International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:8, No:11, 2014

Unified Structured Process for Health Analytics

Supunmali Ahangama, Danny Chiang Choon Poo

***Abstract*—**Health analytics (HA) is used in healthcare systems for effective decision making, management and planning of healthcare and related activities. However, user resistances, unique position of medical data content and structure (including heterogeneous and unstructured data) and impromptu HA projects have held up the progress in HA applications. Notably, the accuracy of outcomes depends on the skills and the domain knowledge of the data analyst working on the healthcare data. Success of HA depends on having a sound process model, effective project management and availability of supporting tools. Thus, to overcome these challenges through an effective process model, we propose a HA process model with features from rational unified process (RUP) model and agile methodology.

***Keywords*—**Agile methodology, health analytics, unified process model, UML.

I. Introduction

HEALTH Analytics (HA) has become popular with its possibility to advance the healthcare system from a volume-based system to a value-based system [1]. Statistical, predictive, quantitative and, other models on healthcare data is used in HA for informed healthcare decision-making. HA applications can be defined as "collections 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 and make better and faster health decisions" [2]. HA applications allow a healthcare system to be more efficient (improved outcomes, improved coordination, reduced time and cost, and better value) while providing constant or better quality care. However, most of the 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. To address these needs, HA applications are required to be introduced in hospitals to improve the performance of the healthcare system.

Application of HA has been deterred by problems such as heterogeneous nature of available data sources and unstructured and ad hoc approach to HA process [3], [4]. Moreover, the accuracy and completeness of the results depend on the skills of the data analyst working on the healthcare data. This is largely due to vague project objectives and uncertain methodologies. Most of the existing HA projects are performed in an ad-hoc manner without addressing proper project management or quality assurance aspects. As HA projects progress and become too complex, the need arises for a standardized process model. Presently, clearly defined HA process models with the inputs, outputs and tasks to convert input to output are lacking. Even though data mining (DM) can be linked with HA having common functions, there are certain differences. While analytics include hidden pattern recognition and reporting of results; in DM it is only identification of hidden patterns. Thus, these can be applied in HA projects only with certain modifications.

These issues or problems could be addressed through a well-designed HA framework. Such a framework will facilitate the performance of all these projects as a structured process [2], with clearly defined objectives, proper project planning and with systematically documented prior knowledge, data, methodologies and results [5]. Numerous examples and possible best approaches could be drawn from DM and software engineering (SE) projects [6]. Many authors have proposed frameworks like CRISP-DM [7], DM-UML [3] and other specific approaches for each DM technique [8]–[10]. Though, Raghupathi and Raghupathi [2] have proposed a health analytics framework, they have not proposed the specific methodologies and relevant steps for each stage of their proposed framework. Due to the diversity of available HA techniques (descriptive, predictive and prescriptive analytics) and the diversity of healthcare ecosystem, seamless application of these approaches in health analytic projects is not feasible. Thus, a unified structured and agile framework is proposed in this research to perform HA at ease, independent of skills of the data analyst and to carry out a systematic and flexible project.

This study is performed using a design science research approach (DSR) [11], [12] which is used to explain the problem and the related theoretical principles for the proposed process model. The findings of this study will have a significant impact on both theoretical discourse and the practical discourse of HA. First, this unified structured framework may be used as a standardized process and as a reference model to provide a better understanding of the flow of the HA process. Second, this will offer a clearer comparison of existing and future models. Third, while this framework 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 tool for various stakeholders in this sector too. Finally, real world illustrations may offer a clearer explanation on the applicability of the framework in actual healthcare projects.

The subsequent sections in this paper will provide the background, problem definition and the proposed unified structured framework along with the discussion.

S. Ahangama and D. C. C. Poo are with the Department of Information Systems, School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417 (e-mail: supunmali@comp.nus.edu.sg, dannypoo@nus.edu.sg).

A. Background

HA (or DM) has been considered by many as an 'art' (creative process) and data analysts followed their own styles when carrying out HA projects [13]. According to a survey conducted to understand the 10 most challenging problems in DM, non-availability of a unifying theory for DM (it is the top priority problem in the list of 10 problems) and issues related to DM process have been identified as two of them [4]. The former refers to the lack of a theoretical framework that unifies different DM tasks (classification, clustering, association, etc.) and DM approaches (databases, statistics, machine learning, etc.) as various techniques are created for individual projects (e.g. for classification or clustering problems). The latter identifies issues such as automating different DM process operations and building a methodology into DM system. As a result, methodology related issues are created where the success of the DM project depends on the skills and the knowledge of the person of the team analyzing the data without giving any prospect for repetition of successful practices in future assignments [14]. Numerous process models are being proposed, to avoid these complexities and to facilitate a standardized approach in performing DM studies.

CRISP-DM (Cross Industry Standard Process for DM) [7] and SEMMA (sample, explore, modify, model, assess) by SAS [15], [16] are two such popular DM process models. Compared to CRISP-DM, SEMMA had failed to provide an adequate attention to rigorous requirements of a complete DM process. SEMMA focuses only on the technical portion of the project (statistical, modeling and data manipulation sections in a DM 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 DM [17]. SEMMA does not consider DM 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 designed approaches associate strongly with the SAS Enterprise Miner Software package [16] 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 DM process.

As a result of limitations of other models (including SEMMA model), CRISP-DM is implied as the de facto standard in DM [18] for several reasons: (1) it is a standardized step by step approach to DM [7], [14], (2) it is based on pre-CRISP-DM models and incorporated some of their substantial features [14], (3) it is used as the foundation for many forthcoming models [18], (4) it is the most frequently used model in DM projects [5], [17], [18] and (5) it is vendor independent [14].

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 DM 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 be the actual representation of the problem. Second, CRISP-DM does not support new data collection during later stages of the process (e.g. data processing and modeling) 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 [19]. Third, it lacks project management processes and an integral process to ensure the project completeness and quality. Fourth, CRISP-DM (even SEMMA) assumes that sufficient knowledge of the requirement is already available [20]. However, in actual settings the clients use a different terminology compared to data analysts making it hard to translate the requirements. Thus, the available tools do not support it.

B. Uniqueness of Medical Data

According to Cios and Moore [21], there are several factors that differentiate medical data from other data. First, Heterogeneity of medical data [22]: medical data is voluminous and is collected from various sources (images, patient interviews, physicians' notes, and biomedical data) [5]. Though the standard HL7 (v3.0, RIM): international health informatics interoperability standards provide 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 [23]. Since medical data (e.g. case notes are unstructured, ambiguous and have different grammatical constructions for each physician) is complex and hard to analyze (when compared to well-structured financial data – e.g. stock exchange data), it is hard to use automated analysis systems [5]. Even though 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 had failed to implement such standards in healthcare sector due to its complexity and failure to have one single standard. 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 or models can be effectively applied in determining the relationships.

Second, ethical, legal and social issues: 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 allow only authorized person to access them) [24] and (3) rigid administrative guidelines (e.g. IRB-Institutional Review Board, privacy rules in HIPAA of USA) [25]. Such administrative policies are normally not required for non-medical DM.

Third, statistical philosophy: the data is collected (or not collected) to use for 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 [26]. Fourth, special status of medicine: 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. Thus, in medical DM, a special consideration should be paid to these specific attitudes in medicine. Finally, but the most important factor in HA is that the decisions should always be supported with valid justifiable explanations [5] as these applications are working in a safety critical context.

By considering these unique features in medical data, it is not possible to directly apply CRISP-DM or any other approach due to problems (in CRISP-DM and SEMMA) highlighted before. Thus, a new structured process model needs to be proposed to HA while making use of the best practices in those models.

C. Software Engineering Approach

During the early years of software development, the main focus was on programming languages and algorithms. The programmers implicitly designed the programs (in their mind rather than documenting the design) and developed them according to their personal style. 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 even though were paid and 68% of software overrunning delivery schedules [27]. These issues in software development and delivery led to 'software crisis' in 1968 [28]. Many of these shortcomings are due to failure to use a standardized procedure and faults in 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. Thereby, software development led to a new discipline called SE 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, rational unified process (RUP) and agile process too are very popular in software development industry. RUP is a SE process used to transfer user requirements to a software system. It can be considered as a generic process framework that could be used in very large-scale application developments. Unified modeling language (UML) is an integral part of RUP 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 RUP [19]. UML 'use cases' are used in the SE projects to capture functional requirements and based on them developers design and develop the system and review the systems (whether it confirms to use cases). Thus, RUP 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 RUP, we believe that we could adopt these two aspects into health analytic projects as well. Thus, HA projects could be iterative and incremental while being a use case driven process.

Agile software development manifesto [29] provides interesting principles that can be adopted in HA projects while handling issues in above DM models when applying in HA context. Welcoming changing requirements even late in the project, business people and developer working together, building projects around motivated individuals (provide the support and the environment to work) and having regular intervals reflecting on how to improve are some of the key principals that could be adopted in HA projects. These factors could be taken into account when developing the process model for HA.

II. Research Method

This process model (called as artifact in DSR) for HA is developed using the DSR approach since the exploratory and confirmatory hypothesis research approach is not suitable [30]. According to Hevner et al. [11] the main notion in DSR is to clarify the goals of the artifact and then to evaluate its utility. We specifically used the DSR approach proposed by Pries-Heje and Baskerville [12] which includes (1) explaining the problem, (2) analyzing various models available in the HA context, (3) designing the process model, and (4) evaluation of the model. In designing the process model we went through several iterations in order to get a clear understanding of the problem, prior to its development.

The model is designed based on case study evidence [31] accessed by working in a leading hospital in Asia. We were able to gather domain knowledge and experience on carrying out HA projects through the consultations with a physician, a data analyst and the other relevant parties. Since the target users of this model are novice users, after designing the first version of the model students enrolled in a master program in business intelligence were interviewed to understand the problems they encounter in using the model and how the first version could be further improved. Subsequently, the model was demonstrated to get the feedback from two academics and PhD students in Computer Science and Information Systems at the National University of Singapore. Based on their feedback and the literature review we decided to consider the agile methodology and RUP in developing the model.

III. Application of Agile Methodology

In the application of agile methodology, there are several factors that need to be considered. First, it uses incremental, iterative and evolutionary approach. An initial design plan (not all content of the analysis) will be made to initiate the project

and to support interaction with stakeholders (to get feedbacks). Thus, the conceptual model built at the beginning will evolve to a physical analytic model with necessary flexibility to allow changes during the project.

Second, in agile based projects user story driven approach (statements in the point of view of users) is preferred over a data driven approach (just in time data modeling – [32]). The former approach allows connecting the business goals with user needs. Later on user stories can be converted to requirements with incremental iterations [33].

Third, unlike in waterfall method and other sequential methods where interaction among stakeholders occur only at the requirement gathering stage (at initial stages of the project) with limited interaction during later on the project, agile method supports continuous collaboration between analyst, sponsors and users of the system.

Fourth, in testing agile projects it is necessary to consider business acceptability (whether it meets the end user expectations) and technical acceptability (whether it does what analyst expects it to do) [34]. In addition, when testing analytic projects it is important to have test cases under version control.

Thus, considering these factors, the proposed process model will be developed as an incremental iterative framework with an evolutional approach (to allow modifications throughout the project). To ensure repeatability, we will also focus on having version control as in SE projects. For example, versions of the training data and testing data will be documented. Moreover, collaboration among stakeholders will be maintained to facilitate capturing of user requirements and meet user expectations at the presentation of results.

Table I provides an outline of how the design of the process model satisfies the design criteria. Basic assumptions considered and the elements that make the satisfied solution are indicated.

TABLE I
SATISFACTION OF DESIGN CRITERIA BY THE MODEL DESIGN

Design criteria 1	Support establishment of a collaborative process
Assumption	To maintain collaboration among stakeholders there should be mutual understanding and a commitment to work together.
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.	
Design criteria 2	Support evolutionary design
Assumption	Not possible to define requirements upfront
This is achieved by 1. Having minimally sufficient upfront design, so that, the team can evolve the design when project progresses 2. Modeling small increments and demonstrating the findings to stakeholders. 3. Refactoring without having an undesirable influence on things done in previous iterations (without breaking previously developed models). 4. Version controlling	
Design criteria 3	Supports self organizing teams
Assumption	The team members value the autonomy, mastery (improve skills) and purpose (value being part of something greater than them) and they must adhere to organization guidelines and regulations as well as have self-discipline.
This is achieved by project manager being a facilitator rather than being the manager from outside the project team. The leaders will be an integral part of the team. The team needs to periodically review and revalidate their objectives and assumptions through communication with customers. The performance will be measured through the frequency of delivery working project steps (iteration of the project).	

IV. Application of Rational Unified Process Methodology

A modeling 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 [3], [8]. By using a universal visual modeling language, the users and analysts can direct their focus to the main objective that is on to HA process. Even if documentation strategies based on UML are proposed they had 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 proposed process model for HA. By using extension mechanism, UML profiles customize the diagrams to a particular domain (for different use) [35]. The extensions are specified through stereotypes, properties and restrictions while respecting original semantics in UML [36]. Thus, in this study we extended the UML profile by proposing a new UML profile to facilitate the HA process proposed by us (USFHA).

The main improvements are made on profiles by defining new stereotypes, new tagged values (to describe basic parameters) and by extending the meta-model with new characteristics [35]. These improvements will be explained through different sections in this paper. We will produce a technical documentation necessary to carry out the complete process. 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 DM [3] and are revised to support HA.

V. Proposed Model: USFHA

Noted dearth of studies on developing a framework for HA, has created a necessity to develop a framework to carry out HA. With the popularization 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 lack a proper consolidated structure representing the complete process of HA. Thus, it is important to develop a well-defined process to perform HA as an engineering process and such a new framework for HA by adopting significant and related components from SE process and DM processes is proposed in this study.

This unified structured framework is developed categorically to carry out HA. It is important to clarify what is referred to as a framework, to begin with the illustration of the

USFHA (unified structured framework for HA). A framework could be distinguished as a description of a complex process with a set of assumptions, concepts, values and practices that constitute the reality (adopted from the American heritage dictionary of the English language). The term 'structured' in the USFHA refers to the arrangement of steps in a highly organized and in a definitive pattern. That is, the proposed framework will be a planned block structure with distinctly defined steps intending to improve the clarity, quality and modeling time. Here, 'unified' stands for consolidated or full representation of an entity Therefore, this framework will be a unified block structure.

The USFHA framework is built as follows. First, the tasks are grouped into several conceptual domains. The first domain defines the "universe of discourse" of HA. Thus, it will focus on defining the general concepts for all HA projects and will be independent of any particular set of objectives. The next domain will focus on the HA process model and its components and intermediate states (i.e. life cycle of the HA process model). The aim of this study is to promote conceptual view of HA while reflecting current literature and extracting a set of key concepts from related fields. The derivations for existing concepts or models are given as and when required. Finally, the unified structured framework is presented with a set of associated terminology and deliverables.

The USFHA is composed of 7 steps. They are domain understanding, data understanding, conceptualization, data preparation, data modeling, validation and presentation. It is an iterative-incremental life cycle model. As shown in Fig. 1, the process iterates in a cycle (data, model cycle) until there is high confidence on the validity of the data prepared and on the model built. Thus, in data cycle the gathered data will be modified going through the loop until there is high confidence on the quality and usefulness of the data. Similarly, the data model will be fine-tuned until the model is validated in the loop in model cycle. This is not a rigid one way cycle as moving back and forth between steps is always possible [7]. Thus, this is considered as an iterative process. Moreover, there is a feedback loop from one cycle to another cycle 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) [7]. Thus, such new projects will be more focused as are planned based on the experience from prior projects.

USFHA begins with access to data source and domain understanding. After getting access to a suitable data source and attaining relevant domain knowledge the process enters the data cycle. The data cycle starts with understanding the data by exploring the dataset and then by extracting 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. The data is structured and constructed to facilitate the data modeling. The initial conceptualization and data preparation will be used as a guide for subsequent data collection and analysis. The initial conceptualization is modified with new constructs and conditions based on the additional data collected iteratively. When the data scientist is sufficiently confident with the conceptualization and the data preparation, he/she can move on to the model cycle. The first step in model cycle is data modeling where an appropriate analytic model will be selected and then 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.

VI. UML Profile Extension

UML models proposed for the USFHA process model are given with their connections in Fig. 2.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. It is important to note that there is dependency between individual stages and the most of the initial designing of the diagrams will be carried out in the domain understanding and data understanding phases. Later on in other subsequent phases the models will be fine-tuned based on the necessary new requirements.

Business use case will be set based on the business goals (Fig. 3). Accuracy of business use cases will be based on their alignment with the business goals of the organization. For every business use case and business goal relationship there will be a HA use case (if the business goals are SMART). The relationship between the HA use case and the business goals are represented in Fig. 3. As indicated by the dependency, HA use case can be evaluated based on the business goals. HA goals will be useful at the validation phase in determining whether the goals are met at the end of the modeling. Thus, it will act as a reference in selecting the techniques and tools for data preparation and modeling and it depends on the HA use case.

In HA use cases; there will be HA goals and actors using the knowledge extracted. The actors could be using HA knowledge as knowledge itself in form of a report or a software system. This is illustrated in Fig. 4. As illustrated below (1) HA actor directly uses the knowledge extracted, (2) HA actor uses HA documents prepared based on the knowledge extracted and (3) HA user accesses a software application developed using the knowledge extracted.

Considering the best practices of RUP and Agile, the proposed process model for HA is developed as an extension to the CRISP-DM model. The constraints in CRISP-DM are attempted to be solved by incorporating RUP and Agile methodologies.

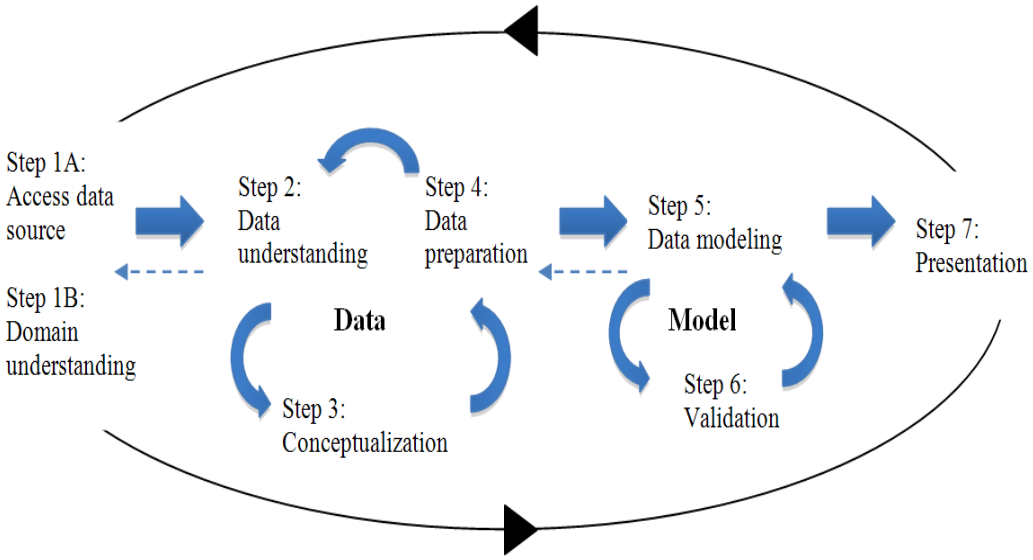

Fig. 1 A structured unified framework for health analytics

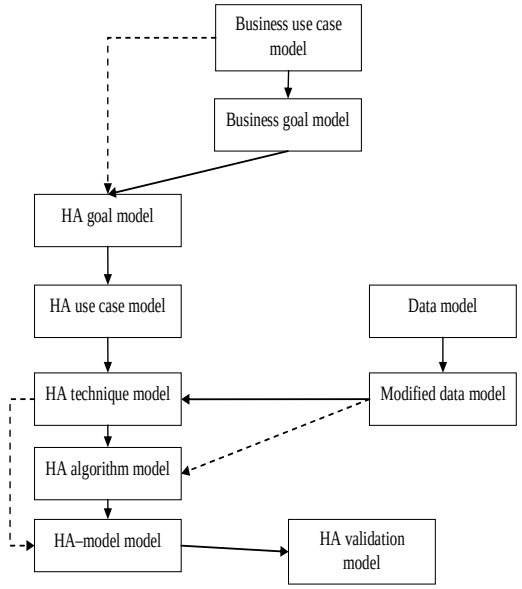

Fig. 2 Overview of the UML models used

VII. Evaluation

According to the DSR approach, the utility of the first version was evaluated by interviewing the novice users performing HA related projects and then by creating summary reports of interview transcripts. This was very useful to understand the areas requiring further amendments. The users indicated the importance of having user collaboration, as it is very hard to understand the user requirements prior to the designing of the project. Furthermore, they mentioned that the model is very useful at the beginning of the project specifically for novice users lacking any experience in commencing a project. They indicated that the use of UML diagrams are easy to understand and it takes less time to read a document compared to reading a text-based document. However, they indicated that the documenting each step is time consuming. One can commence to use the proposed practices with familiarization of the process and by understanding the uses of documentation.

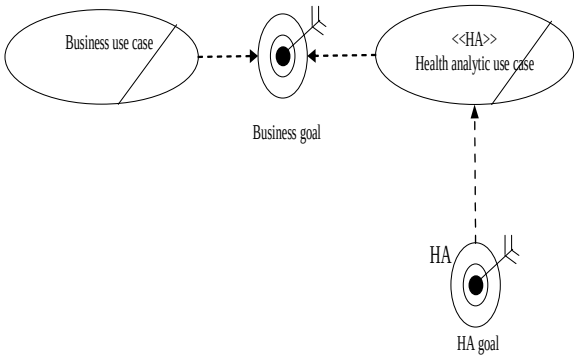

Fig. 3 Dependency between business use case, business goal, health analytic use case and HA goal

As future work, we planned to carry out semi structured interviews and observations to evaluate the efficacy, quality and the utility of the model [11]. It is planned to present the model to more number of novice and experienced users in healthcare industry to get their input for further improvements.

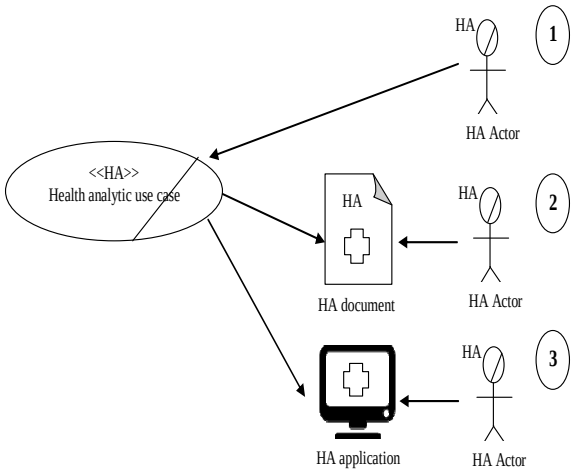

Fig. 4 Health analytics use case with Actors

VIII. Discussion

In the evolving field of HA, there has been a necessity for a simple set of best practices or a standard methodology to deal with diversified and iterative processes in healthcare projects. We believe that this 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 under idealized conditions (by having practical techniques and by illustrating its application in real scenarios) and (3) have a gradual learning curve (by using UML as method to clearly and explicitly document actions carried out). To address these aspects we introduced a process model, consisting of 7 steps starting from gaining access to the data and domain understanding to the presentation of results. This framework will allow implementing HA projects in a coherent manner. Documentations created at each phase in USFHA are summarized in the Table II. There will be dependencies between these documents. That is, we avoided having independent reports in each stage as in CRISP-DM. CRISP-DM has many duplicating documents, making it a burden on the designers and the analysts. To avoid that USFHA will be a compact process using cross-referencing among documents. Moreover, we considered the role of team members responsible for each document that has not been explicitly presented in CRISP-DM [7].

A comparison is carried out on USFHA model against CRISP-DM model in Table III, to determine the differences in the two models. We used CRISP-DM as it is the de facto standard for DM and several authors have used it in medical DM. The comparison is carried out in 7 areas. As indicated below, USFHA model has better flexibility over the healthcare project and it allows iterative cycling among different steps in the model.

TABLE II
USFHA Documentation Support

Step	Document	Role
Data access	Data access report	Project manager
Domain understanding	Project (domain understanding) report, business use case model and description, business goal model, health analytic use case model, health analytic goal model	Project plan – project manager Deployment plan – deployment manager DUR – business analyst
Data understanding	Data de-identification report, data description report, data exploration report, data quality report, health analytic data model, data component model	Data analyst
Conceptualization	Theoretical background report, list of research questions and conceptual model	Data analyst
Data preparation	Data set description, data preparation report, health analytic modified data model	Data designer
Data modeling	Model selection document, model assessment report, health analytic technique model, health analytic algorithm model, health analytic-model model	Data mining engineer
Validation	Model evaluation report, list of future actions, health analytic test model	Final report – project manager, quality assurance manager Evaluation report – business analyst
Presentation	Final report, experience documentation, health analytic deployment model	Project manager

TABLE III
Comparison of USFHA Model against CRISP-DM Model

Factors	CRISP-DM	USFHA
Domain	General process model	Specialized for health domain
Target technique	Support DM (mainly predictive analytics)	Supports descriptive, predictive and prescriptive analytics and exploratory analysis
Aim	Promotes perfect results	Promotes perfect results with comprehensive documentation as a deliverable
Process model structure	Waterfall with feedbacks	Iterative and incremental
Documentation support	Provides user manual to assist documentation	Provides a template to documents with extended UML diagrams at specific steps
Change management	Rigid to changes	Responding to change by incremental and iterative process and customizable phases
Initiation	Business understanding	Gain access to the organization and then to data sources

There are several theoretical and practical implications in this study. First, this unified structured framework could be used as a standardized process and as a reference model to provide a better understanding of the flow of the HA process. This framework is having considerable importance, as this is one of the first steps in developing a suitable process model for HA. Second, this model extends from previous work (SE processes like iterative incremental process and unified process, and DM models like CRISP-DM [7] and engineering process model [6]) and this is proposed as a base model to compare existing and future models. Third, this allows to perform health analytics easily without depending on the skills

of the data scientist and to carry out a systematic documentation as a communication tool for various stakeholders in this sector.

It is considered that for a business to be successful by using advanced analytics, there should be (1) a well-defined problem, (2) appropriate data to solve the problem, (3) proper data preparation and manipulation actions, (4) required skill set and tools, (5) steps to validate results, (5) deployment of the results and (6) follow up to modify the model. We believe that USFHA model is able to support the process of advanced analytics as it is developed paying due consideration to each of these aspects.

There are several limitations in this study. First, exclusion of discovery analytics from this study is a limitation. Model is constructed considering descriptive, predictive and prescriptive analytics. Since it is not very commonly used in HA projects yet and it is still evolving as a technique, we believe that a certain time should be given for it to mature to understand the special requirements and constraints of that. Second, the model does not consider genetic health related data. Even though it may restrict our model to a certain extent (as genetic data is becoming an essential component in healthcare field) [5], [37], considering it at this preliminary stage of the process model development could make it too complicated. However, we provide necessary flexibility to accommodate heterogeneous data sources to the existing model.

While acknowledging that this study proposes a preliminary framework to HA, the model could be made a de facto standard in HA with following recommendations. First, further research could be done into introduction of discovery analytics into USFHA. Discovery analytics is an important component in drug trails and could be beneficial to understand how its uncertainty aspect could be approached. Second, the model could be further validated in different application scenarios using heterogeneous sources (e.g. case notes, user discussions and chats). A likely application scenario is the possibility of using this model in a data source integrating genetic data, personal behavior data, socio-economic data and clinical data [37]. Third, we need to do a case study in future at an actual organization setting to understand the usage of USFHA by normal stakeholders of a project. Thus, interviews and observations could be used to understand how it is effective.

IX. Conclusion

Success of health analytics depends on having a sound process model, effective project management and necessary supporting tools. The benefits of HA process model varies depending on the stakeholder. First, if the analyst has no prior experience in carrying out HA projects, the process model will provide the guidance on performing the study and will advise on activities to be carried out in each phase. Even if the analyst is experienced, still the process model will be a checklist to determine whether any activity is omitted. Second, even the client will have a clear understanding of what is happening in the project and what is to be expected at the end of the project. Moreover, having a general framework will make the client more at ease, as they will be advised on the same approach from various vendors. Finally, for project managers this will ease the task of planning the project. A process model enables users to understand the interactions among the phases and will assist to link various tools, skill sets to implement an effective project.

The whole HA process model should have practical well defined steps to deal systematically with extracting useful knowledge from healthcare data to solve the problem under study. As a means of achieving that, we have proposed the USFHA process model. This will be a complete process, which will be an iterative problem solving cycle (with data cycle and data model cycle). There will be 7 steps starting from gaining access to the organization and the data source to presentation of results to the prospective clients.

References

[1] P. Horner and A. Basu. “Analytics & the future of healthcare,”*Analytics Magazine*, 2012, pp. 11-18.
[2] W. Raghupathi and V. Raghupathi, "An Overview of Health Analytics," *J Health Med Informat*, vol. 4, p. 2, 2013.
[3] Ó. Marbán and J. Segovia, "Extending UML for Modeling Data Mining Projects (DM-UML)," *Journal of Information Technology & Software Engineering*, vol. 3, 2013.
[4] Q. Yang and X. Wu, "10 challenging problems in data mining research," *International Journal of Information Technology & Decision Making*, vol. 5, pp. 597-604, 2006.
[5] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," *International journal of medical informatics*, vol. 77, pp. 81-97, 2008.
[6] O. Marbán, J. Segovia, E. Menasalvas, and C. Fernández-Baizán, "Toward data mining engineering: A software engineering approach," *Information systems*, vol. 34, pp. 87-107, 2009.
[7] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, CRISP-DM 1.0 Step-by-step data mining guide, 2000.
[8] J. Zubcoff and J. Trujillo, "Conceptual modeling for classification mining in data warehouses," in *Data Warehousing and Knowledge Discovery*, ed: Springer, 2006, pp. 566-575.
[9] N. Prat, J. Akoka, and I. Comyn-Wattiau, "A UML-based data warehouse design method," *Decision Support Systems*, vol. 42, pp. 1449-1473, 2006.
[10] S. Luján-Mora, J. Trujillo, and I.-Y. Song, "A UML profile for multidimensional modeling in data warehouses," *Data & Knowledge Engineering*, vol. 59, pp. 725-769, 2006.
[11] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, pp. 75-105, 2004.
[12] J. Pries-Heje and R. Baskerville, "The design theory nexus," *MIS Quarterly*, pp. 731-755, 2008.
[13] C. Westphal and T. Blaxton, *Data mining solutions: methods and tools for solving real-world problems*, 1998.
[14] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," presented at the Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000.
[15] R. Matignon, *Data mining using SAS enterprise miner* vol. 638: John Wiley & Sons, 2007.
[16] SAS. *SAS Enterprise Miner: SEMMA*, 2008 Available: http://www.sas.com/technologies/analytics/datamining/miner/semma.html
[17] O. Marban, G. Mariscal, and J. Segovia, "A Data Mining & Knowledge Discovery Process Model," *Data Mining and Knowledge Discovery in Real Life Applications. IN-TECH*, vol. 2009, p. 8, 2009.
[18] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *The Knowledge Engineering Review*, vol. 25, pp. 137-166, 2010.
[19] I. Jacobson, G. Booch, and J. Rumbaugh, *The unified software development process* vol. 1: Addison-Wesley Reading, 1999.

[20] P. Britos, O. Dieste, and R. García-Martínez, "Requirements Elicitation in Data Mining for Business Intelligence Projects," in *Advances in Information Systems Research, Education and Practice*, ed: Springer, 2008, pp. 139-150.
[21] K. J. Cios and W. G. Moore, "Uniqueness of medical data mining," *Artificial intelligence in medicine*, vol. 26, pp. 1-24, 2002.
[22] M. Kwiatkowska, M. S. Atkins, N. T. Ayas, and C. F. Ryan, "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*, vol. 11, pp. 651-660, 2007.
[23] N. Esfandiary, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge Discovery in Medicine: Current Issue and Future Trend," *Expert Systems with Applications*, vol. 41, pp. 4434–4463, 2014.
[24] X.-B. Li and J. Qin, "A Framework for Privacy-Preserving Medical Document Sharing," in *Thirty Fourth International Conference on Information Systems*, Milan, Italy, 2013.
[25] H. Chen, R. H. Chiang, and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS quarterly*, vol. 36, 2012.
[26] T. J. Eggebraaten, J. W. Tenner, and J. C. Dubbels, "A health-care data model based on the HL7 reference information model," *IBM Systems Journal*, vol. 46, pp. 5-18, 2007.
[27] Jibitesh Mishra and A. Mohanty, *Software Engineering*: Pearson Education India, 2011.
[28] P. Naur and B. Randell, *Software Engineering: Report of a conference sponsored by the NATO Science Committee, Garmisch, Germany, 7-11 Oct. 1968, Brussels, Scientific Affairs Division, NATO*, 1969.
[29] K. Beck, M. Beedle, A. Van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, and R. Jeffries, Manifesto for agile software development, 2001, Available: http://agilemanifesto.org/
[30] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Quarterly*, vol. 37, pp. 337-356, 2013.
[31] S. L. Pan and B. Tan, "Demystifying case research: A structured–pragmatic–situational (SPS) approach to conducting case studies," *Information and Organization*, vol. 21, pp. 161-176, 2011.
[32] S. W. Ambler, "Agile Model Driven Development (AMDD)," in *XOOTIC Symposium 2006*, 2006, p. 13.
[33] M. Cohn, *User stories applied: For agile software development*: Addison-Wesley Professional, 2004.
[34] K. Collier, *Agile analytics: A value-driven approach to business intelligence and data warehousing*: Addison-Wesley, 2011.
[35] J. Zubcoff and J. Trujillo, "A UML 2.0 profile to design Association Rule mining models in the multidimensional conceptual modeling of data warehouses," *Data & Knowledge Engineering*, vol. 63, pp. 44-62, 2007.
[36] OMG, "Omg Unified Modeling Language (OMG UML) Superstructure specification version 2.4. 1," document formal/2011-08-06. Technical report, OMG, 2011.
[37] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *New England journal of medicine*, vol. 363, pp. 301-304, 2010.