DESIGNING A PROCESS MODEL FOR HEALTH ANALYTIC PROJECTS

Supunmali Ahangama, Department of Information Systems, School of Computing, National University of Singapore, Singapore, supunmali@comp.nus.edu.sg

Danny Chiang Choon Poo, Department of Information Systems, School of Computing, National University of Singapore, Singapore, dannypoo@nus.edu.sg

Abstract

The health analytic (HA) process model presented in this paper is developed to conduct HA projects in an orderly manner instead of approaching them in an ad-hoc manner. While a well-documented process model will avoid overdependence on the experience of an analyst it will also assist novice users in learning the best practices to analytics very easily. The existing process models for data mining have failed to meet user requirements and thus have been unsuccessful in diffusing among the data analysts. This study aims to propose a refined HA process model to guide novice users. The process model is developed using design science research approach followed by an ex-ante evaluation. Design criteria to develop the model were based on the challenges in HA identified through literature and by utilizing a data analytic case. The process model is developed using the necessary theoretical foundation, best practices in software engineering and data mining process models along with design science research approach. The theoretical and practical implications of the model too are provided in this paper.

Keywords: process model, health analytics, USAM, design science research.

1 INTRODUCTION

Usage of a unified methodology will improve the process and output of health analytics (HA). Such a process model will facilitate performance of all the inclusive projects as a structured process by dividing a complex process of analytics 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). The unified structured HA process model proposed in this paper 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. Furthermore, it is supposed that the absence of a framework or a methodology is a limitation to the advancement of a field (Dzeroski 2007).

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 they do 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 such the proposed unified structured process model will be developed focusing on healthcare domain.

In considering the above mentioned issues, this study will be carried out to address the research question: What methodological steps are needed to be followed by a novice user in health analytics? User requirements of a HA process was examined using Design Science Research (DSR) approach (Hevner et al. 2004) prior to the development of this process model. The proposed HA process model itself will be a contribution of this study in addition to the methodology used to develop the process model.

In the subsequent sections, the literature related to the HA process model development will be discussed followed by the methodology used to develop the model with an explanation of the design artefact and the research approach used. The problems identified through literature and by working as an external analysts in a HA project and design criteria used to solve those problems will be discussed in proceeding sections along with an explanation of the developed HA process model. A general discussion and conclusion on the study too will be presented at the end of this paper.

2 LITERATURE REVIEW

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). Based on the comparison carried out on history of data mining against software engineering, Marban et al. (2009b), had shown the parallelism between the two and had 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. 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). Various process models are being proposed, to elude these issues and to facilitate a standardized approach in performing data mining studies.

Several process models like 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 developed to allow a standardized approach to data mining projects. It is important to note that data mining and data analytics are commonly used interchangeably in the literature though there are variations in the two terms. As such, we too do not differentiate the two terms in this paper.

2.1 Data Mining Models

According to KdNuggets Polls (KdNuggets.com 2014) CRISP-DM and SEMMA can be considered as two popular approaches to data mining. SEMMA used in the SAS Enterprise Miner Software package (SAS 2008) focusses only on data manipulation and data modelling components in data analytics and it fails to reflect requirement gathering, design and implementation stages. Furthermore, when considering a holistic process model it is important to consider about project management and documentation too (Marban et al., 2009).

Compared to SEMMA, CRISP-DM proposes a holistic process model for data analytics and as such it is considered as the de-facto standard for data mining (Mariscal et al. 2010). However, there are several limitations in CRISP-DM compared to other engineering based process models (for e.g. software engineering process models). The limitations are (1) collecting data at the launch of the project and inability to accommodate new data at later stages (when the project progress and with better exposure to the project requirements new data will be required in real scenarios) (Jacobson et al. 1999); (2) lack of specifications on how to carry out analytics (no guidelines and specific steps given on how to commence a HA project) (Marban et al. 2009a); and (3) omission of project management and documentation components. As such in developing a HA process model it is important to consider these limitations.

2.2 HA Frameworks

HA can borrow best practices from c process models to develop a unified and structured process model. Cios and Moore (2002) and Eggebraaten et al. (2007) proposed a data mining knowledge discovery (DMKD) process for medical applications as an extension to CRISP-DM considering the uniqueness of medical data mining (Cios and Moore 2002). 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 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) had focused mainly on introducing a consistent nomenclature using XML. However, that process does not provide sufficient details of each stage. Though they have mentioned about proposed extensions to CRISP-DM, no distinguishable extensions could be identified. Most importantly, even though they had mentioned about the uniqueness of medical data mining they had failed to incorporate any such specific components into their process model.

In a case study carried out by Catley et al. (2009), they 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. They had proposed a new CRISP-DM model named as CRISP-TDM considering temporal data mining (TDM) and had 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 having 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).

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 engineering based process model and it considers only the documentations. Furthermore, the proposed documentary strategy lacks proper methodological steps (inexplicit) 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).

In most of the HA studies found in published literature, individual approaches of the scientists have been used instead of following a standardized approach. Thus it is hard to manage and repeat successful project steps or identify mistakes in certain steps proposed in these studies and it is hard to translate the findings to specific actionable steps. Furthermore, failure to use a proper research methodology in developing the process model is noted in most of these studies in spite of the many benefits to be gained from using a methodology. Thus the provision of a complete and structured process model supporting their specific needs will be a great assistance to the novice users.

3 METHODOLOGY

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). 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 paper is the DSR approach (Hevner et al. 2004; Pries-Heje and Baskerville 2008).

According to the DSR knowledge contribution framework proposed by Gregor and Hevner (2013), this study 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).

3.1 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 aims 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 uses process management, project management, knowledge management and communication management as focusing constructs (or as the dimensions of the proposed method). It conceptualizes 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 (components and phases in the process model) through this thesis will be presented in Chapter 6.

3.2 Research Method

The research method used in this thesis is illustrated in Figure 1. 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 1 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), having, (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 1, and demonstration and evaluation will be covered by the evaluation in Figure 1. March and Smith (1995) state that "build" and "evaluate" are the two phases in DSR. They are represented by first three phases in Figure 1.

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 1. A design science research method applied to HA process model development

Right hand side of Figure 1 illustrates how the DSR methodology is applied in this paper.

The first step - problem awareness has already being addressed in Chapter 1 where the problems are being defined by research question as What methodological steps are needed to be followed by a novice user in health analytics? A survey was carried out with the aim of understanding the novice user's intention to use a methodology for analytics. Through the survey, it is identified that perceived relative advantage (over other methods), perceived result demonstrability (visibility of the usage and its outcome) and perceived usefulness of knowledge management are important factors affecting the usage intention (Ahangama and Poo 2015).

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 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 is 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 paper is the analytical data model development 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 of evaluation- observational, comprises case studies and field studies. This study uses an 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.

4 DESIGN CRITERIA

Initially, literature was reviewed to understand the existing approaches to understand how the artefact should be implemented. It was 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 feedbacks for further improvement (formative evaluation). The first evaluation is carried out 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).

4.1 Challenges

As an external analyst, one of the authors of this paper worked on a HA project in a hospital. Several shortcomings of existing process models were identified. First, the actual process of data analytics cannot be carried out as a linear process where the project requirements are complex and ambiguous. Specially, new data requirements may arise while carrying out the project and there should be a mechanism to accommodate them.

Second, the communication with the stakeholders is very important in carrying out a HA project. As the data analysts are dealing with projects that are different from their domain knowledge of

understanding, continuous communication with them is important. A project ream 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) and consumers (use the outputs generated by the doers) (Collier 2011). The planners and consumers may not know what they exactly require at the beginning.

Third, maintaining the privacy of medical data is on uniqueness of medical data mining to be considered. There are several standards such as HIPAA (Health Insurance Portability and Accountability Act) to maintain confidentially and security of patient data. As such, it is important to consider how it can be incorporated into the HA process model.

Fourth, the data analysts may be dealing with multiple sets of different versions of the same file and as such they find it hard to keep track of the exact file that should be used at different stages. For example, there will be different versions of scripts, data files and documents and data analysts will find it hard to determine the exact version to revert back to in future references. Specially, if there is an error in the current version, there should be a way to move to an earlier correct version.

Finally, the existing data mining related models do not have a conceptualization stage. It is important to identify the constructs to be used and relationships to explore as it is not advisable to use a "kitchen sink models" where all the possible data are tossed into the data model.

Most of these challenges mentioned in literature were identified while working as an external analyst in the HA project as well. The case study pointed out the importance of using the agile approach where there will be evolutionary and continuous collaboration with users.

4.2 Design Criteria

Based on the challenges found through the case organisation and from the literature, five design criteria were determined.

- Support evolutionary design – Considering the challenges faced in using a linear process model, it is important to use an evolutionary design. Thus, even if the project requirements are complex and the stakeholders do not clearly understand the requirements, the project can progress still by having several milestones and evaluating the results at each of these steps.
- Support establishment of a collaborative process – Since the data analysts are not aware of the domain knowledge and to understand the data and the requirements it is important to have continuous communication with the consumers. Also, the data analysts need to communicate the findings from each milestone to the consumers and planners as such they will understand the current status of the project and will be able to guide the analysts.
- Protect patient data – It is important to de-identify and anonymize patient data while maintaining the richness of medical data.
- Configuration management – It is important to configure the different versions as such the data analysts can easily refer an earlier version if there is a requirement.
- Conceptualization of the problem – It is important to conceptualize the problem where the constructs that are going to be used will be identified after understanding the user requirements and data. Data modeling can be carried out based on the conceptualization.

These abstract design criteria determined through the challenges mentioned in previous section will be used in the development of the process model.

5 UNIFIED STANDARDISED ANALYTIC MODEL

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 project

requirements. Thus, it is important to develop a well-defined process to perform HA as a well-documented process using Designed Science Research (DSR) approach. This "unified structured analytic model" (USAM) is developed specifically targeting novice users carrying out HA projects. While 'unified' stands for consolidated or full representation of an entity, the term 'structured' indicates the well-organized arrangement of the steps involved. 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.

5.1 Satisfying Design Criteria in USAM

Several design criteria are considered in developing the USAM model based on the challenges identified through literature and through case organisation. They are specified below in Table 1:

Criteria 1	Support evolutionary design
Assumption	It is not possible to clearly define requirements at the beginning of the project
This is achieved by 1. Having a simiple requirement at the beginning, so that, the requirement can evolve when the project progresses 2. Breakdown the requirement into several incremental goals 3. Modelling each increment and demonstrate the findings to stakeholders. 4. Configuration management	
Criteria 2	Support establishment of a collaborative process
Assumption	Stakeholders will be having mutual understanding and a commitment to work together.
This is achieved by 1. Define clear guidelines on communication modes, frequency and content to discuss. 2. Frequent discussion among data analysts as well as with data analysts, planners and users to avoid conflicts. 3. Document tasks carried out in each steps of the process and findings (Ahangama and Poo 2014).	
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.
This is achieved by 1. De-identification and anonymization of patient data using the HIPAA standards (See Appendix D) 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	
Criteria 4	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	
Criteria 5	Conceptualization of the problem
Assumption	To use data modelling algorithms the constructs should be determined.
A new phase is introduced in to the CRISP-DM model after domain and data understanding to conceptualize the model.	

Table 1. Design criteria for USAM.

5.2 Process Management

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 2, required confidence on the validity of the data prepared and the model built is achieved by iterating the process in a cycle (data, model cycle).

Collected data will be analysed in the data cycle depending on the nature of the domain. This will be followed by conceptualization of the model by going through the loop until there is an assurance on the quality and usefulness of the data collected relevant to the problem defined. Fine-tuning of the data model will be carried out for validation of the model in the loop of the model cycle.

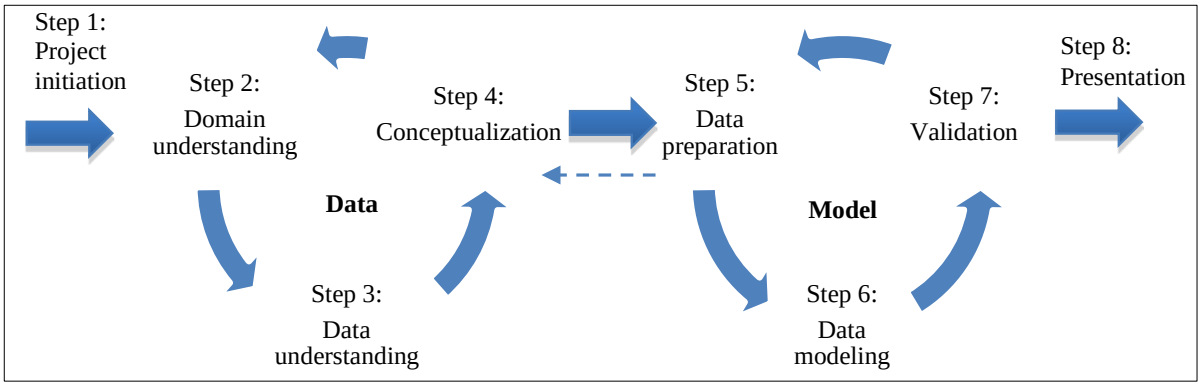

Figure 2. High-level overview of the USAM

This is an iterative and not a rigid process as there is a possibility of alternating between steps. Moreover, in case of an error in a particular step or in a case where expected results are not achieved the necessary corrections can be facilitated by going through the feedback loop from one cycle to another cycle. Thus the whole process can be considered as a life cycle model with continuation of the analytics even after the presentation of the solution. Lessons learnt during the development of the HA process and the results obtained will undoubtedly trigger new projects (Chapman et al. 2000). Planning new projects based on the experience from prior projects will make them more focused on the specific requirements.

At the initiation of the project the problem to be solved needs to be identified. Then the process enters the data cycle. The data cycle starts with attaining relevant domain knowledge. Exploring the dataset and extracting the relevant data are vital to understand the data and to theorize at the preliminary stages. The research questions and relevant hypotheses are developed based on the prior understanding of the domain through literature review and the data collected. Data modelling is facilitated by structuring and constructing the collected data and the subsequent data collection and analysis will be carried out according to the initial conceptualization. Modifications to the conceptualization with new constructs and conditions can be brought about based on the additional data collected iteratively. The data scientist can proceed to the model cycle upon gaining adequate confidence on the conceptualization.

After the data preparation, based on the conceptualization of the problem, an appropriate analytic model will be selected to build the data model. Validation of the developed data model to ensure its expected level of accuracy is carried out using a new set of data. In completion of the project, the model cycle will be followed by the step dealing with documentation of the results and tasks performed.

5.2.1 Project Initiation

A 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).

5.2.2 Domain Understanding

Domain understanding is quite important after gaining access to the data. 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 healthcare domain (Cios and Moore 2002), a greater effort is required to understand the domain requirements and necessary objectives relevant to the field.

This involves determining organisation objectives, stakeholder requirements (identification of stakeholders and their expectations), situation assessment (feasibility), compliant needs (rules and regulations) and project planning (resources, communication, testing and implementation planning).

5.2.3 Data Understanding

This step 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.

5.2.4 Conceptualization

Conceptualization refers to abstract representation of some selected areas in the real world using entities and concepts to illustrate some interesting relationships. A 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 variable and independent variable) 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. A hypothesis should be given for each and every research question.

5.2.5 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. This includes data selection, cleaning, construction, integration and formatting tasks.

5.2.6 Data Modelling

This is the main step of the data analytic process. Data modelling can be performed based on the clear identification of the user requirements and data, and upon the data preparation. Modelling step includes application of the selected modelling techniques where relevant algorithms and parameters are altered to get the optimal results. This includes selection of the modelling technique, data modelling and the assessment of the data model. There are various techniques (e.g. classification, clustering and their different algorithms). Specific algorithm will be selected based on the conceptualization. Since it is not possible to consider each and every algorithm, we will not discuss them in this paper.

5.2.7 Validation

The finalized data model needs to be evaluated in the validation step. 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.

5.2.8 Presentation

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 2 will be adjusted based on the new requirements (to avoid creation of many new reports).

6 DISCUSSION AND CONCLUDING REMARKS

In this paper, a refined process model is proposed for HA projects. The designed process model is composed of 8 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. USAM is developed based on DSR approach using action case research approach. To develop the model, practices from data mining are employed. The evaluation of the model is carried out as an ex-ante evaluation. USAM will be useful to the novice users to HA projects in understanding the necessary actions to be carried out and it will guide them in performing the projects. Even though, the model is developed considering HA application scenarios, it can be generalised to other contexts as well.

Implications

There are several theoretical implications of this project. First, the artefact itself is a contribution of the study. The proposed model can be considered as a contribution due to the limited availability of models in HA and failure to adopt available approaches in real contexts. The model allows repeatability and applicability into various scenarios. Second, the approach used to develop and evaluate the model is also a contribution to the theoretical discourse. We used an action research based approach to develop and evaluate the model. When practical contributions are considered, USAM could be used as guidance by the practitioners when carrying out their projects. This is especially useful to novice users just commencing work in projects. Second, the proposed model considers limitations and proposes an agile based approach to handle issues in HA projects. Continuous collaboration with stakeholders, evolutionary and iterative approach and story driven approaches could be used in HA projects. Finally, the introduction of pair data analytics (like pair programming in software engineering) and handling of concurrent projects are practical contributions from this study. They are useful in knowledge management and project management. It takes considerable time to commence a project after an interruption, thus, having multiple projects will avoid idling time and allows achieving maximum benefits from limited human resources.

Limitations and Future Direction

Different HA project types were not considered in this study. For example, HA projects can be varied based on the complexity of the project and the ambiguity of the project requirements. The aspects such as project management and knowledge management will vary based on the project types, and we plan to extend the model in future and consider the variations to USAM based complexity and ambiguity of

project requirements. Moreover, the model evaluation can be further explored through an experiment. In this study we had used and action case research based approach to evaluate and refine the model. The model's utility can be further studied in future. Finally, we did not consider individual algorithms in data modelling, as it is not useful to consider specific algorithms to achieve the generalization. Also, considering the huge number of data modelling algorithms available, it is not practical to take each individual algorithm into account.

At the initial step of the process model development, the refining of the data model development process was considered and as such, evaluation of the process using an external project is 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.

References

Ahangama, S., & Poo, C. C. D. (2014). Unified Structured Process for Health Analytics. International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering, 8, 744-752.

Ahangama, S., & Poo, C. C. D. (2015). What Methodological Attributes Are Essential for Novice Users to Analytics? – An Empirical Study. In S. Yamamoto (Ed.), HIMI 2015, Part II, LNCS 9173 (pp. 1–12): Springer International Publishing Switzerland. doi: 10.1007/978-3-319-20618-9_8

Argote, L., McEvily, B., & Reagans, R. (2003). Managing knowledge in organizations: An integrative framework and review of emerging themes. Management science, 49(4), 571-582.

Arnott, D. (2006). Cognitive biases and decision support systems development: a design science approach. Information systems journal, 16(1), 55-78.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics, 77(2), 81-97.

Catley, C., Smith, K., McGregor, C., & Tracy, M. (2009). Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. Paper presented at the Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on.

Chan, F. K., & Thong, J. Y. (2009). Acceptance of agile methodologies: A critical review and conceptual framework. Decision Support Systems, 46(4), 803-814.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Cios, K. J., & Moore, W. G. (2002). Uniqueness of medical data mining. Artificial intelligence in medicine, 26(1), 1-24.

Collier, K. (2011). Agile analytics: A value-driven approach to business intelligence and data warehousing: Addison-Wesley.

Dzeroski, S. (2007). Towards a general framework for data mining. In S. Dzeroski & J. Struyf (Eds.), Knowledge Discovery in Inductive Databases (Vol. 4747 of Lecture Notes in Computer Science, pp. 259-300): Springer-Verlag.

Eggebraaten, T. J., Tenner, J. W., & Dubbels, J. C. (2007). A health-care data model based on the HL7 reference information model. IBM Systems Journal, 46(1), 5-18.

Fitzgerald, B. (1996). Formalized systems development methodologies: a critical perspective. Information systems journal, 6(1), 3-23.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. MIS quarterly, 37(2), 337-356.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 28(1), 75-105.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). The unified software development process (Vol. 1): Addison-Wesley Reading.

Johannesson, P., & Perjons, E. (2014). An Introduction to Design Science: Springer.

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

Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., & Goodenday, L. S. (2001). Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial intelligence in medicine, 23(2), 149-169.

Lee, A. S. (2001). Editorial Comments: MIS Quarterly's Editorial Policies and Practices. MIS quarterly, 25(1), iii-vii.

Luján-Mora, S., Trujillo, J., & Song, I.-Y. (2006). A UML profile for multidimensional modeling in data warehouses. Data & Knowledge Engineering, 59(3), 725-769.

Marban, O., Mariscal, G., & Segovia, J. (2009). A Data Mining & Knowledge Discovery Process Model. In J. Ponce & A. Karahoca (Eds.), Data Mining and Knowledge Discovery in Real Life Applications (Vol. 2009, pp. 8): InTech. Retrieved from http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/a_data_mining__amp__knowledge_discovery_process_model.

Marban, O., & Segovia, J. (2013). Extending UML for Modeling Data Mining Projects (DM-UML). Journal of Information Technology & Software Engineering, 3(121). doi: 10.4172/2165-7866.1000121

Marban, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2009). Toward data mining engineering: A software engineering approach. Information systems, 34(1), 87-107.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision Support Systems, 15(4), 251-266.

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(02), 137-166.

Matignon, R. (2007). Data mining using SAS enterprise miner (Vol. 638): John Wiley & Sons.

Mohan, K., & Ahlemann, F. (2011). What methodology attributes are critical for potential users? understanding the effect of human needs. Paper presented at the Advanced Information Systems Engineering.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45-77.

Prat, N., Akoka, J., & Comyn-Wattiau, I. (2006). A UML-based data warehouse design method. Decision Support Systems, 42(3), 1449-1473.

Pries-Heje, J., & Baskerville, R. (2008). The design theory nexus. MIS quarterly, 731-755.

Raghupathi, W., & Raghupathi, V. (2013). An Overview of Health Analytics. J Health Med Informat, 4(132), 2.

SAS. (2008). SAS Enterprise Miner: SEMMA. Retrieved 27 February 2014, 2014, from http://www.sas.com/technologies/analytics/datamining/miner/semma.html

Sawyer, S., & Jarrahi, M. (2014). Sociotechnical approaches to the study of information systems Computing Handbook, Third Edition: Information Systems and Information Technology (pp. 5.1-5.27).

Vaishnavi, V., & 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

Westphal, C., & 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), 470-475.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Paper presented at the Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Zubcoff, J., & Trujillo, J. (2006). Conceptual modeling for classification mining in data warehouses Data Warehousing and Knowledge Discovery (pp. 566-575): Springer.