized into sub goal “determine treatment duration based on the treatment complexity”. As the treatment duration varies based on the tumour size and treatment fraction, it is important to identify the treatment duration based on those factors (represent treatment complexity) rather than setting fixed durations for each treatment.

As depicted in the business goal diagram, business use-case depends on a business goal. For example, the business use-case “schedule patients” is associated with the “treat more patients” and “reduce delays” goals. Therefore, it is associated with two goals. However, “report KPI” use-case is associated with only one goal that is “meet KPI”. Even though, there are many sub goals, to ensure the simplicity in the illustrations, we have illustrated only the important goals here.

As could be seen from the illustrations, the domain could be understood easily when the UML diagrams are developed systematically representing the complete business process related to the problem. Business goals could be linked with use-cases only after their identification. Here, the use-cases will be in line with goals.

Analytic use-case diagram

This analytic use-case model (Figure D. 3) is developed based on the business use-case diagram and the business goal diagram in Figure D. 1 and Figure D. 2. Each business goal is linked to a HA use-case. It is necessary to consider whether a business goal linked to a business use-case can be directly evaluated in business terms. If not, it is necessary to consider a generalized business use-case appearing one

level up as well. This is illustrated in Figure D. 3 where “determine patients per room” goal is considered along with the “treat more patients” goal.

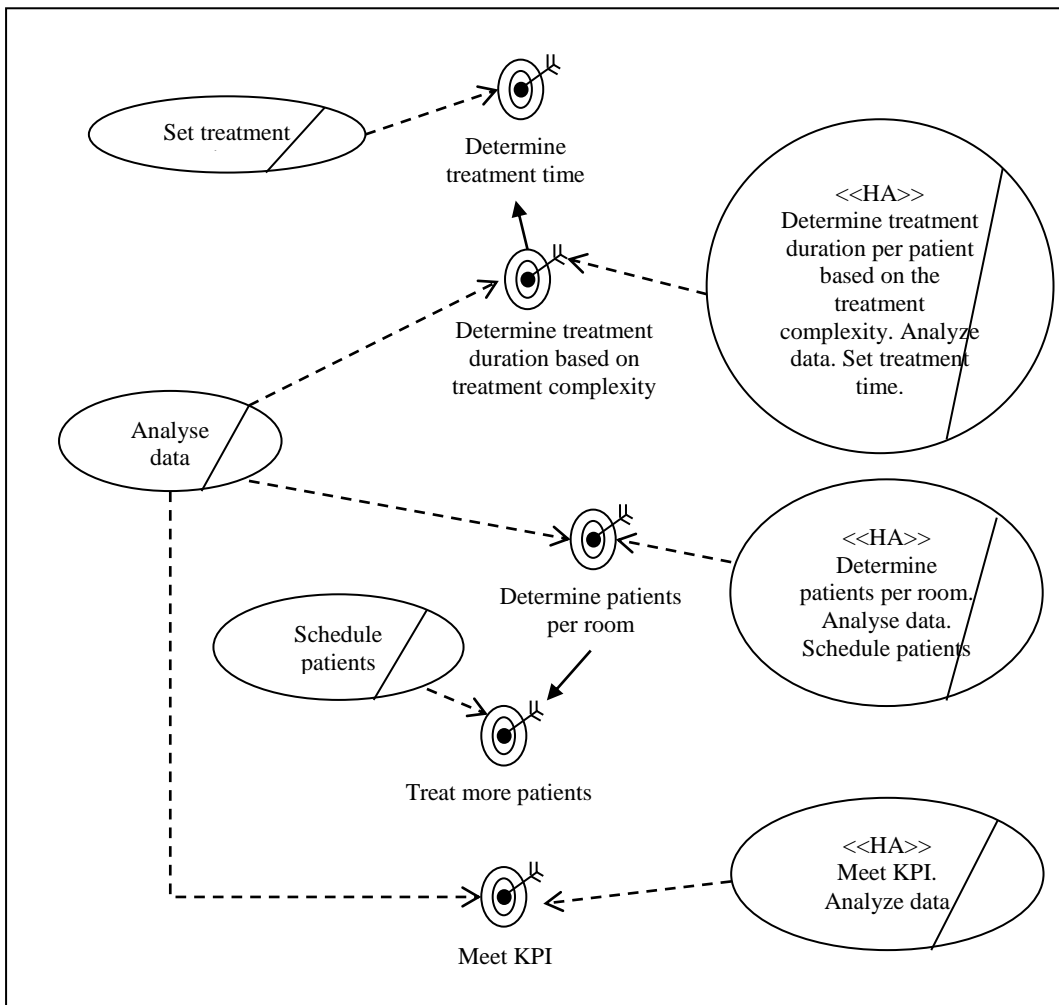


Figure D. 3: Analytics use-case diagram

There are several HA use-cases. First, “determine treatment duration per patient based on the treatment complexity. Analyse data. Set treatment time”. This HA use-case is to decide the total time taken specifically to carry out the treatment based on the required treatment time. This will depend on the type of the equipment used and the type of treatment activity to be carried out depending on the severity of the disease condition.

Second, “determine patients per room” HA use-case deals with identifying the number of patients that could be allocated per room per day. Thus, this will allow the

institute to serve maximum number of patients. It is therefore, important to analyse the available data to decide on the number of patients that could be accommodated in a day and to schedule patients in the waiting list correctly to achieve the business goal.

Third, it is important to analyse the data based on the type of the radiotherapy equipment to meet the expected KPI levels.

Analytic goal diagram

The analytic goal diagram relevant for the current scenario is illustrated in Figure D. 4 indicating the relationship between HA use-cases and HA goals. HA goals are defined for each HA use-case defined in Figure D. 3. Moreover, HA actors are included to depict the users who will be using the data collected from the HA use-cases. As can be seen here, the HA actor is the business analyst. For example, “determine treatment duration per patient based on the treatment complexity. Analyse data. Set treatment time”. This use-case has three HA goals. First is to determine the patient treatment profile as the treatment complexities can vary with the patient’s condition (“identify patient treatment profile”). Second, it is important to determine whether there is an association between the treatment type and patient’s condition (“identify factors influencing treatment duration”). Finally, it is necessary to forecast the treatment duration based on the treatment complexity (“predict treatment duration”).

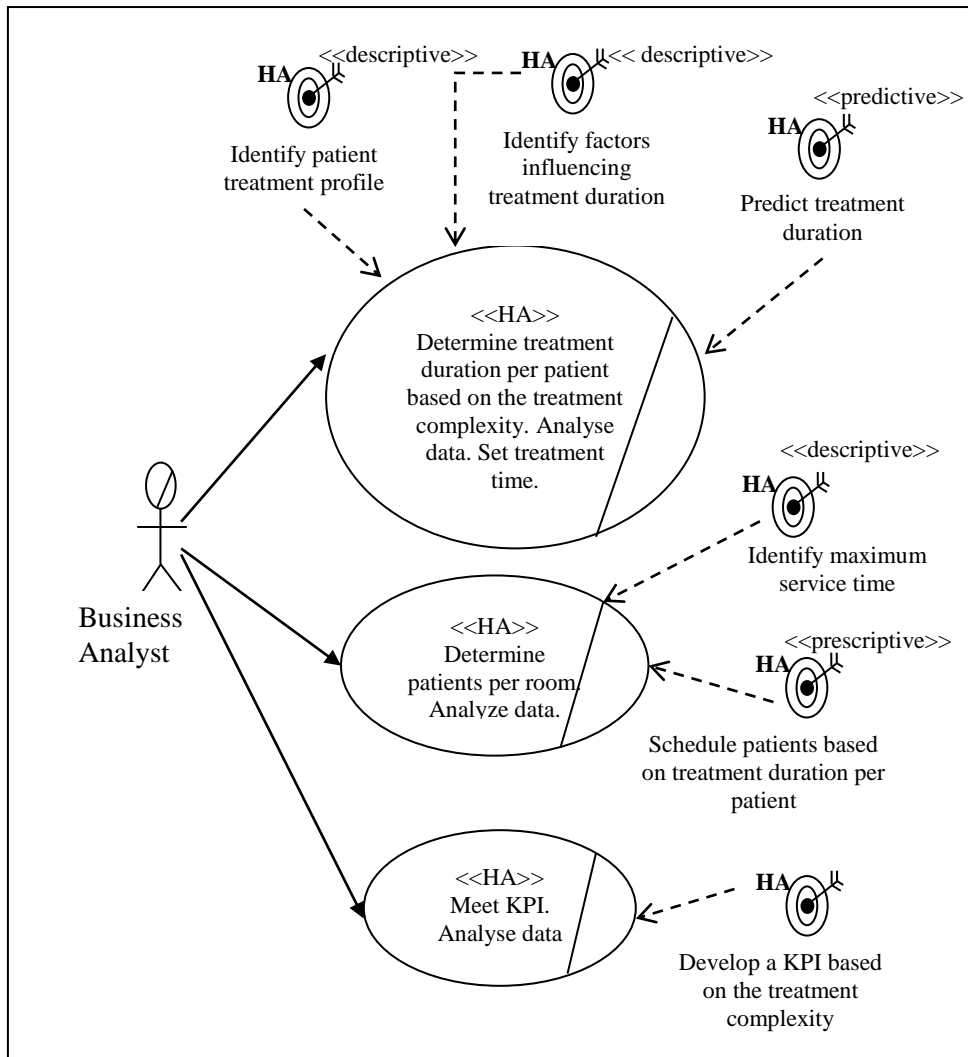


Figure D. 4: Analytic goal diagram

As can be seen in the analytic goal diagram, there are three stereotypes defined to indicate the type of the HA goal. They are <<descriptive>> to represent descriptive analytics, <<predictive>> to represent predictive analytics and <<prescriptive>> to represent prescriptive analytics. “Identify patient treatment profile” is considered under descriptive analytics as it includes summarization and describing characteristics of data. Similarly, the “identify factors influencing treatment duration” is considered under descriptive analytics as it deals with the association among the dependent variable and the independent variables. The HA goal “predict treatment duration” is considered under predictive analytics. The HA

goal “schedule patients based on treatment duration per patient” is an optimization problem and as such it is considered under prescriptive analytics type.

At this stage, we will be able to determine the business goals and the relevant HA goals and the modelling will be carried out based on these use-cases. However, a clear understanding of the data is vital for fine tuning the analytic goal diagram and HA uses cases.

Step 3: Data Understanding

Data diagram

The data diagram is shown in Figure D. 5.

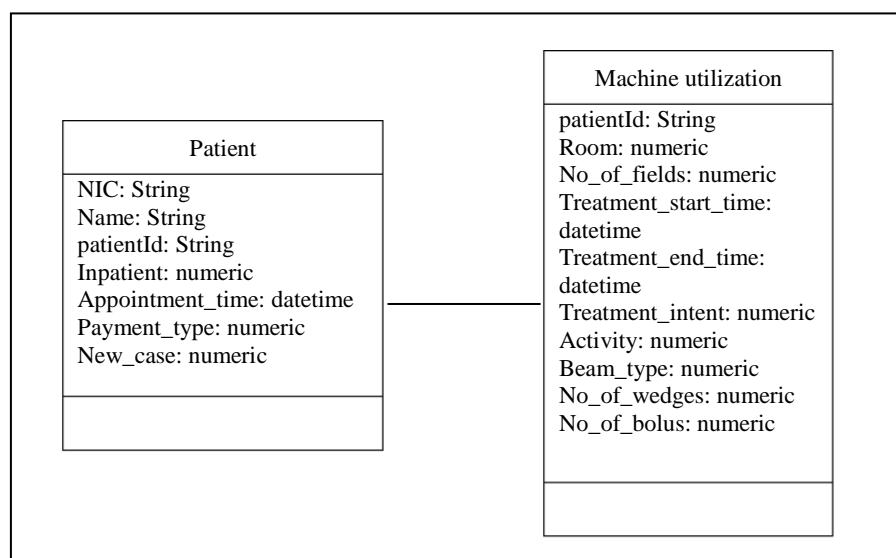


Figure D. 5: Data diagram

Data component model demonstrates the connection between data sources (e.g. image files, flat files, case notes) used for the analysis. However, since the data is received as an excel file, a data component model was not developed. The data component model is very useful when dealing with heterogeneous data sources and when certain data needs to be extracted from those sources (extract content from text documents) as the programs and other steps carried out in data extraction are documented.

Step 5: Data Preparation

Modified data model

During data preparation, even though a modified data diagram can be designed, the UML diagram for this dataset will not be designed since we do not have different tables. As the dataset is in an excel sheet, all the records are in one single sheet after the data integration. However, data processing on the acquired dataset is necessary and some of the derivations and transformations are indicated in Table D. 1.

Table D. 1: Modified dataset

Treatment<<integrated: Machine utilization + patient>>	
Variable	Description
<<derived>> tx_duration = tx_end - tx_start	The difference of start time and end time of treatment is computed
<<transformed>> activity	Group in as IGRT, IMRT, VMET, others and BTE
<<derived>> wedges_count = (if wdg_appl_yesNno = no => 0)	The columns with umber of wedges were kept blank in the dataset if there are no wedges. Thus, if wedges variable is No, then the no. of wedges column is filled with zero.
<<derived>> bolus_count = (if bolus_yesNno = no => 0)	The columns with number of bolus were kept blank in the dataset if there is no bolus. Thus, if bolus variable is No, then the no. of bolus column is filled with zero.
<<PHI>> patient_code	Replaced with a new code. All the unique values are replaced with 'S' and a 6 digit numeric value. E.g. S000001, S000005, S001758

Data preparation activity diagram

The activity diagram is used to indicate the flow of the data preparation task to regenerate the modified dataset from the initial dataset. Figure D. 6 illustrates how the dataset v1 is converted into dataset v2. Branching is indicated for activities that could be carried out concurrently.

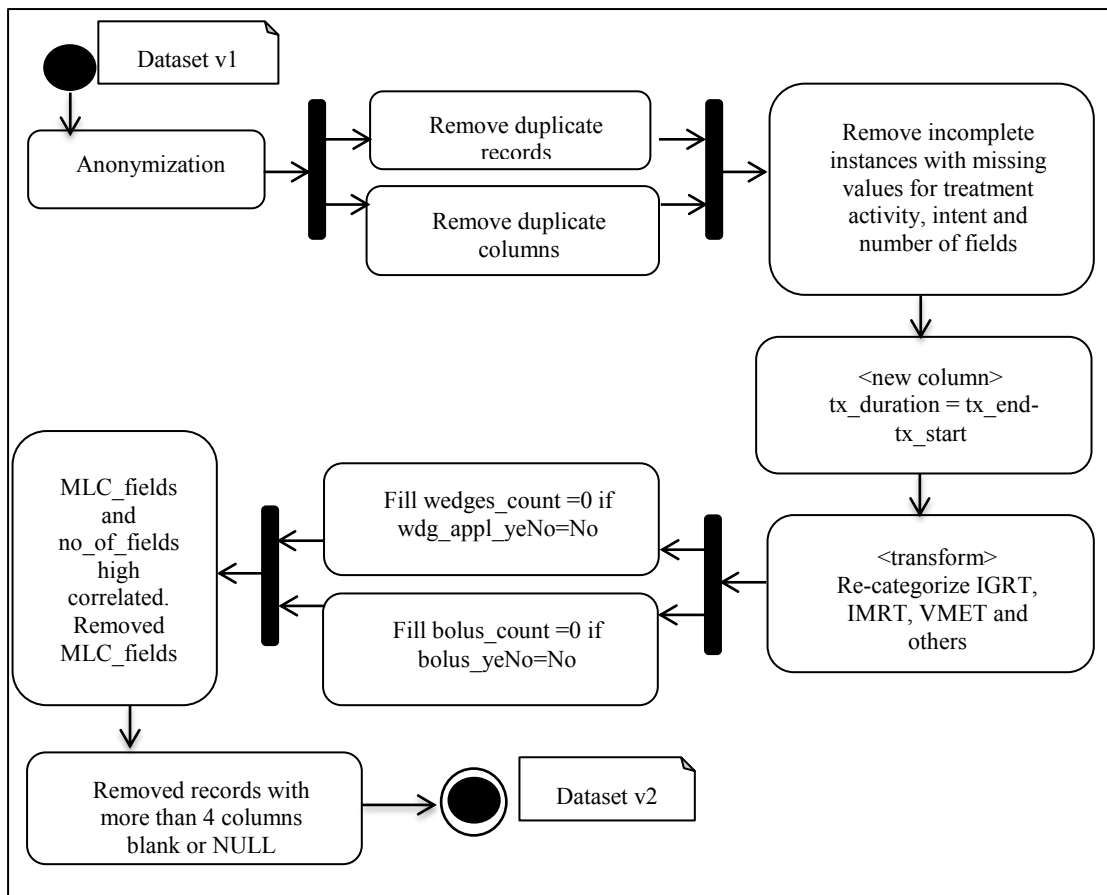


Figure D. 6: Data preparation activity diagram

Step 6: Data Modelling

In data modelling stage there are three UML diagrams, namely, technique diagram, algorithm diagram and analytic model diagram. At this stage various HA techniques are selected with corresponding algorithms and parameters to determine the optimal result.

A. Identify patient treatment profile

Technique diagram

The technique diagram demonstrates how the techniques are used to achieve HA goals. As shown in Figure D. 7, mean is used as the technique to get a

summarization of the treatment profile based on treatment intent and activity. Also the version of the dataset used for the modelling is documented.

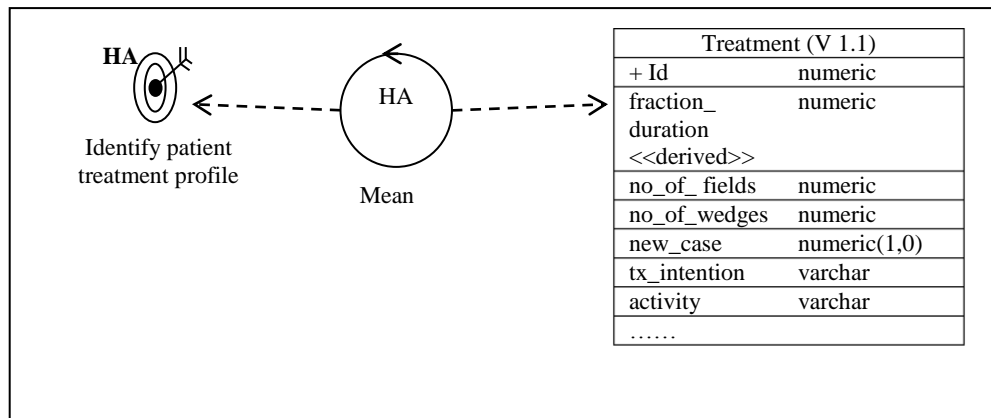


Figure D. 7: Technique diagram for identify patient treatment profile

No algorithm is used and as such algorithm diagram is bypassed in descriptive analytics. Similarly, analytic model diagram is not drawn for descriptive analytics. Only the results will be stored in the documentation with the interpretations (Table 5). Mean values obtained indicates that, the patient undergoing first fraction takes more average time compared to other fractions. Also it could be seen that certain technologies are not used for certain treatments.

B. Identify factors influencing treatment duration

Technique diagram, algorithm diagram and analytic model diagram are given below.

Technique diagram

As shown in Figure D. 8, we used attribute evaluation techniques to select the most influential variables.

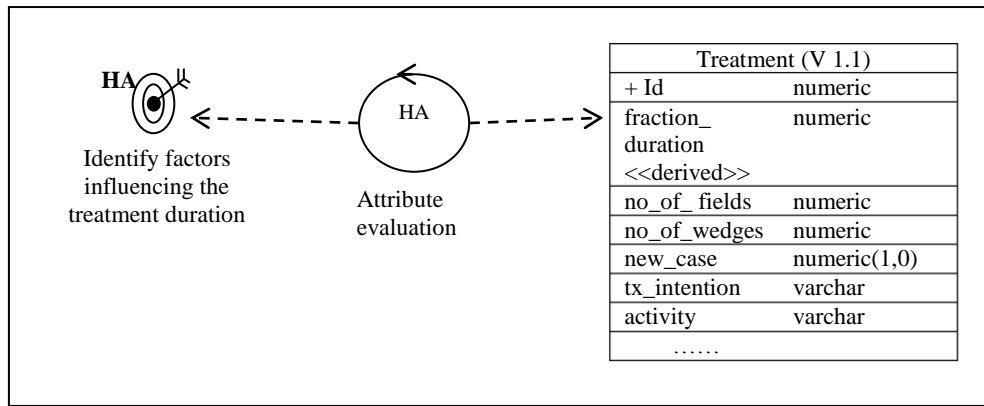


Figure D. 8: Technique diagram for identify factors influencing treatment duration

Algorithm diagram

This is used to represent the HA algorithm selected based on the HA technique. The HA model will be created considering the tool used and the parameters considered. As in Figure D. 9, ordinary least square (OLS) will be used as the HA technique to determine the effect of each factor on fraction duration (attribute evaluation).

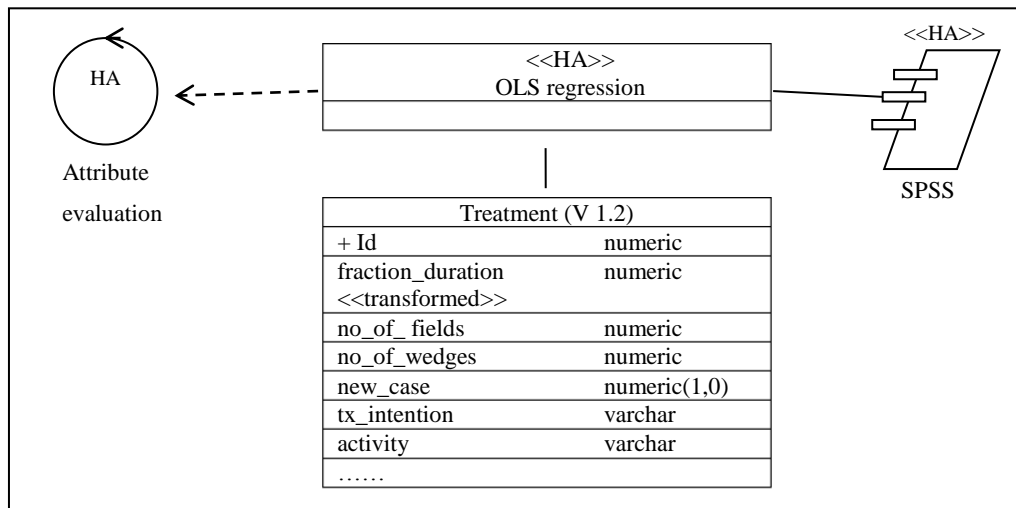


Figure D. 9: Algorithm diagram – OLS regression

Analytic model diagram

As shown in Figure D. 10, the results and the workspace are saved and the file locations are indicated in the HA-model diagram. For OLS regression, we did not

have any specific parameter selection as in GEE regression model (Figure D. 13) and as such they are not mentioned.

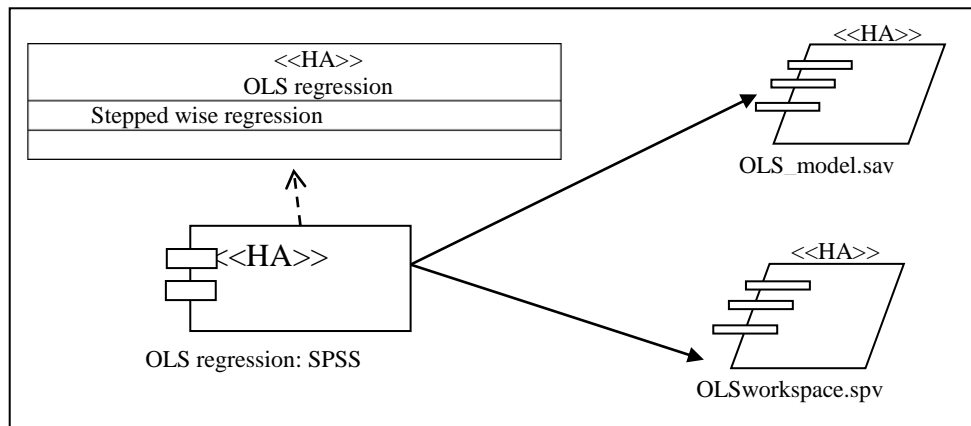


Figure D. 10: Analytic model diagram for identifying factors influencing treatment duration

A. Predict treatment duration

The relevant technique diagram, algorithm diagram and the Analytic model diagram are given below.

Technique diagram

As shown in Figure D. 11, regression is used as the technique for prediction.

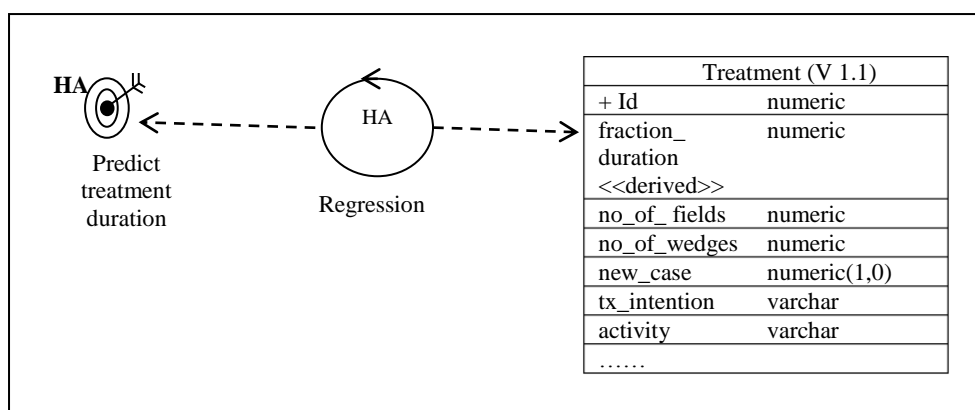


Figure D. 11: Technique diagram – Predict treatment duration using regression model

The same dataset, with certain transformations will be used in decision trees (Figure D. 12).

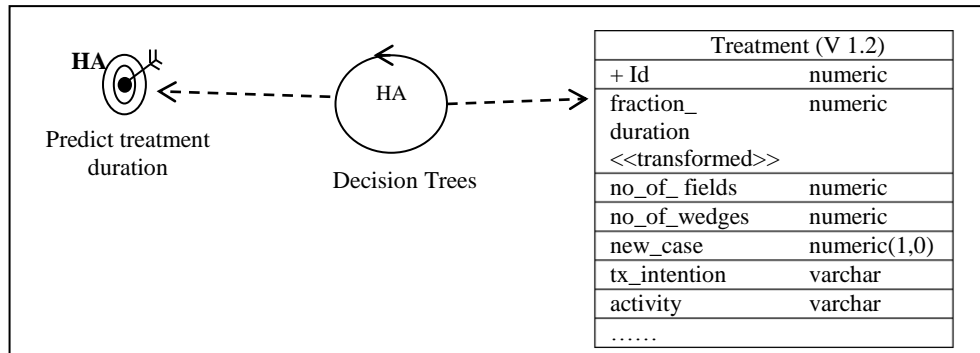


Figure D. 12: Technique diagram – Predict treatment duration using decision trees

Algorithm diagram

As shown in Figure D. 13 we will use generalized estimation equation (GEE) regression model as the algorithm for the prediction technique.

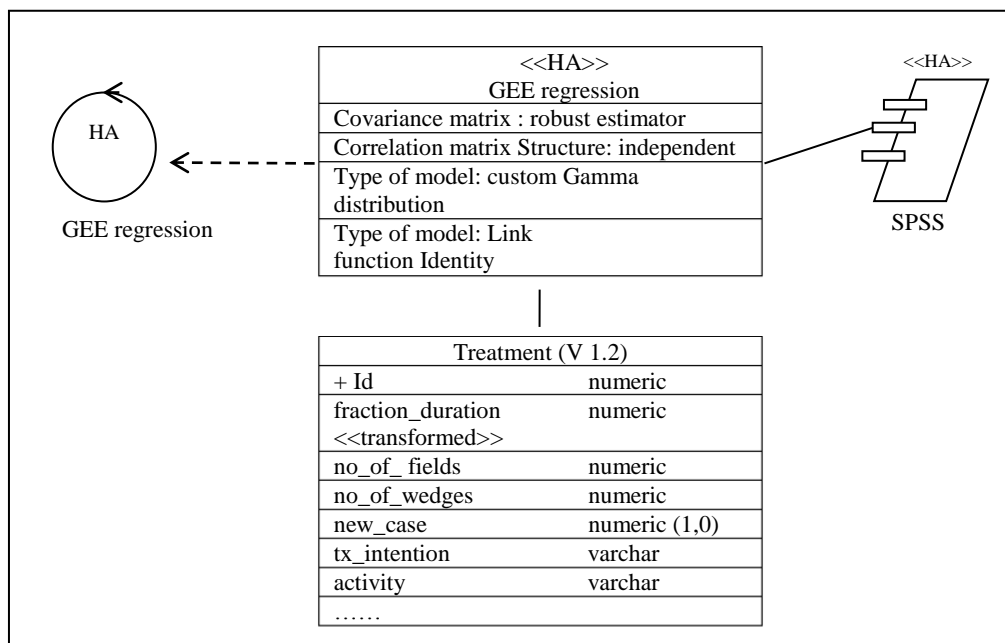


Figure D. 13: Algorithm diagram – GEE regression

Analytic model diagram

The Analytic model diagram in Figure D. 14 indicates the model, HA workspace file (saved) and HA model file (saved). The file names are given with the saved file location of them.

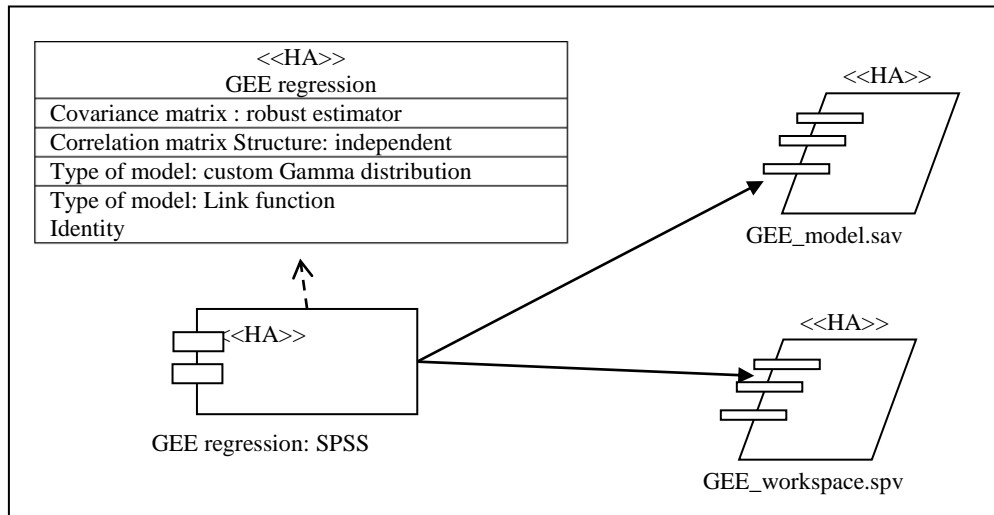


Figure D. 14: Analytic model diagram – GEE regression

The GEE regression model will be saved in GEE model.sav file and location of the file is indicated in the diagram. Then the workspace file location also will be recorded as whenever required one can easily access the previous workspaces.

Step 8: Validation

Validation Diagram

As shown in Figure D. 15, the HA validation diagram will have two datasets.

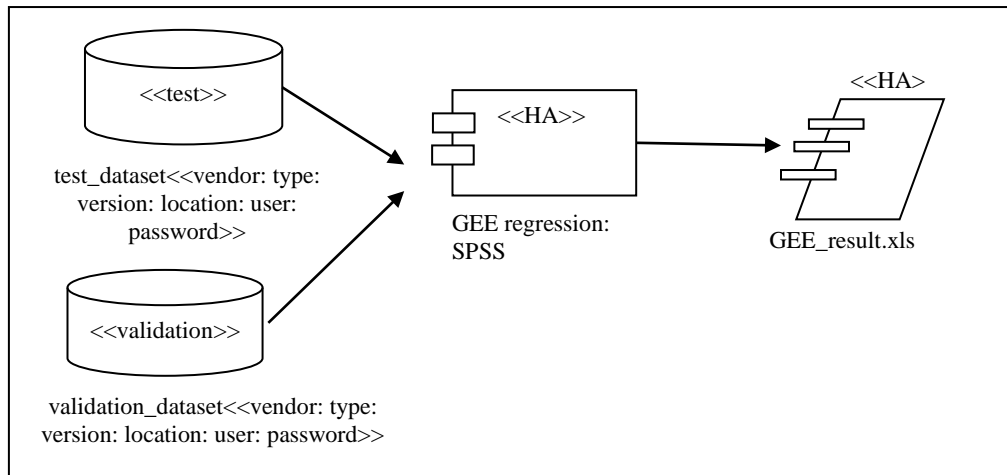


Figure D. 15: Validation diagram