Baltic J. Modern Computing, Vol. 3 (2015), No. 2, 92-109

CRISP Data Mining Methodology Extension for Medical Domain

Olegas NIAKŠU

Institute of Mathematics and Informatics
Vilnius University
Akademijos g. 4, LT-08663 Vilnius, Lithuania

niaksu@acm.org

Abstract. There is a lack of specific and detailed framework for conducting data mining analysis in medicine. Cross Industry Standard Process for Data Mining (CRISP-DM) presents a hierarchical and iterative process model, and provides an extendable framework with generic-to-specific approach, starting from six phases, which are further detailed by generic and then specialized tasks. CRISP-DM defines following data mining context dimensions: application domain, problem type, technical aspect, and tools & techniques. In this study, we propose an extension of the CRISP-DM, called CRISP-MED-DM, which addresses specific challenges of data mining in medicine. The medical application domain with its typical challenges is mapped with CRISP-DM reference model, proposing the enhancements in the CRISP-DM reference model. Furthermore, the model to evaluate compliance to the CRISP-MED-DM is proposed. The model allows evaluating and comparing to what extent different data mining projects are following the process model of CRISP-MED-DM.

Keywords: data mining methodology, data mining application, healthcare, medicine

1. Introduction

Since 1990, a number of domain independent process models, application methodologies, industry standards have been proposed. Cross Industry Standard Process for Data Mining (CRISP-DM), "Sample, Explore, Modify, Model and Assess" (SEMMA) process model, Predictive Model Markup Language (PMML) are the most prominent of them. However, these process models are generic and shall be tailored dependable on the data mining (DM) context.

Medical domain is known for its ontological complexity and constraints in respect with medical data analysis and healthcare process computerization (Cios and Moore, 2002). According to Esfandiari et al. (Esfandiari et al., 2014), the application of DM in medicine lacks standards in the knowledge discovery process. The standards for data pre-processing could unify data gathering and integration, while standards for DM post-processing could unify the models deployment.

The uniqueness of DM and medicine is well analyzed and described in works of K. J. Cios and G. W. Moore (Cios and Moore, 2002), N. Esfandiari et al. (Esfandiari et al., 2014), R. D. Jr. Canlas (Canlas Jr, 2009), R. Belazzi and B. Zupan (Bellazzi and Zupan, 2008). However, there are few known attempts to provide a specialized DM methodolo-

gy or process model for applications in the medical domain. Spečkauskienė and Lukoševičius (Spečkauskienė and Lukoševičius, 2009a) proposed a generic workflow of handling medical DM applications. The 11 steps of the proposed process model presents an iterative approach of defining optimal data set, and finding the best performing DM algorithm by actual trial of each available algorithm; first, in its default configuration, and then changing its parameters. The proposed DM application method concentrates on the optimization of the initial dataset and trying of as much as possible DM algorithms. However, the authors do not cover other important aspects of practical DM application, such as data understanding, data preparation, mining non-structured data, and deployment of the modelling results.

Catley et. al. (Catley, et al., 2009) introduced a CRISP-DM extension for mining temporal medical data of multidimensional streaming data of Intensive Care Unit equipment. The authors provided an example of CRISP-DM activities mapping with the defined application domain, DM problem type, and technical aspect. As such, the results of the work will benefit the researchers of intensive care unit temporal data, but not directly applicable for other medical specialties, data types or DM application goals.

In this study, we propose a novel methodology, called CRISP-MED-DM, based on the CRISP-DM reference model and aimed to resolve the challenges of medical domain.

Overall, but specific to medical domain, DM application methodology would benefit multi-disciplinary process participants for better-aligned collaboration.

2. Cross-industry standard process for data mining

Cross Industry Standard Process for Data mining (CRISP-DM) is a general purpose methodology which is industry independent, technology neutral, and it is said to be de facto standard for DM (Azevedo and Santos, 2008; Chapman, et al., 2000). According to the online poll, conducted by the international DM community KDNuggets in 2014 (Piatetsky-Shapiro, 2014), CRISP-DM is the most referenced and used in practice DM methodology. CRISP-DM, alongside with SEMMA, is an informal methodology, since it does not provide the rigid framework, evaluation metrics, or correctness criteria. However, it provides the most complete toolset to the date for DM practitioners. The ultimate goal of the CRISP-DM founding parties was to create a non-proprietary and freely available standard process model for DM application engineering. The current version includes the methodology, reference model, and implementation user guide. The methodology defines phases, tasks, activities and deliverables outputs of these tasks.

As it shown in Fig. 1, CRISP-DM proposes an iterative process flow, with non-strictly defined loops between phases, and overall iterative cyclical nature of DM project itself. The outcome of each phase determines which phase has to be performed next. The six phases of CRISP-DM are as follows:

1. Business understanding. The preliminary phase highlights the understanding of the objectives of data analysis project and the converting of these requirements, from the perspective of the subject area, and the problem formulated into a definition of DM problem. In this phase it is determined the initial plan of achievement of goals, defining the success criteria.

2. Data understanding. This phase starts with the gathering of initial data and access to the dataset. The problems of data quality must be identified and are created the initial assumptions which datasets can be of interest for further steps.

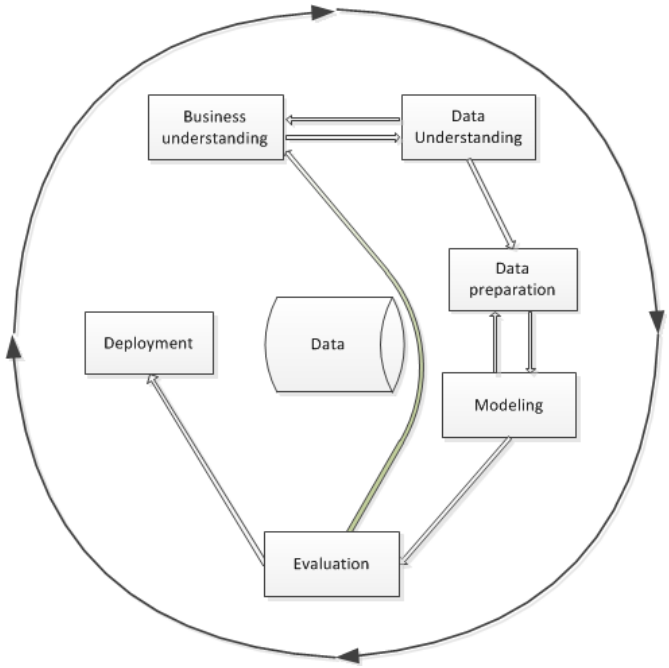

Fig. 1. Phases of the original CRISP-DM reference model

3. Data preparation. The data preparation phase covers all the activities that are required for pre-paring the final dataset. The activities of the data preparation phase heavily depend on the features and the quality of the original raw data. Some of the characteristic tasks of data preparation involve the choosing of table, attribute projections and record, attributes transformation, classification, normalization, noise elimination and sampling.

4. Modeling. In this phase, a suitable selection of modeling techniques, algorithms, or combinations thereof is done. Then, optimal algorithm parameters' values are chosen. Generally, for the same task, there are quite a few possible modeling methods available. Some of the methods have specific data quality constraints or data types. Consequently, this step is often performed in an iterative way until it is achieved the chosen model quality criteria. The model quality it is formally assessed. In order to evaluate the quality of the model, there are used metrics which are popular in DM and statistic: sensitivity, accuracy, specificity and ROC curve. Sensitivity – positive results properly classified as such in the results set. Accuracy – the percentage of properly classified objects. Sensitivity – positive results correctly classified as such in the results set. Specificity – negative results correctly classified as such in the results set. The relationship between sensitivity and specificity may be assessed with the help of ROC curve (Receiver Operating Characteristic) or a numerical expression of the area under the curve (AUC).

5. Evaluation. The evaluation phase has already a technically high-quality formed model (or several models). Prior to the final deployment of the model, it is essential to carefully evaluate it, to review the model construction steps, and make sure that business

objectives are properly achieved. The final result of this phase – the choice whether the DM results may be used in practical settings.

6. Deployment. The model generation is not the last step of the DM project. Despite the cases where the objective of DM project was to learn more about the data available, the acquired knowledge should be structured and presented to the end user in an understandable form. Depending on the set of requirements, the deployment phase may involve, for the simplest case, a report or deployment of repeated DM process. The prediction model resulting, using PMML modeling language can be saved and exported for additional use in healthcare management or clinical decision support systems. Often, it will be the end user, rather than the data analyst who will carry out the deployment activities. It is important that the end user anticipates the actions needed to be carried out in order to get the practical benefits of the generated DM model.

3. Uniqueness of data mining in medicine

The data mining and more generally knowledge discovery challenges in medical domain have been covered in works of R. Belazzi and B. Zupan (Bellazzi and Zupan, 2008), K.J. Cios and G.W. Moore (Cios and Moore, 2002), Canlas Jr, R. D. (Canlas Jr, 2009) and others. As the above mentioned authors emphasized, the practical application of DM in medicine meets a number of barriers: technological, interdisciplinary communication, ethics and protection of patient data. In addition, there are several well-known problems of biomedical data, such as inaccurate and fragmented information. The challenges of medical DM are described in Table 1.

Table 1. Challenges of data mining in medicine

Challenge	Description
Variety of data formats and representations.	Medical data lies in all sorts of data formats and representations. These formats include multi-relational structured data, video and image files, text files and others. Additional data pre-processing, feature extraction activities, or non-standard DM techniques are required to deal with those data. Multi-relational DM, text mining, inductive logic programing, and multi-media data pre-processing – are a few of them to mention.
Heterogeneous data	Analysis of data of several medical specialties raises additional challenges. In medicine, the same concept semantically may have multiple names and different identifiers in different code systems. Before applying DM algorithms, the data have to be integrated, and semantically unified. In the cases, when information systems use standard biomedical classifiers, nomenclatures, and ontologies, the semantic interoperability task is to define a common ontology. However, it is impossible when healthcare institutions use the extended, proprietary or regional versions of code systems, which are not identical to the international versions. In such cases, DM and medical informatics specialists have to create data transformation methods to ensure correct semantic data mapping. The problem of medical information systems interoperability also needs to be addressed. Frequently, departmental

	clinical information systems are not integrated. Medical informatics offers a range of interoperability standards. In theory, modern medical information systems have to support industrial medical data exchange standards, like HL7, HL7 CDA, DICOM, and to rely on inter-national classifiers. In practice, the situation can be opposite. According to the survey (Niakšu and Kurasova, 2012), medical information systems being used frequently do not support data exchange standards. Therefore, successful application of DM methods faces an additional challenge – integration of information systems. The integration of systems should be understood in a broad sense, ranging from data exchange architecture and ending with semantic data integrity.
Patient data privacy	The legislation protecting personal privacy prohibits the use of the patient's clinical in-formation without her consent. This complicates the use of clinical information for research purposes. This problem might be solved by automatic data depersonalization techniques (Vcelak, et al., 2012), which is done by separating clinical data from demographic data, which identify the patient. Datasets used for research must not include patient's name, passport or insurance ID numbers or other identifying attributes.
Clinical data quality and completeness	Another typical challenge in medical DM projects is variable quality of available medical data. Clinical data quality is affected by inaccurate measurements, human or equipment errors. For these reasons, it is essential to consider larger samples of clinical data, and to employ data preprocessing, where outliers can be identified and ruled out.

4. Extension of CRISP-DM data mining methodology for medical domain

A number of papers addressed the uniqueness of DM in health care (Cios and Moore, 2002; Canlas Jr, 2009; Bellazzi and Zupan, 2008). All of those papers suggested additional activities to be considered for effective knowledge discovery process in medicine and healthcare. According to our best knowledge, there is no specific and detailed framework for conducting DM analysis in medical domain.

As it was described in Section 2, CRISP-DM is a hierarchical process methodology, which provides an extendable framework. The methodology proposes 3rd and 4th abstraction layers for mapping generic models to specialized models. According to CRISP-DM classification, mapping for the future type of extension has be used to ensure specialization of the generic process model according to a pre-defined context for future systematic use.

Summarizing Section 3, the following issues shall be considered when applying DM in medical domain:

1. Mining non-static datasets: multi-relational, temporal and spatial data
2. Clinical information system interoperability
3. Semantic data interoperability

4. Ethical, social and personal data privacy constraints
5. Active engagement of clinicians in knowledge discovery process

In order to enhance CRISP-DM, specialized tasks and activities, which address the issues listed above, were introduced.

4.1. CRISP-MED-DM methodology

The CRISP-MED-DM specialized methodology reference model was developed. The changes to the each original CRISP-DM phase are described below. The full list of activities and deliverables is provided in Table 3.

Phases 1-2. Project scope definition.

The CRISP-DM phase 1 "Business understanding" and phase 2 "Data understanding" are the phases, where the DM project is being defined and conceptualized. The rest of the phases are implementation phases, which aim to resolve the tasks being set in the first phases. As in the original CRISP-DM, the implementation phases are highly incremental and iterative. However, the changes in Phase 1 or 2 lead to the change of project objectives and available resources. Therefore any significant change in these phases shall be regarded as an incremental project restart.

The first phase "Business understanding" was renamed to "Problem understanding" to avoid ambiguous meaning within two different perspectives, i.e. clinical application domain, and healthcare management application domain. In addition, the task "Define Objectives" has been split into "define clinical objectives" and "define healthcare management objectives". Addressing the issue of patient data privacy, a new activity under "Assess situation" was introduced: "Assess patient data privacy and legal constraints". Addressing the issue of heterogeneous data source systems, the activity of "Evaluate data sources and integrity" was added. The described tasks and activities are shown in Fig. 2.

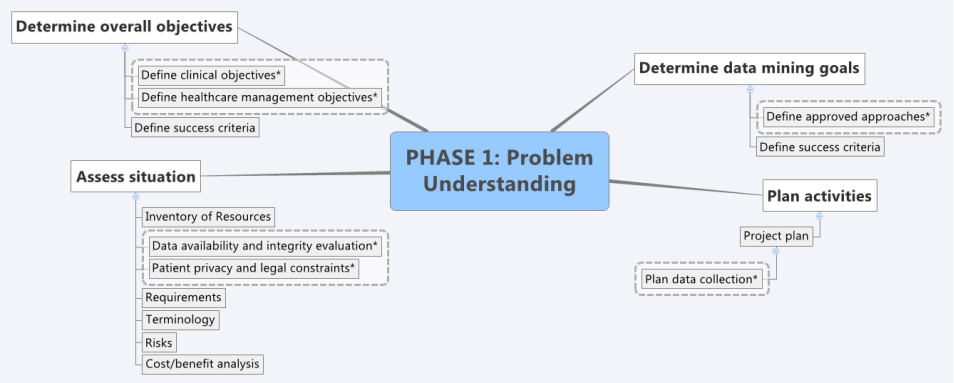

Fig. 2. CRISP-MED-DM 1st phase tasks and activities. Enhanced activities are marked with "*".

In the second phase "Data Understanding", a new general task "Prepare for data collection" was introduced. Issues of transport, semantic and functional interoperability have been considered in this activity. The wealth of medical data formats is considered

through the introduced activity of non-standard data pre-processing design, which includes support of multi-relational data, temporal, unstructured text and media data. Definition of medical nomenclatures, classifiers and ontologies used in data is substantial for further data pre-processing. Finally, definition and analysis of clinical data models and clinical protocols used in data source systems shall be carried out. The described activities are shown in Fig. 3.

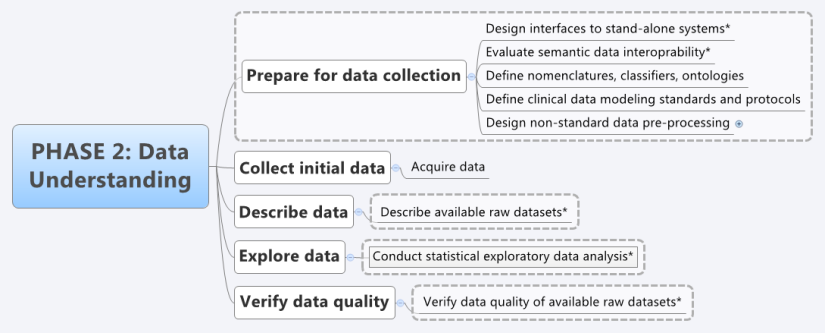

Fig. 3. CRISP-MED-DM 2nd phase tasks and activities. Enhanced activities are marked with "*".

Phase 3. Data preparation.

A vast body of experimental DM literature demonstrates that the most resource intensive step is data pre-processing. According to Q. Yang (Yang and Wu, 2006), up to 90 percent of the DM cost is in pre-processing (data integration, data cleaning, etc.). This is very true in medical domain as well.

The original CRISP-DM task "select data" had limitations for practical application in medical domain. First, it is mostly assumed for single-table static data format. Second, it lacks activities to handle data conversion and unification of the medical terminologies being used, lacks activities to integrate stand-alone medical information systems. The new general task "Prepare data" with the following activities was introduced:

- implement interfaces of stand-alone systems;
- prepare medical terminologies mapping;
- analyze and preprocess data from different sources, based on the agreed clinical data models and protocols.

In addition, a new general task "Extract data" was added to the process model. It includes the activities for unstructured data pre-processing, to facilitate feature extraction and prepare for DM modelling step. The activities of the task as follows:

- text data processing;
- media data processing:
  - image data processing;
  - video data processing;
  - audio data processing;
  - other signal data processing.

The original CRISP-DM task "Select data" was enhanced with Feature selection using statistical and DM techniques and data sampling activities. The activity stipulates the

usage of feature extraction and dimensionality reduction techniques to define possible attribute sets for modeling activities. Predictive DM methods requires separate training, validating and testing datasets, therefore data sampling activity was introduced.

Missing data is very common issue for clinical data. In addition, errors due to faulty sensors and laboratory and monitoring equipment interfaces shall be identified through outliers detection and semantic analysis. Automated semantic error analysis typically is based on business rules, implementing min/max checks, block lists, gender, and age dependency checks. These activities have been reflected under the general task of "Clean data".

Within "Data integration" task, activity of changing data abstraction level was added. This activity is required for temporal data. For example, intensive care units equipment may generate thousands of data items per second. Thus, methods of temporal abstraction have to be used prior to actual DM modelling activities.

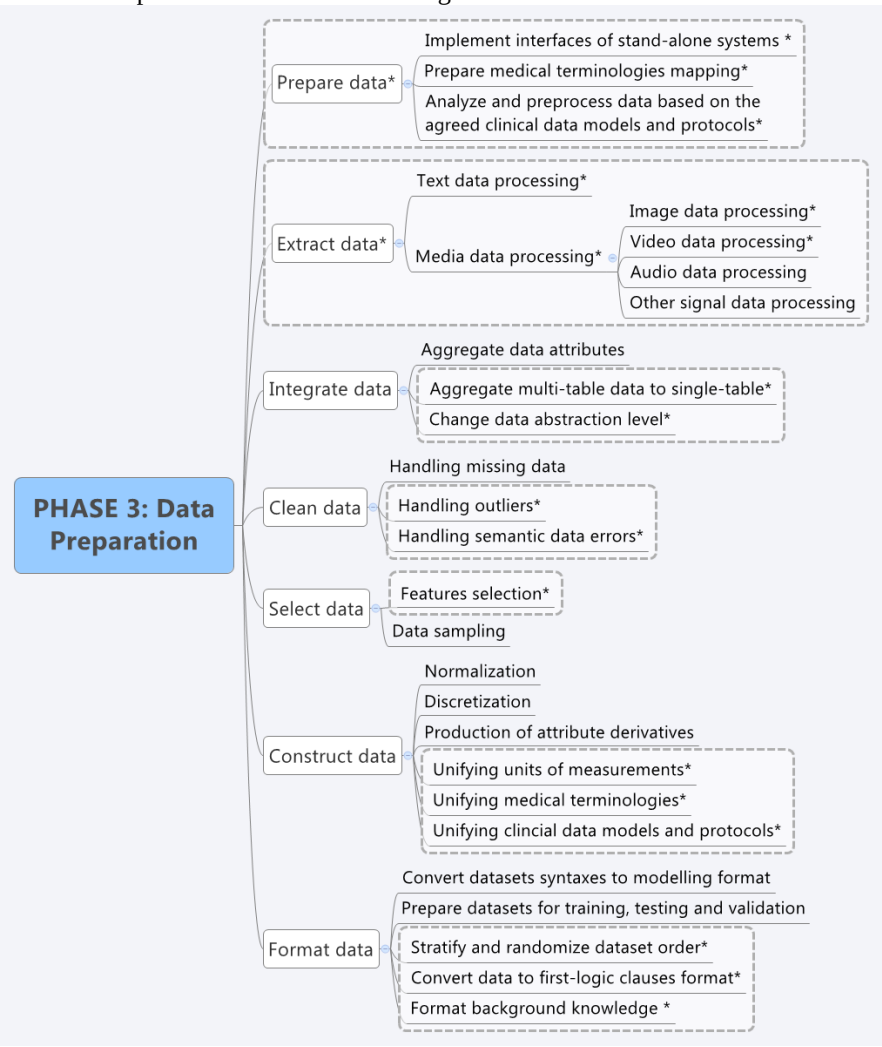

Fig. 4. CRISP-MED-DM 3rd phase tasks and activities. Enhanced activities are marked with "*".

Multi-relational data requires either propositioning of data to single-table format or will imply the use of multi-relational DM techniques, such as inductive logics programming (ILP). In the first case, conversion from multi-table to single-table must take place.

Finally, formatting data tasks, including data formatting for specific DM software environment, and complex conversions to first-logic predicates used in ILP. In addition, data stratification activity was added, because of its importance in predictive DM (Spečkauskienė and Lukoševičius, 2009). The described tasks and activities of the phase 3 are shown in Fig. 4.

Phase 4. Modelling.

According to CRISP-DM, Modelling phase is iterative and recursively returns back to the data preparation phase. In addition, there is iteration within Modelling phase between the task "Build Model" and "assess Model". However, the process flow of these iterations is not defined in the reference model and is not self-evident.

Spečkauskienė and Lukoševičius (Spečkauskienė and Lukoševičius, 2009b) proposed iterative 11-step DM process model, tailored for finding optimum modelling algorithm. Authors proposed the following flow:

1. To collect and access to a series of classification algorithms.
2. To analyze the dataset.
3. To sort out algorithms appropriate for the dataset.
4. To test the complete dataset using a selection of classification algorithms with the standard parameter values.
5. To select the best algorithms for further analysis.
6. To train the selected algorithms with a reduced dataset, eliminating attributes that have proven uninformative while constructing and visualizing decision trees.
7. To adjust standard values of the algorithms using the optimal set of data assembled for each algorithm of the most useful data identified in step 6.
8. To evaluate the results.
9. To mix-up the attribute values of the dataset in a random order.
10. To perform steps 6 and 7 with a new set of data.
11. To evaluate and compare the performance and efficiency of the algorithms.

This approach is resource intensive, but it can be automated by a specialized software support offered by the authors. The proposed method is based on greedy trial of all possible modeling algorithms and their parameters. This might be inefficient or even not feasible with big datasets, streaming data, or unstructured data. Thus, the findings of the authors were partially applied in CRISP-MED-DM. Particularly, iterative selection of a set of feasible modelling techniques, opposed to a few modelling techniques; iterative parameterizing of the selected modelling algorithms; and using predefined quality metrics to identify rejected, accepted, and best performing model (Fig. 5).

According to C. Catley, collaborative DM methods (e.g. method ensembles, method chains) may provide higher performance (Catley, et al., 2009). Accordingly, a new activity "Define optimum model or model ensemble" was introduced.

Finally, in order to prepare for the Deploying phase, the resulting models have to be prepared for the use in external decision support or scoring systems. One of the available possibilities is to export the resulting model or set of models in PMML format. The described tasks and activities of phase 4 are shown in Fig. 5.

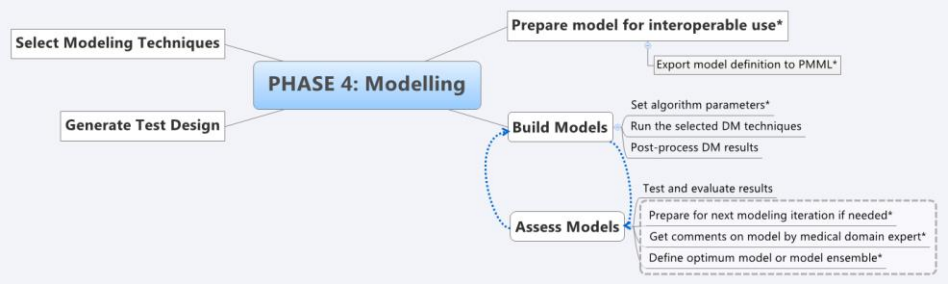

Fig. 5. CRISP-MED-DM 4th phase tasks and activities. Enhanced activities are marked with "*".

Phases 5-6. Evaluation and Deployment.

The activities of the original CRISP-DM Evaluation and Deployment phases are covering well medical domain and can be used for variety of projects and research objectives. Therefore, these phases remain with no significant changes.

Frequently, creating new predictive models for medical domain, the current golden standard exist, against which the outcomes of DM modelling shall be verified and cross-checked. Accordingly, the relevant activity was introduced (Fig. 6).

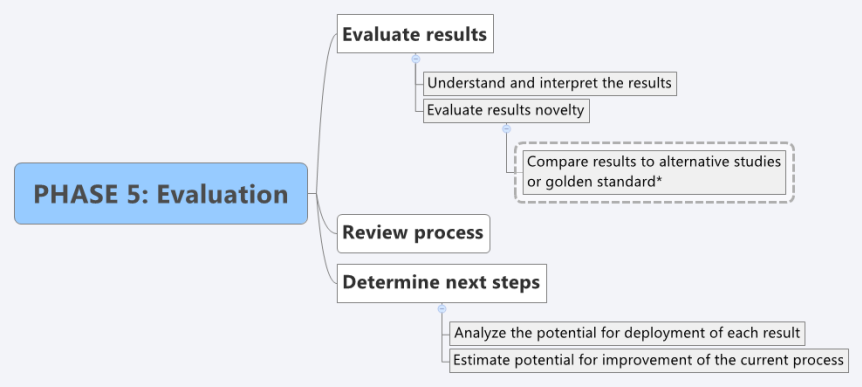

Fig. 6. CRISP-MED-DM 5th phase tasks and activities. Enhanced activities are marked with "*".

Deployment phase remains with no changes, as shown in Fig. 7.

Fig. 7. CRISP-MED-DM 6th phase tasks and activities.

The full list of general tasks, activities and associated deliverables of CRISP-MED-DM is outlined in Table 3.

5. Assessment of conformance to the CRISP-MED-DM

Assessing, monitoring and improving quality of DM processes requires not only well established process model, but also reliable and valid measurement and assessment models. A number of possible evaluation and assessment models respecting CRISP-MED-DM are defined for this purpose.

The goal to assess how compliant is the DM project to the methodology requires that the activities and their outcomes would be measurable. Measurement issues at this level may relate to specific process model activities or deliverables. However, regardless which process measurements are applied, they should support the quality objectives of the whole KDD process.

DM projects are very different with respect to DM goals and methods, data structure complexity, and data volume, thus, it is impossible to define a strict standard for methodology application's evaluation. Bearing that in mind, the proposed assessment model possesses certain flexibility.

The following assumptions are setting the common ground and eligibility for a KDD project, where CRISP-MED-DM methodology could be fruitfully applied and evaluated:

- The DM goals are well defined.
- Project participants have the domain and DM competences.
- Existing DM methods and algorithms will be used, and tools to apply them are available (creation of new DM algorithms or their extension is possible; however it remains beyond the scope of the methodology).
- Research data is legally and technically available to conduct a research.

5.1. Assessment and evaluation model

The DM application project evaluation strategy is proposed. It is based on the presumption that each phase of the process model has the same importance. Exception is made for the last phase "Deployment", which shall is treated as a utilization of the actual DM process results.

The CRISP-MED-DM activities and their related deliverables have different significance to the process: "the required", "required if applicable", "optional" and "conditionally required" - activities shall be distinguished. All but optional activities are valid metrics for quantified evaluation.

Table 2. CRISPM-MED-DM compliance evaluation method

Phase	Number of activities in phase	Activity evaluation points	Evaluation maximum points
Problem understanding	9	1.11	10
Data understanding	9	1.11	10
Data preparation	15	0.67	10
Modelling	9	1.11	10
Evaluation	3	3.33	10
Deployment	4	2.50	10

Each phase except Deployment phase is assigned with 10 commutative points, representing the maximum score achieved when all non-optional activities of CRISP-MED-DM have been successfully completed. Accordingly, each phase's non-optional activity is evaluated with maximum evaluation points divided by number of activities as stipulated in Table 2.

The list of CRISP-MED-DM tasks, activities, deliverables and metrics according to the 1st strategy is provided in Table 3.

5.2. Evaluation of measurement results

CRISP-DM and accordingly CRISM-MED-DM reference model includes many activities not related directly to DM process, but rather to the phases of KDD process, its management and organizational part. These activities are important for larger scale DM engagements, but could become an overhead in smaller ones.

Due to this reason, it is difficult to justify objective fixed threshold for meeting CRISP-MED-DM requirements. In the most conservative approach 100% of non-optional activities shall be performed. In a more flexible evaluation, the range could start from 60% for small projects and up to 90% for the complex ones.

The results of actual DM project's assessment using the proposed evaluation models provide comparable total project score, or scored CRISP-MED-DM phases, which can be visualized with Radar plot as shown in Fig. 8.

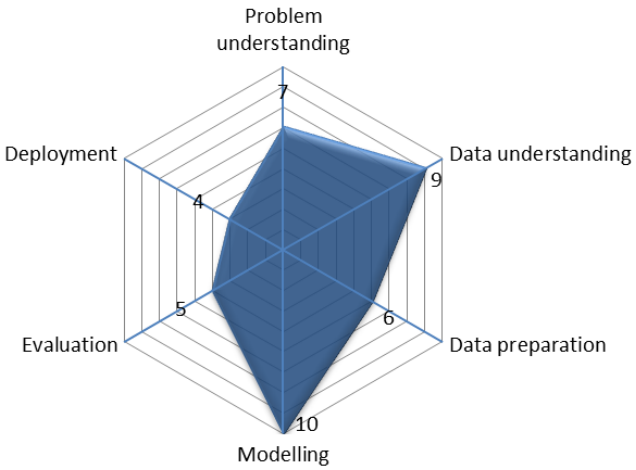

Fig.8. Example radar plot of DM project assessment

5.3. List of tasks, activities and deliverables

CRISP-DM defines a generic task as a task that holds across all possible data mining projects; a specialized task as a task that makes specific assumptions in specific DM context; and a deliverable as a tangible result of performing a task. The introduced

CRISP-MED-DM generic tasks and specialized tasks are marked with "*" and listed in Table 3.

Table 3. Tasks, activities and deliverables of CRISP-MED-DM

Notation: R - Activity is required, R2 – Activity is required if applicable, O - Activity is optional, C - Activity is conditional.

Generic tasks	Specialized tasks	Deliverables	Assessment
1 Phase: PROBLEM UNDERSTANDING			
GT1. Determine overall objectives	Define clinical objectives* Define healthcare management objectives*	Overall objectives	R
	Define success criteria	Overall success criteria or vision statement	R
GT2. Assess situation	Inventory of Resources	Project resource list	R
	Data availability and integrity evaluation*	List of data sources Data access evaluation	R
	Patient privacy and legal constraints*	Evaluation of legal requirements and limitations in data usage	R
	Requirements, assumptions and constraints	DM project resources, costs, timelines assessment	R
	Terminology	Glossary of multi-discipline relevant clinical and DM terminology	O
	Risks	Risks & Contingencies matrix	O
	Cost/benefit analysis	CBA statement or CBA report	O
GT3. Determine data mining goals	Define approved approaches (golden standard)*	Data mining goals	R
	Define success criteria	List or hierarchy of data mining success criteria	R
GT4.Plan activities	Project plan Plan data collection*	Overall plan Data collection plan	R
2 Phase: DATA UNDERSTANDING (total 10 points)			
GT5. Prepare for data collection *	Design required interfaces to the stand-alone systems*	Design the interfaces of IS involved	R2
	Evaluate semantic data interoperability*	Semantic interoperability analysis report	R
	Define nomenclatures, classifiers and ontologies used*	List of medical nomenclatures, classifiers and ontologies used	R
	Define clinical data modeling standards and protocols used*	Mapping of used clinical models, protocols	R2
	Design non-standard data pre-processing*	Prepare strategy and design for handling multi-relational, temporal,	R2

Generic tasks	Specialized tasks	Deliverables	Assessment
		non-structured data (media, text).	
GT6. Collect initial data	Acquire data	Initial data collection report	R
GT6. Describe data	Describe available data sources, and raw datasets*	Data model Clinical data meaning report	R
GT7. Explore data	Conduct statistical exploratory data analysis	Exploratory analysis report	R
GT8. Verify data quality	Verify data quality of available raw datasets*	Data quality report Medical expert data quality assessment	R
3 Phase: DATA PREPARATION			
GT9. Prepare data*	Implement interfaces of stand-alone systems *	Stand alone IS are interfaced	R2
	Prepare medical terminologies mapping*	Medical terminologies mapped	R2
	Analyze and preprocess data from different sources, based on the agreed clinical data models and protocols*	Clinical data models mapped	R2
GT10. Extract structural data*	Text data processing*	Preprocessed data, suitable for the planned text mining	C *if required by DM technique*
	Media data processing and feature extraction*: Image data processing Video data processing Audio data processing Other signal data processing	Dataset ready for further pre-processing and modelling	R2
GT11. Select Data	Features selection using statistical and DM techniques*	Selected features (attributes) for modelling	O
	Data sampling	Prepared data sample feasible for modelling	O
GT12. Clean data	Handling missing data	Data cleaning report Higher quality data set	C – *if required by DM technique*
	Handling outliers*		R2
	Handling semantic data errors*		R2
GT13. Construct data	Normalization	Constructed data	O
	Discretization		O
	Production of attribute derivatives		O

Generic tasks	Specialized tasks	Deliverables	Assessment
	Unifying medical terminologies in datasets *		R2
	Unifying units of measurement in datasets *		R2
	Unifying clinical data models and protocols in datasets *		R2
GT14. Integrate data	Aggregate multi-table data to single-table*	Aggregated, merged data	C – *if required by DM technique*
	Aggregate data attributes		O
	Change data abstraction level* (diagnosis; anatomic parts of body, systems)		O
GT15. Format data	Stratify, randomize datasets*	Balanced datasets ready for selected modelling algorithms	O
	Prepare datasets for model training, testing and validation	Training, testing and validation datasets ready	C- *if required by DM technique*
	Convert datasets syntaxes to modelling format	Datasets ready for the selected DM tools	R
	Convert data to first-logic clauses format*	Data in first-logic clauses format ready for ILP inference	C- *if required by DM technique*
	Format background knowledge *	Facts in first-logic clauses format ready for ILP inference	C- *if required by DM technique*
4 Phase: MODELLING (total 10 points)			
GT16. Select Modeling Technique	Select technique w.r.t.: Techniques appropriate for problem Understandability/interpretation requirements Constraints	Modeling Technique Modeling Assumptions	R
GT17. Generate Test Design	Generate model design w.r.t. testing and evaluation criteria Compare model design with DM goals	Test design	R
GT18. Build Model	Set algorithm parameters*	Parameter settings	R2
	Run the selected DM techniques	Models	R
	Post-process DM results	Ready for evaluation DM	R

Generic tasks	Specialized tasks	Deliverables	Assessment
		model results, e.g. trees, rules Model Description	
GT19. Assess Model	Test and evaluate results w.r.t. evaluation criteria and test design	Model assessment Revised Parameter settings Assessment	R
	Prepare for next modeling iteration if needed*	Revised parameter setting Alternative modelling technique	C- If DM goals are not achieved
	Define the best performing model or model ensemble* Get comments on model by medical domain expert	Best performing model Initial assessment of the model by domain expert	R
GT20. Prepare model for interoperable use*	Export model definition to PMML *	Prediction model in PMML standard	C – If model will be used in scoring IS
5 Phase: EVALUATION			
G21. Evaluate Results	Understand and interpret the results	Assessment w.r.t. Overall Success Criteria	R
	Evaluate results novelty Compare results to alternative studies*		R
G22. Review Process	Review of DM process: Identify failures, misleading steps, possible alternative actions	Review of Process	O
G23. Determine next steps	Analyze the potential for deployment of each result	List of possible actions and rationale for them	R
	Estimate potential for improvement of the current process		O
6 Phase: DEPLOYMENT			
G24. Plan Deployment	Summarize deployable results Develop alternative deployment plans Establish how the model will be deployed within organization's systems Identify possible problems	Deployment plan	C – If relevant

Generic tasks	Specialized tasks	Deliverables	Assessment
G25. Plan Monitoring and Maintenance	Decide how accuracy will be monitored Determine usage limitations and constraints of the result model Develop monitoring and maintenance plan	Maintenance plan	C – If relevant
G26. Produce Final Report	Develop set of final documentation, including executive summary, presentation, and detailed technical report.	Final report & Presentation	C – If relevant
G27. Review Project	Interview people involved in the project Summarize feedback Analyze the process retrospectively Document the lessons learned	Experience Documentation	O

6. Discussion and Conclusion

Researches and practicing data analysts face numerous challenges when applying DM techniques in medical domain. To our best knowledge, there is no well-defined process model or methodology, addressing the problems and constraints of medicine and healthcare. Therefore, a novel methodology called CRISP-MED-DM, based on Cross Industry Standard Process for Data Mining was developed.

The CRISP-MED-DM addresses the specific challenges and issues of DM application in medical domain. In the proposed extension of the industry standard CRISP-DM reference model, 38 generic and specialized tasks have been introduced. These tasks and their related deliverables are aimed to resolve the following issues:

1. Mining non-static datasets: multi-relational, temporal and spatial data
2. Clinical information system interoperability
3. Semantic data interoperability
4. Ethical, social and personal data privacy constraints
5. Active engagement of clinicians in knowledge discovery process

In order to assess compliance to the CRISP-MED-DM, the evaluation method is proposed. The method is flexible to support various levels of formalities, which may differ in small and large complexity DM projects. It gives comparative assessment and a baseline for the evaluation of DM and KDD projects.

Currently, the CRISP-MED-DM undergoes practical approbation in cardiology domain. The future work will include evaluation and critical analysis of the proposed specific activities and the evaluation method.

References

Azevedo, A., Santos, M. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European conference data mining, 182–185.
Retrieved from http://recipp.ipp.pt/handle/10400.22/136

Bellazzi, R., Zupan, B. (2008, Feb). Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform, 77(2), 81-97.
Retrieved from http://www.hubmed.org/display.cgi?uids=17188928

Canlas Jr, R. D. (2009). Data Mining in Healthcare: Current Applications and Issues. [MS in Information Technology thesis].

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. Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on, (pp. 1-5).Retrieved from
http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5255394

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. Retrieved from
https://www.researchgate.net/publication/225070403_CRISP-DM_1.0_Step-by-Step_Data_Mining_Guide

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

Esfandiari, N., Babavalian, M. R., Moghadam, A.-M. E., Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications, 41(9), 4434-4463.

Niakšu, O., Kurasova, O. (2012). Data Mining Applications in Healthcare Theory vs Practice. Local Proceedings of 10th International Baltic Conference on Databases and Information Systems, (pp. 58-70). Vilnius.

Piatetsky-Shapiro, G. (2014, 10). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved from
http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Spečkauskienė, V., Lukoševičius, A. (2009a). A data mining methodology with preprocessing steps. Information Technology and Control, 38(4), 319-324.

Spečkauskienė, V., Lukoševičius, A. (2009b). Methodology of adaptation of data mining methods for medical decision support: Case study. Electronics and Electrical Engineering, 2(90), 25-28.

Vcelak, P., Kratochvil, M., Kleckova, J., Rohan, V. (2012). MetaMed—Medical meta data extraction and manipulation tool used in the semantically interoperable research information system. Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference on, (pp. 1270-1274).

Yang, Q., Wu, X. (2006). 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making, 5(04), 597-604. Retrieved from http://www.worldscientific.com/doi/abs/10.1142/S0219622006002258

Received April 20, 2015, revised May 10, 2015, accepted May 26, 2015