VILNIUS UNIVERSITY

OLEGAS NIAKŠU

DEVELOPMENT AND APPLICATION OF DATA MINING METHODS IN MEDICAL DIAGNOSTICS AND HEALTHCARE MANAGEMENT

Doctoral Dissertation

Technological Sciences, Informatics Engineering (07 T)

Vilnius, 2015

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics of Vilnius University in 2009–2014.

Scientific Supervisor:

Assoc. Prof. Dr. Olga Kurasova (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

VILNIAUS UNIVERSITETAS

OLEGAS NIAKŠU

DUOMENŲ TYRYBOS METODŲ, SKIRTŲ MEDICININEI DIAGNOSTIKAI IR SVEIKATOS APSAUGOS VADYBAI, VYSTYMAS IR TAIKYMAS

Daktaro disertacija

Technologijos mokslai, informatikos inžinerija (07 T)

Vilnius, 2015

Disertacija rengta 2009–2014 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinis vadovas:

doc. dr. Olga Kurasova (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Acknowledgments

I would like to thank my doctoral dissertation scientific supervisor Assoc. Prof. Dr. Olga Kurasova for the support and guidance. I want to thank Prof. Habil. Dr. Gintautas Dzemyda for the motivation and support.

I greatly appreciate the time and effort of Prof. Dr. Paulius Miškinis for being a true example of a scientific researcher, and for his help and advice in my research.

Special thanks for collaboration and support to Prof. Dr. Elona Juozaitytė, Dr. Jurgita Gedminaitė, Dr. Giedrė Balčiūnaitė and Prof. Dr. Jonas Žaptorius.

I would like to thank reviewers Prof. Habil. Dr. Laimutis Telksnys and Dr. Viktor Medvedev for their valuable advice and effort, which helped to improve the quality of this work.

Finally, I wish to thank my family for their support and patience during the preparation of this thesis.

Olegas Niakšu

Abstract

The research area of the thesis is the application of data mining in healthcare and medicine. When applying data mining in medicine, additional problems such as varied information representation formats, semantic interoperability and patient privacy have to be resolved. The object of the dissertation research is the process and methods of data mining in medicine. The following topics are directly associated with this subject: medical data preprocessing methods, medical images processing, and multi-relational data mining. The key goal of the thesis is to develop and explore methodology for the application of data mining methods in medicine and healthcare, which would increase the efficiency of data analysis. Achieving this goal, the following tasks have been completed: analysis of the existing methodologies and process models, creation and trial of the data mining application methodology for the medical domain, developing the supporting diagnosing models and medical data processing methods.

In this thesis, a new application methodology CRISP-MED-DM was developed, which is based on the industry standard Cross-Industry Standard Process for the Data Mining. The CRISP-MED-DM was successfully applied for predictive modeling in cardiology and oncology domains. In addition, a new blood flow echocardiography image processing technique was developed, which enables semi-automation of aortic valve stenosis degree diagnostics. In addition, a new similarity measure for multi-relation clustering was proposed. The research results of the work revealed new opportunities in the application of data mining methods in the medical domain.

Reziumė

Disertacijos tyrimų sritis yra duomenų tyryba medicinoje ir sveikatos apsaugos sistemoje. Duomenų tyrybos metodų taikymus medicinoje lemia keletas papildomai spręstinų uždavinių, tokių kaip medicininės informacijos pateikčių įvairovė, semantinio sąveikumo problemos, pacientų duomenų apsaugos lemiami apribojimai. Pagrindinis darbo tyrimų objektas yra duomenų tyrybos taikymo medicinoje procesas. Su šiuo objektu tiesiogiai susiję specializuoti medicininių duomenų apdorojimo metodai, medicininių vaizdų apdorojimas, multireliacinė duomenų tyryba. Pagrindinis disertacijos tikslas yra sukurti ir įvertinti duomenų tyrybos taikymo metodiką, kuri leistų padidinti duomenų tyrybos medicinoje ir sveikatos apsaugos srityse efektyvumą. Siekiant šio tikslo buvo iškelti ir sprendžiami šie uždaviniai: egzistuojančių metodikų ištyrimas, specializuotos duomenų tyrybos metodikos sukūrimas ir aprobavimas sukuriant diagnostinius modelius ir reikalingus medicininių duomenų apdorojimo metodus.

Šioje disertacijoje sukurta Cross-Industry Standard Process for the Data Mining metodikos versija, pavadinta CRISP-MED-DM. Ji buvo sėkmingai panaudota taikant prognostinę duomenų tyrybą kardiologijoje ir onkologijoje. Pasiūlyta kraujo srauto echokardiografijos vaizdų apdorojimo metodika, kuri leidžia iš dalies automatizuoti aortos vožtuvo stenozės laipsnio diagnostiką. Taip pat sprendžiant multireliacinių duomenų klasterizavimo uždavinį buvo pasiūlyta nauja panašumo metrika. Darbe atliktų tyrimų rezultatai atskleidė naujas duomenų tyrybos metodų taikymo galimybes medicinoje.

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

AV	Aortic Valve
BRCA	Breast Cancer Susceptibility Gene
C4.5	Decision tree algorithm, which extends ID3 algorithm
CDA	The HL7 Clinical Document Architecture is a XML-based markup standard intended to specify the encoding, structure and semantics of clinical documents for exchange.
CLARANS	Clustering Algorithm based on Randomized Search - partitioning algorithm based on PAM for large datasets
DM	Data Mining
EHR	Electronic Health Record
FURIA	Fuzzy Unordered Rule Induction Classification Algorithm
HIS	Hospital Information System
HL7	Health Level Seven International is an ANSI-accredited standards developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information.
HPO	Healthcare Provider Organization
ICD-10	ICD-10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization.
ID3	Decision tree algorithm proposed by R. Quinlann
IHE	Integrating the Healthcare Enterprise is a initiative by the healthcare industry to improve the way computer systems share information
ILP	Inductive Logic Programming
IR	Information Retrieval
LIS	Laboratory Information System
LVOT	Left Valve Output Tract
MCDM	Multiple-criteria Decision Making
MESH	The medical subject headings is a controlled vocabulary, which is used for indexing, cataloging, and searching for biomedical and health-related information and documents.
MIS	Medical Information System
MPI	Master Patient Index is a software component that is used across a healthcare enterprise to maintain consistent index of essential patient data

MRDM	Multi-relational Data Mining
OR	Operational Research
PAM	Partitioning Around Medoids is a partitioning clustering algorithm
PDM	Priority distribution method
PIX	Patient Identifier Cross-Referencing is an IHE profile, which is used for the cross-referencing of patient identifiers from multiple Patient Identifier Domains
PMML	Predictive Model Markup Language
PSV	Peak Systolic Velocity
TED	Tree Edit Distance
TM	Text Mining
T-SQL	Transact SQL is a version of Structured Query Language used in MS SQL DBMS
UML	Unified Modeling Language
UMLS	Unified Medical Language System is a generalizing biomedical ontology
XDS	Cross-Enterprise Document Sharing IHE profile is used for managing the sharing of documents between any healthcare enterprise
XML	Extensible Markup Language

Introduction

Research Context and Motivation

The healthcare domain is known for its ontological complexity and variety of medical data standards and variable data quality (Cios & Moore, 2002; Chen, et al., 2006; Bodenreider, 2008; Esfandiari, et al., 2014). With the addition of patient data privacy issues, making an effective and practically usable medical knowledge discovery is of ongoing importance over recent decades. Modern clinical practices also undertake a transformation not only in diagnosis and treatment methods, but also in the understanding of health and illness concepts, moving from disease-oriented problem solving to a patient-centric approach, where computer-aided knowledge discovery methods play an important role (Rudnick, 2004).

Although data mining methods and tools have already been applied in various domains for more than 40 years, their applications in healthcare are relatively young. R.D. Wilson et al. (Wilson, et al., 2004) have started to classify and collect medical publications where knowledge discovery and data mining techniques were applied or researched from 1966 until 2002.

Starting from the twentieth century, many countries have chosen e-Health as a prioritized national program, which in essence proposes to benefit from the standardized aggregation of patients' clinical information and healthcare services rendered by providing instant access to this information for healthcare professionals as well as to patients themselves (Castro, 2009; Stroetmann, et al., 2011). According to the strategic plans of EU member states, the USA and many other nations from all continents, a considerable amount of investments are allocated to enable the global computerization of healthcare data. Taking a linear progression would propose that in 10 years all new medical encounters will be thoroughly digitalized, at least in the developed countries. For the first time in history, the research community is going to get a full set of a person's medical

history from the birth date until the decease date. This anticipated scenario forecasts a tremendous potential for machine learning and in particular for data mining applications in healthcare.

Problem Statement

The application of data mining in healthcare raises additional challenges which require specific methods, tools, and methodology. Moreover, cross-domain knowledge is of key importance to achieve practical results. The rapid progress in the computerization of the healthcare industry gave a vast amount of heterogeneous, both structured and unstructured, data available for research and secondary use. There are hundreds of algorithms implemented to classify, cluster, and find hidden patterns in data. However, domain specific issues of healthcare are still to be resolved. As it was discussed by Cios et al., Bellazi et al., Špečkauskienė et al. (Cios & Moore, 2002; Bellazzi & Zupan, 2008; Špečkauskienė & Lukoševičius, 2009), specific problems shall be resolved to successfully apply data mining methods. According to their studies, without resolving depersonalization, multi-relational and media data pre-processing, clinical data heterogeneity, and quality issues, data mining application is sub-optimal or impossible.

The surveys conducted by the data mining community KDNuggets (Piatetsky-Shapiro, 2014) in 2009 and 2014 have revealed the most widely used data mining application methodology is the Cross-Industry Standard Process for the Data Mining (CRISP-DM). However, due to its generic purpose CRISP-DM is not well suited for applications in the medical domain. Furthermore, a survey of university hospitals (Niakšu and Kurasova, 2012) has revealed that frequently data mining research projects remain theoretical, have no clinical follow-up, and rarely go beyond the institutions directly involved in the research. In order to apply data mining methods for clinical data, the researchers shall additionally resolve the problems related to patient privacy, semantic interoperability, heterogeneous data sources, and unstructured data presented in text or media formats.

Thus, there is a need for a methodology with a data mining process model to tackle the problems of the medical domain. Such methodology shall address the following issues: methodological application of predictive and explorative data mining methods for computer-aided medical diagnosis, unstructured data pre-processing, feature extraction, and multi-relational data mining.

Tasks and Objectives of the Research

The main goal of this thesis is to develop and evaluate a medical domain specific methodology for predictive and explorative data mining in medicine and healthcare. The methodology shall address the issues typical for data mining in medicine, by defining the activities and the deliverables to tackle them. In addition, an evaluation model is needed to provide the compliance assessment to the methodology.

In order to achieve this goal, the following objectives and corresponding tasks have been formulated:

1. To analyze the existing data mining application methodologies by investigating data mining as part of a knowledge discovery process model.
2. To propose a novel, specific to the medical domain, data mining application methodology, which resolves the issues of the existing methodologies.
3. To evaluate the proposed data mining methodology in several medical specialty domains by creating the required medical data, such as diagnostic images, multi-relational data, analysis and processing methods.
4. To propose a multi-relational clustering method implementation for mining data in a multi-relational format.

Practical Significance of the Results

The practical significance of the thesis is as follows:

- The proposed CRISP-MED-DM methodology facilitates a data mining process in the medical domain by proposing the improved reference model and the compliance evaluation method.
- The proposed *BRCA1* gene mutation prediction model can be used as a decision support tool, to indicate the gene mutation risk before an expensive genetic test is carried out.
- The proposed echocardiography image analysis and feature extraction method and its software implementation allows automating the labor-intensive manual systole tracing performed by cardiologists when assessing aortic stenosis. The created aortic valve stenosis predictive model can be used as a decision support tool.
- The proposed and implemented distance metric can be applied to any exploratory analysis problem in a multi-relational environment, which cannot be reduced to a "single-table" form without a significant loss of information. The developed software calculates the distance matrix for multi-relational objects.

Research Methods

The exploratory research and systematic literature review were used to analyze and apply the results of other research. Various methods of statistical analysis, operation research, data mining, and image processing techniques were applied. Experimental research was used to evaluate the proposed methods and compare them with alternative approaches.

Statements to be Defended

1. The data mining methodology CRISP-DM can be specialized and extended to improve data mining performance in the medical domain.

2. Applying CRISP-MED-DM to create the breast cancer susceptibility gene *BRCA1* prediction model improves the model's accuracy.
3. The application of CRISP-MED-DM with the novel echocardiography image transformation techniques results in a highly accurate aortic valve stenosis prediction model sufficient for aortic valve stenosis grading.
4. The partitioning clustering with the proposed multi-relational similarity measure is more precise in multi-relational settings where data generalization to one-table format leads to information loss.

Scientific Novelty and Results

The scientific novelty and results of the thesis are as follows:

1. A novel data mining application methodology CRISP-MED-DM is created. It defines tasks and activities to resolve the issues typical to the medical domain. Application of the CRISP-MED-DM allowed the *BRCA1* gene mutation risk prediction model's accuracy to improve from 0.88 to 0.94, sensitivity from 0.67 to 0.83.
2. A novel method for cardiology echocardiography image analysis, transformation and feature extraction is created, allowing the prediction of the aortic valve stenosis grade. The proposed method implements semi-automated systole cycle tracing and provides the cardiologists a time saving of up to two minutes per patient. The derived aortic valve stenosis predictive model has 100 % sensitivity and specificity for the research dataset.
3. A novel similarity (distance) measure for multi-relational data is created. The proposed metric when compared with a propositionalized dataset clustering and multi-relational clustering with RTED metric showed higher clustering accuracy with silhouette values of 0.21–0.31 against 0–0.16 (propositionalized) and 0.15–0.23 (RTED).

Approval of the Research

The main results of the thesis were presented and approved at the following scientific conferences.

International conferences:

1. Data mining applications in healthcare: research *vs* practice. *10th International Baltic Conference on Databases and Information Systems (DB&IS 2012)*, July 8-11, 2012, Vilnius, Lithuania;
2. Data mining approach to predict BRCA1 genes mutation.*16th International Scientific Conference of the Lithuanian Computer Society "Computer Days - 2013"*, September 19-21, 2013, Šiauliai, Lithuania.
3. Calculating distance measure for MRDM clustering. *16th Multi-conference on Information Society, Conference on Data Mining and Data Warehouses (SiKDD-2013),* October 7th, 2013, Ljubljana, Slovenia;
4. A systematic literature review of Data Mining applications in healthcare. *14th International Scientific Conference „Web Information Systems Engineering (WISE - 2013)*, October 13-15, 2013, Nanjing, China;
5. Applying Operational Research and Data Mining to Performance Based Medical Personnel Motivation System. *2nd Scientific Conference eHealth Summit Austria "Outcomes Research: The benefits of health IT"*, May 22-23, 2014, Vienna, Austria.
6. Mining Aortic Valve Stenosis Data using CRISP-MED-DM Methodology. *10th Annual South East European Doctoral Student Conference (DSC2015)*, September 18-19, Thessaloniki, Greece.

Regional conferences:

1. Challenges of data analysis and data mining in healthcare domain. *3rd National Young Scientists Conference of the Lithuanian OR Society "Operational research for business and social processes" (LOTD–2010)*, October 1st, 2010, Vilnius, Lithuania;

2. Application of multi-relational data mining in medicine. *1st Young Scientists Conference of Academy of Sciences of Lithuania „Interdisciplinary research in physical and technological sciences“*, February 8th, 2011. Vilnius, Lithuania, Award for the best paper;
3. Mathematical modelling of time-related blood velocity changes in human aorta. *54th Conference of Lithuanian Mathematical Society*, June 19-20, 2013, Vilnius, Lithuania;

List of Publications

Articles in the reviewed scientific periodical publications:

1. Niakšu, O.; Balčiūnaitė, G.; Kizlaitis, R. J.; Treigys P. Semi-automation of Doppler Spectrum Image Analysis for Grading Aortic Valve Stenosis Severity. *Methods of Information in Medicine.* 2015 (accepted), ISSN: 0026-1270 (IF: 2.248).
2. Niakšu, O. CRISP Data Mining Methodology Extension for Medical Domain. *Baltic Journal of Modern Computing.* 2015. Vol. 3, 2: 92-109, ISSN: 2255-8942.
3. Niakšu, O.; Gedminaitė, J. & Kurasova, O. Data mining approach to predict BRCA1 gene mutation, *Computational Science and Techniques*, 2013, vol. 1, 155-170, ISSN: 2029-9966.
4. Miškinis, P.; Niakšu, O.; & Valuntaitė, V. Mathematical Modelling of Time-Related Blood Velocity Changes in Human Aorta. *Laboratorinė medicina*. 2013, 15(4), 182 – 187, ISSN: 1392-6470.
5. Niakšu, O. Duomenų tyryba medicinoje: taikymas, problemos ir galimybės. *Visuomenės sveikata.* 2014, vol. 4(67), 9-19, ISSN: 1392-2696.
6. Niakšu, O., & Žaptorius, J. Applying operational research and data mining to performance based medical personnel motivation system. *Studies in health technology and informatics*, 2014, vol. 198, 63-70, IOS Press, Inc., ISSN: 0926-9630.
7. Niakšu, O.; Skinulytė, J. & Duhaze, H. G. A Systematic Literature Review of Data Mining Applications in Healthcare. *Workshop*

proceedings of Web Information Systems Engineering Conference – WISE 2013, Springer 2014 Lecture Notes in Computer Science, 2014, 313-324, ISBN 978-3-642-54369-2.

Articles in other peer-reviewed editions:

1. Niakšu, O. & Kurasova, O. Data Mining Applications in Healthcare: Research vs Practice, *Databases and Information Systems BalticDB&IS*, Local Proceedings, 2012, 58-70, ISSN: 1613-0073;
2. Niakšu, O. Calculating distance measure for MRDM clustering. *Proceedings of the 16th International Multi-conference "Information Society – IS 2013"*, 2013, vol. A, 192-194, ISBN: 978-961-264-066-8.

Outline of the Dissertation

The text of the thesis consists of 3 chapters, conclusions, references, list of publications and appendixes. Each chapter is provided with an introduction (except introduction and conclusions). The total scope of this thesis is 154 pages (without annexes), 44 figures, 22 tables.

Chapter 1 outlines in detail the issues of data mining in medicine and healthcare. In addition, the results of a literature analysis and university hospitals' survey are provided.

A novel process model for data mining and knowledge discovery in the medical domain is proposed in *Chapter 2*. Further, the theoretical part of the proposed process model's evaluation in the fields of Oncology and Cardiology is described. Moreover, a novel multi-relational clustering method, supporting the multi-relational nature of medical data, is proposed.

Chapter 3 provides experimental results of the proposed methods. The first two sections present use-cases of the applied methodology for predictive data mining in the Oncology and Cardiology domains. The third section illustrates the usage of multi-relational clustering.

The *Conclusions* section presents the main conclusions of the thesis. The *Annexes* section provides the outcome of data mining application in Cardiology domain — the aortic valve stenosis predictive model in PMML format.

CHAPTER 1

Data Mining in Healthcare and Medicine: overview and analysis

1.1. Introduction

The tendency in recent decades to computerize the process of disease treatment ensures a more rapid accumulation of medical information. Information technologies are actively used in the sector of health protection. National electronic health records systems and medical imaging archives are implemented all over the world. Health care institutions implement and deploy hospital information systems (HIS), radiological picture reviewing and archiving systems (PACS), laboratory information systems (LIS), and others. Medical information systems (hereinafter – MIS) accumulate a structured medical history of a patient which includes classified attributes, such as diagnosis, patient demographic data, vital functions, test results, and unstructured data, such as images and video files. Analysis and mining of this data are strategically significant to the health sector and important to each patient. An intellectual analysis of the accumulated data offers new instruments for the following tasks: faster patient diagnosis, selection of optimal treatment, prediction of treatment duration and its outcome, determination of complication risks, and optimization of healthcare facility resources.

Compared to other science and engineering disciplines, data mining

(DM) is in its infancy. Over the past decade, the application of DM in biomedicine has also been actively investigated. A noticeable increase in the number of publications and presentations at conferences indicates the relevance of this topic. Although it is not the first decade that methods of DM are being applied in medicine globally, practical application beyond research is still considered to be innovative and challenging.

In this chapter, a definition, tasks and methods of DM are described, with a focus on applicability in medicine. Section 1.4 gives a perspective of advanced DM techniques, such as multi-relational data, streaming data and text mining, with respect to the heterogeneous nature of medical data. Section 1.5 provides literature analysis on the uniqueness of DM in medicine. Section 1.6 outlines the outcomes of the field survey, on the practical usage of DM methods in university hospitals in a set of developed and emerging economy countries.

1.1.1. Literature Analysis Methods

The references were selected according to the following search criteria: “data mining in medicine”, “data mining in healthcare”, “biomedical data classification”, “biomedical data analysis”, “medical statistics”, “analysis of medical information systems data”, “medical ontology” as well as combining the terms listed.

A great part of this information was selected from the databases *ScienceDirect* and *MedLink*. Priority was given to publications after the year 2005. Earlier publications, which review the fundamental aspects of data analysis and mining, were also used.

Information from healthcare provider organizations was obtained while working in cooperation with the Vilnius University Hospital Santariškės Clinics, the Hospital of Lithuanian University of Health Sciences Kauno Clinics, the Klaipėda University Hospital, and university hospitals of Germany, Switzerland, South African Republic and Albania.

1.2. Defining Data Mining

While DM methods have been applied since the sixties, the term "data mining" first appeared circa 1990. There are numerous definitions of DM. U. Fayyad, G. Piatetsky-Shapiro et al. defined DM as an endeavor of finding useful patterns in data (Fayyad, et al., 1996) and complements the notion of the term knowledge discovery in databases (KDD). Since the nineties, the terms have been used interchangeably by statisticians, data analysts, and information systems experts. Encyclopedia Britannica defines DM as "the process of discovering interesting and useful patterns and relationships in large volumes of data" (Clifton, 2010) with the subfields of predictive modeling, descriptive modeling, pattern mining and anomalies mining.

The notion of DM used in this thesis is a part of the knowledge discovery process that uses data analysis methods such as statistics, machine learning and artificial intelligence to attain new non-trivial knowledge, e.g. prediction values, hidden patterns, and dependencies. According to this definition, the aim of DM is extracting new knowledge and deeper insight into a given dataset (often a large-scale dataset), which may continue to be used for decision making.

DM overlaps with other disciplines, particularly statistics. There is no strict demarcation line, as both disciplines partly employ the same methods. However, a few differences can be named. In practice, statistical methods are commonly used for primary data analysis, and DM – for secondary data analysis.

The most important differences are:

- In statistics, a formulated hypothesis is tested by means of statistical methods. DM allows using induction methods while formulating hypotheses from available data.
- Statistics usually examines a sample of the population. DM, on the contrary, often analyses data of the whole population.
- Formal mathematical methods are used in statistics and the use of imprecise heuristic methods is avoided. DM methods are based on mathematics; however, heuristic, local solution search and other approximate methods which focus on the tasks

containing large volumes of data, categorical variables or poor quality of the data under investigation are widely applied.

K. Waljee et al. (Waljee, 2013) differentiate predictive research from exploratory research in medicine. According to the authors, explanatory research typically applies statistical methods to validate an initially raised hypothesis, whereas predictive research applies statistical methods and data mining techniques without an *a priori* hypothesis. Approaching a problem without a specific predefined hypothesis, helps research from overlooking unexpected predictor attributes and may lead to less biased results.

The first international conference on the subject of DM was organized by ACM in the USA in 1995.The concept "data mining" was registered in the Medical Subject Headings (MeSH) term dictionary in 2009.

Application of the techniques related to DM in the medical field initially had a slow growing pace. However, recently an exponential amount of publications in the subject can be observed. The number of DM related publications has increased from five publications per year in the early nineties to around 879 in the year 2013.

Medical information can be expressed as static information, recording an instantaneous state of a patient, e.g. test results, diagnosis; dynamic – echocardiogram data, graphic – radiographs, three-dimensional graphic – computed tomography 3D models. Mining of these multiple data requires the adaptation of specialized methods, standards and tools ensuring interoperability, data warehousing, and more generally a detailed application methodology.

1.3. Data Mining Tasks and Methods

Depending on their purpose of application, DM methods are divided into two groups: methods for prediction and methods for data characterization. Characterization tasks are aimed at finding patterns and associations, while prediction tasks are meant to predict certain events or certain unknown values within the relevant sphere of interests. The main methodological difference is that prediction requires a specific variable (class) to be included into the primary

data. The solution can be numeric or categorical; respectively, DM methods for prediction are divided into regression and classification.

A full list of DM tasks has not been well established yet. However, typically, information sources (Chen, et al., 2006) distinguish the following tasks: classification, clustering, prediction, association analysis, visualization, and link analysis. DM methods can be divided into three main categories:

- supervised learning technique;
- unsupervised learning technique;
- other.

The first category of "supervised learning technique" includes the tasks of classification and prediction. The second category of "unsupervised learning technique" is assigned to the tasks of clustering and association rules mining. Visualization, outlier detection and link analysis are not classified as "supervised learning technique" or "unsupervised learning technique".

A list of the most popular DM methods and techniques according to a survey conducted by KDNuggets (Piatetsky-Shapiro, 2014) is as follows:

- decision trees and decision rules (classification);
- regression;
- clustering;
- descriptive statistics;
- visualization;
- link analysis;
- sequence mining;
- neural networks (classification);
- support vector machine;
- Bayes classification.

Solving DM tasks includes the selection of appropriate algorithms. Both the selection of DM method and algorithm, and parameterization of the optimal algorithm shall depend on the task objectives of the analysis and characteristics of the available data.

Over the past decade, a certain experience of applying DM methods in medicine has been gained. Esfandiari, Babavalian et al. (Esfandiari, et al., 2014) systematically reviewed publications addressing DM applications in medicine for structured data analysis. According to the authors, the most popular DM tasks and methods applied to the medical domain included classification methods: decision trees, neural networks, decision rules, SVM; clustering methods: k-means and hierarchical clustering; association mining: *A priori* association rules mining. A description of the most often used DM tasks is provided below.

According to Houston et al. (Houston, et al., 1999), for diagnostic purposes, neural networks, decision trees, decision rules are widely applicable. Methods of association rules mining are applied to the cost analysis (Silver, et al., 2001) and combinations of various prediction algorithms are widely used in order to predict the patient's condition and probability of recovery (Bellazzi & Zupan, 2008).

Classification methods are used to assign objects to the predetermined classes. The class role is played by a selected attribute in the data set determining an object. In statistics, the attribute is called the dependent variable. While classifying the objects, the algorithm creates a classification model, which can be further adapted to new data. For example, the diagnostic model of breast cancer developed by means of the classification algorithm may continue to be applied to the decision-making support system for the purpose of the diagnosis of a patient whose data were not used to create the prediction model. Classification is a two-step process consisting of training and testing steps. During the training step, the algorithm analyses the data meant for learning and creates a classification model. During testing, the accuracy of the model is checked using another data set. The most popular classification methods: decision trees, Bayes classifications and artificial neural networks.

Clustering is defined as an unsupervised learning technique. This means that in the process of clustering *a priori* knowledge of the group (cluster) the object it belongs to is unnecessary. By applying heuristic techniques, the clustering algorithm classifies objects into a default number of groups according

to the similarity of data. The similarity measure can be selected considering attributes characterizing the object. In order to evaluate the similarity of the objects, distance metrics are often used: Euclidean, Manhattan, Jacquard, etc. The most common clustering methods: hierarchical and partitional clustering.

The association rules mining method was proposed by Piatetsky-Shapiro in 1991. Sometimes called a market basket analysis method, it allows finding non-trivial patterns in the data. Association rules define the relationships between the data elements. A typical example: during the analysis of the market basket a pattern has been established showing that buyers who bought bread and butter also bought milk. The most common *a priori* algorithm provides two input parameters: rule support and confidence.

Association rule confidence is the proportion of the data set to which this rule applies. For example, an 80 % rule confidence would mean that 80 % of customers who bought bread and butter also bought milk. Association rule support is the proportion of data set which provides the condition for the rule. For example, 20 % rule support would mean that a total of 20 % of customers bought bread and butter.

1.3.1. Data Mining as a Part of Knowledge Discovery Process

The process of intellectual data analysis using DM methods is iterative (Azevedo & Lourenco, 2008). The DM community has proposed a variety of DM process models: CRISP-DM, introduced by consortium of private companies and supported by European Commission project ESPRIT (Chapman, et al., 2000), process model by U. Fayyad et al. (Fayyad, et al., 1996), process model by P. Cabena et al. (Cabena, et al., 1998), process model by K. J. Cios et al. (Cios & Moore, 2002), and SEMMA introduced by SAS Institute, Inc (Matignon, 2007).

The mostly frequently used and referred methodology for the DM process is CRISP-DM. It defines a process model, which decomposes DM into six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. The methodology provides each phase with an input, output, and strategy of execution. CRISP-DM treats the DM process as a

classic project, which has a defined goal and key constraints – time, resources and scope. Since the project already has a formulated goal, CRISP-DM does not focus on the task formulation. However, as emphasized by Baylis (Baylis, 1999), DM in medicine begins precisely from the correct formulation of the task, when clinicians along with data analysis experts formulate the problematic area and by analyzing the scope of activity and data available in medical information systems, formulate the problem as well as the technical task.

A detailed CRISP-DM description and the introduced novel extension CRISP-MED-DM for the medical domain are provided in Chapter 2.

While dealing with the tasks formed by healthcare professionals, the main objective of the DM specialist is to find and apply appropriate DM methods able to discover relationships among attributes, and to develop an appropriate model.

DM has a wide range of instruments and a few methods can be equally well suited for the same purpose. For this reason, it may be impractical to consider all alternative methods, and the choice of a particular method is determined not only by the results of objective analysis, but, as Bellazzi et al. stated (Bellazzi & Zupan, 2008), also by the intuition of the DM expert.

1.4. Advanced Data Mining Techniques

Most data mining methods have been developed to deal with a structured dataset with the attributes typed as numeric, nominal or bit data. The healthcare domain provides a wide choice of data types and representations. Medical information systems store structured data in relational databases and non-structured data, such as text, image and video files in binary object repositories. Following the adoption of recent interoperability standards, such as HL7 CDA and HL7 FHIR, there is a tendency to store structured clinical data in XML objects or plain XML text files. To mine data stored in those representations, non-standard DM techniques are required. In this section, multi-relational data mining, data stream mining, text mining, multimedia mining, and the intersection of DM and Operational Research are described.

1.4.1. Multi-relational Data Mining

In contrast to the traditional DM approach, multi-relational data mining (MRDM) looks for patterns in multiple tables (relations). MRDM methods include the relational association rule discovery, decision tree induction, clustering, regression, and other classification tasks. MRDM is successfully used in the area of bioinformatics, chemistry, marketing and others (Dehaspe, et al., 1998; Dzeroski, 2010).

In a straightforward approach of projecting a MRDM task into single-table DM task, loss of information or meaning can occur, but MRDM has the capacity to take into account background knowledge (domain knowledge represented in first-order statements). The DM approaches that discover patterns in a certain single table are referred to as attribute-values or propositional learning approaches (Dzeroski, 2010). A more comprehensive approach is to upgrade propositional DM algorithms to a multi-relational case. The example of such an upgrade for multi-relational clustering is presented in Sections 2.7 and 3.3.

Historically, the advances of MRDM are related to the techniques of Inductive logic programming (ILP), which are based on a subset of first-order predicate calculus (Dzeroski, 2010; Muggleton, 1991). The task of ILP can be defined as concept learning from positive and negative examples and background knowledge. Relational patterns are stated in more expressive language than the patterns that are defined on a single data table. Relational patterns are expressed in subsets of first-order logic, called differently as predicates or relational logic.

Most commonly, ILP has focused on learning in the normal (or strong or explanatory) ILP setting: given background knowledge B, a set of examples $E = P \cup N$, where P – postivie examples, N – negative examples. The objective is to find a hypothesis H, such that $\forall e \in P: B \wedge H \models e$ (posterior sufficiency) and $\forall e \in N: B \wedge H \not\models e$ (prior satisfiability), meaning H is complete and consistent with respect to the set of training examples P and N and given background knowledge B.

An example of the classification rule for predicting influenza, could be "IF 80 % of differential conditions = True THEN Influenza = True" (Table 1). This rule written in logical program syntax is as follows:

INFLUENZA(Pn, High Fever, Sore throat, Muscle pain, headache, eye pain, fatigue, dry cough) ← sum(positive_symptoms(Pn, High Fever, Sore throat, Muscle pain, headache, eye pain, fatigue, dry cough) > 5.6

Table 1. Patients influenza symptoms

ID	Sex	Age	Fever >38 C	Sore throat or running nose	Body aches and muscle pain	Headache	Pain when you move your eyes	Fatigue	Dry cough	Vomiting	Fever	Influenza
P1	M	5	Y	Y	N	N	Y	N	Y	Y	39	**Y**
P2	F	30	N	N	Y	N	N	N	Y	Y	37	**N**
P3	M	32	N	N	N	N	N	N	N	Y	37	**N**
P4	F	45	Y	Y	Y	N	N	N	N	Y	29.5	**Y**
P5	M	50	N	Y	N	N	N	Y	Y	N	37.5	**N**

Considering additional data available in relation "Family relationship" shown in Table 2, a more complex predictive rule, based on multi-relational data is induced:

INFLUENZA(Pn, High Fever, Sore throat, Muscle pain, headache, eye pain, fatigue, dry cough) ← sum(positive_symptoms(Pn, High Fever, Sore throat, Muscle pain, headache, eye pain, fatigue, dry cough) > 5.6 ∪ (Familly_members(Pn, Pk) ∩ INFLUENZA(Pk) = True)

Table 2. Next of kin relationship

ID1	ID2	Relation type
p1	p2	Son
p1	p3	Son
p2	p3	Wife
p4	p5	Husband

The example illustrates in simple terms the potential of multi-relational

inference. On the other hand, aggregating multiple tables through joins or generalization, can cause information loss, which was discussed in detail by Kramer, Lavrac et al. and Knobbe, Haas et al. (Kramer, et al., 2001; Knobbe, et al., 2001).

Van Laer and De Raedt (Van Laer & De Raedt, 2001) presented a generic approach for upgrading propositional algorithms to relational ones. The idea is to upgrade only the key notions, while keeping as much of the original algorithm as possible. For partitioning clustering, the key notions are distance measures, for rule induction – the refinement operator. Taking this approach, the MRDM algorithm represents a special case of a propositional algorithm for a multi-relational environment. Sections 2.7 and 3.3 illustrate this approach in more detail.

Extending the key notions to multi-relational data (e.g., defining distance measures for portioning clustering methods) requires considerable insight and creativity (Kramer, et al., 2001). Efficiency concerns are also very important, as it is often the case that even testing a given relational pattern for validity is computationally expensive, let alone searching a space of such patterns for valid ones.

1.4.2. Intersection of Data Mining and Operations Research

The ultimate goal of Operations Research (OR) is to optimally solve decision problems (Olafsson, et al., 2008). In order to induce optimal decisions, we need to understand the structure of the application system by modelling it and by providing algorithmic solutions for deriving decisions. OR, like DM, is a multi-disciplinary field. It employs mathematical modelling, optimization, statistical analysis, and computer science. As a formal discipline OR is older than DM, with the onset in the 1950s. However, the initial works of C. Babbage (Babbage, 1832) on transport cost optimization are originally dated back to the 1840s.

There is a growing interest in the intersection of OR and DM. A number of publications, revealing successful application and integration of both approaches highlight the benefits from the integration (Corne, et al., 2012;

Meisel & Mattfeld, 2010; Olafsson, et al., 2008).

DM can be perceived as an extension of the OR problem solving methodology, which is supported by the research of DM algorithms within the OR community (Olafsson, et al., 2008; Smith & Gupta, 2000). The practical cooperation of OR and DM usage is visualized in Fig. 1 . OR uses information about the structure as an input and delivers a derived decision. DM compliments this process by delivering the information, given the application data. According to S. Meisel (Meisel & Mattfeld, 2010), there are three principle types of synergies, where OR and DM can benefit each other. Under the first category is OR increasing DM efficiency; under the second, DM increasing OR effectiveness by replacement, and finally DM increasing OR effectiveness by refinement.

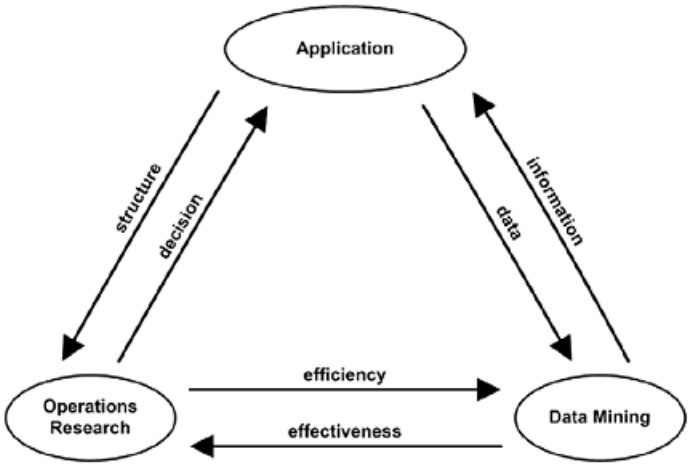


Fig. 1. Synergy of Operation research and DM application (Meisel & Mattfeld, 2010)

First, the efficiency of certain steps of the DM process can be increased by employing OR methods. OR can be considered for the optimum pre-processing method or a method to algorithmically optimize DM modelling. Second, DM is merely replacing OR. Finally, the last, and most promising method is refining OR with the support of DM. This method results in more effective decision making by refining initial decision models, by employing DM. In a closed-loop system, the information provided by DM is used for decision solution refinement with respect to measurement and decision attributes.

Use-case application of this method, where DM is not replacing OR, but

instead, it is integrated into the broader process model is described by Niakšu and Žaptorius (Niakšu & Žaptorius, 2014), where the authors proposed a method for creating a performance based motivation system for healthcare provider organizations personnel. The method aims to ensure cost-effectiveness, employee motivation and social balance. A multi-criteria decision support method is proposed for healthcare provider organizations performance related remuneration model creation. In addition, DM methods for the determination of performance indicators to be used, and for the subsequent monitoring of the achieved results, are proposed.

1.4.3. Text Mining

Text-mining (TM), known also as Knowledge Discovery from Text (KDT), is considered the method of extracting interesting patterns from a large text database for discovering knowledge (Dorre, et al., 1999). Text-mining shares the same analytical functions as data-mining, and also applies analytic functions from natural language (NL) and information retrieval (IR) techniques (Hotho, et al., 2005).

A Text Mining system is composed of three major components:

a) *Information Feeders* connects to any web site, streamed source or internal document collections and enables the relationship between different textual collections and the tagging modules.
b) *Intelligent Tagging* is responsible for text reading and selecting relevant information. It applies to any type of tagging on the documents such as statistical tagging, semantic or structural tagging.
c) *Business Intelligence Suite* consolidates the information from disparate sources, allowing for simultaneous analysis of the entire information background.

According to their tasks, TM can be divided into two major categories, algorithms and formal frameworks. The algorithms are task-oriented pre-processing approaches that visualize the process of creating a structured

document representation. This involves a preparatory target or problem that needs to be solved. The formal frameworks are pre-processing approaches that rely on methods deriving from analyzing complex phenomena. This can also be applied to natural language texts. Such approaches involve classification schemes, probabilistic models, rule-based system approaches, and other methodologies.

There is a significant difference between Text Mining (TM) and Text Retrieval (TR) or Information Retrieval (IR). TM discovers new information from text, through searching for patterns across datasets. The results of the TM process are patterns, connections, profiles or trends, and in order to find the information we do not necessarily have to read the documents. On the contrary, IR helps users find documents that go with their information necessities. The outcome of the information retrieval process is the documents, and in order to understand it, we still need to read the documents.

Known applications of TM and IR in the medical domain are the summarization and tagging of unstructured clinical documentation, e.g. referral letter, discharge summary, radiology report.

1.4.4. Multimedia Mining

Multimedia DM is an emerging active research area, where the necessity for finding tools to extract hidden useful knowledge that is embedded within the multimedia collections is essential for many applications. A vast amount of clinical data is represented in image, video or audio format, which can be used for automated information retrieval.

In a database system, there is always a database management system to administer all the data in the database. However, when the data are unstructured, we do not have a management system, but a collection of multimedia data. Multimedia mining methods are used to index unstructured data, by retrieving its descriptive information. Thus, the multimedia indexing and retrieval involves the indexing and retrieval of a single, non-text modality of data, such as an image, video, or audio.

An echocardiography image processing and mining use case is described in Sections 2.6 and 3.2.

1.4.5. Data Streams Mining

Traditionally, DM is concerned with static data, residing in databases and repositories. However, the emergence of continuous electronic data from sensors, sensor networks, web logs and others requires a different DM approach. Such data are called data streams. Linear and sub linear techniques, producing approximated DM results have been proposed to handle streaming data, and the following techniques have emerged: projection, sampling, group testing, tree method, and robust approximation. According to M. M. Gaber et al. (Gaber, et al., 2005), notable data stream mining tasks are as follows:

- *Managing the continual flow of data streams:* The flow of data in data streams is of a continuous nature. As a result, it requires novel management and analysis techniques that can deal with the constant, rapid flow of data elements. Examples of those are mining of Intensive Care unit data (Catley, et al., 2009; Ramon, et al., 2007).
- *Unrepressed memory requirements*: Data stream sources require the uninterrupted flow of data. Sensors and handheld devices do not have sufficient memory to run traditional DM techniques which require the results be continuous in the memory during the time of data processing. An example of those is mining data from home care systems, which aggregate data from the medical devices and sensors used by patients in their daily environment.
- *Altering detection and modeling of mining results over time*: Keeping track of the change of DM results is considered to be more important in this field rather than the DM results alone. Considering the high volume of data coming from various sources, the modeling of the change is not trivial.

Mining of streaming data utilizes different solutions using well-

established computational and statistical approaches. These solutions are categorized into task-based and data-based ones (Aggarwal, 2007). The data-based solutions search for only a subset of the whole dataset or convert the data horizontally or vertically into a smaller size data representation, whereas the task-based solutions utilize techniques from computational theory in order to reach the most efficient solution.

The data-based techniques consist of gathering the whole dataset or selecting a subset of the incoming stream for analysis. The most popular data-based data stream mining techniques are as follows:

- *Sampling*: choosing a subset of a dataset for analysis using probability theory.
- *Load Shedding*: overlooking a continuous amount of streaming data.
- *Sketching*: indiscriminate projection of a set of features to be analyzed.
- *Synopsis-Data Structure*: transformation of the incoming stream into a compressed form.
- *Aggregation*: calculating statistical measures that capture the characteristics of data.

The task-based techniques are based on existing DM algorithms, upgrading them to address the computational challenges of data stream processing. The most popular task-based data stream mining techniques are as follows:

- *Approximation algorithms*. Data stream mining is assumed to be a hard computational problem, and an approximate solution with the error bounds is proposed. Approximation for frequent pattern discovery in a data stream have been used in Pragarauskaitė's work (Pragarauskaitė, et al., 2013).
- *Sliding window techniques*. The key idea is to concentrate data analysis on the most recent data (window) and a summarized version of the old ones.

1.5. Application of Data Mining in Medicine and Healthcare

1.5.1. Historical Perspective

Collection of medical information and its routine statistical analysis has been carried out since the middle ages. The first known medical publication analyzing medical statistics was published in 1662 in London (Graunt, 1939). In 1863, F. Nightingale (Nightingale, 1863), a pioneer of modern medical care, had complained in her notes about the lack of health records and their desultory storage in hospitals, which used to impede analysis of treatment cost and effectiveness. In 1977, the Congress of the United States published a study "Policy Implications of Medical Information Systems" (Office of Technology Assesment. Congress of the United States., 1977). The study proposed that medical information systems could be used for educational purposes to help medical professionals in the provision of health services and increase the efficiency of treatment and HPO (health provider organization) activity. The study authors stated that eventually such systems would provide information and knowledge that was previously inaccessible to researchers and healthcare governing bodies. Since the year 2000, regional and national electronic health records systems, which aim to collect all relevant patient medical records, have been actively deployed worldwide.

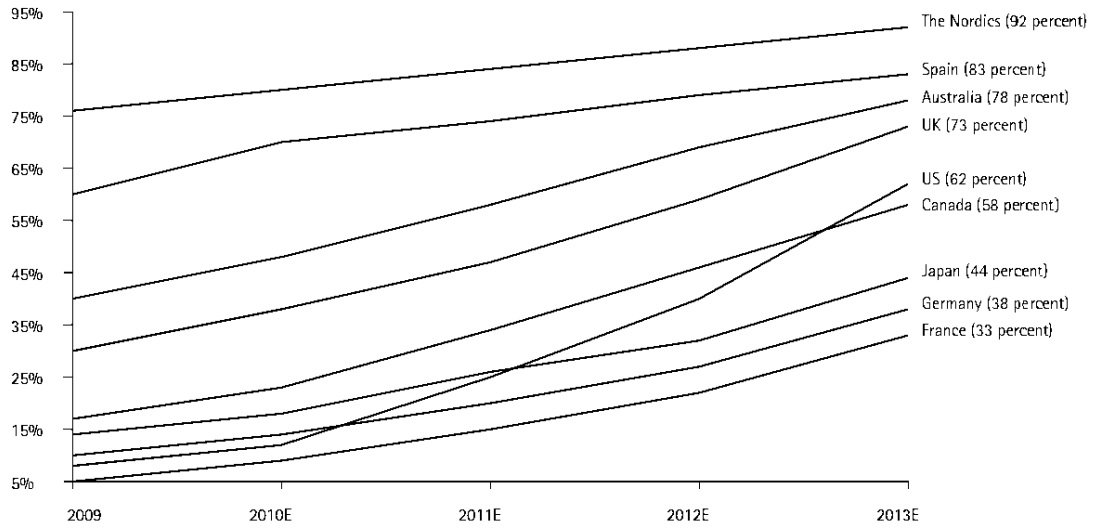


Fig. 2. Estimated hospital based EHR adoption rate (Accenture, 2010)

According to an Accenture market study (Accenture, 2010), the annual EHR and EMR growth is estimated by 5–8 % per year. The estimated hospital-based Electronic Medical Records systems adoption is shown in Fig. 2.

1.5.2. Aims and Objectives of DM in Medical Domain

As it was shown, the importance of stored medical information was understood a few centuries ago. When data is collected electronically, in addition to the standard calculation of statistical indicators inherent in healthcare, healthcare can benefit from DM methods.

As stated by Baylis (Baylis, 1999), the key to the successful mining of medical data is the correct identification of HPO activity or clinical problem. In the literature overview works of Belazzi and Zupan, Cios and Moore (Bellazzi & Zupan, 2008; Cios & Moore, 2002), the DM methods usually perform biomedical data regression, clustering, classification and visualization tasks, in order to facilitate decision-making for healthcare professionals. According to the most recent and sound systematic literature overview, performed by Esfandiari et al. (Esfandiari, et al., 2014), four main application areas of DM application in medicine can be defined:

1. Increasing the efficiency and elimination of the human factor: deals with tasks for diagnosis of certain diseases where accuracy is essential.
2. Reduction of time and cost: applicable when conventional diagnostic methods take a long time or are very expensive.
3. Medical decision support system: uses a multi-process automation, e.g. prediction models and expert systems; applied as assistance for less experienced or lower-skilled medical staff.
4. Knowledge extraction: used for the extraction of new knowledge or hypotheses.

The authors concluded that DM application studies aiming for efficiency improvement, hidden knowledge extraction, and medical decision support are equally popular, scoring 27–28 % each.

Fewer dissemination gained studies are devoted to resource optimization (reduction of time and cost). An example of DM and Operations Research methods' application for healthcare management tasks is described in the study of Niakšu and Žaptorius (Niakšu & Žaptorius, 2014), where the authors proposed a method for creating a performance based motivation system for healthcare provider organizations personnel.

Another approach to determine the objectives of DM in healthcare is to relate them to the process of patient treatment, as the decision supporting activities. The clinical treatment process model as described by J. M. Wehlou (Wehlou, 2014) includes a patient's history assessment, diagnostic activities, actual treatment, and patient's condition follow-up (Fig. 3).

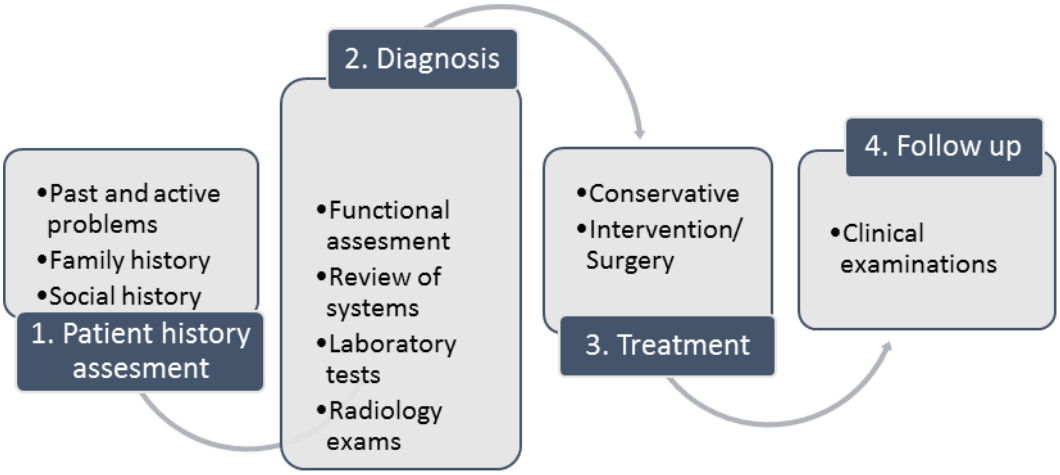


Fig. 3. Clinical process model. The 3rd and the 4th phases are iterative.

In addition, standalone medical tasks of screening, monitoring, and chronic disease management can be defined. Horizontal managerial tasks, such as resource optimization, quality control, adherence to the protocols and treatment plans are also to be considered.

Generalizing both approaches, the objectives of DM usage in medicine can be generalized into two main groups: treatment resources optimization (healthcare management domain), and treatment quality improvement (medical treatment and research domains). We present objectives of DM applications in the medical domain grouped by the goals in Table 3.

Table 3. Goals and objectives of DM in medicine

Goals	Objectives
Optimization of treatment resources	• Identification of opportunities for potential cost reduction and revenue enhancement (Silver, et al., 2001); • Dependence of length of stay (LOS) at a hospital on the patient‘s demographic data, anamnesis, selected method of treatment, and other factors (Baylis, 1999; Yeh, et al., 2011); • Prediction of rehospitalisation; • Prediction of postoperative complications and their probabilities; • Prediction of medical staff efficiency indicators (Niakšu & Žaptorius, 2014); • Identification of unnecessary medical interventions; • Identification of improper prescriptions
Improvement of treatment quality	• Early diagnosis of diseases (screening) • Evaluation of probable complications; • Modelling the progress of disease (Tanwani, et al., 2009); • Determining associations of specific clinical attributes in order to adjust the diagnosis or select a plan of treatment; • Generalizing multidimensional biomedical data recorded in real-time in order to facilitate decision-making (Stacey & McGregor, 2007); • Analysis of the quality of biomedical datasets (Bellazzi & Zupan, 2008): • Determination of dataset completeness; • Determination of dataset fragmentation; • Diagnosis setting or adjustment of diagnosis; • Formation of medical expert systems database; • Prediction of medication effectiveness; • Micro-array analysis for task solving; • Early diagnosis of diseases; • Selection of individual treatment; • Determination of the probability of disease occurrence.

1.5.3. Issues and Challenges of Data Mining in Medicine and Healthcare

The DM and more general knowledge discovery problems in medicine have been well covered by the works of R. Belazzi & B. Zupan, K. J. Cios & G.W. Moore, I. Yoo et al. and others (Bellazzi & Zupan, 2008; Canlas Jr, 2009; Cios & Moore, 2002; Koh & Tan, 2005; Yoo, et al., 2012). The domain specific problems distinguish the process of DM in medicine and set it apart from other subject areas. R. Bellazi and B. Zupan stated that if DM were a simple process, the problems of information management would have been solved long ago.

As the listed 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. Examples of inaccurate information: measurements of vital functions were performed when the patient was not in a rest position; test sample required for testing was taken in non-sterile conditions; lab equipment calibration errors. Fragmented information includes cases when the available patient data is non-sufficient for definitive results.

Another unique characteristic of DM in medicine is the usage in making decisions critical to human life. Therefore, as K. Cios and W. Moore noted (Cios & Moore, 2002), the results of a selected DM method must be descriptive, i.e. presented with explanations, so that medical experts can understand how these results were obtained. In terms of explicitness, some DM methods, such as decision trees, are more preferable than others, e.g. neural networks.

Analysis of the data within the framework of several medical specialties raises additional challenges. In medicine, semantically the same concept may have multiple names and different identifiers in different code systems. Let us consider a hypothetical example. The department of anatomical pathology in a hospital uses Anatomical Pathology Laboratory Information Systems, in which SNOMED-CT nomenclature is used. While the cardiology department uses a cardiology information system, which has the ICD-10 and ICD-PT

classifications installed and expanded according to the needs of the cardiologists. In addition, cardiologists are the users of radiological systems. Radiological systems are computerized but not integrated with the cardiology system. Radiology images and patient data are stored in the DICOM format, whereas for disease classification the ICD-10 code system is used. The prediction task requires the patient's clinical information, combining information accumulated in all the related departments. At this point we are facing a data interoperability problem. Before applying DM algorithms, the data have to be semantically mapped using common ontology. In order to use radiology test results, it is necessary to use computerized image processing algorithms, and possibly to apply text mining for the textual description of the findings, results and complications sections in the unstructured parts of clinical documentation.

In the cases where information systems use standard biomedical classifiers, nomenclatures, and ontologies, the semantic interoperability task projects to the definition of a common ontology. However, it is impossible, when healthcare institutions use the extended, proprietary or regional versions of classifiers, 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.

Returning back to the example of a hypothetical hospital, the problem of medical information systems' interoperability also needs to be addressed. As the information systems of cardiologists, pathologists and radiologists are not integrated, the integration of these systems is required. Medical informatics offers a range of interoperability standards. Lithuanian eHealth Strategy stipulates medical data exchange standard HL7 version 3 for this purpose. However, the standard has not been applied in practice yet. In theory, modern medical information systems have to support industrial medical data exchange standards and profiles, like HL7, HL7 CDA, DICOM, IHE PIX, IHE XDS and to rely on international classifiers and nomenclature. In practice, the situation can be opposite. According to our survey (Niakšu & Kurasova, 2012), the medical information system being used in Lithuania does 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.

The interoperability issues become more complex while integrating systems and data from different countries. The researchers and practitioners are provided with a set of overlapping standards, some of which are more common in Europe and others in the United States and Australia.

Other common issues of DM in medicine are ethics and patient confidentiality. The legislation protecting personal privacy prohibits the use of patients' clinical information without their consent. This complicates the use of clinical information for research purposes. This problem is solved by data depersonalization (Nitzlnader & Schreier, 2014). This is done by separating clinical data from demographic data identifying the patient. Datasets used for research must not include patient's name, passport or insurance ID numbers or other identifying attributes.

In some countries, where legislation of equal opportunities is fully developed, ethical problems are not limited to data depersonalization. For instance, the decisions of the United States for the provision of services cannot be made in accordance with the criteria of race, gender or age. As these demographic characteristics of the patient are very important and are often used in DM research projects, the application of the research results, which had included such data, may become problematic.

Another widely spread issue is the incorrect and fragmented information problem. The selection of efficient algorithms and construction of accurate prediction models with high specificity and sensitivity requires assessing the completeness and reliability of clinical data. Overall data quality is affected by inaccurate measurements, human or equipment errors. For these reasons, it is essential to consider larger samples of clinical data. Thus, by ignoring atypical data outliers, the impact of imprecise data on the results of the analysis is reduced.

Most of the problems in the medical domain correlate to the generic issues and challenges of the KDD process described in the seminal work of Fayaad et al. (Fayyad, et al., 1996). By analyzing and comparing them to the problems and issues outlined above, a structural mapping of Fayaad's generic issue list and specific medical domain problems is provided in Table 4.

Table 4. Mapping Fayaad's and medical domain issues and challenges

General Fayaad's DM issues and challenges	**Medical domain specific issues and challenges**
• Massive datasets and high dimensionality	• Medical data interoperability • Incorrect and fragmented information problems
• User interaction and prior knowledge	• Patient data privacy
• Overfitting and assessing statistical significance	
• Missing data	• Incorrect and fragmented information problems
• Understandability of patterns	• Requirement for DM modelling outcomes interpretability
• Managing changing data and knowledge	• Multidisciplinary collaboration
• Integration	• Medical data interoperability • Incorrect and fragmented information • Patient data privacy
• Nonstandard, multimedia, and object-oriented data	• Medical media data pre-processing and feature extraction • Medical data interoperability

1.5.4. Conceptualization and Exchange of Clinical Data

Since the beginning of the twentieth century, hospitals have been executing routine monitoring and recording of their performance. International and national health care organizations issue metrics and indicators, which reflect preoperative hospitalization time, postoperative complications rate, bed turnover rate, lethality, patient flow, etc. The metrics necessary for the calculation of the indicators are collected in paper or electronic forms. In the case of electronic

data collection, the HPO can analyze the information stored in real time and make appropriate management decisions (Paulus, et al., 2008). Thus, one of the aims of making information available electronically is to support management decisions of the HPO. For example, the Kaiser Permanente Hospital in the United States is continuously measuring both the indicators of the patient's health and medical personnel work efficiency, which are further analyzed and used for the treatment process improvement, evaluation of the personnel work quality, for comparative analysis and research (Paulus, et al., 2008).

Clinical data modelling and medical ontologies are emerging disciplines. Formalized clinical modelling are addressed by open standard specification openEHR (Beale, et al., 2006), its derivative European standard CEN 13606, and HL7 standard ISO 13972 (Schloeffel, et al., 2006). HL7 provides a two-layer validation process, with HL7 Reference Information Model, and its introduced Detailed Clinical Models as described in ISO 13972 standard. HL7 employs UML version 2 as a graphical modelling representation with ISO 21090 defined data types.

The openEHR related standards imply a framework for universal clinical data modelling, based on archetypes and templates, created by a consensus of the clinical community.

Archetypes are the shareable specifications of clinical information. OpenEHR defines archetype as a re-usable, formal definition of domain level information, defined in terms of constraints on an information model (Beale, et al., 2006; Kalra, et al., 2005). An example of the archetype “blood pressure” is shown in Fig. 4 (Heard, 2008).

An archetype template is a set of related archetypes with a common usage scenario. For example, the archetypes for blood pressure, weight and blood sugar may be used to record a routine screening of diabetic person.

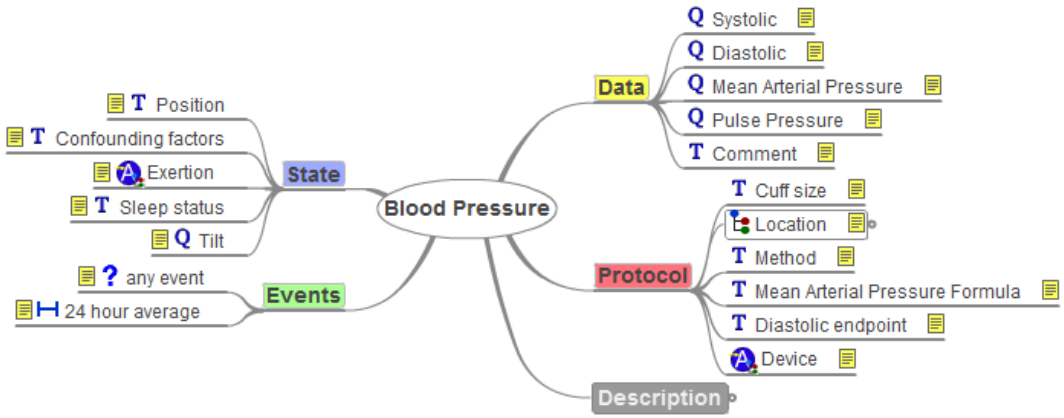


Fig. 4. OpenEHR archetype "Blood Pressure" (Heard, 2008)

All clinical modelling standards rely on medical terminology. Every term in openEHR archetype or detailed clinical model can refer to a standard terminology. Coded terminologies in healthcare are used to ensure semantic interoperability and decision support.

Much attention has been paid to the semantic analysis of existing formal, semi-formal and informal ontologies and investigation of their integrity (Bodenreider, 2008). The principle problems being solved are the creation of medical decision support systems and medical information system interoperability with respect to various terminologies used.

O. Bodenreider indicates (Bodenreider, 2008) the following most commonly used biomedical terminologies: ICD, LOINC, SNOMED, FMA, GO, RxNorm, MeSH, NCI Thesaurus, and UMLS. Table 5 shows the number of concepts within the listed thesaurus, classifiers, nomenclatures and ontologies:

Table 5. The most popular biomedical term thesaurus and ontologies

Title	**Number of concepts**
International Classification of Diseases, ICD	12.318
Logical Observation Identifiers, Names and Codes, LOINC	46.406
SNOMED Clinical Terms, SNOMED CT	310.314
Foundational Model of Anatomy, FMA	~72.000
Gene Ontology, GO	22.546

Title	**Number of concepts**
Catalogue of Clinical Drugs - RxNorm	93.426
Medical Subject Headings, MeSH	24.767
National Cancer Institute, NCI Thesaurus	58.868
Unified Medical Language System, UMLS	1,4 M

Since 1986, the United States National Library of Medicine has been generating an integral biomedical ontology UMLS (Unified Medical Language System).

The UMLS is a generalizing biomedical ontology. The UMLS provides tools for healthcare professionals and researchers to select and integrate information from different electronic biomedical sources, starting from patients' electronic health records systems, and ending with knowledge sources (databases). The UMLS version 11 contains over 1.4 million concepts and 6 million links. Conceptions unify synonyms from 100 different classification systems, such as MeSH, ICD-10, ICD-9-CM and SNOMED.

The UMLS consists of the following components:

- Metathesaurus – UMLS database, which consists of a glossary of concepts and terms, lists of relationships, as well as links to external vocabularies (classifications);
- Semantic Network – a set of categories and their relationships, which are represented in the definitions of entries in the Metathesaurus;
- Specialist Lexicon – a lexicographic vocabulary used in natural language processing;
- Software Tools.

The UMLS is particularly useful when integrating multiple data sources it is necessary to deal with the problem of multiple terms for a single concept. An example is the concept of a disease – "Addison's disease". This concept, depending on the classification system, may also be referred to as "Primary

hypoadrenalism", "Primary adrenocortical insufficiency", "E27.1", etc.

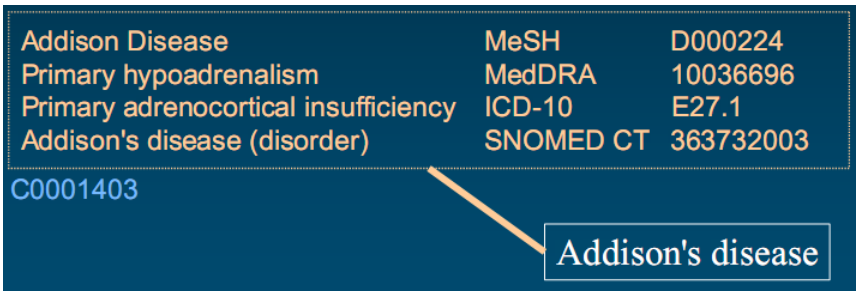


Fig. 5. Mapping model of UMLS concept "Addison's disease" (Bodenreider, 2012)

All names and codes represented in Fig. 5 are completely or partially identical to the concept "Addison's disease".

Currently, Lithuania has not yet implemented national medical records, which would enable data exchange in unified and standardized formats. However, national information systems administered by the National Health Insurance Fund control and provide HPOs with the following classifications:

- ICD-10-AM - International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification;
- ACHI - Australian Classification of Health Interventions;
- List of large and small operations;
- Diagnose related groups (DRG);
- National Social Insurance reimbursed services;
- Classifier of Lithuanian medical doctors;
- Classifier of Lithuanian health care institutions.

In order to successfully transmit data between different information systems, medical data transmission standards are used. HL7 is a standard system defining clinical and administrative patient data exchange and integration. A standard HL7 version 2, which is the most commonly used structured patient data exchange format, enables electronic messaging for the exchanging of administrative, financial and clinical information of a patient. As HL7 version 2 was created in the 1980s, its syntax is not designed for modern service oriented architecture. In 2005, the subsequent HL7 version 3 was proposed, and in 2014

a draft specification of HL7 FHIR was introduced. HL7 version 3 is based on a formal HDF methodology and refers to an object-oriented paradigm. HL7 Fast Healthcare Interoperability Resources is the most recent to date standards framework provided by HL7. FHIR combines previous versions and introduces the latest web standards to support RESTful architectures, mobile phone apps, cloud communications, and EHR-based data sharing.

It is planned to build a Lithuanian national eHealth system interoperability backbone based on the HL7 FHIR standard. Another important clinical data structuring and capturing standard, constituting a part of HL7 version 3, is HL7 CDA. The HL7 Clinical Document Architecture is an XML syntax standard specifying the coding, structure and semantics of clinical documents and ensuring unified clinical document structure. The HL7 CDA standard solves the problem of clinical document exchange between HPO, regional and national medical information systems.

1.6. Data Mining Uptake in Healthcare Facilities

Heterogeneous information systems, a variety of terminologies, semantic interoperability, data quality issues, and patient data privacy constraints are the most important reasons limiting deployment and use of DM in healthcare provider organizations.

In order to evaluate the practical usage of DM in healthcare, we have conducted a survey of tertiary hospitals in five countries (Niakšu & Kurasova, 2012). Countries from diverse economic development regions were selected to represent hospitals with unlike economic potential. The qualitative assessment of the survey results has been compared with quantitative literature analysis in the field of DM applications in the medical domain.

Surveying tertiary level healthcare facilities, which conduct scientific and commercial research studies, allowed us to draw conclusions on actual DM application activities, and to understand what the gap was between the data analysis experts' community and healthcare practitioners and scientists. The outcome of this comparative analysis suggested that a relatively low percentage

of academic research efforts result in practical DM applications in healthcare.

Thomson Reuters Web of Science (Thomson Reuters Web of Science, 2014), Google Scholar (Google Inc., 2014) and PubMed (National Center for Biotechnology Information, 2009) databases were used to analyze the number and distribution of scientific publications related to DM in medicine in the last decade.

The PubMed database is comprised of more than 21 million citations for biomedical literature from Medline, life science journals, and online books. PubMed is operated by the National Healthcare Library of U.S. and indexes all publications classifying its content with the help of MESH structured vocabulary (National Library of Medicine, MeSH, 2015). Using MESH vocabulary terms as a search parameter in the PubMed database guaranties that not only search wording matching publications will be found, but also its matching synonymic wording or previously used terms. MESH term, named as "data mining" is mapped to other similar concepts like "text mining". The term "data mining" has been appended to the vocabulary only in 2010 and the former terms e.g. "Information Storage and Retrieval" are mapped to the latest one. A search criterion "data mining" was used to retrieve a number of publications and books within the medical domain with the assigned MESH heading and MESH term "data mining". The first publication is dated 1984, however the second one appears only after a 10 year interval in 1994. This search resulted in 424 publications.

The Web of Science has been providing access to more than 12,000 journals in a variety of subject areas. It also includes citations to conference proceedings. The advanced search filter allows the use of logical operations, search restricted to the selected subject areas, and the search scope. The following search query was used for further analysis:

> (TS=(DM) AND TS=(medic* OR clinical OR healthcare)) AND Document Types=(Article OR Abstract of Published Item OR Proceedings Paper)
>
> Refined by: [excluding] Web of Science Categories=(OPERATIONS RESEARCH MANAGEMENT SCIENCE OR TELECOMMUNICATIONS)

Timespan=1996-2012. Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH. Lemmatization=On

This search resulted in 238 publications.

Google Scholar provides a scholarly literature search service across many disciplines and sources, including theses, books, abstracts and articles. However, it is not limited to scientific publications only. Google Scholar indexes content items published since 1993. Google's search filter allows the use of logical operations AND, OR, NOT, a restricted search only in the selected subject areas, and search scope (title of the publication or whole text). The following search query was used for further analysis:

Search in the title: "data mining" AND (medical OR clinical OR medicine OR healthcare)

The choice of searching in the whole article text was rejected due to the fact that many DM centric publications have keywords "medicine" or "healthcare" in the text with a purpose to illustrate DM usage.

The distribution of publications found in Web of Knowledge, Google Scholar and PubMed databases starting from 1996 to 2012 is shown in Fig. 6.

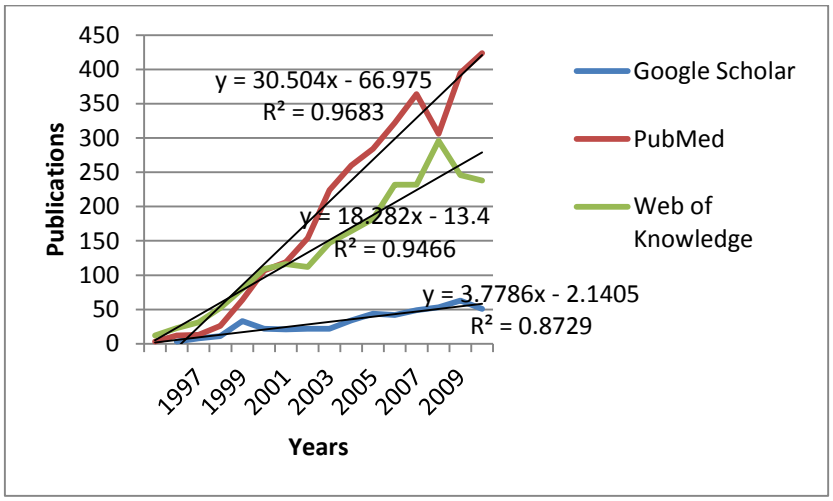


Fig. 6. Trend lines of DM applications in medicine related publications

As it is seen from the explanation of the queries searched in the databases, the results are not directly comparable and are used to illustrate a constant increasing interest of the academic society in DM applications in the medical domain.

1.6.1. Surveying DM Applications in Healthcare Facilities

As shown above, the volume of medical related DM research increases from year to year. Hypothetically, one can suppose that DM usage penetration is increasing accordingly.

Due to the fact that the healthcare sector is very diverse and its entities as well as actors have different objectives and activities, they employ different methods and tools in their operations. The initial experience of interviewing healthcare institutions suggested that the highest probability of DM usage is in tertiary hospitals, which have tight relations with the academic society and participate in scientific and commercial research on a regular basis. Therefore, the analysis scope was limited to tertiary hospitals, representing different levels of economic development, and having a different magnitude of electronically available patient related data. Hospitals from the following countries participated in the survey: South African Republic, Lithuania, Switzerland, Albania, and Germany.

1.6.2. Preparation and Conducting the Survey

The respondents of the survey were briefed on the purpose and terms used in the questionnaire. Early feedback indicated that hospitals' IT departments could better identify actual applications of DM deployed in IT systems used by the organization. Whereas medical representatives were minimally knowledgeable about what exactly DM is and how or if it is used in the institution. Taking this diverse interviewing audience into consideration, questions were formulated in a comprehensible way for a broader range of respondents with a medical or IT background. The final questionnaire, which was distributed to the hospitals is as follows:

1. Have you heard about practical applications of DM in medicine?
2. Do you know any research projects in your hospital using DM methods?
3. Have you or your colleagues been involved in a DM research project, aiming to identify new patterns or finding new rules for patient

diagnostics, prediction of treatment results or other. If yes, please provide a brief summary of research aim and the results.
4. If DM methods have been used, was your experience successful? Please comment.
5. Has the clinical decision support IT system been used in your hospital?
6. Please specify which clinical specialties could benefit by using DM methods on collected patient clinical data in your hospital (choose from the list).
7. What type of clinical research is your hospital involved in?
8. Are you or your colleagues potentially interested in the benefits DM could provide to you?
9. How many years has patient data been collected in IT systems in your organization?
10. Please specify what clinical patient information is stored in IT systems (HIS, EHR, EMR, RIS, etc.). Select from the list: Observations, Lab results, Radiology reports, Anamnesis, Surgery reports, Discharge summary, Visit summary, Nursing data (vitals), Medication used (for inpatients).
11. Mark medical IT systems used in your organization. Select from the list: EMR / EPR, HIS.
12. RIS/PACS, LIS, Specific clinical information systems, Emergency IS, OP clinic information system, Blood bank information system, Clinical decision support system, Pathology information system.
13. Specify what standard nomenclature is used in your organization (e.g. ICD9, ICD10, SNOMED-CD, LOINC). Select from the list: Patient diagnosis, Pathologic diagnosis, Procedure coding, Laboratory coding.
14. Are you interested in international clinical DM research projects?
15. Specify the clinical specialty or problem you are interested in.

1.6.3. Method of Survey

The survey was conducted according to methodical guidelines of the Centre for Health Promotion of University of Toronto (Centre for Health Promotion of University of Toronto, 1999). A call for survey was openly published in the eHealth news portal eHealthServer.com (eHealthServer.com, 2012). The survey was prepared in both online questionnaire and offline forms. Hospitals were asked that at least two respondents from each institution should fill out the

questionnaire; a person in charge of medical services, e.g. medical superintendent, director of medicine, head of the clinical department and a person in charge of Information Technology e.g. chief of the IT department. Initially, the enquiries were sent to the officials of the hospitals in eight countries.

1.6.4. Analysis of Survey Results

Out of fourteen respondents, twelve confirmed that they had heard about practical applications of DM. However, after response validation, only nine positive answers could be qualified. However, even out of the remaining nine respondents with positive answers only four were familiar with practical examples of such usage, totaling 29 % of the respondents. The survey answers suggested that the majority of medical respondents have no information about DM research initiatives and applications in their own facilities. The selected method of surveying two or more representatives from each facility has shown that typically medical specialists, who are not engaged in DM projects in their own institution, would have no information about it. The summary of the survey results is presented in Table 6. In any case where 70 % or more respondents answered positively, then the answer is averaged as "Yes"; if more than 40 %, but less than 70 % - the answer is averaged as "Differs", and the remaining – averaged as "No".

Table 6. Summary of survey answers.

Summarized questions	**Hospitals of developing countries**		**Hospitals of emerging countries**		**Hospitals of western countries**	
	IT	**Clinical**	**IT**	**Clinical**	**IT**	**Clinical**
Understanding, practical usage and interest						
Good understanding of DM concept	Yes	Differs	No	No	Yes	Differs
Awareness of practical use	No	No	No	No	Differs	No
Hands-on DM applications	Differs	No	No	No	Yes	No
Interest in the topic	Yes	Yes	Yes	Yes	Yes	Yes
Clinical specialties	All	All	All	All	All	All

Availability of electronic data for research

Number of years data is electronically captured	4-13 years	1-3 years	5-15 years
Variety of medical information systems used to capture and operate with patient related data	patient demographics, radiology images, partly lab results, partly detailed clinical data, billing data	patient demographics, limited radiology images, partly billing data	patient demographics, radiology images, lab results, detailed clinical data, billing data

Evaluating the benefits of gained DM experience, 50 % of respondents, who declared a personal involvement in DM projects, were satisfied with the results achieved and 50 % had a neutral opinion of the project success.

The interest in getting additional information on potential DM benefits was expressed by 86 % of respondents, regardless of their initial experience with DM.

The analysis of electronic data availability for DM purposes showed us the correlation between depicted years of clinical data collection in a facility with the level of the institutions' economic state. Each surveyed hospital is presented as a separate column in Fig. 7. The data collection timeframe values spread is 1–15 years with the mean values: 1 year in developing countries, 8.6 years in emerging countries, and 10 years in western countries.

All respondents have identified that a hospital information system is in use; electronic medical record systems are used in 60 % of facilities and radiology imaging systems in 83 %.

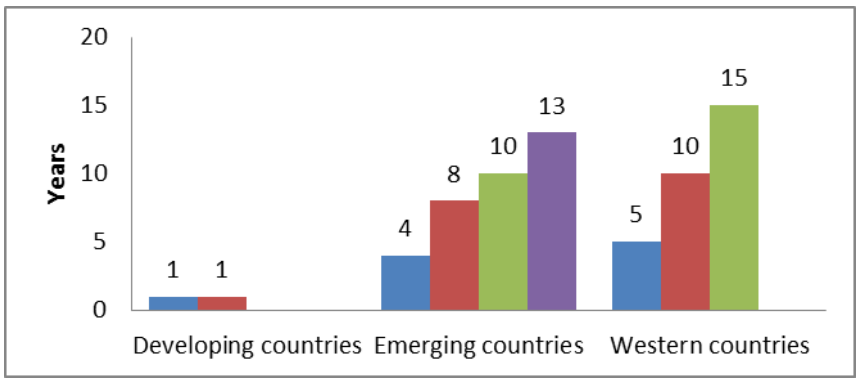


Fig. 7. Clinical patient data collected electronically in hospitals

The usage of standard terminologies varies depending on national legislation. The usage of ICD-9 and ICD-10 is common for coding diagnoses. However, other nomenclatures, critical for DM applications to code procedure/intervention, laboratory tests, and pathology diagnoses, are either partly implemented or not implemented at all.

1.6.5. Summary of the Survey Results

All the respondents confirmed that they are aware of practical DM applications in medicine. However, only 29 % of respondents were able to provide an example of such usage.

There was a noticeable confusion in differentiating DM and statistics concepts among healthcare professionals, and very rarely was DM treated by them as a practically valuable tool for clinical purposes.

The respondents from healthcare facilities with a relatively recent adoption of IT in the patient treatment process tended to mix statistical reporting and DM, hospital information systems, Electronic medical record systems and decision support systems.

Regardless of understanding and experience of DM, 86 % of respondents expressed their interest in the DM topic and 93 % confirmed their intent to participate in international DM research projects and to be informed about DM applications in the future.

1.7. Generalization and Conclusion

DM application in the healthcare domain is known for its complexity, which is due to data heterogeneity, overlapping clinical data exchange and modelling standards, complex data structures, ambiguous semantics, ethical, social and legal constraints. That makes the effective medical knowledge discovery an evolving subject with a growing interest of academics and medical practitioners.

The analysis of publications in the field of DM application in the medical domain has shown a steady growth since its accountable beginning. In the early nineties, up to five publications were produced during one year, and more than

400 publications in 2013. A tremendous growth of interest and scientific advancement took place in the last decade. Therefore, literature analysis has shown that DM is used to support patient treatment quality improvement and healthcare management optimization objectives. A variety of DM methods including classification, clustering, association analysis, visualization, and link analysis have been successfully used in different clinical specialties.

The survey of university hospitals revealed that the majority of their medical personnel has minimal awareness of DM practical usage and its possibilities. All the respondents from the largest surveyed university hospitals confirmed to be familiar with DM applications in healthcare, however only 29 % were capable of providing any example of practical DM application. The survey identified a considerable potential for further DM application penetration due to an increasing amount of patient clinical data collected in healthcare provider organizations.

It was shown that DM perception and deployment in healthcare facilities is beyond its steady growth in the academic research field. More attention should be paid to the domain specific issues of DM application in healthcare.

In the next chapter, a systematic approach of DM application in the medical domain is investigated. A novel methodology, which improves and specializes the industry standard methodology CRISP-DM for the medical domain is proposed. Furthermore, it is thoroughly applied for predictive DM in oncology and cardiology and for descriptive DM in meta-analysis of PubMed database publications.

This page is intentionally left blank.

CHAPTER 2

Systematic Application of Data Mining and Data Analysis Methods in Medical Domain

2.1. Introduction

To transform the DM endeavor from a unique artisan act to a systematic engineering process, a robust methodological framework is required. There are few standard generic process models describing the steps of DM analysis. Since 1990, a number of domain independent process models, application methodologies, and industry standards have been proposed. The Cross Industry Standard Process for DM (CRISP-DM), "Sample, Explore, Modify, Model and Assess" (SEMMA) process model, and Predictive Model Markup Language (PMML) are the most prominent (Piatetsky-Shapiro, 2014). However, there are no established standards or methodologies for DM applications in medicine.

N. Esfandiari et al. have analyzed (Esfandiari, et al., 2014) 291 papers with the aim of extracting knowledge from structural medical data published between 1999 and 2013 in 90 journals. The authors concluded that 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. Finally, an overall DM process methodology specific to the medical

domain would benefit multi-disciplinary process participants for better-aligned collaboration.

In the following sections, a review of the existing knowledge discovery process models and available DM standards is provided.

The CRISP-DM, SEMMA and Fayyad process models are described. The CRISP-DM process model, as the most comprehensive available framework, is analyzed in more detail. Then, an extension of this methodology is proposed, called CRISP-MED-DM, which is based on the analysis of the needs and unique features of DM in the medical domain. Finally, a few case studies are provided in Cardiology, Oncology and medical literature meta-analysis domains.

2.2. Methodologies for Data Mining Applications

The CRISP-DM methodology, the SEMMA set of core DM activities, and Fayaad's knowledge discovery in databases process model, are the most popular systematic knowledge discovery process handling guidelines for the DM analyst.

2.2.1. Fayaad's Knowledge Discovery in Databases Process Model

The interactive and iterative KDD process model was introduced by Fayyad, Piatetsky et al. (Fayyad, et al., 1996).

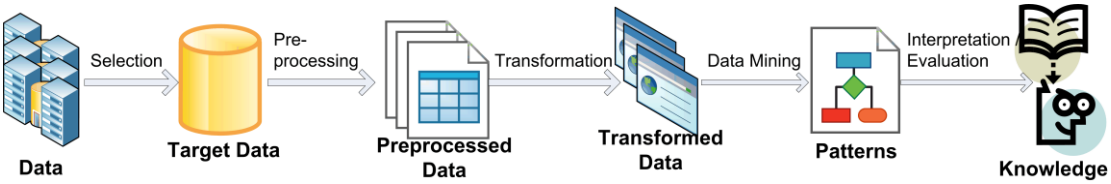


Fig. 8. Summarized KDD process steps according Fayyad, Piatetsky et al.

Fayaad's process model shown in Fig. 8 includes nine steps:

1. Understanding the application domain: it involves pertinent prior knowledge and the objectives of the application.
2. Constructing a target dataset: consists of choosing a dataset or focusing on a subset of variables or samples of data on which the discovery is to be carried out.

3. Data clean-up and pre-processing: consists of basic operations, such as eliminating noise or outliers if necessary, gathering the necessary information to model or account for noise, coming to a decision on strategies for treating missing data fields, and accounting for time sequence information and changes known, as well as making a decision on DBMS issues, such as schema, data types and mapping of unknown and missing values.
4. Data trimming and projection: consists of finding practical features to represent the data, depending on the objective of the task, and using transformation methods or dimensionality reduction to decrease the effective number of variables that are being considered, or to get invariant representations for the data.
5. Selecting the function of DM: involves choosing the purpose of the model derived by the algorithm of DM (e.g., classification, summarization, clustering and regression).
6. Selecting the DM algorithm: involves choosing the method to be used for searching for patterns in the data, such as choosing which parameters and models may be appropriate, (e.g., the categorical data models are different from models on vectors over reals) and matching a certain DM method with the general criteria of the KDD process (e.g., the user might be more interested in understanding the model than in its predictive capabilities).
7. Data mining: involves looking for patterns of interest in a certain representational form or a set of similar representations, including regression, classification rules or trees, clustering, dependency, sequence modeling, and line analysis.
8. Interpretation: involves interpreting the found patterns and possibly returning to any of the prior steps, as well as the possible visualization of the patterns extracted, removing the irrelevant or unnecessary patterns, and translating the useful ones into terms comprehensible by users.

9. Utilizing discovered knowledge: involves incorporating this knowledge into the system's performance, taking actions based on the knowledge, or merely documenting it and reporting it to the interested parties, as well as inspecting and resolving potential conflicts with previously supposed (or extracted) knowledge.

The strength of the process model is in its explicit simplicity, which makes it generically applicable to all possible knowledge discovery domains. However, the authors of the process model have provided a generic guideline, with no formal methodology or accompanying toolset. Nevertheless, this process model is one of the most referenced and used for general KDD purposes, and it became the base model for other more detailed models.

2.2.2. SEMMA Process Model

The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the proprietary generic DM process model, proposed by the SAS Institute Inc. (Azevedo & Lourenco, 2008). SEMMA was initially created to support the software application SAS Enterprise Miner. Later on, its usage stretched beyond the boundaries of SAS software. SEMMA is limited to the core activities of DM processes and does not cover other phases of KDD, such as business understanding and deployment. Moreover, it is also poorly supported with documentation, and implementation guides.

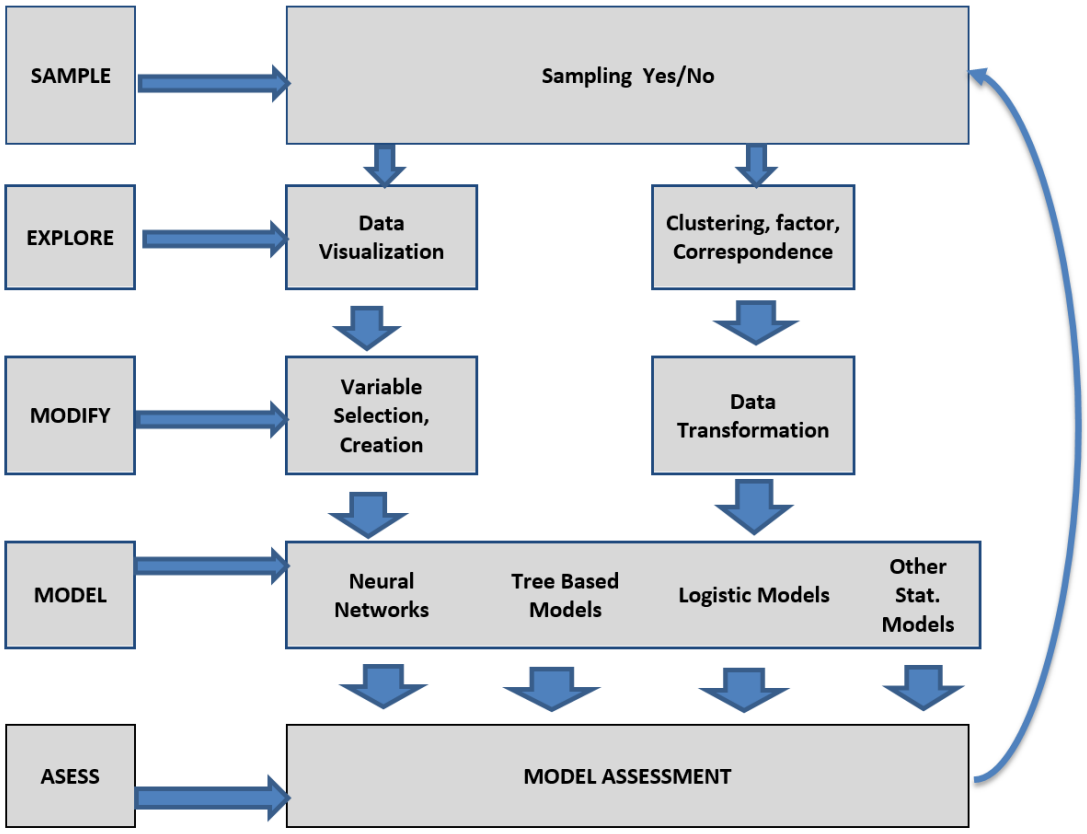


Fig. 9. SEMMA process model (original from SAS Institute)

As it is shown in Fig. 9, the SEMMA process model consists of five main phases:

1. Sample. This process begins with the sampling of data, e.g., choosing the data set for modelling. The data set must be large enough to hold sufficient information to retrieve, yet small enough to be utilized efficiently. Data partitioning is also dealt with in this phase.
2. Explore. This phase covers the data understanding through determining the predictable and unpredictable relationships between variables, as well as the abnormalities, with the help of visualization data.
3. Modify. This phase includes methods to choose, create and transform the variables in preparation for the data modeling.
4. Model. This phase focuses on applying various modeling techniques (DM) on the prepared variables for creating models

that tend to provide the desired outcome.

5. Assess. This phase involves the evaluation of the modeling results, which demonstrates the reliability and effectiveness of the created models.

SEMMA does not provide a reference model, detailed guideline or compliance evaluation model. Therefore, it cannot be considered as a methodology, but rather as a process model, focused on the core tasks of DM.

2.2.3. Cross-Industry Standard Process for Data Mining

According to the online poll conducted by the international DM community KDNuggets in 2014 (Piatetsky-Shapiro, 2014), the most referenced and used in practice DM methodology is CRISP-DM. According to G. Piatetsky-Shapiro: "CRISP-DM remains the most popular methodology for analytics, DM, and data science projects, with a 43 % share in the latest KDnuggets poll…" (Piatetsky-Shapiro, 2014; Azevedo & Lourenco, 2008).

The 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 the de facto standard for DM (Chapman, et al., 2000; Olson & Delen, 2008). Notably, CRISP-DM is an informal methodology, since it does not provide the rigid framework, evaluation metrics, or correctness criteria. However, the methodology provides the most complete toolset to date for DM practitioners.

The first version of the CRISP-DM specification was developed by a consortium of European and American private companies in 1996, aiming to create a non-proprietary and freely available standardized process model and a toolset for DM application engineering. The current version includes the methodology, reference model, and implementation user guide. The methodology defines phases, tasks, activities and deliverables.

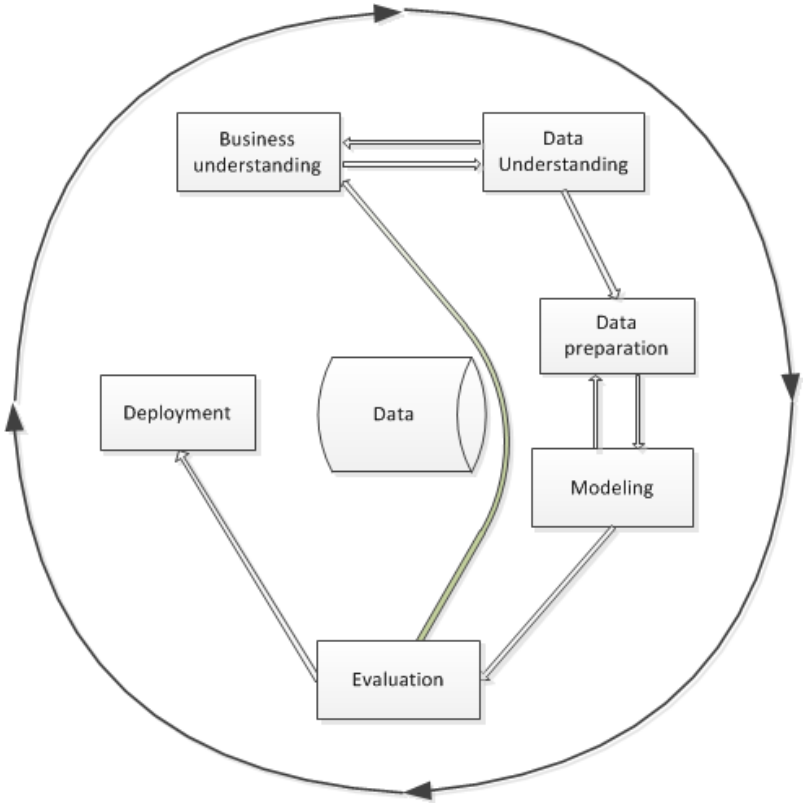


Fig. 10. Phases of the original CRISP-DM reference model

As is shown in Fig. 10, the CRISP-DM proposes an iterative process flow, with non-strictly defined loops between phases, and an overall iterative cyclical nature of the 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.

Phase 1: Business understanding

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

Phase 2: Data understanding

This phase starts with the gathering of initial data and access to the dataset. Problems of data quality must be identified and the initial assumptions of which datasets can be of interest for further steps are made.

Phase 3: Data preparation

The data preparation phase covers all the activities that are required for

preparing 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 choosing the tables, attributes projections and records, attributes transformation, normalization, noise elimination and sampling.

Phase 4: Modelling

In this phase, a suitable selection of modelling techniques, algorithms, or combinations thereof is done. Generally, for the same task, there are a few possible modelling 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 the chosen model quality criteria is achieved. The model quality is formally assessed. In order to evaluate the quality of the model, there are metrics used which are popular in DM and statistics, e.g. sensitivity, accuracy, specificity, ROC curve, and cumulative gain chart. Sensitivity – positive results properly classified as such in the results set. Accuracy – the percentage of properly classified objects. Specificity – negative results correctly classified as such in the results set. The relationship between sensitivity and specificity may be assessed with the help of a ROC curve (Receiver Operating Characteristic) or a numerical expression of the area under the curve (AUC). Cumulative Gain charts display the percentage of positive responses predicted by the models versus the percentage of the population.

Phase 5: Evaluation

The evaluation phase already has 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 of whether the DM results may be used in practical settings.

Phase 6: Deployment

The deployment phase stipulates the utilization of the results of DM. Model generation is not the last step of the DM project. Despite the cases where the objective of a 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 a repeated DM process. 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.

The CRISP-DM reference model describes the generic tasks and deliverables for each phase. The implementation guide details the task to the activities level, offering additional warnings, remarks and hints. The generic tasks and deliverables of the CRISP-DM reference model are outlined in Table 7.

Table 7. Generic tasks and outputs of the original CRISP-DM reference model

Generic tasks	**Deliverables**
Business understanding phase	
Determine Business Objectives	Background Business Objectives Business Success Criteria
Assess Situation	Inventory of Resources Requirements Assumptions, and Constraints
Determine DM Goals	DM Goals DM Success Criteria
Produce Project Plan	Project Plan Initial assessment of Tools and Techniques
Data Understanding phase	
Collect Initial Data	Initial Data Collection Report
Describe Data	Data Description report
Explore Data	Data Exploration report
Verify Data Quality	Data Quality Report
Data Preparation phase	
Select Data	Rationale for Inclusion/ Exclusion
Clean Data	Data Cleaning Report
Construct Data	Derived Attributes Generated Records
Integrate Data	Merged Data
Format Data	Reformatted Data
Dataset	Dataset Description

Modelling phase	
Select Modelling Techniques	Modelling Technique Modelling Assumptions
Generate Test Design	Test Design
Build Model	Parameter Settings Models Model Descriptions
Assess Model	Model Assessment Revised Parameter Settings
Evaluation phase	
Evaluate Results	Assessment of DM Results Approved Models
Review Process	Review of Process
Determine Next Steps	List of Possible Actions Decision
Deployment phase	
Plan Deployment	Deployment Plan
Plan Monitoring and Maintenance	Monitoring and Maintenance Plan
Produce Final Report	Final Report Final Presentation
Review Project	Experience Documentation

The authors of the CRISP-DM reference model (Chapman, et al., 2000) stated "…future extensions and improvements are both desirable and inevitable…". In the next section, the enhanced version CRISP-MED-DM for the application in medicine is proposed, where the unique features of the domain are addressed.

2.3. Standards and Technologies in Data Mining

Usage of industry standards help to achieve interoperability and leverage the reuse of DM project results. PMML, XMLA, and JavaDM API are the currently available standards for technological DM activities.

2.3.1. Predictive Model Mark-up Language

PMML refers to predictive DM markup language maintained by the Data Mining Group (DMG, 2014), an independent, vendor led consortium. PMML is an open

standard for defining and sharing predictive models of DM. PMML allows for formalizing predictive models and provides the means for interoperability of DM software. It allows researchers to create a predictive model using a single software package and then apply this model in another information system, such as a hospital information system or clinical decision support system.

The latest version 4.2.1 was introduced in February of 2014 (DMG, 2014), in addition to predictive modelling it supports text mining and clustering.

PMML uses XML syntax. The structure of a model is defined by an XML schema. A few DM models can be described in a single PMML XML document. In the PMML document each model shall be uniquely identified by name or by *functionName*, which DM method (classification, clustering), and *algorithmName*.

PMML includes features to expose the quality parameters of the models. These descriptive parameters are useful for the evaluation of a model's validity. An optional attribute *isScorable* enables indicating if the model should be deployed and processed normally or, if the attribute is set to false, then the model is intended for information purposes only and should not be used to generate results.

2.3.2. XMLA - XML for Analysis

XML for Analysis (XMLA) is an open industry standard for data access and analysis, introduced and maintained by XMLA Council with Microsoft, Hyperion and SAS (Microsoft, Hyperion, SAS, 2001).

XMLA provides a set of XML Message Interfaces, based on SOAP, to define the data access interaction between a client application and an analytical data provider (OLAP and DM). XMLA consists of two SOAP methods: *Execute* and *Discover*. The *Execute* method executes commands in MDX, DMX or SQL with properties provided in XML syntax. The *Discover* method obtains information and metadata from a Web service.

2.3.3. Java Data Mining API

Java DM API (JDM) is a technology which enables integration of DM

techniques into Java applications. The Public Draft of the standard was released in 2002. Currently, it is maintained under Java community Process (Oracle, 2011). The standard leverages other industry standards, i.e. CWM, SQL/MM, JCX, and PMML. The JDM specifies mining functions: classification, regression, attribute ranking, association, clustering, and feature extraction. The list of supported algorithms is growing. Currently JDM supports the following DM methods: K-Means, Decision Trees, SVM, and Feed-Forward Neural Networks. JDM have the following main objectives:

- model building;
- scoring using models;
- creation, storage, access and maintenance of data to support DM results;
- providing seamless interface for DM tasks.

In addition, JDM methods support import/export to PMML, model testing, batch and real-time scoring.

JDM provides test metrics for model performance evaluation. For classification models, accuracy, confusion-matrix, lift, and receiver-operating characteristics are available to access the model performance; R-squared and RMS errors are provided for regression models.

2.4. Data Mining Application Methodology for Medical Domain

There is a lack of specific and detailed framework for conducting DM analysis in medicine. The DM application issues in medicine and healthcare are described in Section 1.5.3. A number of papers addressed the uniqueness of DM in health care (Bellazzi & Zupan, 2008; Canlas Jr, 2009; Koh & Tan, 2005; Cios & Moore, 2002). All of those papers suggested the need for additional activities to be considered in the knowledge discovery process within the medical domain.

2.4.1. Related Work

There are few known works contributing to the topic. Catley et al. (Catley, et al.,

2009) proposed an extension of the CRISP-DM process model to support temporal data abstraction. The enhancements proposed by the authors adopts CRISP-DM to the needs of DM application for data streams generated by intensive care unit equipment.

Špečkauskienė and Lukoševičius (Špečkauskienė & Lukoševičius, 2009) proposed a generic KDD process model for the medical domain. The proposed methodology and software tool emphasis optimization of a dataset and selection of the best performing DM algorithm and its parameterization. For datasets optimization, the authors do recommend techniques such as feature extraction, sampling and data set stratification to balance it with regards to the class attribute. In terms of the selection of the DM algorithm, it is proposed to test all available algorithms for the specific task and data type and to compare their results with the metrics of sensitivity, specificity, ROC AUC, and F-measure. The process flow of the methodology is shown in Fig. 11.

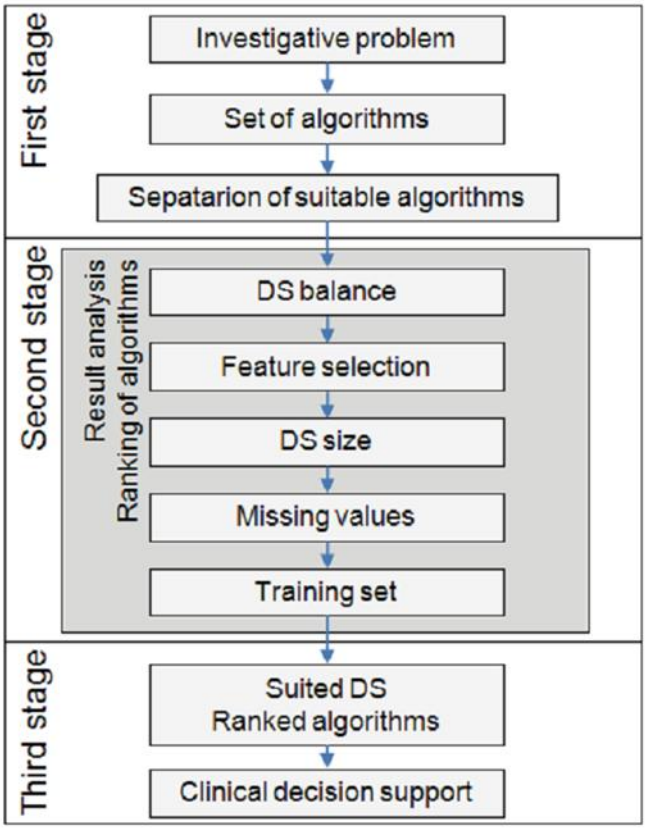


Fig. 11. Špečkauskienė and Lukoševičius methodology (Špečkauskienė & Lukoševičius, 2009)

However, the proposed methodology is limited to overall recommendations on the dataset preprocessing and providing an iterative algorithm selection and parameterizing process flow. It does not address the specific issues of data mining applications in the medical domain listed in Section 1.5.3 *Issues and Challenges of Data Mining in Medicine*. Therefore, the process flow proposed by the authors can be treated mostly as a greedy DM application approach, which allows semi-automation of typical classification or regression tasks for structured datasets, whilst selecting the winning model with the best pre-defined model's evaluation criteria values.

Furthermore, the authors did not aim to provide either a complete and detailed process flow, or a formal evaluation model.

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

The CRISP-DM is a hierarchical process methodology, which provides an extendable framework going from generic to specific. The methodology proposes third and fourth abstraction layers for mapping generic models to specialized models (Fig. 12). The medical domain is proposed as the application domain context for the mapping. According to CRISP-DM specification (Chapman, et al., 2000), mapping for the future type of extension shall be used to ensure specialization of the generic process model according to a pre-defined context for future systematic use.

The following recommendations for extending or specializing CRISP-DM have been used:

1. Analyze specific context.
2. Remove any details not applicable to the context.
3. Add any details specific to the context.
4. Specialize generic contents according to concrete characteristics of the context.
5. Rename generic contents to provide more explicit meanings in the context.

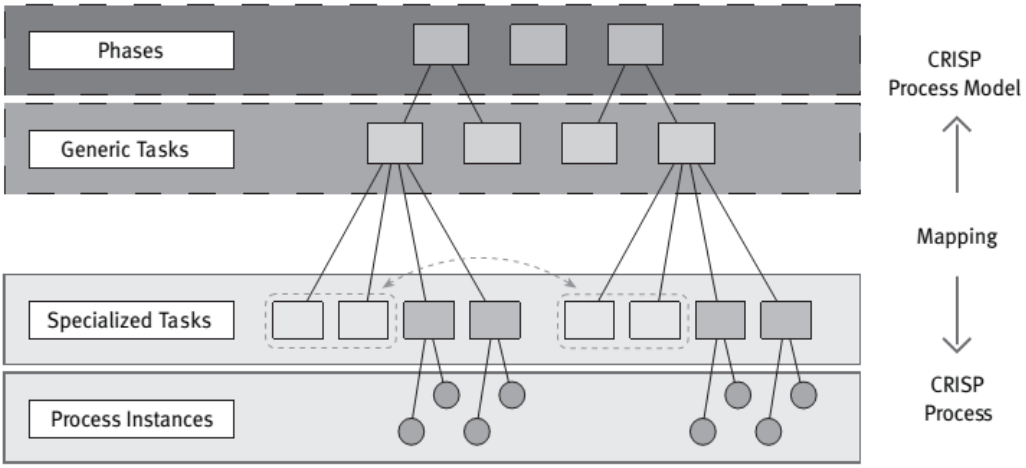


Fig. 12. Hierarchical breakdown structure of CRISP-DM methodology (Chapman, et al., 2000)

As it was described in Section 1.5.3, the following issues shall be considered when applying DM in the medical domain:

1. heterogeneous medical information systems;
2. semantic data interoperability;
3. mining complex datasets: multi-relational, stream data, text and multimedia;
4. incomplete and fragmented data;
5. ethical, social and legal constraints.

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

Another shortcoming of CRISP-DM is a lack of metrics or evaluation criteria, which allow for assessing the compliance of a DM project against the requirements of the methodology. To address these issues, a compliance assessment method is proposed. The method enables the evaluation of a DM project in a simple and practical way, determining to which extent it was handled in the frame of the CRISP-DM.

To address the outlined limitations of CRISP-DM and propose specific tasks for the medical domain a novel CRISP-MED-DM specialized methodology reference model is proposed.

Phase 1 “Business understanding” and Phase 2 “Data understanding” are

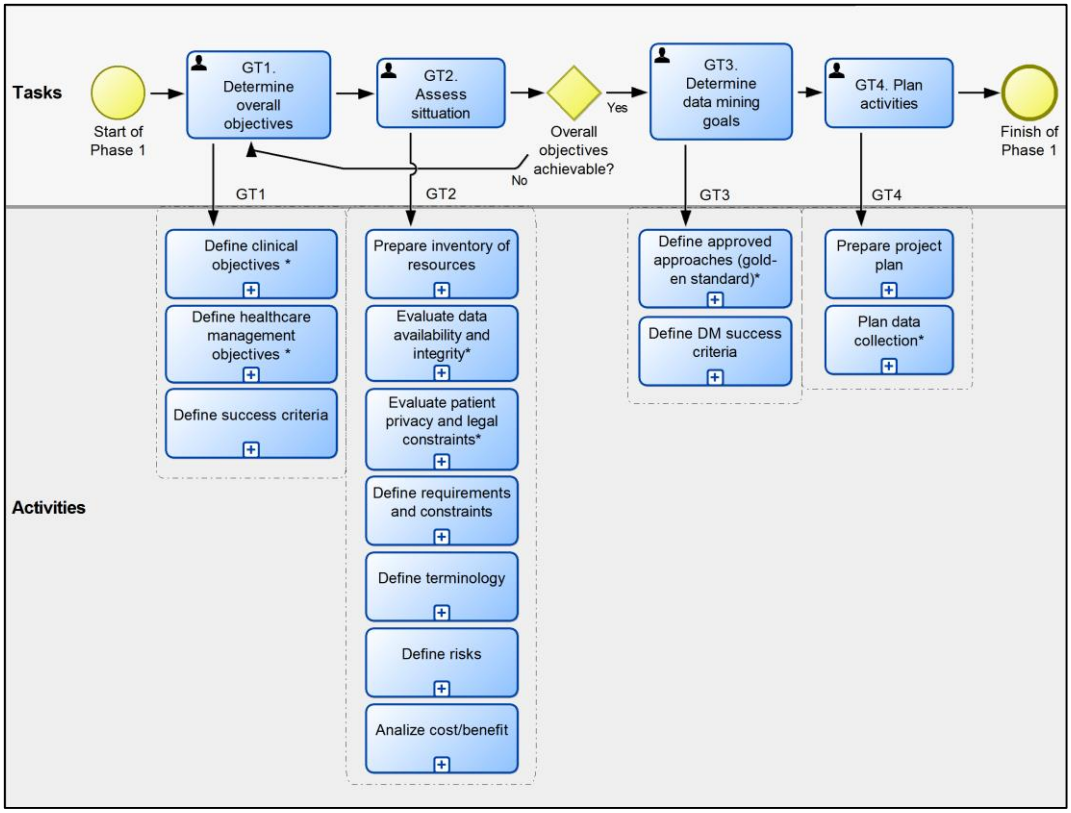
the phases where the DM project is being defined and conceptualized. The rest are implementation phases, which aim to resolve the tasks being set in the first phases. As in the original CRISP-DM, the implementation phases are highly incremental and iterative. However, the changes in Phase 1 or 2 lead 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 introduced general tasks, activities and deliverables of CRISP-MED-DM are outlined in the following sections.

2.4.3. CRISP-MED-DM Phase 1 and Phase 2

The first phase "Business understanding" was renamed "Problem understanding" to avoid ambiguous meanings 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 general tasks (GT) and activities of Phase 1 are shown in Fig. 13.

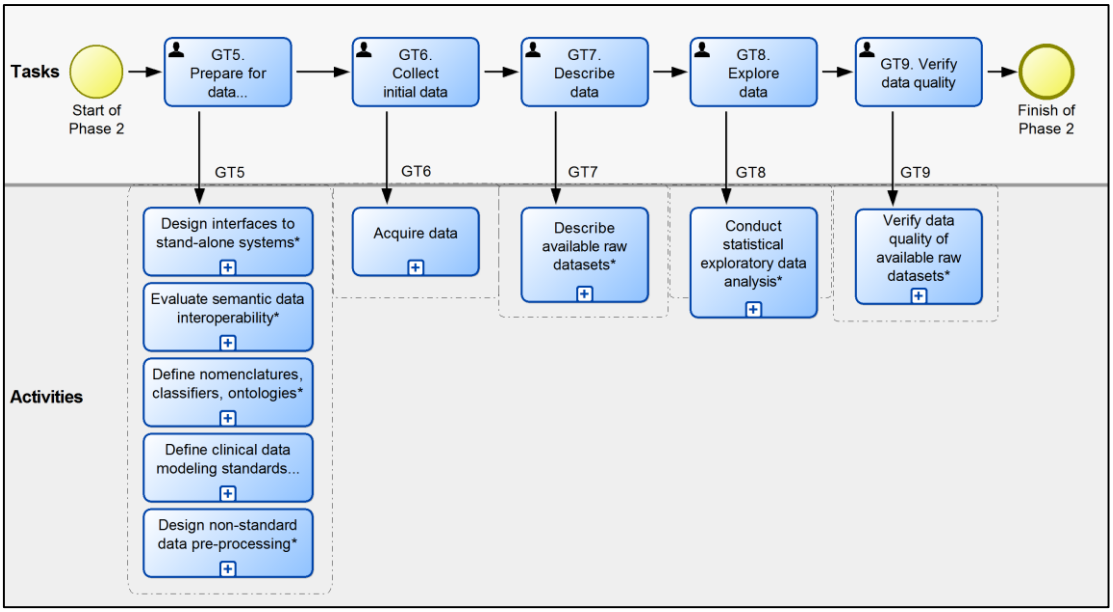
In the second phase "Data Understanding", a new general task "Prepare for data collection" was introduced. Issues of transport, semantic and functional interoperability were considered. The wealth of medical data formats are considered through the introduced activity of non-standard data pre-processing design, which includes the support of multi-relational data, temporal, unstructured text and media data. The definitions of medical nomenclatures, classifiers and ontologies used in data is substantial for further data pre-processing.



Enhanced activities are marked with *

Fig. 13. Phase 1 "Problem Understanding" general tasks and activities.

Finally, the definition and analysis of clinical data models and clinical protocols used in data source systems shall be carried out. The general tasks and activities of Phase 2 are shown in Fig. 14.



Enhanced activities are marked with *

Fig. 14. Phase 2 "Data Understanding" general tasks and activities.

2.4.4. CRISP-MED-DM Phase 3 and Phase 4

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

The original CRISP-DM task "select data" had limitations for practical application in the medical domain. First, it is mostly assumed for a single-table static data format. Second, it lacks activities to handle data conversion and unification of the medical terminologies being used, and 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 the DM modelling step. The activities of the task are 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 require separate training, validating and testing datasets, therefore

data sampling activity was introduced.

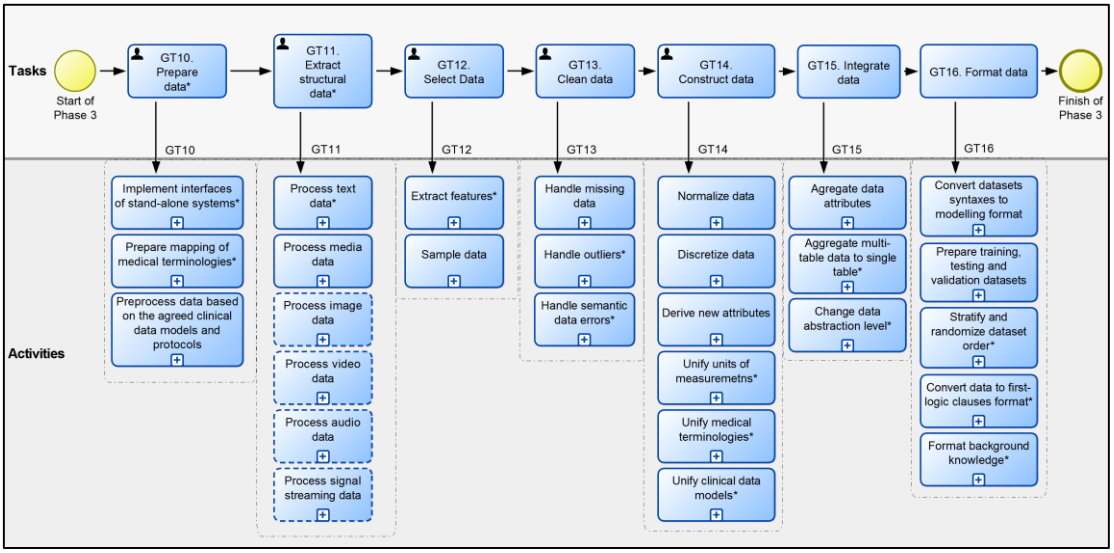
Missing data is a very common issue for clinical data. In addition, errors due to faulty sensors and laboratory and monitoring equipment interfaces shall be identified through outliers' detection and semantic analysis. Automated semantic error analysis is typically based on business rules, implementing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

Within the "Data integration" task, the activity of changing the 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 modeling activities.

Multi-relational data requires either propositionalization of data to a 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 the 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 & Lukoševičius, 2009). The described general tasks and activities of phase 3 are shown in Fig. 15.

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



Enhanced activities are marked with *

Fig. 15. Phase 3 "Data Preparation" general tasks and activities.

Špeckauskiene et al. (Špečkauskienė & Lukoševičius, 2009) proposed an iterative 11-step DM process model, tailored for finding the optimum modelling algorithm. The authors proposed the following flow:

1. To collect and access 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 the 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 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

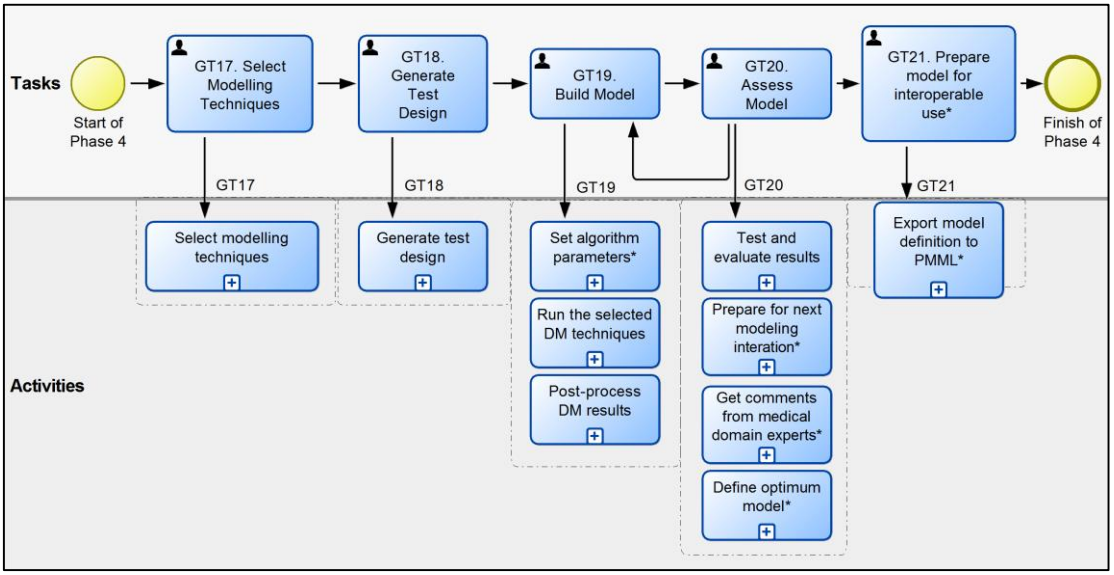
of the algorithms.

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 a greedy trial of all possible modelling 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, the iterative selection of a set of feasible modelling techniques, opposition to a few modelling techniques; iterative parameterizing of the selected modelling algorithms; and usage of predefined quality metrics to identify rejected, accepted, and the best performing model (Fig. 16).

According to C. Catley (Catley, et al., 2009), collaborative DM methods (e.g. method ensembles, method chains) may provide a higher performance. 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 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. Moreover, an exported interoperable model definition may be used by other researchers for further external validation. As was shown by D. G. Altman (Altman, et al., 2009), validation studies are of great importance to crosscheck and ensure overall objective performance of the derived predictive models.

The described general tasks and activities of phase 4 are shown in Fig. 16.



Enhanced activities are marked with *

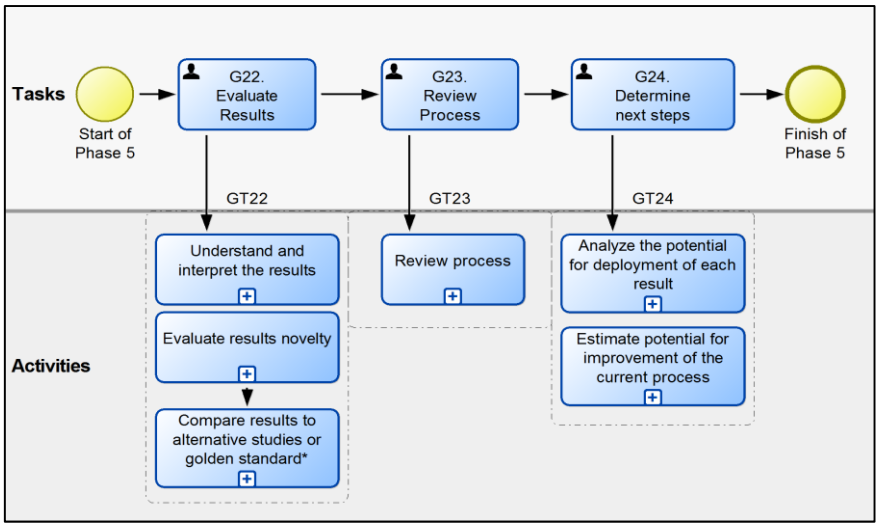
Fig. 16. Phase 4 "Modelling" general tasks and activities.

2.4.5. CRISP-MED-DM Phase 5 and Phase 6

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

Frequently, creating new predictive models for the medical domain, the current golden standard exists, against which the outcomes of DM modelling shall be verified and crosschecked.

The general tasks and activities of Phase 5 are shown in Fig. 17.



Enhanced activities are marked with *

Fig. 17. Phase 5 "Evaluation" general tasks and activities.

Deployment phase remains with no changes, as shown in Fig. 18 .

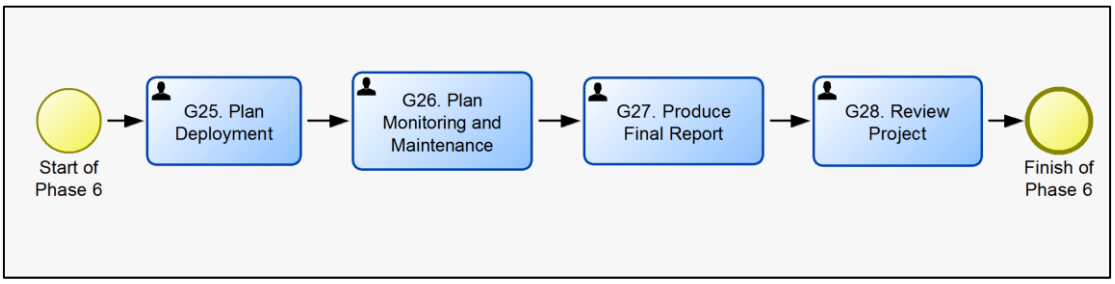


Fig. 18. Phase 6 tasks "Deployment" general tasks and activities.

The developed CRISP-MED-DM methodology has been used and assessed in predictive DM research projects in the oncology and cardiology domains. These case studies are presented in Chapter 3.

2.4.6. CRISP-MED-DM Compliance Assessment Model

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

The goal to assess how compliant the DM project is to the methodology requires that activities and their outcomes shall be measurable. Measurement issues at this level may relate to specific process model activities or deliverables. However, regardless of 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 the methodology compliance assessment. Bearing that in mind, the proposed assessment model possesses certain flexibility.

The following assumptions set 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 research.

2.4.6.1. Assessment and evaluation model

Two evaluation strategies are proposed. The first one is based on the presumption that each phase of the process model has the same importance. An exception is made for the last phase "Deployment", the activities of which shall be treated as a utilization of the actual DM process results.

The second proposed approach is based on the sequential nature of the CRISP-DM process model. The outcomes of each phase define and shape the consequent phase. To accommodate that principle in the evaluation model, we assign gradually decreasing weights from the first to the last phase. For the weight distribution between phases, an arbitrary quadratic function formula was used.

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.

In the first evaluation strategy, each phase except the Deployment phase is assigned with 10 commutative points, representing the maximum score achieved when all non-optional activities of CRISP-MED-DM have been completed. Dependent on the amount of activities per phase, each phase's non-optional activity is evaluated with certain points as stipulated in Table 8.

In the second evaluation strategy, each phase is evaluated differently. Accordingly, each phase's non-optional activity is evaluated with the points derived from the cumulative score of the phase divided by the number of activities in it.

Table 8. Scoring CRISP-MED-DM activities

Phase	**Number of activities in phase**	**1st strategy**		**2nd strategy**	
		Activity evaluation points	Unweighted evaluation max points	Activity evaluation points	Weighted progressive evaluation max points
Problem understanding	**9**	1.11	10	4.00	36
Data understanding	**9**	1.11	10	2.78	25
Data preparation	**15**	0.67	10	1.07	16
Modelling	**9**	1.11	10	1.00	9
Evaluation	**3**	3.33	10	1.33	4
Deployment	**4**	2.50	10	0.25	1

The list of CRISP-MED-DM tasks, activities, deliverables and metrics according to the 1st strategy is provided in Table 9.

2.4.6.2. Evaluation of compliance assessment

CRISP-DM and accordingly the CRISM-MED-DM reference model include many activities not related directly to the DM process, but rather to the phases of the knowledge discovery process, especially its management and organizational part. These activities are important for larger scale DM engagements, but could become an overhead in smaller ones. The assessment examples are provided in Sections 3.1.8.1, 3.2.9.1, and 3.3.8.1.

Due to this reason, it is difficult to justify an 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 complex ones.

The results of an actual DM project's assessment using the proposed evaluation models provides a comparable total project score, or scored CRISP-MED-DM phases, which can be visualized with a Radar plot as shown in Fig. 19.

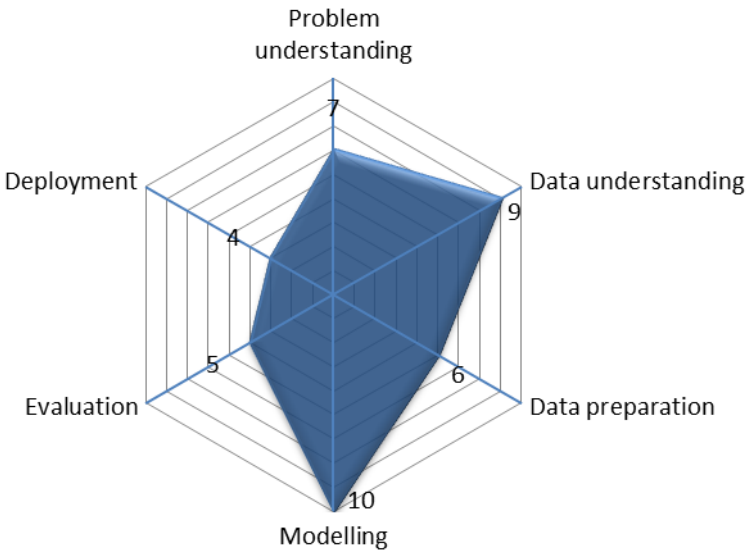


Fig. 19. Example radar plot of DM project assessment

2.4.6.3. 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 with related activities that makes specific assumptions in a 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 9.

Table 9. CRISP-MED-DM tasks, activities, deliverables and metrics

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

Generic tasks	**Specialized tasks and activities**	**Deliverables**	**Part of assessment**	**Points**
Phase 1: PROBLEM UNDERSTANDING (total 10 points)				
GT1. Determine overall objectives	Define clinical objectives* Define healthcare management objectives*	Overall objectives	R	1.11
	Define success criteria	Overall success criteria or vision statement	R	1.11
GT2. Assess situation	Inventory of Resources	Project resource list	R	1.11
	Data availability and integrity evaluation*	List of data sources Data access evaluation	R	1.11
	Patient privacy and legal constraints*	Evaluation of legal requirements and limitations in data usage	R	1.11
	Requirements, assumptions and constraints	DM project resources, costs, timelines assessment	R	1.11
	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 DM goals	Define approved approaches (golden standard)*	DM goals	R	1.11
	Define success criteria	List or hierarchy of DM success criteria	R	1.11
GT4.Plan activities	Project plan Plan data collection*	Overall plan Data collection plan	R	1.11

Phase 2: 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	1.11
	Evaluate semantic data interoperability*	Semantic interoperability analysis report	R	1.11
	Define nomenclatures, classifiers and ontologies used*	List of medical nomenclatures, classifiers and ontologies used	R	1.11
	Define clinical data modeling standards and protocols used*	Mapping of used clinical models, protocols	R2	1.11
	Design non-standard data pre-processing*	Prepare strategy and design for handling multi-relational, temporal, non-structured data (media, text).	R2	1.11
GT5. Collect initial data	Acquire data	Initial data collection report	R	1.11
GT6. Describe data	Describe available data sources, and raw datasets*	Data model Clinical data meaning report	R	1.11
GT7. Explore data	Conduct statistical exploratory data analysis	Exploratory analysis report	R	1.11
GT8. Verify data quality	Verify data quality of available raw datasets*	Data quality report Medical expert data quality assessment	R	1.11
Phase 3: DATA PREPARATION (total 10 points)				
GT9. Prepare data*	Implement interfaces of stand-alone systems *	Stand alone IS are interfaced	R2	0.67
	Prepare medical terminologies mapping*	Medical terminologies mapped	R2	0.67
	Analyze and preprocess data from different sources, based on the	Clinical data models mapped	R2	0.67

	agreed clinical data models and protocols*			
GT10. Extract structural data*	Text data processing*	Preprocessed data, suitable for the planned text mining	C	0.67
	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	0.67
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	0.67
	Handling outliers*		R2	0.67
	Handling semantic data errors*		R2	0.67
GT13. Construct data	Normalization	Constructed data	O	
	Discretization		O	
	Production of attribute derivatives		O	
	Unifying medical terminologies in datasets *		R2	0.67
	Unifying units of measurement in datasets *		R2	0.67
	Unifying clinical data models and protocols in datasets *		R2	0.67
GT14. Integrate data	Aggregate multi-table data to single-table*	Aggregated, merged data	C	0.67

	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	0.67
	Convert datasets syntaxes to modelling format	Datasets ready for the selected DM tools	R	0.67
	Convert data to first-logic clauses format*	Data in first-logic clauses format ready for ILP inference	C	0.67
	Format background knowledge *	Facts in first-logic clauses format ready for ILP inference	C	0.67
Phase 4: MODELLING (total 10 points)				
GT16. Select Modelling Technique	Select technique w.r.t.: • Techniques appropriate for problem • Understandability/interpretation requirements • Constraints	Modelling Technique Modelling Assumptions	R	1.11
GT17. Generate Test Design	Generate model design w.r.t. testing and evaluation criteria Compare model design with DM goals	Test design	R	1.11
GT18. Build Model	Set algorithm parameters*	Parameter settings	R2	1.11
	Run the selected DM techniques	Models	R	1.11
	Post-process DM results	Ready for evaluation DM model results, e.g. trees, rules Model Description	R	1.11

GT19. Assess Model	Test and evaluate results w.r.t. evaluation criteria and test design	Model assessment Revised Parameter settings Assessment	R	1.11
	Prepare for next modelling iteration if needed*	Revised parameter setting Alternative modelling technique	C	1.11
	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	1.11
GT20. Prepare model for interoperable use*	Export model definition to PMML *	Predictive model in PMML standard	C	1.11
Phase 5: EVALUATION (total 10 points)				
G21. Evaluate Results	Understand and interpret the results	Assessment w.r.t. Overall Success Criteria	R	3.33
	Evaluate results novelty Compare results to alternative studies*		R	3.33
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	3.33
	Estimate potential for improvement of the current process		O	
Phase 6: DEPLOYMENT				
G24. Plan Deployment	• Summarize deployable results • Develop alternative deployment plans • Establish how the model will be deployed within	Deployment plan	C	N/A

	organization's systems Identify possible problems			
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	N/A
G26. Produce Final Report	Develop set of final documentation, including executive summary, presentation, detailed technical report.	Final report & Presentation	C	N/A
G27. Review Project	Interview people involved in the project Summarize feedback Analyze the process retrospectively Document the lessons learned	Experience Documentation	O	N/A

2.4.6.4. CRISP-MED-DM practical approbation

The developed CRISP-MED-DM methodology were used and assessed in predictive DM research projects in the oncology and cardiology domains and descriptive DM for publications meta-analysis research. The case studies are presented in Chapter 3. The subject matter information and novel data analysis methods are described in the Sections 2.5–2.7.

2.5. Breast Cancer Gene BRCA1 Prediction

2.5.1. Background

Breast cancer (BC) is the most common cancer in women worldwide. It is also the major cancer mortality reason among women. According to D. M. Parkin et al. (Parkin, et al., 2005), about 89 % of women diagnosed with BC are still alive five years after their diagnosis in Western countries, which is due to advances in detection and treatment. Survivability is a major concern and is highly related to early diagnosis and an optimal treatment plan.

The process of breast cancer treatment typically starts from the identification of malignant tumors. Hence, information about the tumor from examinations and laboratory and radiology diagnostic tests are gathered. The stage of a cancer is one of the most important factors to define an optimal treatment plan. In general, the staging defines the spread of the cancer and its metastasis in the body. For breast cancer the TNM staging system is typically used, where T – Tumor, N – Nodes and M –Metastasis. First, the patient's T, N, and M category values are determined using gathered examinations and laboratory test data, then this information is combined to determine a disease stage ranging from stage I to stage IV. The stage called *carcinoma in situ*, is an initial cancer stage, indicating a high probability to develop an invasive form of cancer in a short period of time.

Despite significant efforts, scientists still do not know the exact causes and triggering mechanisms of breast cancer; however, some of the risk factors are known, i.e. genetic risk factors, family history, aging, alcohol abuse, and obesity. Therefore the current state of oncology research is highly dependent on genetic, clinical and treatment data collection and its analysis. The growing amount of heterogeneous data being collected in clinical settings highlights the importance of proper DM techniques and application methodology.

In our research, we deal with the issue of cancer suppression genes *BRCA1* mutations. Patients with a pathological mutation of BRCA genes have a 65 % lifelong breast cancer probability. We propose a new approach for the prediction of BRCA1 gene mutation carriers by methodically applying knowledge discovery steps and utilizing DM methods. A novel *BRCA1* gene mutation risk assessment model has been created utilizing a decision tree classifier model. A systematic approach, following CRISP-MED-DM methodology, has been applied through the knowledge discovery process steps.

2.5.1.1. BRCA genes

The gene named BRCA stands for breast cancer susceptibility gene. BRCA are human genes that belong to a class of genes known as tumor suppressors.

In normal cells, BRCA genes help ensure the stability of the cell's genetic material (DNA) and help prevent uncontrolled cell growth. Mutation of these genes has been linked to the development of hereditary breast and ovarian cancer. However not all mutations have a proven breast cancer prognostic effect. A woman's risk of developing (pathogenic) breast and/or ovarian cancer is greatly increased only if she inherits a deleterious (pathogenic) *BRCA* gene mutation. Men with these mutations also have an increased risk of a breast cancer. Both men and women who have harmful *BRCA* mutations may be at increased risk of other cancers.

The identification of patients having a risk of *BRCA* mutations is of great importance. In general, individuals with at least a 5–10 % chance of having a mutation in either gene are considered good candidates for genetic testing. Identifying patients with a *BRCA* mutation allows for the application of risk-reducing preventive medical interventions, which are proven to be life-saving (National Cancer Institute, 2013).

2.5.2. Related Work

BC diagnosis is a medical domain, which has a recognizable footprint in DM applications. A number of articles (Bellaachia & Guven, 2006; Choi, et al., 2009; Delen, et al., 2005) investigate the utilization of various DM methods: support vector machines, artificial neural networks, genetic algorithms, regression, etc. The most popular DM models in the BC domain are diagnostic models that aim to distinguish a benign from malignant tumor, or prognosis models, where patients' survival period is predicted. However, less attention is paid to the more specific topics in the BC domain.

Different risk models are currently used to calculate the likelihood of carrying a *BRCA* mutation. The BRCAPRO, Penn II, Myriad II, FHAT and BOADICEA models calculate risk on the basis of the inclusion of different cancer diagnoses within a family (Panchal, et al., 2008). All models incorporate a family history of breast and ovarian cancer as a main prediction factor. The Penn II model, provided by the Abramson Cancer Center of the University of

Pennsylvania, has the best Sensitivity (0.93) among all mentioned risk models (Panchal, et al., 2008).

2.5.3. Application of DM Methods for BRCA1 Prediction

Firstly, the statistical analysis methods, which are typically applied in medical research, have been applied. The correlation of gathered attributes in the dataset were tested using χ^2 criterion with α=0.95. The cancer reoccurrence survival analysis was performed with the Cox regression model and Kaplan–Meier. The statistical analysis revealed a few statistically significant attribute dependencies. *BRCA1* mutation has statistically significant dependency on family history ($p - value = 0.001$), age ($p - value = 0.001$), tumor grade degree ($p - value = 0.004$), progesterone receptors ($p - value = 0.03$), and triple negative BC ($p - value = 0.001$).

Further analysis of the collected research data was performed applying a set of DM techniques. The details and results of experimental research are provided in Chapter 3, Section 3.1.

2.6. Echocardiography Images Data Analysis

2.6.1. Background

Unstructured and image data are the most common clinical patient data accumulated electronically. As was shown in the introduced CRISP-MED-DM methodology, data pre-processing is a key phase in the whole DM process life cycle. In this section, we propose the techniques and methods for echocardiography images (echocardiograms) pre-processing and feature extraction.

The aorta is the main artery that delivers blood from the heart to the rest of the body. The aortic valve serves as a gateway between the heart and the aorta. When the aortic valve orifice narrows due to calcification or other processes, the left ventricle has to work harder to create more pressure to pump blood out through the valve, and the blood supply might become insufficient. The described condition is called valvular aortic stenosis, or aortic stenosis (AS).

There are three main causes of AS: calcific aortic stenosis, rheumatic aortic stenosis, and congenital aortic stenosis. Calcific stenosis is the most common type.

According to the latest data (Ren, et al., 2014), AS is related to aging and is present in 29 % of individuals older than 65 years and in 37 % of individuals older than 75 years.

The transthoracic two-dimensional Doppler echocardiography permits doctors to diagnose and estimate the severity of aortic stenosis in the majority of cases. When the aortic valve orifice becomes narrower, a pressure gradient develops between the left ventricle and aorta, indicating the aortic stenosis.

The methodology for stenosis severity evaluation was developed in the late 80s and has been in use since then (Hatle, et al., 1980; Skjaerpe, et al., 1985). To assess aortic stenosis severity, a cardiologist has to measure the peak transaortic jet velocity in the aortic valve (AV) and in the left ventricular outflow tract (LVOT), and LVOT diameter. Afterwards, by tracing AV and LVOT blood flow spectrograms, to calculate time velocity integrals, peak and mean gradients, aortic valve area (AVA), and velocity ratio, using a simplified Bernoulli equation and Gorlin's formula (Otto, 2012).

According to the joint guidelines of the American College of Cardiology and American Heart Association, aortic stenosis severity shall be classified with the parameters AV peak jet velocity, AV mean gradient, AVA, and velocity ratio. Severe aortic stenosis is defined as clinical conditions when AVA < 1 cm^2 and mean gradient > 40 mmHg or Jet velocity > 4 m/s. The full list of parameters and their values are presented in Table 10.

Table 10. ACC/AHA guidelines to determine aortic stenosis severity

	No stenosis	Mild	Moderate	Severe
Peak jet velocity (m/s)	< 2.6	2.6 - 3	3 - 4	> 4
Mean gradient (mmHg)	-	< 30	30 - 50	> 50
AVA (cm^2)	-	> 1.5	1 -1.5	< 1
Indexed AVA (cm^2/m^2 BSA)	-	> 0.9	0.6 - 0.9	< 0.6
Velocity ratio	-	> 0.5	0.25 - 0.50	< 0.25

2.6.2. Related Work

Medical image analysis as a topic of its own has developed successfully in the last decades. There are a number of papers using signal processing techniques to structure and compare echocardiograms and electrocardiogram (ECG) signal images (Kurgan, et al., 2001; Sacha, et al., 2000; Shalbaf, et al., 2013; Sani, et al., 2014), where Fourier and wavelet transformation have been applied.

The approach and implementation for the semi-automated categorization of medical images was proposed by T. M. Lehmann, M.O. Guld, et al. (Lehmann, et al., 2005). Instead of using commonly applied features describing color and shape, the authors proposed applying texture measure and resized representations of the images, like coarseness, contrast, directionality, and properties of edges within an image as global feature vectors.

T. Tak et al. (Tak, et al., 1996) analyzed the intensity of the regurgitant signal obtained by continuous-wave Doppler to indicate the severity of aortic regurgitation. The methods applied by the authors included the calculation of mean pixel intensity and statistical analysis of the grouped image sets.

To the best of our knowledge, there is limited research addressing processing and analysis of blood flow Doppler echocardiography images, aiming to extract features required for the diagnosis of cardiovascular diseases.

2.6.3. Mathematical Modelling of Time-Related Blood Velocity Changes

The time-dependent blood flow velocity measurements allow us to construct dynamic models of the processes (Noordergraaf, 2011). The ultrasound visualization of blood flow, whether color flow or spectral Doppler, is obtained by measuring the substance movements detected by a captured reflected ultrasonic beam. In a generic ultrasonic scanner, a series of pulses are transmitted to detect blood movement. Echoes from stationary tissue are the same from pulse to pulse. In contrast, echoes from moving particles exhibit a shift in time for the signal to be returned to the receiver. These differences are usually measured in terms of the phase shift from which the Doppler frequency is obtained.

In our experiments, blood flow velocity was measured using the Color Doppler and Pulse Way Doppler modes (PWDM). A noise filter with default cut-off values was used. The size of the wave gate in the PW Doppler was chosen in relation to the diameter of the measured artery. The average ultrasonic velocity was 1540 cm/s. The measured blood flow had a real-time graphical visualization with a waveform echocardiogram (Fig. 20).

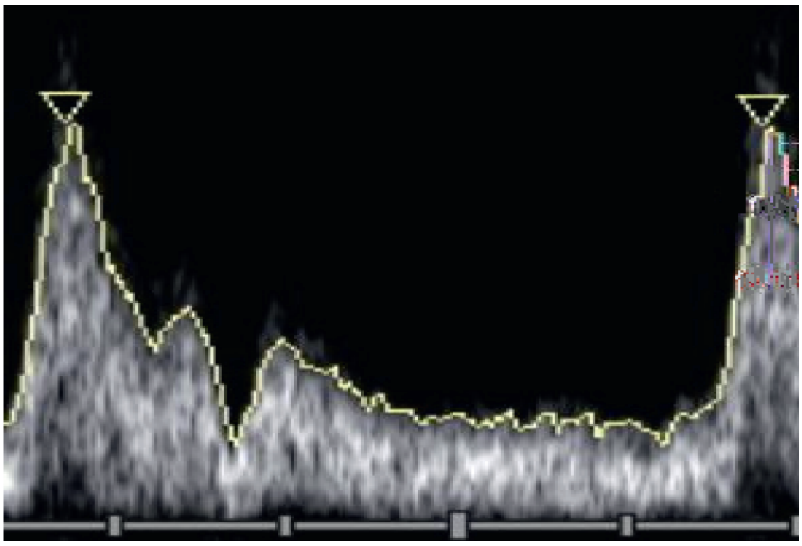


Fig. 20. Echocardiogram visualizing blood Doppler measurements

2.6.3.1. Measurement errors

The measurement results depend on the features of the ultrasonic signal, the body reaction to the signal, the ultrasound system settings, the type of transducer used, and the skills of the operator. If measurement is performed correctly, according to the specification of ultrasound medical device, the measurement errors varies approximately within 2–5 %.

2.6.3.2. Approximation

As is shown in the echocardiogram (Fig. 20) and the graph (Fig. 21), in the course of one cycle the blood velocity is not a monotonous time function. The echocardiogram reflects the complicated blood compression and decompression activity:

- during one heart period (marked by triangles), the blood in the artery is both influenced by the heart muscle and is flowing freely without the impact of an external moving force. Such a

complicated blood flow is reflected by numerous local minima(1–17) and maxima(1'–17') in the blood velocity distribution (Fig. 20);

- the maximum value of the measured blood velocity is 73.7 cm/s, which corresponds to the first maximum 1', and the minimum value is 12.1 cm/s, which corresponds to the minimum value of 15 in the blood flow Doppler echocardiogram (Fig. 20 and Fig. 21). The mean blood velocity value equals to $v = 28.8 \pm 15.8$ cm/s.

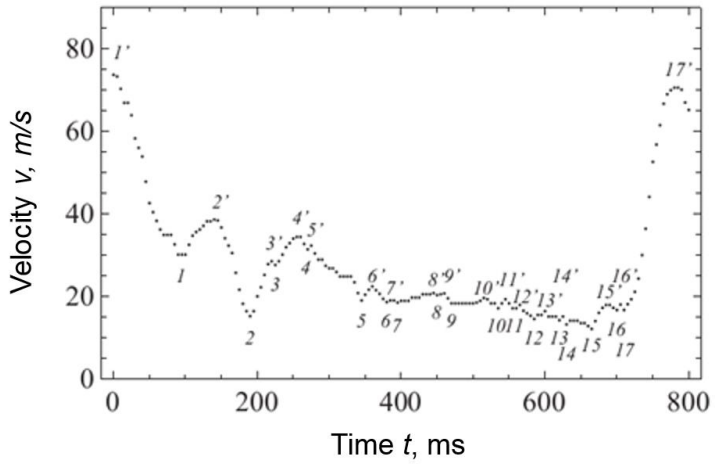


Fig. 21. Blood flow velocity measurement results and approximation

We see that the time interval between the 2'and 4'maxima is related with heart compression, whereas the intervals between the 15' and 17'maxima reflect heart decompression. To pick out a free blood flow velocity interval independent of heart activity, we selected the time interval between 260 ms and 665 ms, i.e. between the 3' maximum and the 15' minimum. It comprises 134 values of velocities from 260 ms to 665 ms at the time of measurement. The distribution of these velocity values is not very regular. Its only regular feature is the general decrease of the Vn values in the sequence of velocities.

The maximum value of blood velocity in this interval is 34.4 cm/s, which corresponds to the initial value and the minimum value is 12.1 cm/s, which corresponds to the end of the interval in Fig. 20. The mean value of the blood

velocity equals to n = 19.9±4.9 cm/s.

Let us take instead of sequence $V(n)$, the sequence $V(n)/V(1)$ in which the initial value ($V(1) = 34.43$ cm/s) and represent it in a double logarithmic scale (Fig. 22).

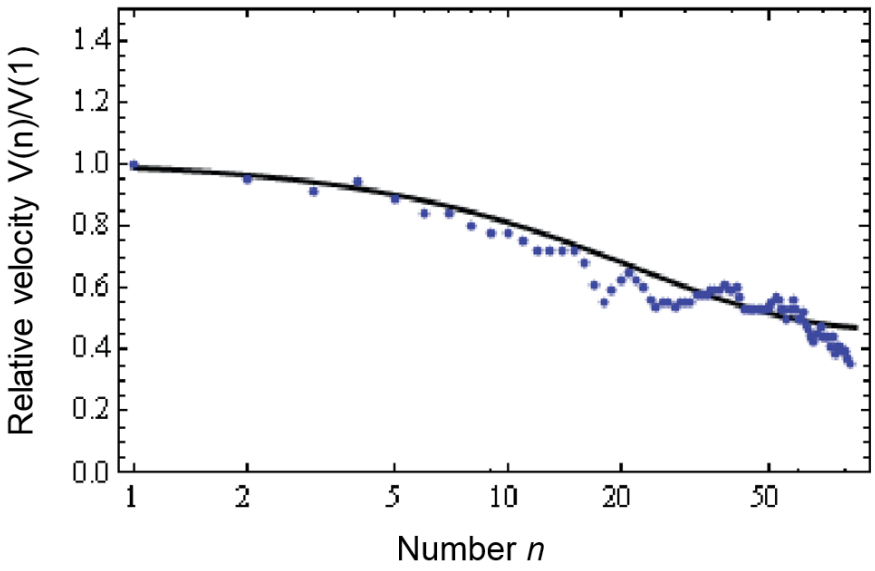


Fig. 22. The distribution of the same velocities in a double logarithmic scale

This dependence looks much more regular. Blood velocity logarithms are approximated by the exponential function $V(n) = ae^{-\beta n} + \gamma$ in which the dimensionless coefficients are:

$$a = 0.5289, b = 0.04521, \gamma = 0.4408 \quad (1)$$

The dependence of velocity distribution should be approximated by the function $V(t) = ae^{-bt} + c$, where

$$a = 18.209\ cm/s, b = 9.0426 \cdot 10^{-3} s^{-1}, c = 15.177\ cm/s \quad (2)$$

Indeed, the summation of dependence V_n and the approximating function $V(t)$ gives a rather good correspondence (Fig. 20). Thus, the velocity distribution should be approximated by the ordinary differential equation:

$$V = -bV + d, V(0) = V_0, \quad (3)$$

where $c = d/b$.

For a quantitative assessment of the obtained approximation, we shall

find the deviations of the obtained function from the real values of the velocities. The deviation interval – standard deviation $\sigma = 5119$ cm/s from the mean value $< \delta V >= 1.051 \cdot 10^{-3}$ cm/s. Most of the deviation values fit within the interval of two standard deviations.

2.6.3.3. The Reynolds number of the blood flow

Blood velocity measurements allow us to estimate the turbulence level of the blood flow. Let us consider the blood flow in the artery as a liquid flow in a pipe. In this case, the Reynolds number characterizes the different flow regimes, such as laminar or turbulent (Landau & Lifshitz, 1987):

$$Re = \frac{VD_H}{v}, \tag{4}$$

where V is the velocity of the liquid fluid in the pipe, D_H is the hydraulic diameter of the pipe, and v is the kinematic viscosity (Landau & Lifshitz, 1987).

The laminar flow occurs at low Reynolds numbers (Re < 2300) when viscous forces in the liquid are dominant as compared to the inertial ones, and is characterized by a smooth, constant fluid motion. The turbulent flow occurs at high Reynolds numbers (Re > 4000) and is dominated by inertial forces, which tend to produce chaotic eddies, vertices and other flow instabilities.

For a circular pipe, the hydraulic diameter D_H is exactly equal to the internal pipe diameter D that in our case is equal to the diameter of the measured artery. The diameter of the artery of our patient is estimated as 1.2 cm. Substituting the mean value of the blood velocity from measuring the interval $\bar{v} = 19.9$ cm/s and the kinematic velocity of the blood $v = (2.8 - 3.8) \cdot 10^{-6}$ cm^2/s (according to (Lenz, et al., 2008)) into expression (4), we obtain the Reynolds number Re = 628–853. This value is much less than the bordering value 2300, meaning that in the measured artery a pure laminar blood flow is present.

The possible correlation to the Non-Newtonian character of the blood flow was discussed in (Liu, et al., 2011). The expression of the Reynolds number (4) allows us to answer the following question: at which value of the velocity in

the artery we have a purely laminar flow. Substituting the same values of the kinematic viscosity of the blood, the diameter of the measured artery and the bordering value of the Reynolds number for the laminar flow (Re < 2300), we obtain the mean value of the blood velocity $v_{mean} = 54 - 73$ cm/s.

On the other hand, the analysis of fluctuations with respect to the trend leads to a deviation from the Gaussian distribution. The insignificant influence of turbulent flows on the ultrasound measurements means a qualitatively different character of the blood flow at the micro level.

The practical value of the proposed model (2) is its usage for the identification of turbulent flow in human body vessels. Although there are no examples of turbulent flow in our collected dataset of echocardiography images, the formula can be used on larger datasets for the feature extraction from echocardiography images to predict clinical conditions comprising turbulent flows.

2.6.4. Echocardiogram Image Data Analysis Method

In general, systolic peak extraction from the diagnostic image involves four steps as visualized in Fig. 23.

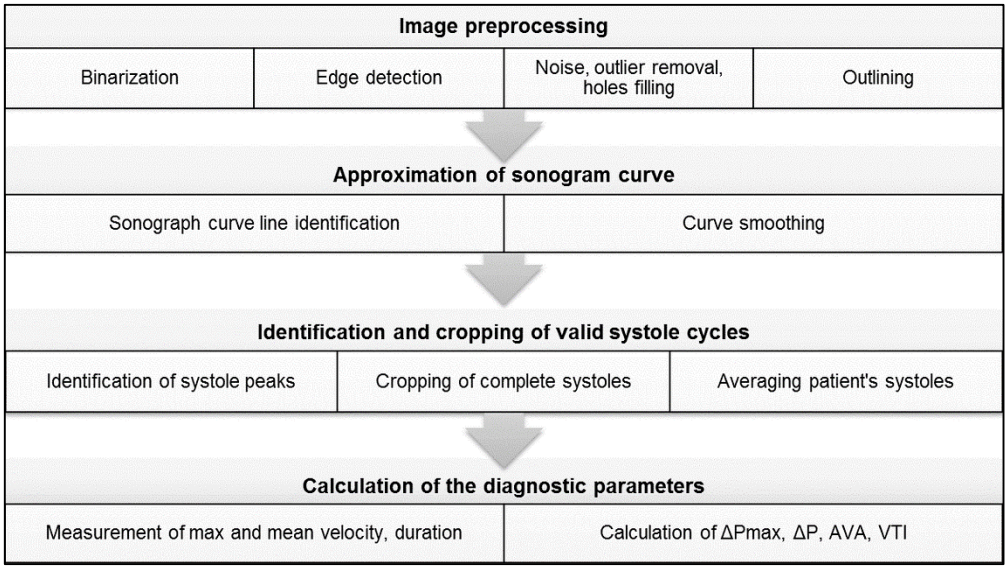


Fig. 23. Semi-automatic aortic stenosis evaluation methodology

The essential precondition was that images, containing valid AV and LVOT echocardiograms, had to be preselected by the cardiologist. Automatic

image type recognition was beyond the scope of our study. The image data pre-processing methods were implemented in R (R Core Team, 2014). ImageJ (Abramoff, et al., 2004; Schneider, et al., 2012) library functions were used for the first step of the image pre-processing tasks.

Step 1 - Image pre-processing

First, echocardiography images were converted to black and white images, using a binary filter. The threshold level was determined using an Isodata algorithm (Ridler & Calvard, 1978), which resulted in brightness cut-off values between 90 and 255 (shown in Fig. 24b). G. Landin's implementation (Landin, 2006) of flood filling algorithm (Soille, 2013) was used for the initial edges smoothing (shown in Fig. 24c). Then, Sobel's edge detector (Ziou, et al., 1998) was used to separate the echocardiogram curve from the rest (shown in Fig. 24d). Two 3 by 3 convolution kernels were used to generate vertical and horizontal derivatives, to produce the final image. An outline filter, which generates a one pixel wide outline of the objects in the image, was applied as an alternative solution. Our experiments showed that the outline filter gave more precise and consistent results on AV and LVOT echocardiography images.

Furthermore, to improve image pre-processing results, outliers, noise, and artefacts such as holes had to be removed. The method based on a flood filling algorithm was used to fill the holes in the 1-bit images. To reduce noise artefacts, typical for echocardiographic images, we used a depeckling filter, which replaces a pixel by the median of the 3x3 surrounding pixels when it deviates by certain threshold. The results of the described pre-processing steps of binarization, holes filling, outlining and depeckling are presented in Fig. 24.

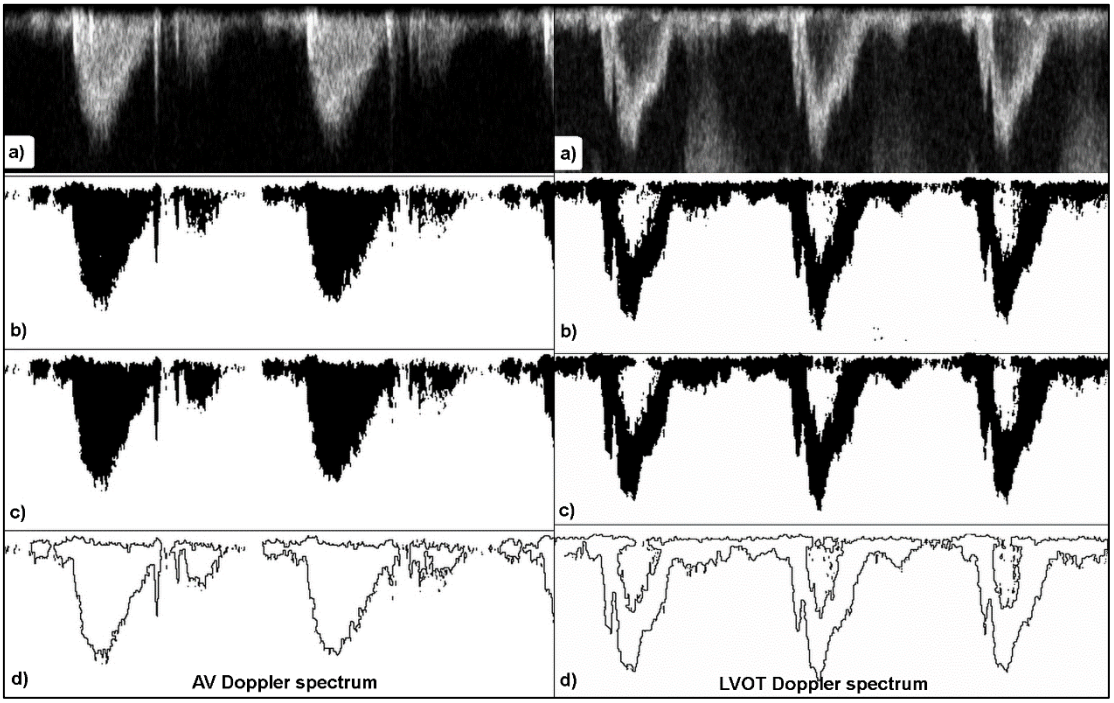


Fig. 24. Initial pre-processing steps of blood flow echocardiography images.

AV flow images on the left. LVOT flow images on the right. Pre-processing steps in horizontal layers from top to bottom: a) – original image, b)- binarized image, c) images with filled holes, d) outlined after despeckling filter images.

Step 2 - Approximation of the blood flow curve

The resulting image from the image pre-processing step represents an outline of the systole, matching the one measured by ultrasound equipment. However, systoles have a full closed contour and in some cases of LVOT images an inside closed contour (e.g. the bottom-right image in Fig. 24). These excessive data were ignored by considering only the 10^{th} decile of data.

Another issue encountered were random notches, which were a result of the Doppler measurement signal noise, captured in the images. We addressed this issue, smoothing the curve with the help of local polynomial regression fitting (Cleveland & Loader, 1996). In this method, fitting is done locally by weighted least squares. Fitting values of a data point X (representing a curve in our case) is made using neighboring points, weighted by their distance from X. The size of the neighborhood is set by parameter φ. The degree of smoothing was tuned empirically and the best results were achieved, with $\varphi \in [0.1;\ 0.2]$, and the 2^{nd} degree of the polynomials. The illustration of the steps

is shown in Fig. 25.

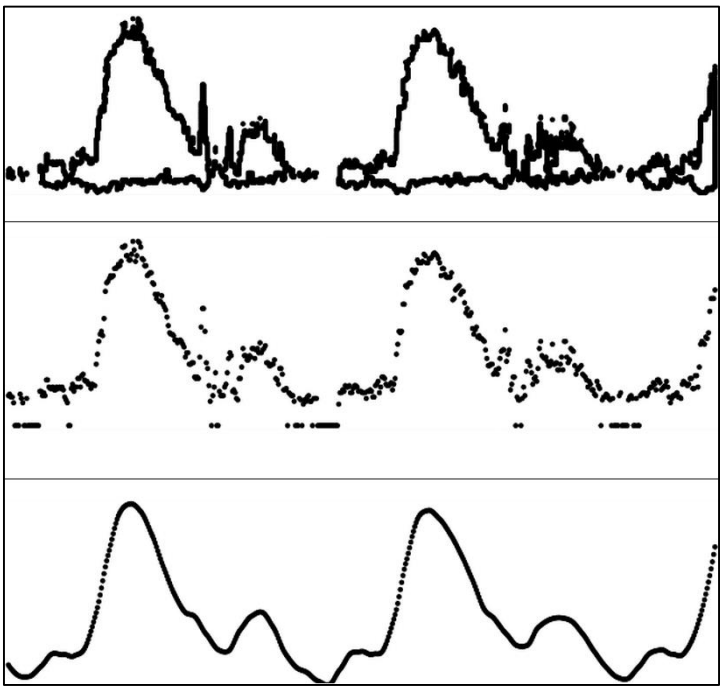


Fig. 25. Smoothing steps of AV blood flow curve.
The top image – resulting image of the pre-processing step; the middle – 10^{th} decile; the bottom – interpolated by local polynomial.

The final result of this step is a clean, smooth blood flow velocity echocardiogram, which is usable for measurements and calculations.

Step 3 - Identification and cropping of the valid systole peak cycles

Since echocardiograms might contain a varying number of peak systolic cycles (in our study – from one to three), a method for identifying and cropping a complete systole cycle was created. Considering the range of possible peaks frequency, and possible minimum and maximum peak values, we have eliminated smaller peaks. The incomplete systoles have been rejected as well. Continuing the example of AV blood flow curve pre-processing (Fig. 25), the identified systole cycles are presented in Fig. 26.

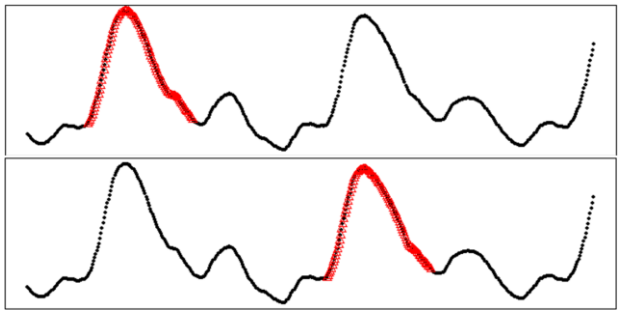


Fig. 26. The identified full AV systoles.

Two complete systoles in the blood flow curve are identified. The top image: the first complete systole cycle. The bottom image: the second complete systole cycle.

The pseudo-code of the simplified systole identification algorithm is as follows:

```
#Finding systole cycles
Peaks = func_Find_all_peaks(image)
For each Peaks[i] {
 If (Peaks[i] > min_peak_const AND EXIST(Peaks[i-1] AND Peaks[i+1]) {
    If (High_Peaks[i]
       AND Peak[i-1] < min_systole_bottom_const
       AND Peak[i+1] < min_systole_bottom_const
    AND duration(Peak[i+1]- Peak[i-1]) BETWEEN min_duration_const
AND max_duration_const) {
              full_cycle = full_cycle + peak[i-1;i+1]
    }
    next
 }
}

Function func_Find_all_peaks(image) {
 For x=2:length(image) {
    If ((image[x-1] < image[x] AND image[x+1] > image[x]) OR
          (image[x-1] > image[x] AND image[x+1] < image[x])) AND
          |image[x-1]-image[x]| > min_threshhold_const AND

          |image[x+1]-image[x]| > min_threshhold_const AND)
       Peaks = peaks + image[x]
 }
}
# where min_peak_const, min_threshhold_const,
min_systole_bottom_const, min_duration_const arbitrary parameters,
defined by the clinical domain experts and applied to a certain image type.
```

By applying the algorithm described above, a set of AV and LVOT systole cycles have been captured for each patient. Between one and five systole cycles per patient have been captured by the algorithm for further processing.

Step 4 - Calculation of the diagnostic parameters.

Before performing calculations, images have been scaled to the predefined Doppler ultrasound images axis values. All diagnostic echocardiograms had a fixed duration of 2 seconds on abscissa, and variable velocity value on ordinate. The average systole cycle was derived by local polynomial regression fitting (Fig. 27). Finally, the parameters - duration and peak systolic velocity ($Vmax$) - were directly calculated, as the cropped parabola's length on abscissa, and its height on ordinate, respectively. For VTI calculation, the curve was fitted with a 2nd degree polynomial and its definite integral was calculated. Our experiments showed that higher order polynomials tend to overfit and have scalability problems.

Other required parameters were calculated using formulas provided in Table 21.

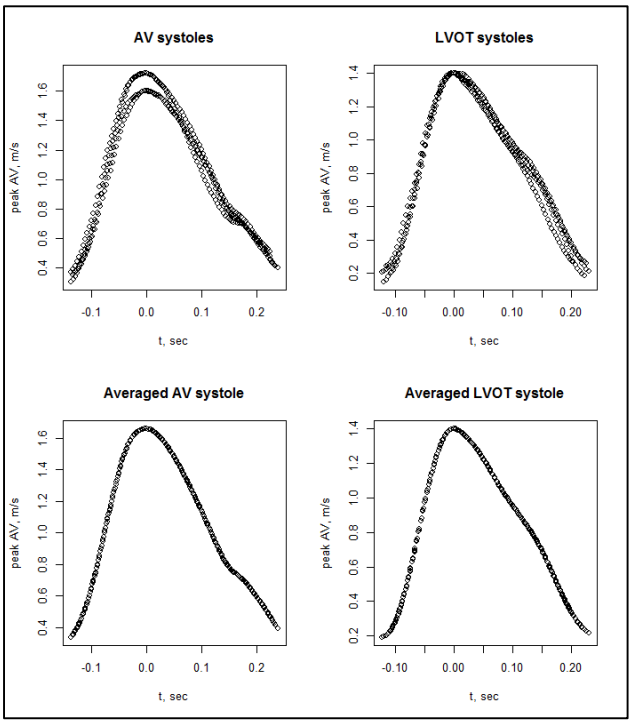


Fig. 27. Resulting AV and LVOT systoles

2.6.5. Predictive Data Mining for Grading Aortic Stenosis

Current diagnostic practice requires a sheer amount of manual image processing. This is a labor-intensive process and is also prone to human-factor errors. Therefore, a semi-automatic image data analysis tool, which utilizes echocardiography images analysis described in Section 2.6.4, and DM classification methods, was created to help medical practitioners minimize time-consuming image processing tasks and propose computer-aided diagnosis.

According to the CRISP-MED-DM process model, after Phase 3 "Data preparation", the modelling activities are to be started. The iterative process flow of optimum algorithm selection, described by Špečkauskienė et al. (Špečkauskienė & Lukoševičius, 2009) was used. The details and results of the predictive DM methods application are provided in Chapter 3, section 3.1.

2.7. Multi-relational Clustering

2.7.1. Background

Clustering methods have been studied for decades in statistics in DM. Clustering can be defined as a DM task, where objects are being *unsupervisedly* subdivided into groups, in such a way, that objects of each group are more similar to each other than in comparison to the objects in other groups. Clustering algorithms represent one of the following clustering method groups: hierarchical methods, e.g. EM clustering (Dempster, et al., 1977), partitioning methods, e.g. K-means, Partitioning Around Medoids (Kaufman & Rousseeuw, 1987) , density-based methods, e.g. DBSCAN (Ester, et al., 1996), model-based methods (Fraley & Raftery, 2002), spatial clustering (Ng & Han, 2002), and fuzzy clustering (Bezdek, et al., 1984). However, the majority of these clustering methods have been created to process data in a "single table" format. Therefore, standard clustering algorithms underperform in high-dimensional and multi-relational data.

For multi-relational clustering, Partitioning Around Medoids (PAM) as a base clustering algorithm was used. PAM was proposed by L. Kaufman and

P. J. Rousseeuw in 1987 (Kaufman & Rousseeuw, 1987), and is regarded as a follower of a k-means algorithm. The choice in favor of PAM has been made due to its high scalability, resistance to outliers, applicability in non-Euclidean space and its underlying feature, allowing us to use a distance matrix as input data. In the PAM algorithm, arbitrary data points as initial centers, called medoids are set. Then, the algorithm minimizes the sum of the dissimilarities between each object and its corresponding reference point and reassigns each object to the nearest medoid.

A basic PAM algorithm flow is as follows:

1. Randomly select k of the n data points as the medoids.
2. Associate each data point to the closest medoid, using a preselected distance measure (similarity measure).
3. For each medoid m:
   3.1. For each non-medoid data point o:
      3.1.1. Swap m and o and compute the total cost of the configuration.
4. Select the configuration with the lowest cost.
5. Repeat steps 2 to 4 until the solution is stable.

The objects similarity measure is of key importance. In this section, a novel similarity measure suited for multi-relational data is proposed. It reflects the relational features of the input data, i.e. attributes in multiple entities, and one-to-many joins between them. Using the introduced compound similarity measure, based on Gower and Ochiai metrics, the distance matrix is calculated, and later used with partitioning clustering methods.

Use-case application of the medical publications meta-analysis using the proposed multi-relational clustering technique is described in Section 3.3. It is worth noting that the algorithm can also be used for a wide range of multi-label classification tasks.

2.7.2. The Similarity Measure in Multi-Relational Settings

Relationally connected data structures, having numeric and nominal values, are hardly represented in Euclidean space. In this case, the classical distance measures, being used in distance based clustering methods, like Manhattan, Minkowski or Euclidean distances, are not suitable. For mixed data types, Gower's general coefficient of similarity (Gower, 1971) can be used as a base. Gower's coefficient of similarity s_i is defined as follows:

$$s_{i,j} = \frac{\sum_k w_k s_{ijk}}{\sum_k w_k}, \tag{5}$$

where: s_{ijk} denotes the contribution provided by the k_{th} variable dependant on its data type, and w_k is the assigned weight function. In other words, the similarity measure of the two objects i & j, is a sum of normalized weighted similarities of each object's variable k (attribute of the entity).

The calculation of s_{ijk} depends on the data type as described below. For nominal variables:

$$s_{ijk} = 1, \text{ iff } x_{ik} = x_{jk}, \text{ and } s_{ijk} = 0, \text{ when } x_{ik} \neq x_{jk} \tag{6}$$

For numeric variables:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k}, \tag{7}$$

where: r_k is a difference between max and min values of k'th variable. As in the case with nominal variables, s_{ijk} equals to 1, when $x_{ik} = x_{jk}$. And s_{ijk} equals to 0, when x_{ik} and x_{jk} represent maximum and minimum values of the variable.

Binary data is treated as a nominal data type. In this case, $s_{ijk} = 1$, iff the compared values are equal to 1. Additionally, it shall be stated, that for the cases where all variables are of a binary type, another similarity measure might be more preferable, like the Jaccard similarity coefficient (Jaccard, 1901).

Furthermore, to compare two value lists in the case of comparing objects with one-to-many relations, we propose to use Ochiai (also called Ochiai-Barkman) coefficient, as proposed by A. Ochiai (Ochiai, 1957). Hence,

when comparing objects (entities), consisting of the nominal attributes, and having other *one-to-many* related entities, s_{ijk} is defined as:

$$s_{l1,l2} = \frac{n(l_1 \cap l_2)}{\sqrt{n(l_1) \times n(l_2)}}, \tag{8}$$

where $l1, l2$ – nominal value lists, $n(l)$ – the number of elements in l.

In a relational data structure, the compared objects are represented by a number of relations and relational joins. For each attribute of a relation, denoted as a variable k, which is considered to be a part of the selected search space, atomic similarities s_{ijk} have to be calculated using the Gower similarity for a specific data type, value lists using Ochiai coefficient extended by Gower similarities for numeric and binary data types. Finally, the overall similarity measure between two objects is calculated as a weighted sum of s_{ijk} according to (5).

A relational data model always has to be treated with care, and certain pre-processing, de-normalization has to be applied. Considering the whole available relational data might be impractical. Hence, only valuable entities and attributes have to be selected. There are various recommendations on the relational feature selection, e.g. as described in works of R.T. Ng and J. Han (Ng & Han, 2002).

The selected entities of the data model shall be analyzed for de-normalization possibility, assuming their relational join type. Entities with one-to-one type joins can typically be easily merged. For the entities connected with one-to-many joins, Ochiai with Gower coefficient for numeric, binary data types shall be used. Many-to-many related entities in many cases can be de-normalized to a one-to-many relationship.

2.7.2.1. Edit distance for multi-relational data

Another approach to measure distance between multi-relational objects is to represent relations and their joins as ordered labelled trees and apply a tree edit distance (TED) measure to calculate dissimilarity of the given trees. The tree edit distance problem is well studied and a number of algorithms have been

proposed (Tai, 1979; Demaine, et al., 2007; Pawlik & Augsten, 2011), with state-of-the art algorithms running in $O(n2m(1 + \log \frac{m}{n}))$ time and $O(mn)$ space (Demaine, et al., 2007).

The tree edit distance between ordered labeled trees is defined as a minimum set of a tree node edit operations that transforms one tree into another. The following edit operations are performed:

1. insert a node;
2. delete a node;
3. rename the label of a node.

The cost of each edit operation can be weighted, thus allowing parameterization of the TED algorithm.

The modern algorithms run in polynomial time, utilizing dynamic programming and robust algorithms, e.g. trees are recursively divided into sub-trees, and then sub-solutions are cached and later reused for comparing compound trees.

General tree edit distance algorithm

Given: trees A and B. Task is to compute the distance between the trees.

First, a path in one of the trees is chosen, and the distances between the relevant sub-trees of both sub-trees are computed. Those distances will be cached and later reused. Second, the distance between the trees is computed in a bottom-up manner computing the distances between the relevant sub-forests of A and all corresponding sub-forests of B, utilizing all cached distances.

In general, TED algorithms use a left path, right path or heavy path strategy to compute the distance. To do so, the algorithm performs the following steps:

1. For a given pair of trees A, B look up the path in the path strategy.
2. If the path is in A do the following steps, otherwise reverse the trees (B, A) and continue from the step (1):
    a. Run the algorithm for every relevant sub-tree A' in A, and the tree B.

b. Compute the single-path function for A, B trees according to the path's type (left, right, heavy path).

The proposed algorithms (Tai, 1979; Demaine, et al., 2007) have unequal performance in balanced and imbalanced trees. However, a robust TED algorithm called RTED, proposed by Pawlik and Augsten (Pawlik & Augsten, 2011) performs well with various tree shapes, and is currently one of the best performing algorithms.

The utilization of RTED and the proposed multi-relational similarity measure implementation for clustering tasks is compared in Section 3.3.

2.7.3. Applying Multi-relational Clustering for Exploratory analysis

In multi-relational clustering, the distance between two objects is computed by not only relying on the attributes of the objects, but also including the objects related to them as well as considering the type of the relation, i.e. one-to-one, one-to-many, many-to-many, and the semantic strength of the relation.

A few studies propose clustering methods for multiple relations (Kirsten, et al., 2001; Neville, et al., 2003; Yin, et al., 2005; Yin, et al., 2006). In principle, two generic approaches have been used: problem transformation (data reducing to propositional form), and algorithm modification to multi-relational form by updating the key notion. The second group of methods extends traditional DM algorithms or uses specific techniques, e.g. first-order logic and Inductive Logic Programming in order to handle multi-relational data.

In this study, the feature selection task is omitted, relying on the existing knowledge of the domain experts. Hence, we undergo strong user guidance in feature selection and concentrate on the heuristics for similarity measure corresponding to the given relational data structure calculation. A method outlining how to apply semi-automated feature selection with multi-relation clustering, is described by X.Yin, J. Han, et. al (Yin, et al., 2005).

As proposed by Van Laer et al. (Van Laer & De Raedt, 2001), we have upgraded the propositional algorithm to the first-order learners type, retaining

much of the original algorithm, and changing only the key notion, which in our case is the distance measure or its direct derivative - similarity measure.

Following T. Horvath et al. (Horvath, et al., 2001) we store all available objects, comprising aggregated distance measures. Later, we compare each object with neighboring objects, using the PAM algorithm.

2.8. Generalization and Conclusion

In this chapter, we showed that the DM methodology CRISP-DM is limited in support of the issues and constraints specific to medical domain. The introduced methodology CRISP-MED-DM extends the CRISP-DM in the following phases: business understanding, data understanding, data preparation and modelling. The proposed additional tasks and activities reflect the uniqueness of medical DM, described in Chapter 1: heterogeneous data, semantic interoperability, missing values, patient data privacy and legal constraints. Moreover, three use-case studies and their supporting theories were described.

First, the CRISP-MED-DM methodology was applied for *BRCA1* gene mutation predictive modelling in the oncology domain. The iterative process of data pre-processing and the selection of a classification algorithm was used to improve modelling performance. Experimental results are described in Section 3.1.

Second, the medical image data pre-processing technique for cardio echocardiography images data analysis was proposed. The methodology and image processing methods resulted in feature extraction of the blood flow echocardiogram for grading or diagnosing aortic stenosis. Experimental results are described in Section 3.2.

Finally, a method for multi-relational clustering with a novel distance metric was introduced. The proposed compound similarity measure, based on Gower's similarity coefficient and the Ochiai-Barkman coefficient, is suitable for applications with multi-relational data. The proposed clustering experimental approbation is described in Section 3.3.

CHAPTER 3

Approbation of the CRISP-MED-DM and Data Analysis Methods

3.1. Predictive Data Mining: BRCA1 gene mutation predictive model

3.1.1. Introduction

In this section, a new approach for the prediction of *BRCA1* mutation carriers by methodically applying DM methods according to CRISP-MED-DM methodology is described. Background information on breast cancer, BRCA genes and the related work is provided in Section 2.5.

The conducted research aimed to create a novel *BRCA1* mutation risk assessment model, which meets the requirements for interpretability and external validation and has better accuracy than already existing risk assessment models (Panchal, et al., 2008).

As defined in CRISP-MED-DM, the iterative approach of data pre-processing, the selection of an optimal algorithm and its optimal parameterization has to be applied. We have applied an iterative procedure to stratify the initial dataset and transform it into the optimal dataset (ODS) for each classification task; afterwards we found the best performing classification algorithm; subsequently, its parameters were optimized, and finally, the results were validated with the participating oncologists.

3.1.2. Problem Understanding

3.1.2.1. Overall objectives

The overall goal formulated by the clinicians is to investigate a few questions of clinical interest that have not been answered by exploratory analysis by classical statistical methods:

- Do patients with *BRCA1* pathogenic mutation have any specific clinical, morphological manifestations?
- What other patient features or feature groups can serve as predictors of pathogenic *BRCA1* mutation?
- Are there any predictive factors of breast cancer reoccurrence?
- Is there an impact of *BRCA1* mutation on the time of tumor reoccurrence?

3.1.3. Data Understanding

The original medical research was carried out in the Oncology Institute of Lithuanian University of Health Sciences from 2010 till 2013. The study group consisted of 83 women, who were diagnosed with I–II stage breast cancer with the following tumor morphology: T1 N0, T2 N0, T3 N0, T1 N1, T2 N1. The list of observed clinical, morphological features (attributes), as well as interventions and therapies applied is provided in Table 11, together with attribute types and the number of distinct values of each nominal attribute.

The research duration was determined by considering the number of patients and no less than a two year period of disease progress monitoring. As the cancer stage is a strong predictive factor, only the early (I–II stage) breast cancer were chosen in order to reduce the factors influencing the variation.

After laboratory confirmation of pathologic *BRCA* mutation, all patients were divided into two groups: (1) carriers – patients with pathologic *BRCA* gene mutation, and (2) non-carriers – patients without *BRCA* mutation.

Table 11. The full list of attributes of initial dataset

#	Attribute	Attribute type*	#	Attribute	Attribute type*
1	Age	Continuous	18	Triple neg. BC	Nominal (2)
2	Histology type	Nominal (5)	19	Family history type	Nominal (3)
3	cT	Nominal (5)	20	Prostate cancer fam. Hist.	Nominal (2)
3	pT	Nominal (6)	21	Pancreatic cancer fam. hist.	Nominal (2)
4	Multifocality	Nominal (2)	22	Colorectal cancer fam. Hist.	Nominal (2)
5	cN	Nominal (3)	23	Surgery type	Nominal (4)
6	pN	Nominal (2)	24	Chemotherapy type	Nominal (3)
7	G	Nominal (3)	25	Herceptin	Nominal (2)
8	L	Nominal (2)	26	Cht. complications	Nominal (4)
9	V	Nominal (2)	27	Reoccurrence	Nominal (2)
10	ER	Nominal (4)	28	Metastases	Nominal (2)
11	PR	Nominal (4)	29	Time to diseased	Continuous
12	HER2	Nominal (2)	30	Is Diseased	Nominal (2)
13	BRCA mutation	Nominal (6)	31	Monitoring period	Continuous
14	Bilateral BC	Nominal (2)	32	Time to reoccurrence	Continuous
15	Tumor size	Continuous	33	Adjuv. ST	Nominal (2)
16	CHEK2 mutation	Nominal (4)	34	Adjuv. HT	Nominal (5)
17	Affected l_m number	Continuous			

* For nominal attributes, a number of distinct values is given in brackets

3.1.3.1. Exploring data

The collected research data, which formed the initial dataset, had a very imbalanced structure. As shown in Table 12, the carriers made up 14 %, and non-carriers – 86 % of the whole patient group.

Table 12. The distribution of prediction class attributes

Attribute	**Positive attribute value**		**Negative attribute value**	
	Number of patients	Percentage of the whole group	Number of patients	Percentage of the whole group
BRCA1 mutation	12	14 %	71	86 %
BC reoccurrence	22	27 %	61	73 %
Diseased patients	2	2 %	81	98 %

A set of 19 nominal attributes were used for the *BRCA1* mutation classification task. The attributes values frequency tables are visualized by *BRCA1* class colored histograms in Fig. 28. The black color marks items with no *BRCA1* mutation, and the gray color – items with a *BRCA1* mutation. The nominal attributes value distribution, as can be visually seen does not indicate a trivial single nominal attribute value dependency on dependent class variable.

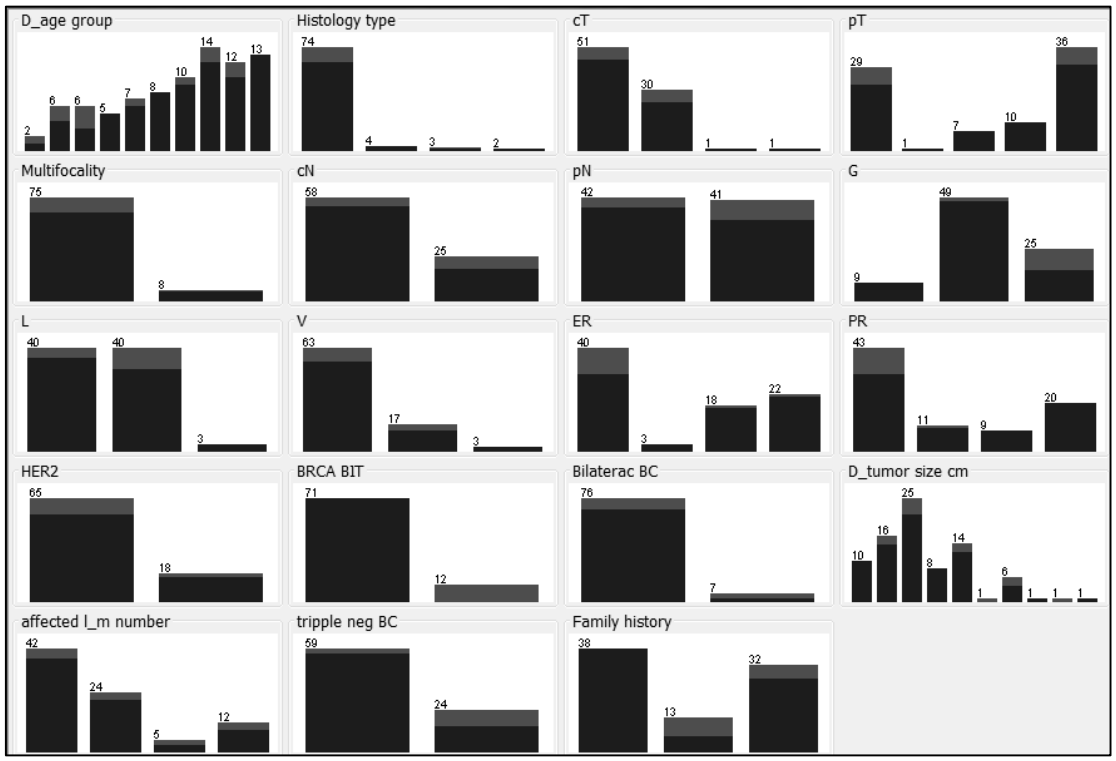


Fig. 28. Histograms of nominal attributes values for *BRCA1* classification task

3.1.3.2. Data quality verification

The dataset was checked for outliers and missing values, using descriptive statistics and scatter plot visualization. Data were of high quality, with no outliers or missing data.

3.1.4. Data Preparation

Continuous attributes *Age, Tumor size, and Time to reoccurrence* were discretized to get results that are more meaningful for clinical interpretation.

Feature selection algorithms Principal Component Analysis, Particle Swarm Optimization based attribute search, and Chi Squared attribute evaluation were used to reduce the dataset's dimensionality.

The initial dataset consisting of 83 items was iteratively optimized to achieve a better performance of DM algorithms. An iterative Optimal Dataset (ODS) was formed by selecting modelling attributes, and stratifying the dataset in respect to the class attribute.

3.1.5. Modelling

According to CRISP-MED-DM methodology, the resulting datasets were iteratively used with a set of classification methods: classification trees, classification rules, multi-layer perceptron, logistic regression, Naïve Bayes classifier, Ada boost and Bagging classifiers. In addition, association rules were used to identify hidden dependencies between dependent and independent variables. Time series analysis was carried out to evaluate if *BRCA1* mutation influences the time to reoccurrence or patient's decease date.

DM software packages WEKA (Hall et al. 2009), Orange (Curk et al. 2005) and Tibco Spotfire Mining (Tibco Software Inc. 2010) were used. The following classification algorithms were compared:

- Classification trees – J48[1], Random Forest, Random tree, tree ensemble;
- Classification rules – ZeroR[2], OneR[3], and FURIA[4];
- Artificial neural networks – Multi-layer Perceptron, SOM[5];
- Regression – logistic regression;
- Bayes – Naïve Bayes;
- Meta – Ada Boost[6], Bagging.

We performed two major iterations of predictive modelling, which are described below.

The first modelling iteration

In the first iteration, the classification algorithms were evaluated on the unbalanced dataset. In addition, the classification results were improved by

[1] J48 – WEKA implementation of C4.5 algorithm
[2] ZeroR – WEKA implementation of classification algorithm, using 0-R classifier
[3] OneR – WEKA implementation of classification algorithm, using 1-R classifier
[4] FURIA - fuzzy unordered rule induction classification algorithm
[5] SOM – self organizing map, a data visualization algorithm
[6] Ada Boost – boosting classification algorithm using Ada Boost M1 method

changing default algorithm parameters. The algorithms parameterization was performed as follows. The Fuzzy Unordered Rule Induction Algorithm (*FURIA*) showed overall performance improvement after changing the uncovered rules handling parameter to "vote for the most frequent class". Main algorithm parameters have been set as follows: T-Norm equals to Product T-norm, error rate $> \frac{1}{2}$ as stopping criterion, two optimization runs, three folds for pruning, random seed equals to one, minimal weight of the instances in a rule equals to three. See the modelling results for the *BRCA1* classification problem in Table 13.

Table 13. FURIA algorithm optimization results

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**ROC AUC**
Furia initial	0.916	0.667	0.958	0.80
Furia optimized	0.940	0.667	0.986	0.81

The adaptive boosting meta-algorithm *AdaBoost* is known for good results with weak classifiers and is more resistant to overfitting. *AdaBoostM1* WEKA implementation was used. We achieved sensitivity improvement from 0.5 to 0.67 by using *DecisionStump* as a basis classifier and increasing the iteration number from 10 to 30. Main algorithm parameters have been set as follows: reweighting resampling was not used, weight threshold for weight pruning equals to 100, random seed equals to one. However, Specificity and ROC area under curve values have reduced. See the modelling results for the *BRCA1* classification problem in Table 14.

Table 14. AdaBoost algorithm optimization results

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**ROC AUC**
AdaBoostM1 initial	0.891	0.5	0.958	0.802
AdaBoostM1 optimized	0.892	0.667	0.930	0.790

The meta-algorithm bootstrap aggregating (bagging) results were improved by choosing *J48(C4.5)* as a base classification algorithm. The main algorithm parameters have been set as follows: one execution slot, ten iterations, random seed equals to one, out-of-bag error was not calculated, bag size percentage equals to 100 %. See the modelling results for the *BRCA1* classification problem in Table 15.

Table 15. Bagging algorithm optimization results

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**ROC AUC**
Bagging with RepTree	0.855	0	1	0.705
Bagging with J48	0.880	0.417	0.958	0.853

The overall result of the first iteration is shown in Table 16 and Table 17.

Table 16. *BRCA1* classifier models performance

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**ROC AUC**
J48 (C4.5)	0.880	0.667	0.915	0.825
Random Forest	0.855	0.167	0.972	0.774
Random tree	0.819	0.333	0.901	0.696
ZeroR	0.854	0.000	1.000	0.428
OneR	0.807	0.000	0.944	0.472
Furia	0.940	0.667	0.986	0.81
Multilayer perceptron	0.819	0.667	0.845	0.805
Multilayer perceptronCS	0.916	0.667	0.958	0.865
Logistic regression	0.795	0.500	0.845	0.738
AdaBoostM1	0.892	0.667	0.930	0.790
Bagging with J48	0.880	0.417	0.958	0.853

Table 17. Breast cancer reoccurrence classifier models performance

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
J48 (C4.5)	0.734	0.000	1.000	0.457
Random Forest	0.71	0.091	0.934	0.516
Random tree	0.639	0.227	0.787	0.484
ZeroR	0.735	0.000	1.000	0.457
OneR	0.675	0.000	0.918	0.459
Furia	0.747	0.091	0.984	0.633
Multilayer perceptron	0.687	0.455	0.770	0.576
Multilayer perceptronCS	0.687	0.455	0.770	0.596
NaïveBayes	0.639	0.136	0.820	0.508
Logistic regression	0.663	0.591	0.689	0.675
AdaBoostM1	0.651	0.000	0.885	0.319
Bagging with J48	0.687	0.045	0.918	0.546

Dimension reduction techniques including Principal Component Analysis, Particle Swarm Optimization based attribute search, Chi Squared attribute evaluation and Correlation Attribute evaluation were used. The methods resulted in different attribute sets. In our experiments, the Particle Swarm Optimization algorithm for the attribute search has shown the best results. However, most of them had significantly worse classification accuracy compared to the dataset with the full set of attributes. See Fig. 29 and Fig. 30 for different classifier models' performance comparison. The only possible advantage of the dimension reduction is a shorter classification model building time, which was not applicable due to the small research dataset.

The second modelling iteration

In the second iteration, we have changed the dataset by incrementally equaling the proportion of dependent binary (class) attribute values until it reached 50 % to 50 % distribution. The balancing of the dataset influenced the performance of most of the classification algorithms. The classifiers derived from the balanced ODS showed 0.90 accuracy, 0.95 Sensitivity, 0.85 Specificity

and 0.96 ROC area value with meta algorithm *Bagging*, and 0.88 accuracy, 0.93 Sensitivity, 0.83 Specificity and 0.85 ROC area value with *J48* tree algorithm.

Further, the initial unbalanced dataset was used as a test dataset for the validation of the model. The comparison of the classifier models' performances was done using ROC and Gain charts. The predictive models with the highest ROC values tested on the unbalanced dataset are presented in Table 18 and Table 19.

Table 18. *BRCA1* prediction classifier

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
Bagging	0.867	0.833	0.873	0.81

Table 19. Breast Cancer reoccurrence prediction classifier

Algorithm	Accuracy	Sensitivity	Specificity	ROC AUC
Bagging	0.711	0.955	0.623	0.65

Compared to the first iteration results, higher Sensitivity was achieved, but in turn Specificity has decreased, resulting in lower ROC performance with value 0.81 for the *BRCA1* classifier which is weaker compared to the performance of the first iteration classifier where ROC AUC was 0.85.

Visual comparison of the best performing *Bagging* algorithm classifiers is provided in Fig. 29 for the *BRCA1* class model and in Fig. 30 for the Reoccurrence class model. The best model for the data is the one with the highest curve above the straight diagonal line.

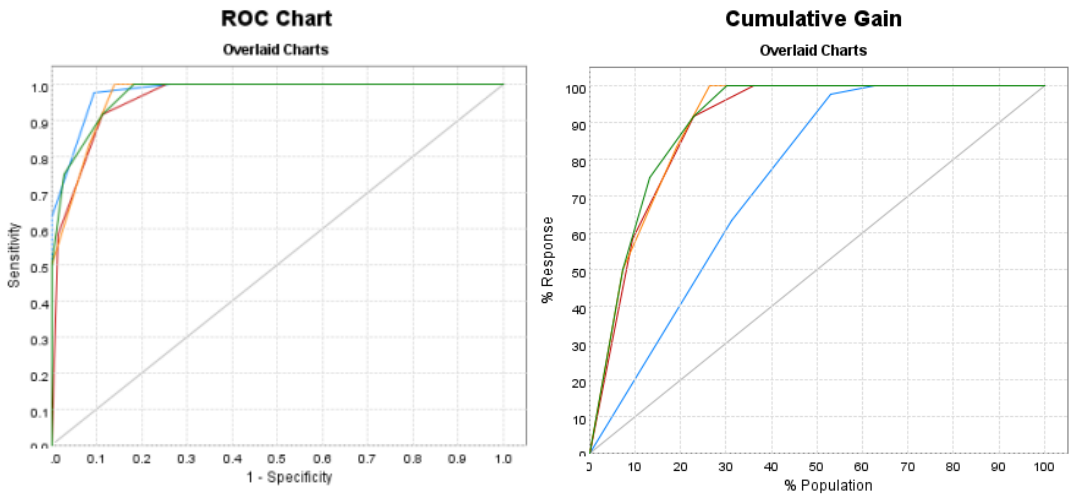


Fig. 29. BRCA mutation predictive models performance charts

In the *BRCA1* mutation classification ROC chart (Fig. 29), the best performance (blue line) is achieved by the *"stratified data"* classifier trained and tested solely on balanced dataset, then the performance gradually decreases as follows: ODS with all 29 attributes (green line), ODS after dimension reduction with 5 attributes (orange line), *"stratified data"* classifier tested on initial dataset (red line). The same color notation is used for the cumulative gain chart. The Gain ranking of the models is as follows: orange, green and red lines have similar Gain values, and then the blue line which represents the *"stratified data"* classifier trained and tested solely on the balanced dataset has a lower Gain value.

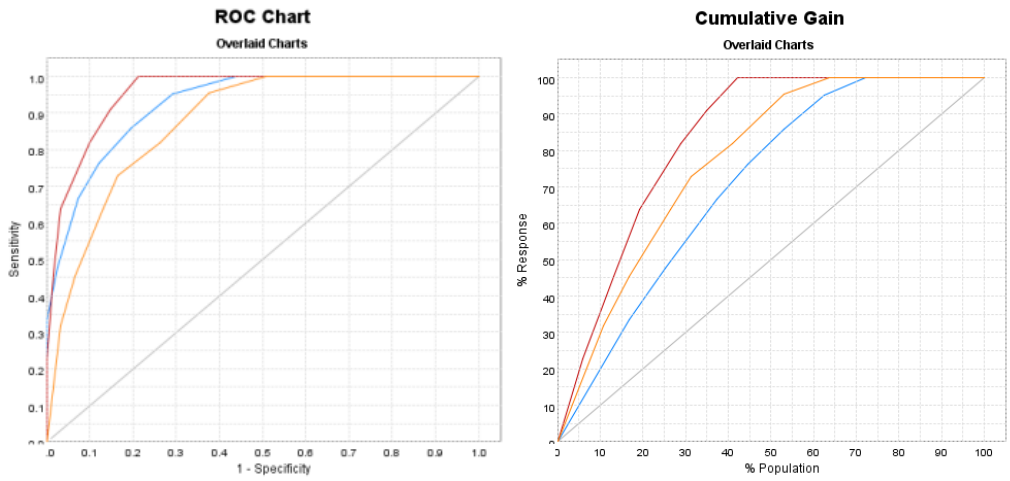


Fig. 30. BC reoccurrence predictive models performance charts

In the BC reoccurrence classification ROC chart (Fig. 30), the best

performance (red line) is achieved by ODS with all 34 attributes. The performance gradually decreases as follows: *"stratified data"* classifier trained and tested solely on the balanced dataset (blue line), and finally *"stratified data"* classifier tested on initial dataset (orange line). The same color notation is used for the cumulative gain chart (Fig. 30). The Gain ranking of the models is as follows: ODS with all 34 attributes (red line), *"stratified data"* classifier tested on initial dataset (orange line), and finally *"stratified data"* classifier trained and tested solely on the balanced dataset (blue line).

3.1.5.1. Hidden patterns analysis

Association rules discovery algorithms were applied to find non-trivial dependencies. *Apriori*, *PredictiveApriori*, and *HotSpot* algorithms. Generic and class specific rules with a minimum support in the range of [0.01; 0.2] with confidence greater than 0.75 were searched.

Three sets of rules were discovered iteratively: class independent, class dependent with the *BRCA mutation* class attribute, and class dependent with *Reoccurrence* class attributes. The search space was incrementally increased: by decreasing the minimum support and confidence values, by increasing the maximum number of antecedents from two to five, and by increasing the associated attribute set from five (attributes found in the 1st iteration by dimension reduction techniques) to the full set of 35 attributes. In the largest search space within our experiments, association rules search has found from 46 thousand to 78 thousand rules. Such an amount of rules is due to the selected lower support and confidence value. The generated rule items were filtered and then analyzed by the oncologists. More than a hundred association rules describing cancer metastases in lungs were discovered by the *Apriori* algorithm. However, all of them were rejected by the clinician expert as being trivial or being possibly resulted by algorithm over-fitting.

3.1.5.2. Time-series analysis

Survival and time-series analysis was performed to find any impact of *BRCA1* mutations to the time of BC reoccurrence or to the time of death. In the cases of

systematic reoccurring BC, we have researched possibilities to predict the localization of metastases. However, neither statistical linear regression or Cox regression, nor DM methods provided satisfactory results. The received results were statistically insignificant or without reasonable accuracy.

3.1.6. Evaluation and Results

The *BRCA1* classifier model with the best ROC AUC value was created using *Multilayer Perceptron* algorithm modification (*MultilayerPerceptronCS* in WEKA) with overall accuracy 0.92, Sensitivity 0.67, Specificity 0.96 and ROC AUC 0.87. However higher classifier Sensitivity and explicit interpretability of a model was required by the clinicians. Therefore decision tree *J48* and decision rules *Furia* classifiers were used for the interpretation by the domain experts. Accordingly, their performance is as follows: overall accuracy 0.88 and 0.94, Sensitivity 0.67 in both cases, Specificity 0.92 and 0.99, and ROC AUC 0.83 and 0.87.

To increase the Sensitivity value, the dataset was balanced and the best classifier results were achieved with the *Bagging* algorithm. Its performance on the test dataset (10-fold cross-validated initial dataset): overall accuracy 0.87, Sensitivity 0.83, Specificity 0.87 and ROC AUC 0.81.

The optimal breast cancer reoccurrence classifier models were created in the second iteration, when the initial dataset was balanced, which significantly improved Sensitivity with remaining similar levels of Specificity and ROC AUC. The achieved performance of the *Bagging* algorithm classifier: overall accuracy 0.71, Sensitivity 0.96, Specificity 0.62 and ROC AUC 0.65. The highest Specificity was achieved applying the *Furia* decision rules algorithm: overall accuracy 0.75, Sensitivity 0.09, Specificity 0.98 and ROC AUC 0.63. The clinical interpretation of the resulting predictive models is presented in Section 3.1.8.2.

3.1.7. Deployment

According to CRISP-MED-DM, the resulting predictive models for *BRCA1* mutation and breast cancer reoccurrence prediction were exported to the PMML

format. PMML models can be used in clinical decisions support systems or generic scoring software in clinical settings.

Following the recommendations of A. K. Waljee et al. (Waljee, 2013), the derived PMML models are provided for further validation of the predictive models with external datasets.

3.1.8. Discussion and Compliance to CRISP-MED-DM

Corresponding to the questions raised by oncologists, three DM problems were formulated and resolved using eleven DM algorithms in accordance with the CRISP-MD-DM process model. The research questions raised were formulated as classification problems. Classification models for the prediction of a *BRCA1* carrier with the dependent variable *BRCA1 mutation,* and for the prediction of BC reoccurrence with the dependent variable *Reoccurrence* were created.

The biggest challenge was the very small size and imbalanced nature of the dataset provided by the participating clinicians. However, iterative optimization of the initial dataset, optimal algorithms selection and their parameterization has resulted in higher classifier model performance, with acceptable prediction accuracy for clinical usage.

By analyzing breast cancer patient data, we have realized the importance of a systematic approach in the knowledge discovery process. The study has shown a high importance of forming an optimal dataset for classification accuracy. A dataset with balanced class attribute values was of key importance. Experimental results have not shown the positive impact of dimension reduction for the model accuracy.

Artificial neural networks have shown the best performance for *BRCA1* gene mutation carrier prediction, but due to its lack of expressivity, decision tree and decision rules methods were preferred by the clinicians.

3.1.8.1. Compliance to CRISP-MED-DM

To evaluate compliance of the undertaken application of DM activities for *BRCA1* predictive modelling, we applied the first evaluation strategy, proposed in subsection 2.4.6.2.

As it is shown in Fig. 31, the core phases Data understanding, Data Preparation, Modelling and Evaluation show good compliance. However, the Problem understanding phase scores only 5.6 out of a maximum 10 points. In the problem understanding phase, the activities related to formal project management, such as activities planning, risk planning, cost/benefit analysis were not performed, due to the exploratory nature of our research project. For the same reason, formal success criteria were not initially formulated by the main stakeholders.

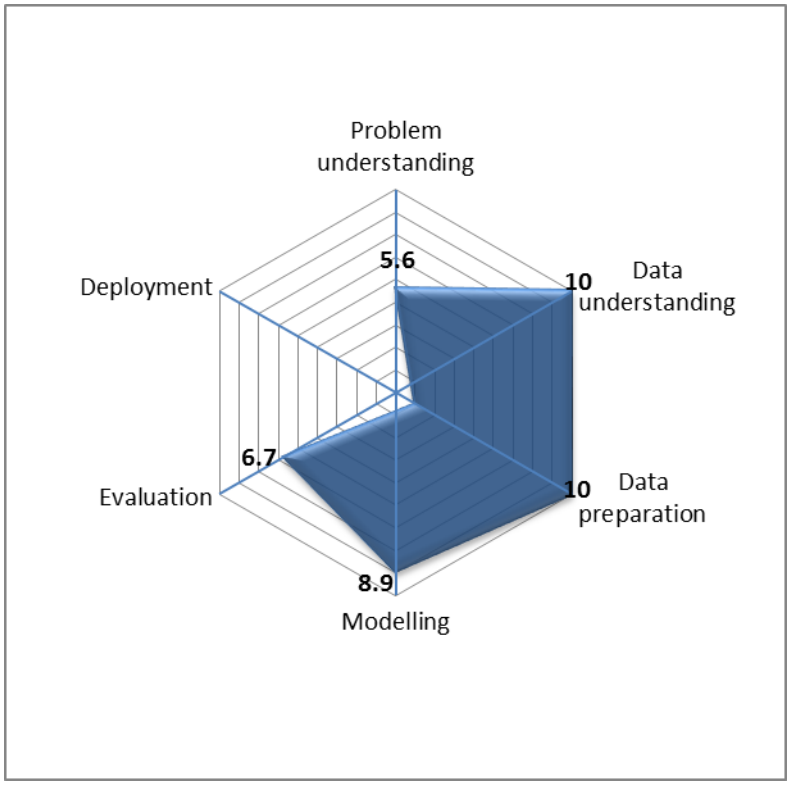


Fig. 31. CRIP-MED-DM compliance radar for BRCA1 prediction

Furthermore, the resulting model is not currently deployed in the healthcare facility; therefore, the Deployment phase scored zero points.

3.1.8.2. Clinical evaluation and conclusions

Remarkably, currently used publicly available clinical BRCA risk evaluation models are based purely on the patient's family history, whilst our classifier models provide similar and in some cases better accuracy by including clinical and morphological patient features.

The created breast cancer reoccurrence models have not included BRCA

mutation as a possible predictor for a patient group with a recurrent tumor. This finding supports the research (Robson et al. 2004) implying the importance of a tumor's clinical-morphological features and diminishing the impact of BRCA mutation to the breast cancer reoccurrence. Though, other research has reported on the lower survival rate for BRCA carriers (Brekelmans et al. 2009). Our predictive models have reconfirmed criteria that is already used in clinical practice. The family history attribute has high predictive value, especially when combined with clinical and morphological features such as bilateral BC, high grade tumor, medullary carcinoma, and triple negative BC. Interestingly, classification tree models highlighted negative expression of progesterone receptors as a possible *BRCA1* mutation predictor, which is a significantly narrower discrimination condition compared to triple negative BC, which additionally includes estrogen R(-) and HER2(-) features.

Another finding is higher *BRCA1* mutation probability for patients with tumor size greater than 1 cm or when more than one axillary lymph node is affected. This can be explained by higher grade of *BRCA1* associated tumors and higher proliferation.

BC reoccurrence classifier reconfirmed the prognostic features approved in previous clinical researches: higher tumor grade, primary tumor size, negative progesterone receptors, young patient age, and type of chemotherapy used.

After additional validation on a larger dataset, the created predictive models can be used as clinical decision support systems.

3.2. Predictive Data Mining: aortic valve stenosis predictive model

3.2.1. Introduction

Since the seminal work of L. Hatle et al. in 1980 (Hatle, et al., 1980), the golden standard for AS diagnostic is to rely on echocardiography measured by Doppler ultrasound. However, a number of pitfalls the clinicians are struggling with exist - difficulty in getting good quality images, localizing the measuring area in

continuous-wave or pulse-wave Doppler modes, and time consuming manual tracing of the images, just to name a few.

Nearly all ultrasound machines provide the required diagnostic parameters after manual tracing of the aortic systolic flow, which requires manual interaction and may lead to human error. According to the recommendations of professional cardiologists associations (Otto, 2012), a well-defined set of parameters is used to differentiate aortic stenosis (AS) and its severity. Some of these can be measured invasively or non-invasively, and others are derived from the first ones using the defined formulas

In our research, we addressed the outlined difficulties by employing image data analysis techniques described in Section 2.6. The successful automatic image processing and predictive modelling using DM methods would support clinicians in routine operations of systolic flow tracing, and furthermore, provide a third opinion in aortic stenosis severity grading.

3.2.2. Problem Understanding

3.2.2.1. Overall objectives

The overall goal formulated by the clinicians is to create a computer aided decision support system, which will be able to accurately:

1. Support the diagnosis of aortic valve stenosis.
2. Grade the severity (Low, Mid, High) of aortic valve stenosis.

To achieve this goal, the following objectives were formulated:

1. To achieve overall accuracy and ROC AUC values more than 90 %
2. The created method shall illuminate the manual systole tracing by the clinicians
3. The resulting predictive model shall be easily understandable and interpretable by clinicians by providing explicit predicting models
4. The resulting predictive model shall be easily integrated in a

Clinical Decision Support System, utilizing interoperable model description language PMML.

3.2.2.2. Assess situation

The following data source information systems have been identified:

1. VUSK centralized PACS system.
2. VUSK centralized HIS/EHR system.

The following relevant data entities have been identified:

1. Echocardiography images in DICOM format.
2. Cardiologists' measurements and evaluations.
3. Patient's demographics: age, gender.
4. Patient's encounter data: primary diagnosis, secondary diagnosis.

3.2.3. Data Understanding

The research data was acquired in Vilnius University Hospital Santariskiu Klinikos, by manually selecting consecutive patients with an equal distribution of AS severity. Studies of 18 patients with demographical, clinical and Doppler echocardiographic data were preselected by the participating cardiologist. In accordance to patient data privacy regulations, the research data were depersonalized and de-identified.

The selection criterion for the second-use data was the severity of aortic stenosis. Of these patients, five – had no clinical signs of aortic stenosis, five – had mild AS, four – moderate AS, and four – manifested severe AS.

Clinical data included age, gender, hypertension, cholesterol level, coronary heart disease, and additional risk factors, such as diabetes and history of smoking. Some data were incomplete, and could not be used in the full extent for statistical analysis.

The list of observed clinical, demographic features (attributes) is provided in Table 20, together with attribute types and the number of distinct values of each nominal attribute.

Table 20. The list of initial dataset attributes

#	*Attribute*	*Attribute type**	#	*Attribute*	*Attribute type**
1	Age (years)	Continuous	9	VTI, cm	Continuous
2	Diagnosis (ICD-10)	Nominal (13)	10	LVOT D, cm	Continuous
3	Gender	Nominal (2)	11	AVA, cm^2	Continuous
4	Heart rate, BPM	Continuous	12	LVOT Vmax, m/s	Continuous
5	AV Vmax, m/s	Continuous	13	LVOT Vmean, m/s	Continuous
6	AV Vmean, m/s	Continuous	14	LVOT PGmax, mmHg	Continuous
7	AV PGmax, mmHg	Continuous	15	LVOT PGmean, mmHg	Continuous
8	AV PGmean, mmHg	Continuous	16	AV Stenosis	Nominal (4)

The acquired image data included several diagnostic sessions for each patient, exported from ultrasound diagnostic equipment in DICOM format. The data set consisted of 36 AV and 35 LVOT echocardiography images.

3.2.3.1. Measuring Doppler Echocardiographic Data.

In our experiments, blood flow velocity was measured with a 5-chamber view continuous-way Doppler for AV flow, and pulsed-way Doppler for LVOT flow. A noise filter with default cut-off values was used. The transducer's alignment with the blood stream across the valve was checked with Color Doppler. The measured blood flow had a real time graphical visualization, with waveform echocardiogram (Fig. 32) which was stored in the hospital's PACS system, and then exported in DICOM format.

A normal aortic valve spectral Doppler trace has a single rounded systolic (S) wave below the baseline, as blood flow is away from the transducer. The S wave is enclosed within the AV opening (OC) and closing clicks (CC).

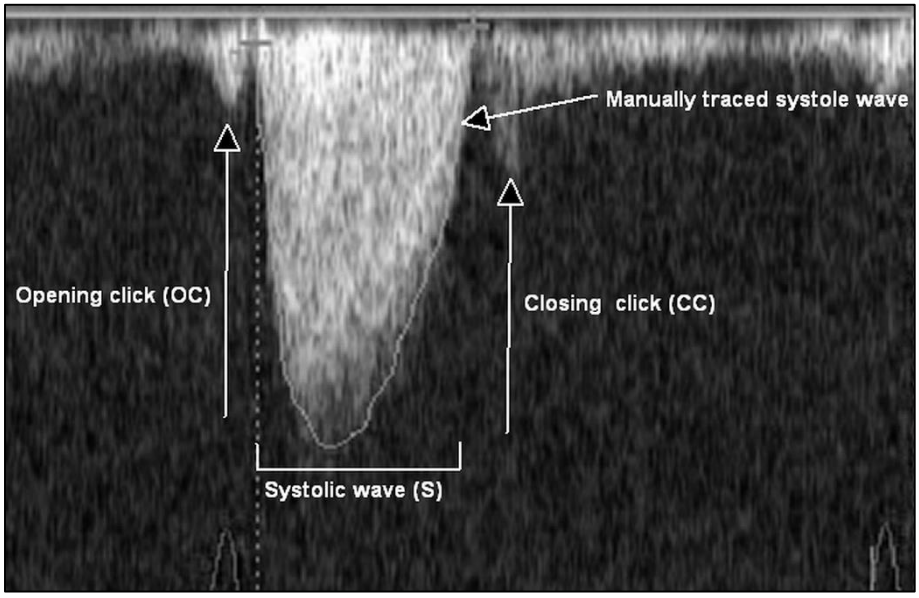


Fig. 32. The Doppler spectrum of AV systolic flow

According to the guidelines of the European Association of Echocardiography, the American Society of Echocardiography and the European Society of Cardiology (Otto, 2012), the Doppler echocardiographic parameters, used to diagnose AS (Table 21), include: AV peak systolic velocity, mean pressure gradient, aortic valve area, and velocity index. Peak pressure gradients are derived from the simplified Bernoulli equation, which is based on the conservation of energy in a closed system (Otto, 2012). Aortic valve area value is derived using the standard continuity equation. The mean gradient is calculated by integrating the pressure gradient over the entire systole.

Table 21. The list of measured and calculated echocardiographic parameters

Parameter	***Type***	***Units***	***Formula***
$Vmax$ (peak systolic velocity)	Velocity	m/s	-
T (duration time)	Time	S	-
$Vmean$ (mean systolic velocity)	Formula	m/s	$Vmean = \frac{VTI}{T}$
$PGmean$ (mean pressure gradient)	Formula	mmHg	$PGmean = \frac{\sum 4v^2}{N}$
$PGmax$ (peak pressure gradient)	Formula	mmHg	$PGmax = 4Vmax^2$
AVA (aortic valve area)	Formula	cm^2	$\frac{\pi \times LVOT\ D^2 \times LVOT\ VTI}{4 \times AV\ VTI}$
VTI (Velocity time integral)	Formula	Cm	$\int Vmax$
VI (Velocity index)	Formula	-	$VI = \frac{VTI_{LVOT}}{VTI_{AV}}$

3.2.3.2. Exploring data

The nature of the original dataset was multi-relational, since the patient entity relates to multiple diagnosis, multiple AV and LVOT measurements. However, we assumed a simpler structure, where each patient has one determining measurement, and one main diagnosis. This was achieved using propositioning of the data (Kramer, et al., 2001). For that, an averaging of the measurement parameters was used.

3.2.3.3. Data quality verification

The provided dataset was manually checked by the participating cardiologist: there were no missing data or manual data entry errors.

3.2.4. Data Preparation

Echocardiography images pre-processing methods described in Section 2.6 were applied. The study drew on the second-use patient data, retrieved from the Santariskiu Klinikos hospital information system and picture archiving and communicating system. The clinical measurements, required for AS diagnosis, were provided by the participating cardiologist, and were used as the golden standard in the study. The initial data set consisted of 18 patients with 71 echocardiography images. By applying our method, the initial image set was transformed to the traced 71 AV and 68 LVOT blood flow velocity complete systole cycles.

In order to evaluate the effectiveness of the proposed method, we compared the manual measurement performed by the cardiologist (M) with the automatic measurement results of the proposed method (A). The performance of the proposed method is reported with the Pearson correlation coefficient and Bland-Altman limits of agreement. Introduced by J. M. Bland and D. G. Altman (Martin Bland & Altman, 1986), the limits of agreement (LoA) are acceptable prediction limits for the difference between the measurements of the two methods on a randomly chosen item. The Bland Altman model is formulated as a two-way analysis of variance model. For the future measurement prediction, the Bland-Altman model stipulates the difference of the new values, obtained

with each of the two compared methods, is within the limits of agreement with 95 % probability. In most cases, we observed a good agreement between the two methods.

The values of parameters directly derived from the processed images relate to the compared manually obtained values as follows:

1. Values of $AV\ Vmax$ and $AV\ VTI$ measured by the two methods were strongly and significantly correlated. For $AV\ Vmax$: R^2= 0.999, p-value < 0.0001; $AV\ VTI$ R^2 = 0.988, p-value < 0.0001. However, $LVOT\ VTI$ measurement showed a lower degree of correlation: R^2=0.68, p-value <0.0001.
2. Bland-Altman plots for the parameters $AV\ Vmax$, $AV\ VTI$, and $LVOT\ VTI$ (Fig. 33), outline Limits of Agreement, and the means of the differences A-M: $AV\ Vmax\ \bar{d} = 0.02\ m/s$, $AV\ VTI\ \bar{d} = 0.16\ cm$, $LVOT\ VTI\ \bar{d} = 3.43\ cm$.

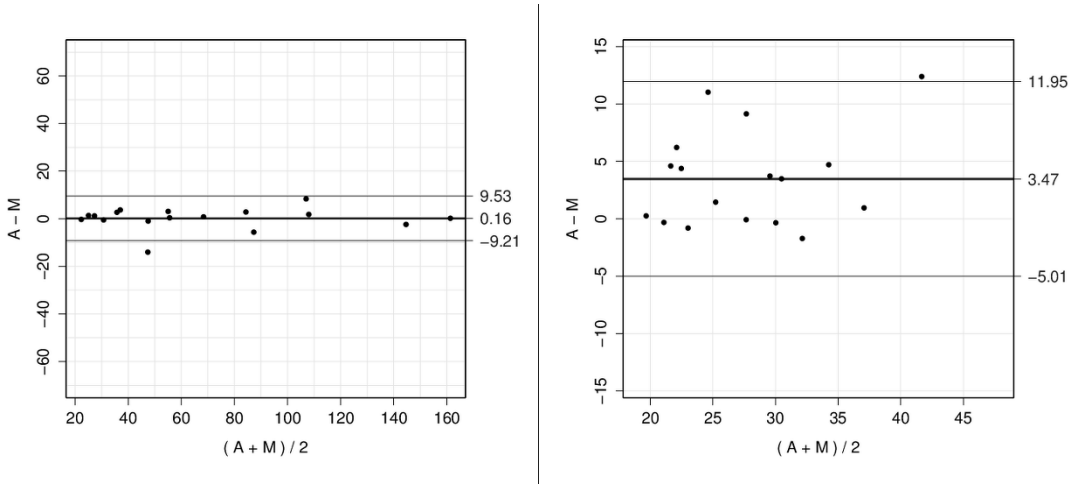


Fig. 33. Bland-Altman plots for the parameters produced by manual (M) and automated (A) measurement methods.

The left image - aortic valve velocity time integral ($AV\ VTI$); the right image - left ventricle output tract velocity time integral ($LVOT\ VTI$).

Of the highest importance for aortic stenosis diagnosis, the remaining calculated parameters - mean pressure gradient ($PGmean$) and aortic valve area (AVA) - relate to the corresponding values of manual measurements as follows:

- $PGmean$ $R^2 = 0.994$, p-value < 0.0001, $d(M - A) \in [-13.37,\ 5.20], \bar{d} = 4.09\ mmHg$;
- AVA $R^2 = 0.894$, p-value < 0.0001, $d(M - A) \in [-0.33,\ 0.70], \bar{d} = 0.19\ cm^2$.

In addition, we compared manual measurements values with the values of averaged systole cycles, calculated in step 4 of the method (automated averages – AA). The Comparison generally showed lower values of the Pearson coefficient and wider Limits of Agreement:

- $AV\ Vmax$ $R^2 = 0.999$, p-value < 0.0001, $d(M - AA) \in [-0.29, 0.15], \bar{d} = 0.07\ m/s$;
- $AV\ VTI$ $R^2 = 0.988$, p-value < 0.0001, $\mathrm{d(M - AA)} \in [-32.40, 49.10], \bar{\mathrm{d}} = 8.35$ cm;
- $LVOT\ VTI$ $R^2 = 0.68$, p-value < 0.0001, $\mathrm{d(M - AA)} \in [-6.83, 23.24\ ], \bar{\mathrm{d}} = 8.20$ cm;
- $PGmean$ $R^2 = 0.9868$, p-value < 0.0001, $\mathrm{d(M - AA)} \in [-20.55, 6.64], \bar{\mathrm{d}} = 6.96$ mmHg;
- AVA $R^2 = 0.759$, p-value < 0.0001, $d(M - AA) \in [-0.56, 1.03], \bar{d} = 0.24\ cm^2$.

The results of blood flow echocardiography images analysis are of reasonably high quality; therefore, they can be used for further DM predictive modelling activities, in accordance with CRISP-MED-DM Phase 4.

3.2.5. Modelling

According to CRISP-MED-DM methodology, the resulting datasets have been iteratively used with a set of DM methods and algorithms:

- Classification trees – J48, tree ensemble.
- Artificial neural networks – Multi-layer Perceptron.

DM software packages R (R Core Team, 2014), and Tibco Spotfire Mining (Tibco, 2010) were used.

Averaging of the systole parameters extracted from diagnostic images have shown considerably lower accuracy compared to the results obtained by the methods on the preselected images.

Accuracy decrease was proportional to the deviations of the patients' image sets. However, the automation of systoles tracing during the pre-processing step also meant reducing the pre-validation of the images by an experienced clinician. In the case of AS that could mean that non-optimal images, when the transducer was not finally aligned with the blood stream, were considered in the measurements. Straightforward averaging of the image features was suboptimal in some cases and wrong in others. Therefore, for the predictive models results evaluation, we used patient diagnostic images, preselected by the cardiologists.

We have conducted comparative experiments comparing classification modelling on the parameters measured and validated by cardiologists, and the parameters extracted by applying the introduced echocardiography images analysis method. The resulting classification trees are shown accordingly in Fig. 34 and Fig. 35. In both cases 100 % accuracy was achieved. The back propagation neural network had lower overall accuracy of 98.9 % for the dataset derived after feature extraction from the images.

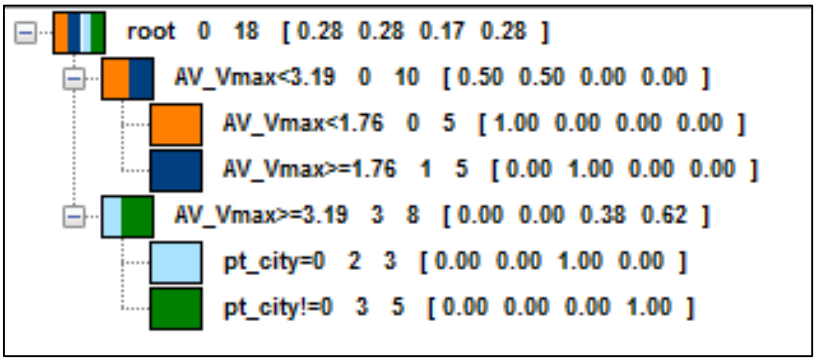


Fig. 34. Aortic stenosis grading decision tree based on cardiologist measurements

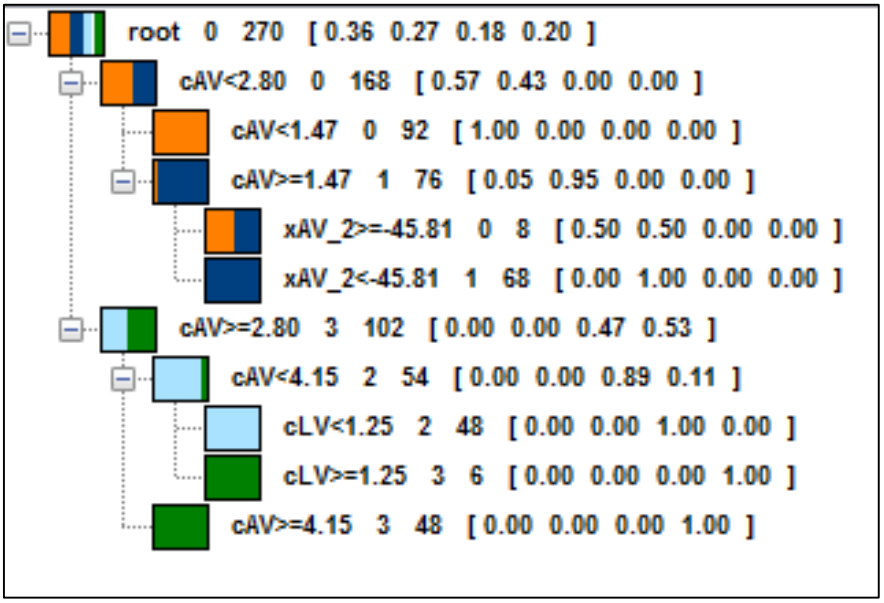


Fig. 35. Aortic stenosis grading decision tree based on feature extraction algorithm

In addition, exploratory analysis was performed by employing association rules (Apriori algorithm) and clustering (k-means algorithm). However, there were no interesting rules identified by clinical experts. K-means clustering algorithm with three clusters grouped the data into groups with the distribution close to low-mid-high severity stenosis, with a 14 % error rate.

3.2.6. Software Implementation

To conduct the experimental trial of the described echocardiography images analysis and further application of predictive data mining methods, software was developed. Image processing routines and predictive modelling were implemented in the R environment, using *ImageJ* library for the standard image processing tasks. The component diagram of the developed software is shown in Fig. 36.

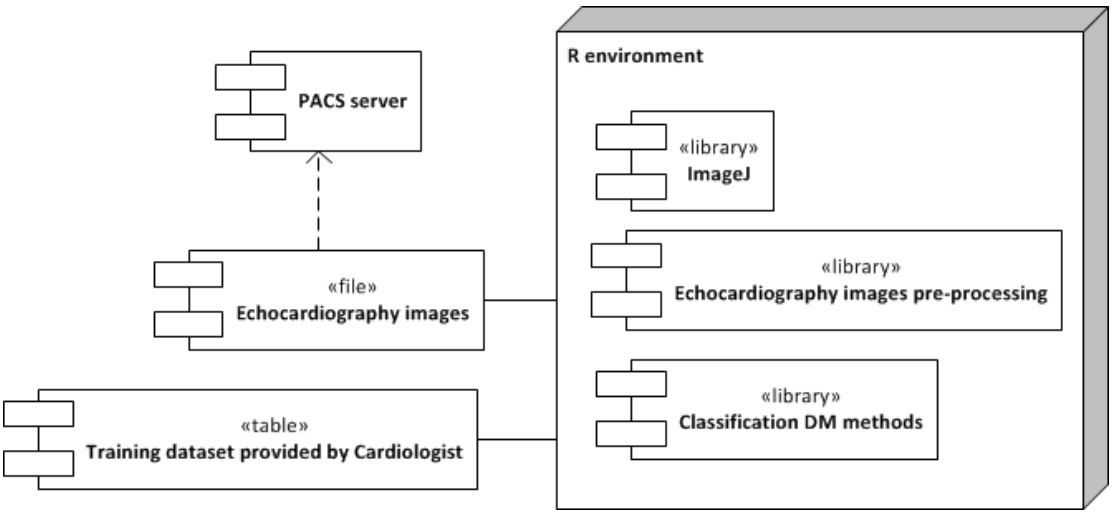


Fig. 36. Blood flow echocardiography images analysis and data mining component diagram

Patient's image processing for the following predictive modelling have the steps as follows:

1. Export of echocardiography images from PACS server for offline processing in JPEG format.
2. Importing of JPEG images into R environment.
3. Application of the implemented echocardiography images processing methods to extract full systole cycles.
4. Calculation of aortic valve stenosis diagnostic parameters.
5. Application of DM classification methods to build the aortic valve stenosis grading predictive model.

The described activities are shown in Use Case diagram in Fig. 37.

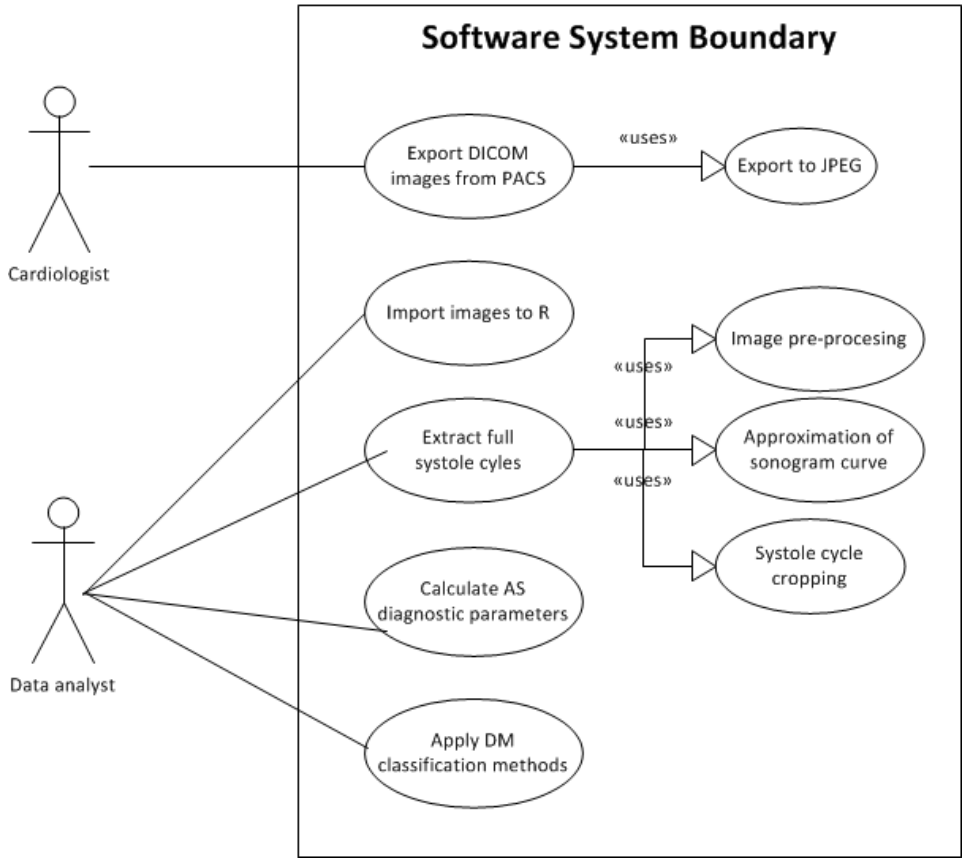


Fig. 37. Blood flow echocardiography images analysis and data mining use case diagram

3.2.7. Evaluation and Results

Echocardiography images analysis allowed for extracting high quality parameters, which were used for aortic stenosis grading. It is notable to mention, the calculated parameters of mean gradient and the aortic valve area provided by the ultrasound machine and our method are not directly comparable, as the diagnostic equipment vendors use proprietary calculation algorithms. For comparison purposes, we used formulas derived from simplified Bernoulli and continuity equations on the measurements provided by ultrasound machines (as outlined in Table 21). Our calculated AVA values showed strong correlation with the proprietary ultrasound's AVA values, with $R^2 = 0.799$, p-value < 0.0001 and a mean methods' difference of 0.16 cm; the calculated $PGmean(Vmax)$ measurements compared to ultrasound's $PGmean(Vmax)$ showed $R^2 = 0.99$ and p-value < 0.0001 with a mean methods difference of 1.84 mmHg.

The created AS stenosis severity predictive models with classification

algorithms of decision trees, random forests and neural networks had the best performance with up to 100 % overall accuracy. Then we compared the performance of the models with identical ones on the data acquired from the ultrasound modalities. The accuracy of both methods was more than 98 %, though the latest has demonstrated a highest accuracy of 100 % using various classification algorithms.

3.2.8. Deployment

According to CRISP-MED-DM, the resulting decision tree predictive model with 100 % overall accuracy was exported to the PMML format (Annex A), which can be used by clinical decision support systems and generic scoring software in clinical settings.

Following the recommendations of A. K. Waljee et al. (Waljee, 2013), the derived PMML model are provided for further validation of the predictive models with external datasets.

3.2.9. Discussion and Compliance to CRISP-MED-DM

The implementation of semi-automated blood flow echocardiograms tracing was carried out. The experimental results of the proposed method were compared to the measurements, acquired within the current clinical practice, relying on manual blood flow echocardiograms tracing in Doppler ultrasound modality user interface. Correlation coefficients for an aortic valve area of 0.77, and for aortic valve maximum jet of 0.99 were found. There was a good agreement between the two methods, resulting in means' differences of 0.19 cm and 0.02 m/s, respectively.

Comparison of the time needed to perform measurements and calculations using method (M), and method (A) had the following results: with a flow of 20 patients per day, the total amount of measurements will count up to 120 systole cycles, which sums up to 20 minutes of net measurement time and additional 20–30 % overhead for systole selection and manual comparing. Summarizing, the time spent for manual tracing and processing of the measurements is 24–28 minutes per cardiologist (20 patients per day). The

running time of the echocardiography image analysis method implementation on the consumer type personal computer was between 1–2 seconds per spectrogram image. Thus, the total projected timesaving per cardiologist is around 22–26 minutes.

Predictive modelling for AS grading resulted in the predictive classification models with 100 % overall accuracy.

The achieved results suggest that the proposed predictive model based on the proposed method for tracing the blood flow echocardiograms and calculation of hemodynamic parameters is reliable and can be used as a supplementary tool for AS severity grading.

3.2.9.1. Compliance to CRISP-MED-DM

To evaluate compliance of the undertaken application of DM activities for predicting and grading of Aortic Valve stenosis, we applied the first evaluation strategy, proposed in chapter 2.4.6.2.

As shown in Fig. 38, the core phases Data understanding, Data Preparation, Modelling and Evaluation show good performance. However, the Problem understanding phase scores only 3.3 from a maximum 10 points.

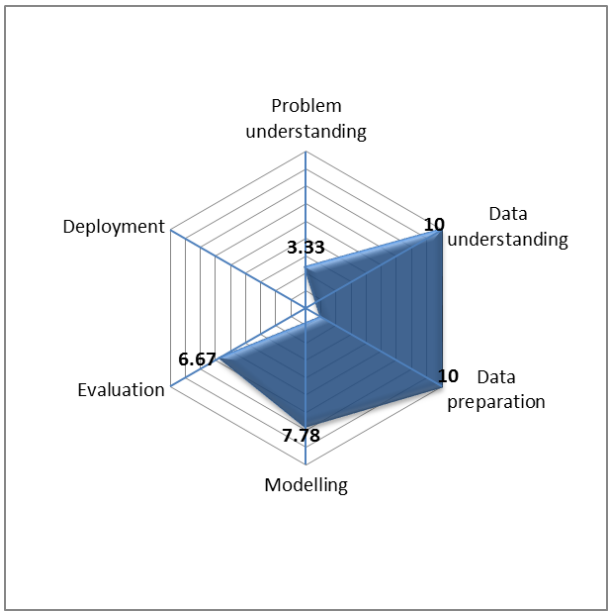


Fig. 38. CRIP-MED-DM compliance radar for AV stenosis prediction

The activities related with formal project management, such as scheduling, risk

planning, and cost/benefit analysis have not been performed due to the exploratory nature of our research project. Furthermore, the resulting model is not currently deployed in the healthcare facility. Actual deployment in clinical settings will require the integration or adaptation of ultrasound modalities used for cardio-echography.

3.2.9.2. Clinical evaluation and conclusions

The selected patients had regular heart rate. However, the blood flow of each cycle may vary depending on the length of the diastole. In practice, the cardiac output depends on the duration of the cycle. The longer the diastole is, the heart is filled with a larger blood volume. When evaluating echocardiograms, the cardiologist may neglect smaller differences. Moreover, the parameters of each cycle might be different due to the technical measurement reasons, e.g. natural movements of a patient's body. The described reasons illustrates the differences of echocardiography images analysis by method A with the blood flow curve averaging and cardiologists manual measurements.

The high accuracy of the derived predictive models, based on the semi-automated blood flow echocardiograms analysis, suggests its good applicability in clinical practice. The application of the described methods would help to save time and avoid possible errors.

3.3. Descriptive Data Mining: PubMed publications meta-analysis

3.3.1. Introduction

As was described in Section 1.4.1, simplification of relational data structures may lead to information loss and consequently to poor knowledge discovery results. In this section, a multi-relation clustering approach, based on a concept of algorithm's key-notion upgrade is addressed. A novel method for a distance matrix calculation in multi-relational settings was introduced in Section 2.7. The method has been tested by analyzing publications indexed in the PubMed database (National Center for Biotechnology Information, 2009).

Clustering based on partitioning around medoids was used for the identification of the most popular topics among the PubMed publications with the "data mining" keyword. The algorithm implements a greedy approach and is suitable for small data sets with a limited number of one-to-many relational joins. The distance matrix calculation algorithm was implemented in R language.

3.3.2. Problem Understanding

The overall research goal was formulated as follows: to investigate the most prevalent clinical topics, disease groups, and DM techniques applied and described in the PubMed research papers.

3.3.3. Data Understanding

In our experiment, the PubMed database was used, as the biggest medical database, having an explicit publications hierarchical semantic tagging system, called MeSH (National Center for Biotechnology Information, 2009). The Medical Subject Headings (MeSH) is a controlled vocabulary, which is used for indexing, cataloging, and searching for biomedical and health-related information and documents.

A simplified hierarchical MeSH terminology structure is presented in Fig. 39. The relational data representation includes one-to-many relational joins between the entities Keyword and MeSH Concept, Keyword and MeSH Descriptor, and between MeSH Semantic Type and MeSH Concept.

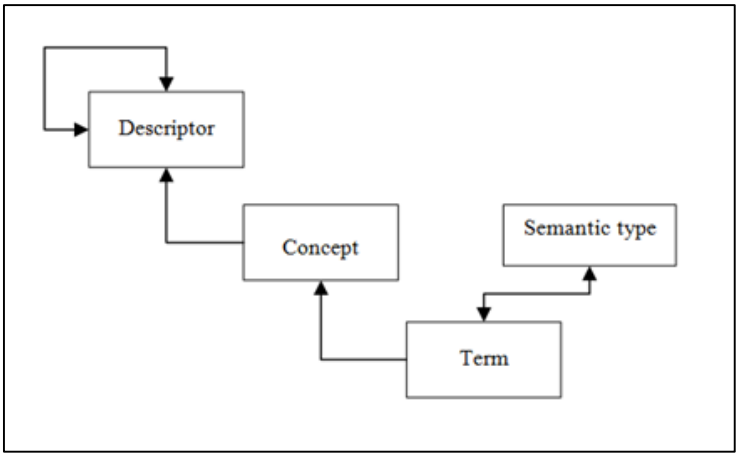


Fig. 39. Simplified MeSH entities entity-relationship diagram

Mesh definitions of the entities are as follows. Descriptors, also known as Main Headings, are used to index citations in the PubMed database for the cataloging of publications. Most Descriptors indicate the subject of an indexed item, such as a journal article. MeSH Descriptors are organized in 16 categories, each of them is further divided into subcategories. Within each subcategory, Descriptors are arrayed hierarchically in twelve levels, from the most general to the most specific.

A Descriptor is broader than a Concept and consists of a class of concepts. Concepts, in turn, correspond to a class of Terms, which are synonymous with each other.

Thus, MeSH has a three-level entity structure: Descriptor → Concept → Term. Every Term is assigned to one or more Semantic Types, which gives a broader meaning to a Term.

The described MeSH vocabulary structure allowed us to extract additional information for the keywords assigned to the articles. The Hierarchical structure of the Descriptors, represented in the MeSH tree, allows grouping by disease groups, anatomy concepts, chemical and drug groups, phenomena and processes group, and computer science categories.

3.3.4. Data Preparation

The complete search result dataset with available attributes has been exported from PubMed to XML format, and then transferred to a relational database. MeSH controlled vocabulary data are freely available and are provided by the National Library of Medicine. Finally, the imported XML data has been transformed to relational format, as shown in ERD diagram in Fig. 40.

Having MeSH vocabulary and the exported publications dataset in one database schema, allowed us to leverage underlying semantic concept aggregation in MeSH and to group articles on a higher abstraction layer using the compound similarity measure introduced in Section 2.7.2.

3.3.5. Modelling

3.3.5.1. Applying multi-relational clustering for publications meta-analysis

As seen in the Entity Relational Diagram (ERD) in Fig. 40, the relations Concept, Descriptor, and Semantic Type, which represent respective MeSH entities, are indirectly joined with the central entity Article. These relations were chosen based on their semantic value and relevance to our analysis.

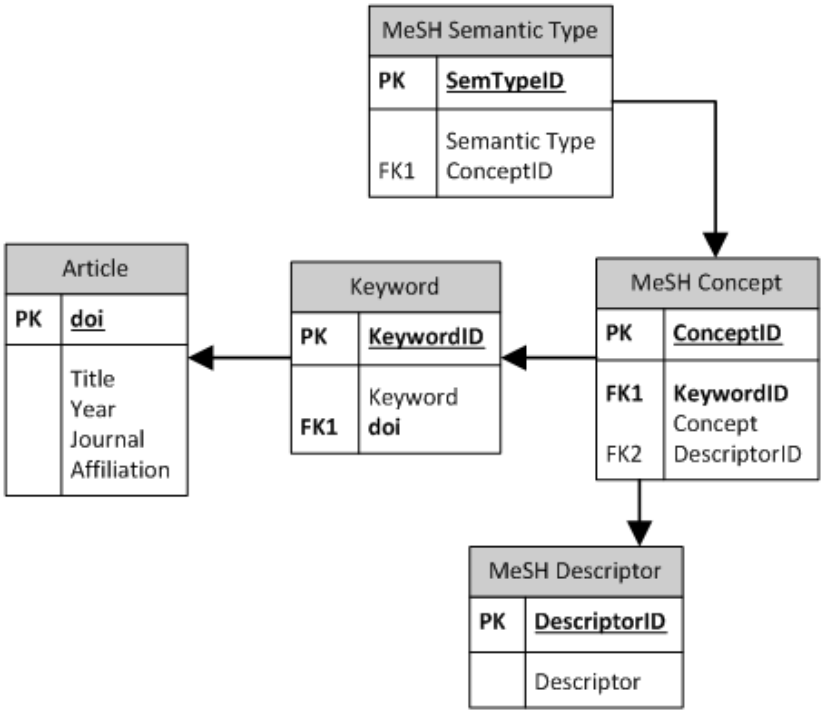


Fig. 40. Entity-relationship diagram of the relational dataset

Let us consider the example of the two articles A and B, with two sets of keywords (MeSH Concepts as in Fig. 40) S_A and S_B. Suppose $S_A \cap S_B \in \emptyset$, however there exists a keyword 'Benpen' $\in S_A$, and 'Benzylpenicillin'$\in S_B$. Relying on single-table paradigm a normalized distance between objects A and B should be: $D_{A,B} = 1$. But in a relational environment, we consider that both 'Benpen' and 'Benzylpenicillin' has a "belong to" join relationship with the higher semantic hierarchy category Descriptor 'Penicilin G'. This leads us to a justified conclusion, that a certain similarity between S_A and S_B exists and hence, the distance should be: $0 < D_{A,B} < 1$.

Following the first-order logics notation, instances of Articles are

represented by the predicate A, and the ground atoms C for MeSH Concept, D for MeSH Descriptor, and S for MeSH Semantic type.

For example a dataset's instance I_1 and I_2, representing articles with keywords "Alkalescens-Dispar Group" and "Diffusely Adherent Escherichia coli" accordingly, can be written:

$$I_1 = A(art1), I_2 = A(art2)$$

where background knowledge is defined by:

$$C(art1, Alkalescens Dispar\ Group),$$
$$D(art1, Escherichia\ coli),$$
$$S(Bacterium, Escherichia\ coli)$$

and

$$C(art2, Diffusely\ Adherent\ Escherichia\ coli),$$
$$D(art2, Escherichia\ coli),$$
$$S(Bacterium, Escherichia\ coli).$$

The definition of predicate A is extended by the background predicates: C, D and S, with the arguments: $A(a1: name)$, $C(a1: name, a2: discrete)$, $D(a1: name, a2: discrete)$, $S(a1: name, a2: discrete)$. The structure of ground atoms repeats a subset of the relational data model (Fig. 40). The entities Keyword and Article could be joined due to their initial one-to-one relationship.

In the example, the background knowledge derived from MeSH vocabulary suggests the similarity of the two articles I_1 and I_2 is greater than zero, since both articles address the bacteria, belonging to the same class "Escherichia coli". In order to define and calculate the similarity we used a similarity measure, introduced in Section 0.

Applying formulas (5–8) to the predicates C, D and S, the similarity measure for comparing two instances of article A, considering the defined background knowledge is derived:

$$sim_{A1,A2} = \frac{w_c simC + w_d simD + w_s simS}{\mathrm{w_c + w_d + w_s}}, \quad (9)$$

where

$$simC = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}(concept_i(A_1), concept_j(A_2))}{\sqrt{m \times n}},$$

$$simD = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}(descriptor_i(A_1), descriptor_j(A_2))}{\sqrt{m \times n}},$$

$$simS = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}(semantictype_i(A_1), semantictype_j(A_2))}{\sqrt{m \times n}}.$$

In essence, *simC*, *simD*, and *simS* calculate the similarity of the value lists (accordingly Concepts, Descriptors, and Semantic Types) which are relationally joined to the central entity Article. Another known approach for this task is described by Horwath, Wrobel et al., where the authors proposed to calculate influence function, the cost of which equals the effort of the lists' equalization. The implementation of this approach is described in Section 3.3.5.3.

As seen in the (9) formula, the similarity value is sensible to the assigned weight values. Hence, a fine-tuning of weight parameters w_c, w_d, and w_s has been performed. T. Horwath and S. Wrobel propose a simplified approach, by not using weights at all. This simplification in our case is unadjusted because of the uneven nature of the data. In our experiment, two approaches were used: the statistical one, where weights are proportional to the number of tuples of the relevant entities; and expert based, where weights have been experimentally adjusted and normalized by the expert.

In the first case, weights have been calculated as follows:

$$w_c = \frac{n_c}{n_c+n_d+n_s}, w_d = \frac{n_d}{n_c+n_d+n_s}, w_s = \frac{n_s}{n_c+n_d+n_s} \quad (10)$$

The described weight distribution is reasonable in the cases where we want to level the importance of each list value variable according to the relative number of tuples in each entity.

In other examples, having a more diverse set of variables, this statistical approach might be appended or changed by the domain expert's knowledge and empirical experiments. If that is the case, for the calculation efficiency, it is

important to store all s_{ijk} values, for further experiments with different wk values. In the opposite case, if only the resulting s_{ij} are preserved, in order to change the weights, the whole distance matrix shall be recalculated from scratch.

According to our experiment results, the described similarity measure derives stable values, meaning that small changes to a term do not cause big changes in distance values. The experiments with real data have shown that in some cases it is even too stable and lacks some responsiveness to the data changes. However, this is solvable by fine-tuning weight parameters w_c, w_d, and w_s.

Finally, the distance (dissimilarity) value was calculated as follows:

$$\mathrm{dist}_{\mathrm{A1,A2}} = 1 - \mathrm{sim}_{\mathrm{A1,A2}} \tag{11}$$

The algorithm, calculating the full distance matrix for the set of articles, has been implemented in R. The output of the algorithm provides values of *simC*, *simD*, and *simS* for each pair of articles. The distances (dissimilarities) values have been derived by applying formulas (9) and (11) with the calculated weight values (10). Then alternatively, the weight values w_c, w_d, w_s were manually assigned based on the intuitive semantic value of the MeSH Concept, Descriptor and Semantic type; giving the highest value to w_d, then lower value to w_c, and the lowest to w_s.

The calculated versions of distance matrixes were applied with PAM algorithm to group articles into clusters.

R libraries "cluster" and "fpc" were used to try different implementations of PAM algorithm. Clustering algorithms have been iteratively applied and compared for the numbers of clusters from two to fifty with a step of five. The clustering results are described in Section 3.3.7.

3.3.5.2. Alternative approach – propositionalization of dataset

According to the theory of multi-relational data mining (Dzeroski, 2010; Knobbe, et al., 2001; Kramer, et al., 2001), the propositionalization of data structure to a single-table format allows applying standard DM algorithms,

instead of upgrading to a multi-relational version.

Data preparation

The structure of a given data model (Fig. 40) consists of 1 one-to-one and 3 one-to-many relationships. Therefore entities *Article* and *Keyword* can be joined without information loss. As for the entities *Descriptor*, *Concept* and S*emantic type*, a propositionalization approach shall be selected.

All semantically valuable attributes of these entities are of nominal type, therefore mathematical aggregation functions cannot be used. Instead, top five *Keywords* and their first related *Concepts* were used. In addition, three most frequent *Descriptors* and two more frequent *Semantic types* been preserved in the entity *Article.* The resulting entity is shown in Fig. 41.

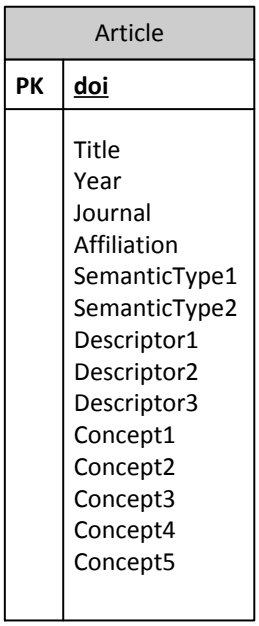


Fig. 41. Propositionalized entity “Article”

Modelling

The same implementation of the PAM algorithm of R libraries “cluster” and “fpc” were used. The number of clusters was increased iteratively from two to fifty. After each iteration, the quality of clusters was evaluated with a silhouette value. There were no high-quality clusters identified. The clusters’ silhouette values varied between 0 and 0.16.

3.3.5.3. Alternative approach – clustering with Edit distance

Another possibility to handle data in relational format is to represent it as a labelled tree. As proposed in Chapter 2.7.2.1, tree edit distance (TED) can be used as a distance measure to group the trees into clusters.

TED algorithm implementation of M. Pawlik and N. Augsten (Pawlik & Augsten, 2011) was used to calculate the dissimilarities between each pair of articles. The CRIPM-MED-DM steps were completed as follows.

Data preparation

In order to process data with a TED algorithm, the data in the relational structure shall be converted to labelled trees. The conversion tool was implemented using a T-SQL procedure for MS SQL server. Following the example from Chapter 3.3.4, representations of art1 and art2 are shown in Fig. 42 and Table 22.

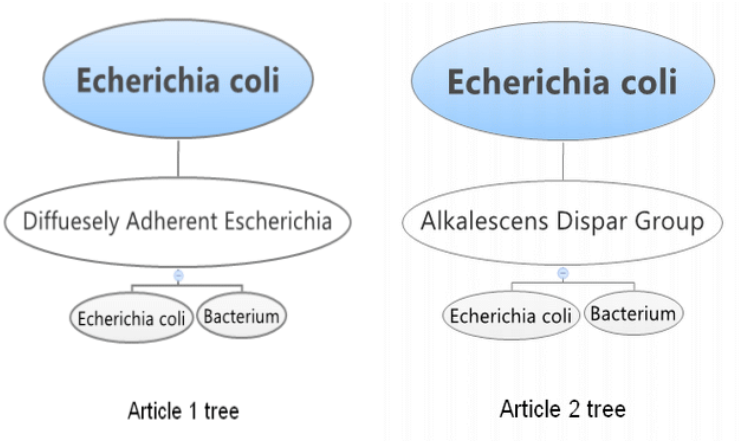


Fig. 42. Trees of art1 and art2 MESH terms

Table 22. Art1 and Art2 data representation for RDET algorithm

`Art1{Escherichia coli{AlkalescensDispar Group{Bacterium}{Escherichia coli}}}`
`Art2 {Escherichia coli{Diffusely Adherent Escherichia coli{Bacterium}{Escherichia coli}}}`

An example of the tree representation of one the extracted article's MESH concepts, descriptors and semantic types is shown below:

```
Article23{{C15072{T028}}{C15313{T170}}{C18531{T032}}{C19072{T0
```

86}}{C19525{T091}}{C3060{T066}}{C5581{T091}}{C5593{T032}}{C650
4{T016}{T098}}{C6748{T083}}{C6946{T058}{T078}}{C8126{T057}{T07
3}{T170}}}

For computational efficiency, all literals are substituted with ID codes, which uniquely specify the concepts, descriptors and semantic types.

Modelling

The same implementation of the PAM algorithm of R libraries "cluster" and "fpc" were used. We iteratively were increasing the number of clusters from two to fifty. After each iteration, the quality of clusters was evaluated with a silhouette value. Summarizing the results of the iterations, as when applying the proposed multi-relational similarity measure, there were no high-quality clusters identified.

3.3.6. Software Implementation

To conduct the experimental trial of the described multi-relational clustering, a software module was developed. Initial pre-processing of publication metadata exported from PubMed was implemented in a MS SQL relational database management system. First, XML data was transferred to a relational format using the data transformation services of MS SQL. Afterwards, relational data was cleaned, aggregated and normalized to achieve the entity-relationship structure shown in Fig. 40. Structured data processing was developed in T-SQL. The derived structured multi-relational data was used to calculate distances matrixes. The proposed multi-relational similarity measure calculation was implemented as a set of functions in R language. Distance matrix calculation was run with a *parallel* package enabling multi-core support to run calculations in parallel. Then, an *fpc* package was used to apply PAM clustering algorithms. Finally, with a *methComp* package the silhouette values of the clusters were calculated to evaluate clustering results. The components of the developed software are shown in the component diagram (Fig. 43).

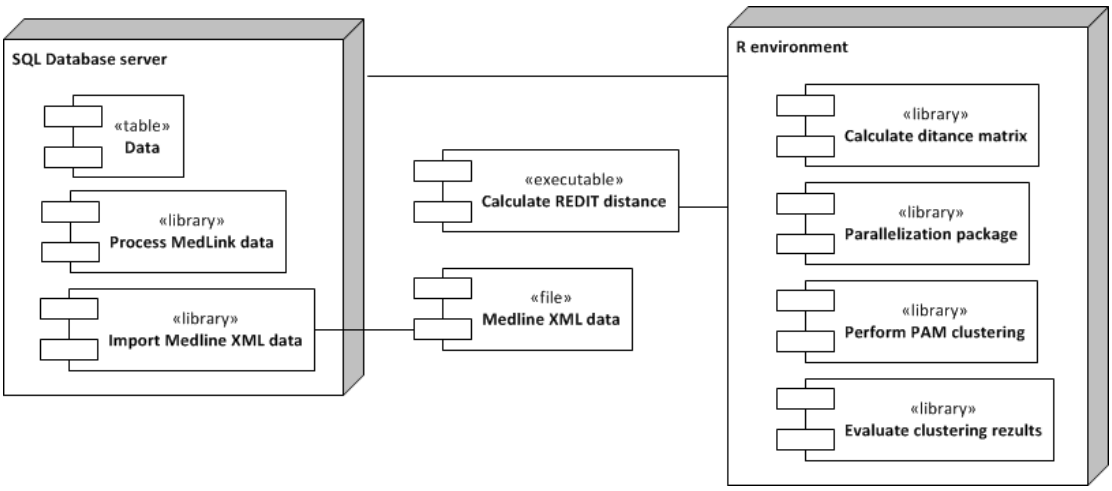


Fig. 43. PubMed publications multi-relational clustering component diagram

In total, 2.284.453 similarity values were calculated with each of the methods. Due to the large search space of multi-relational data, the algorithms require vast computational power.

In the case of the proposed multi-relational distance algorithm, multiple iterations of distances between each object and its selected related compound entities resulted in the algorithm complexity of $O(n^2 L_c^2 L_d^2 L_s^2)$, where L_c – the length of list of *Concepts*, L_d – the length of list of *Descriptors*, L_s – the length of list of *Semantic types*. After code optimization, the overall achieved performance of an average size dataset for 100 similarity values was in the range of 40-60 seconds on one core of Intel i7 CPU. The parallelization had a huge effect, since each similarity measure is independent and thus can be calculated in parallel. However, data exchange between nodes required by the parallelization had a negative impact and reduced the positive effect of the parallelization.

Tree edit distance implementation RTED showed better performance. According to Pawlik, RTED runs in $O(n^3)$ time. In practical terms, the RTED algorithm performed ~10 times faster on the same hardware, and thus did not require parallelization.

3.3.7. Evaluation and Results

For the evaluation of the overall clustering quality, a cluster's silhouette value was used. The silhouette value depicts the quality of each object's cluster. A cluster's silhouette value is derived in the following way. Let $a(i)$ be the average distance between object i and all other objects of the cluster A, to which it belongs. For another cluster C_1, let $d(i, C_1)$ equal to average distance of i to all objects of cluster C_1. Then, calculate $d(i, C)$ for all the remaining clusters $C_{2..n}$ and assign the smallest of these $d(i, C)$ to $d\_min(i)$. The silhouette value of an object *i* is defined as follows:

$$silh_i = \frac{d\_min(i) - a(i)}{\max\{a(i), d\_\min(i)\}} \quad (12)$$

The cluster's silhouette value is an average silhouette value of all its members. Values near 1 mean that the object *i* is assigned to a correct cluster. In contrast, values close to -1 mean that it is likely that an object is assigned to a wrong cluster. The silhouette value around 0 means that the object *i* can be equally assigned to the selected or the nearest cluster.

In our case, trying a number of clusters from two to fifty, the maximum achieved silhouette values were in the range: 0.21–0.30 for the proposed multi-relational similarity measure, 0–0.16 for propositional clustering, and 0.15–0.23 for TED distance. Objectively, the achieved results indicate the overall clustering results are low, and shows that the found clusters poorly describe the data set as a whole. However, considering the non-trivial task of scientific publications' semantic grouping, the whole exercise was not fruitless, and gave us some interesting insights.

The application of clustering with the described similarity measure on relational data of PubMed and MeSH showed that the research topics are evenly distributed, and the research within the DM application in the area of healthcare is very diverse. However, the clustering outcomes have revealed a couple of clusters with a higher research interest. Among them: DM applications within protein structure analysis, specific patient profile search, text mining of medical text, public health legislation documents mining, commerce practices (fraud

detection), chronic disease diagnostics, survival prediction, information retrieval, and image data analysis.

Our findings are comparable to the manual systematic literature analysis performed by N. Esfandiari et al. (Esfandiari, et al., 2014). Although the authors of the former limited the scope of their research to the publications related to knowledge extraction from structural medical data. According to Esfandiari, medical diagnosis is the prevalent medical DM task, then screening, prognosis, treatment, monitoring and management are equally distributed.

3.3.8. Discussion and Compliance to CRISP-MED-DM

A compound similarity measure calculation algorithm for multi-relational data has been created, implemented and tested with a real world data clustering task. The proposed similarity measure aggregates the Gower similarity coefficient and Ochiai-Barkman coefficient and is applicable for relational data models with nominal attributes and lists.

The activities of the CRISP-MED-DM Modelling phase were applied with different clustering methods iteratively changing the number of targeted clusters. In addition, the calculation of the distance matrix, according to the proposed method, was updated adjusting weight values of the calculation parameters. The Modelling phase iterations resulted in higher cluster quality, which was increased from 0.16 silhouette value — the best result of propositional clustering — to 0.31 silhouette value — the best result of the proposed multi-relational clustering approach.

Though the greedy PAM algorithm used was suitable for our case study due to a relatively small dataset, in other cases large data clustering algorithms CLARA or CLARANS (Ng & Han, 2002) shall be used instead of PAM.

Remarkably, multi-relational clustering with the proposed similarity measure is not specific for the medical domain and therefore can be applied in other domains with mixed data type datasets structured in a multi-relational format.

3.3.8.1. Compliance to CRISP-MED-DM

To evaluate compliance of the undertaken application of DM activities for PubMed database publications meta-analysis, we applied the first evaluation strategy, proposed in Section 2.4.6.2.

In contrast with predictive DM, descriptive DM has an exploratory nature and in many cases does not require the formulation of measurable success criteria or a deployment plan. The problem understanding phase scores only 3.3 from a maximum 10 points. Due to the exploratory nature of the research, overall success criteria, vision statement and the activities related with formal project management have not been performed. As shown in Fig. 44, the core phases Data understanding, Data Preparation, Modelling and Evaluation show good performance.

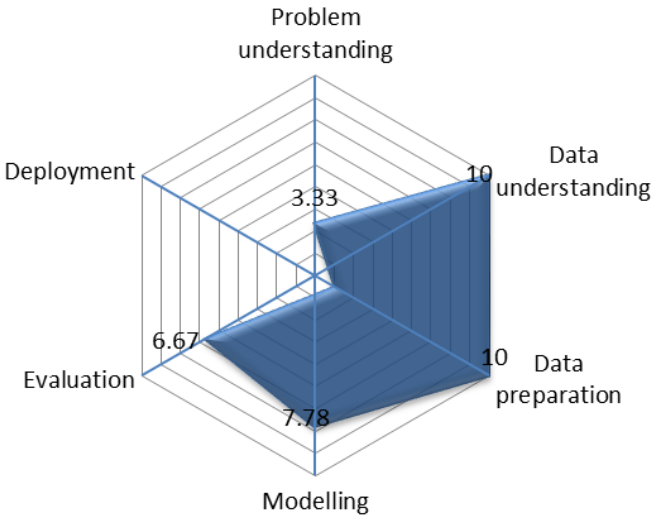


Fig. 44. CRIP-MED-DM compliance radar for BRCA1 prediction

The deployment phase activities are not relevant for a descriptive DM project.

3.4. Generalization and Conclusion

In this chapter, we showed the experimental use of the methods described in Chapter 2 “Systematic application of data mining and data analysis methods in medical domain”.

Predictive DM case studies in the Oncology and Cardiology domain, presented in Section 3.1 and 3.2, followed CRISP-MED-DM methodology, which improved the performance of the predictive models.

Experimental application of echocardiography image data analysis methods, showed high accuracy results, closely matching the manual measurements of professional cardiologists.

A novel multi-relational clustering approach was tested for medical publications meta-analysis. In comparison to traditional clustering methods on a single-table dataset, multi-relational clustering resulted in more informative clusters, suggesting a better understanding of the most popular research topics.

This page is intentionally left blank.

Conclusions

A constantly increasing amount of medical data is produced and captured electronically in everyday clinical practice. Automated knowledge discovery techniques, which employ data mining and machine learning techniques are capable of providing decision support for clinicians and discover new relevant patterns in silos of electronic patient data. However, the collected heterogeneous medical data lacks structural, functional and semantic interoperability. In addition, issues of patient data privacy and data ownership prevent effective usage of data mining methods.

Thus, the specialized data mining methodology CRISP-MED-DM, which addresses these issues, was proposed and experimentally tested in the oncology, cardiology and healthcare management domains.

The topics investigated and experimentally proved in the thesis allow us to conclude that:

1. The created data mining application methodology CRISP-MED-DM extends the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology with the following distinguished features:
    - The CRISP-MED-DM methodology, in comparison with other data mining applications methodologies, for the first time outlines the detailed process model specific to the issues and constraints of the medical domain. To achieve that, the initial CRISP-DM process model's five phases were extended with thirty three activities, addressing the issues of medical data pre-processing, semantic interoperability, and patient data privacy protection.
    - The created compliance evaluation model allows for performing the formal assessment of the data mining application projects' compliance to CRISP-MED-DM. The model provides metrics and evaluation formulas to assess the overall quality of application projects and allows for their comparison.

- The CRISP-MED-DM has been successfully tested in predictive modelling research projects in the oncology and cardiology domains.

2. The accuracy of the created breast cancer susceptibility gene *BRCA1* mutation predictive model has been increased by applying the CRISP-MED-DM methodology:
   - The improvement of the *BRCA1* gene mutation predictive model is as follows: overall accuracy from 0.88 to 0.94, sensitivity from 0.67 to 0.83, specificity from 0.85 to 0.97, ROC AUC from 0.70 to 0.81.
   - The improvement of breast cancer reoccurrence predictive model is as follows: overall accuracy from 0.73 to 0.75, sensitivity from 0.59 to 0.96, specificity has not changed, ROC AUC from 0.63 to 0.65.
3. The developed blood flow echocardiography image analysis technique saves the cardiologist time spent for systolic cycle tracing by extracting a systole cycle curve from standard Doppler ultrasound images and extracting features for further application of predictive data mining methods:
   - The developed software implementation of the proposed echocardiography images analysis technique, compared to the manual measurement of the professional cardiologists resulted in high accuracy for the main aortic valve stenosis diagnostic parameters: maximum aortic valve systolic velocity $AV\ V_{max}$ Pearson coefficient $r(16) = 0.999$ (p-value<0.0001); aortic valve time integral $AV\ VTI\ -\ r(16) = 0.988$ (p-value<0.0001); mean peak gradient $\Delta P_{max} - r(16) = 0.994$ (p-value<0.0001); aortic valve area $AVA - r(16) = 0.894$ (p-value<0.0001).
   - Applying CRISP-MD-DM with the proposed echocardiography image pre-processing techniques showed that the resulting accuracy is sufficient for practical decision support usage for aortic stenosis grading and diagnosis. The resulting predictive model had 100 % sensitivity and specificity on the research dataset.

4. Partitioning the clustering method with the proposed novel similarity measure allows clustering multi-relational data without its de-normalization and generalization to one-table format:
    - Application of the created similarity measure for PubMed database articles meta-analysis allowed for grouping multi-relational data into clusters with silhouette values 0.21–0.31, which showed better results in comparison with Tree Edit Distance measure results 0.15–0.23, and propositional approach results 0–0.16.
    - The calculation of the distance of each multi-relational object pair is independent and therefore can be successfully parallelized. The developed software implementation of multi-relational clustering supporting parallel calculation allows decreasing similarity measure calculation time in proportion to available processor nodes.

This page is intentionally left blank.

References

Abramoff, M. D., Magalhaes, P. J. & Ram, S. J., 2004. Image processing with ImageJ. *Biophotonics international,* 11(7), pp. 36–43.

Accenture, 2010. *Overview of International EMR/EHR Markets,* s.l.: s.n.

Aggarwal, C. C., 2007. *Data streams: models and algorithms.* s.l.:Springer Science & Business Media.

Altman, D. G., Vergouwe, Y. & Royston, P., 2009. Prognosis and prognostic research: valdiating a prognostic model. *BMJ,* 338(b605).

Azevedo, A. & Lourenco, I. R., 2008. KDD, SEMMA and CRISP-DM: a parallel overview.

Babbage, C., 1832. *On the economy of machinery and manufactures.* s.l.:s.n.

Baylis, P., 1999. Better health care with data mining. *SPSS White Paper, UK,* pp. 1–8.

Beale, T., Heard, S., Kalra, D. & Lloyd, D., 2006. OpenEHR architecture overview. *The OpenEHR Foundation.*

Bellaachia, A. & Guven, E., 2006. Predicting breast cancer survivability using data mining techniques. *Age,* 58(13), pp. 10–110.

Bellazzi, R. & Zupan, B., 2008. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform,* Feb, 77(2), pp. 81–97.

Bezdek, J. C., Ehrlich, R. & Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. *Computers \& Geosciences,* 10(2), pp. 191–203.

Bodenreider, O., 2008. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics,* p. 67.

Bodenreider, O., 2012. *Medical Ontology Research.* [Online]. Available at: https://mor.nlm.nih.gov/pubs/pres/20120222-IBM_Watson.pdf. [Accessed 01 04 2015].

Cabena, P., Hadjinian, P., Stadler, R. & Verhees, J. &. Z. A., 1998. *Discovering Data Mining: From Concept to Implementation.* s.l.:Prentice-Hall, Inc..

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

Castro, D., 2009. Explaining international IT application leadership: Health IT. *Available at SSRN 1477486.*

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.* s.l., s.n., pp. 1–5.

Centre for Health Promotion of University of Toronto, 1999. *Conducting Survey Research,* s.l.: s.n.

Chapman, P. et al., 2000. CRISP-DM 1.0 Step-by-step data mining guide.

Chen, H., Fuller, S. S., Friedman, C. & Hersh, W., 2006. *Medical informatics: knowledge management and data mining in biomedicine.* s.l.:Springer.

Choi, J. P., Han, T. H. & Park, R. W., 2009. A hybrid Bayesian network model for predicting breast cancer prognosis. *Journal of Korean Society of Medical Informatics,* 15(1), pp. 49–57.

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

Cleveland, W. S. & Loader, C., 1996. Smoothing by local regression: Principles and methods. In: *Statistical theory and computational aspects of smoothing.* s.l.:Springer, pp. 10–49.

Clifton, C., 2010. *Definition of Data Mining,* s.l.: s.n.

Corne, D., Dhaenens, C. & Jourdan, L., 2012. Synergies between operations research and data mining: The emerging use of multi-objective approaches. *European Journal of Operational Research,* 221(3), pp. 469–479.

Curk, T. et al., 2005. Microarray data mining with visual programming. *Bioinformatics,* 21(3), pp. 396–398.

Dehaspe, L., Toivonen, H. & King, R. D., 1998. *Finding Frequent Substructures in Chemical Compounds.*. s.l., s.n., p. 1998.

Delen, D., Walker, G. & Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med,* Jun, 34(2), pp. 113–127.

Demaine, E., Mozes, S., Rossman, B. & Weimann, O., 2007. An optimal decomposition algorithm for tree edit distance.. *Automata, languages and programming. Springer Berlin Heidelberg,* pp. 146–157.

Dempster, A. P., Laird, N. M. & Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological),* pp. 1–38.

DMG, 2014. *PMML Standard v.4.2.1.* [Online] Available at: http://www.dmg.org/pmml-v4-2-1.html. [Accessed 20 05 2015].

Dorre, J., Gerstl, P. & Seiffert, R., 1999. *Text mining: finding nuggets in mountains of textual data.* s.l., s.n., pp. 398–401.

Dzeroski, S., 2010. *Relational data mining.* s.l.:Springer.

eHealthServer.com, 2012. *Survey on Application of Data Mining to Support Clinical Decisions.* [Online] Available at: http://www.ehealthserver.com/research-and-development/935-survey-on-application-of-data-mining-to-support-clinical-decisions. [Accessed 11 02 2014].

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), pp. 4434–4463.

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X., 1996. *A density-based algorithm for discovering clusters in large spatial databases with noise.*. s.l., s.n., pp. 226–231.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. From data mining to knowledge discovery in databases. *AI magazine,* 17(3), p. 37.

Fraley, C. & Raftery, A. E., 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association,* 97(458), pp. 611–631.

Gaber, M. M., Zaslavsky, A. & Krishnaswamy, S., 2005. Mining data streams: a review. *ACM Sigmod Record,* 34(2), pp. 18–26.

Google Inc., 2014. *Google Scholar.* [Online] Available at: http://scholar.google.com/. [Accessed 01 05 2015].

Gower, J. C., 1971. A general coefficient of similarity and some of its properties. *Biometrics,* pp. 857–871.

Graunt, J., 1939. *Natural and Political Observations made upon the Bills of Mortality (1662).* London: The Johns Hopkins Pres.

Hall, M. et al., 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter,* 11(1), pp. 10–18.

Hatle, L., Angelsen, B. & Tromsdal, A., 1980. Non-invasive assessment of aortic stenosis by Doppler ultrasound.. *British heart journal,* 43(3), pp. 284–292.

Heard, S., 2008. *OpenEHR archetypes and terminology.* [Online] Available at: https://openehr.atlassian.net/wiki/display/healthmod/Archetypes+and+Terminology. [Accessed 04 06 2015].

Horvath, T., Wrobel, S. & Bohnebeck, U., 2001. Relational instance-based learning with lists and terms. *Machine Learning,* 43(1–2), pp. 53–80.

Hotho, A., Nurnberger, A. & Paas, G., 2005. *A Brief Survey of Text Mining.*. s.l., s.n., pp. 19–62.

Houston, A. L. et al., 1999. Medical data mining on the internet: Research on a cancer information system. *Artificial Intelligence Review,* 13(5–6), pp. 437–466.

Jaccard, P., 1901. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura.* s.l.:Impr. Corbaz.

Kalra, D., Beale, T. & Heard, S., 2005. The openEHR foundation. *Studies in health technology and informatics,* Volume 115, pp. 153–173.

Kaufman, L. & Rousseeuw, P., 1987. Clustering by means of medoids.

Kirsten, M., Wrobel, S. & Horvath, T., 2001. Distance based approaches to relational learning and clustering. In: *Relational data mining.* s.l.:Springer, pp. 213–232.

Knobbe, A. J., De Haas, M. & Siebes, A., 2001. Propositionalisation and aggregates. In: *Principles of Data Mining and Knowledge Discovery.* s.l.:Springer, pp. 277–288.

Koh, H. & Tan, G., 2005. Data mining applications in healthcare. *J Healthc Inf Manag,* 19(2), pp. 64–73.

Kramer, S., Lavrac, N. & Flach, P., 2001. *Propositionalization approaches to relational data mining.* s.l.:Springer.

Kurgan, L. A. et al., 2001. Knowledge discovery approach to automated cardiac SPECT diagnosis. *Artificial intelligence in medicine,* 23(2), pp. 149–169.

Landau, L. & Lifshitz, E., 1987. *Fluid mechanics.* s.l.:Butterworth-Heinemann.

Landin, G., 2006. *BinaryFill library.* s.l.:s.n.

Lehmann, T. M. et al., 2005. Automatic categorization of medical images for content-based retrieval and data mining. *Computerized Medical Imaging and Graphics,* 29(2), pp. 143–155.

Lenz, C. et al., 2008. Blood viscosity modulates tissue perfusion: sometimes and somewhere. *Transfus Altern Transfus Med,* 4(9), p. 265–72.

Liu, X., Fan, Y., Deng, X. & Zhan, F., 2011. Effect of non-Newtonian and pulsatile blood flow on mass transport in the human aorta.. *Journal of Biomechanics,* 6(44), p. 1123–1131.

Martin Bland, J. & Altman, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet,* 327(8476), pp. 307–310.

Matignon, R., 2007. *Data Mining Using SAS Enterprise Miner: A Case Study Approach,* s.l.: s.n.

Meisel, S. & Mattfeld, D., 2010. Synergies of operations research and data mining. *European Journal of Operational Research,* 206(1), pp. 1–10.

Microsoft, Hyperion, SAS, 2001. *MSDN - XMLA.* [Online]. Available at: https://msdn.microsoft.com/en-us/library/ms977626.aspx. [Accessed 21 05 2015].

Muggleton, S., 1991. Inductive logic programming. *New generation computing,* 8(4), pp. 295–318.

National Cancer Institute, USA, 2009. *BRCA1 and BRCA2: Cancer Risk and Genetic*

Testing. [Online]. Available at: http://www.cancer.gov/cancertopics/factsheet/Risk/BRCA. [Accessed 02 04 2015].

National Center for Biotechnology Information, 2009. *PubMed - database of references and abstracts on life sciences and biomedical topics.*. [Online]. Available at: http://www.ncbi.nlm.nih.gov/pubmed. [Accessed 01 04 2015].

National Library of Medicine, MeSH, 2015. [Online]. Available at: http://www.nlm.nih.gov/mesh/meshhome.html. [Accessed 01 04 2015].

Neville, J., Adler, M. & Jensen, D., 2003. *Clustering relational data using attribute and link information.* s.l., s.n., pp. 9–15.

Ng, R. T. & Han, J., 2002. CLARANS: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on,* 14(5), pp. 1003–1016.

Niakšu, O. & Kurasova, O., 2012. Data Mining Applications in Healthcare Theory vs Practice.. *DB&IS Local Proceedings*, pp. 58–70.

Niakšu, O. & Žaptorius, J., 2014. Applying operational research and data mining to performance based medical personnel motivation system.. *Stud Health Technol Inform,* Volume 198, pp. 63–70.

Nightingale, F., 1863. *Notes on hospitals.* London: Longman, Green, Longman, Roberts, and Green.

Nitzlnader, M. & Schreier, G., 2014. Patient Identity Management for Secondary Use of Biomedical Research Data in a Distributed Computing Environment. *EHealth2014--Health Informatics Meets EHealth: Outcomes Research: The Benefit of Health-IT,* Volume 198, p. 211.

Noordergraaf, A., 2011. *Blood in motion.* New York,: Springer.

Ochiai, A., 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bull. Jpn. Soc. Sci. Fish,* 22(9), pp. 526–530.

Office of Technology Assesment. Congress of the United States., 1977. *Policy Implications of Medical Information.* Washington: U.S. Government Printing Office.

Olafsson, S., Li, X. & Wu, S., 2008. Operations research and data mining. *European Journal of Operational Research,* 187(3), pp. 1429–1448.

Olson, D. L. & Delen, D., 2008. *Advanced data mining techniques.* s.l.:Springer Science & Business Media.

Oracle, 2011. *JSR 247: Data Mining 2.0.* [Online]. Available at: https://www.jcp.org/en/jsr/detail?id=247. [Accessed 21 05 2015].

Otto, C. M., 2012. *The practice of clinical echocardiography.* s.l.:Elsevier Health Sciences.

Panchal, S. M., Ennis, M., Canon, S. & Bordeleau, L. J., 2008. Selecting a BRCA risk assessment model for use in a familial cancer clinic. *BMC medical genetics,* 9(1), p. 116.

Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P., 2005. Global cancer statistics, 2002. *CA Cancer J Clin,* Mar-Apr, 55(2), pp. 74–108.

Paulus, R. A., Davis, K. & Steele, G. D., 2008. Continuous innovation in health care: implications of the Geisinger experience. *Health Affairs,* 27(5), pp. 1235–1245.

Pawlik, M. & Augsten, N., 2011. RTED: a robust algorithm for the tree edit distance. *Proceedings of the VLDB Endowment,* Volume 5.4, pp. 334–345.

Piatetsky-Shapiro, G., 2014. *CRISP-DM, still the top methodology for analytics, data mining, or data science projects.* [Online]. Available at: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html. [Accessed 01 04 2015].

Pragarauskaitė, J. et al., 2013. *Frequent pattern analysis for decision making in big*

data, s.l.: s.n.

R Core Team, 2014. *R: A Language and Environment for Statistical.* [Online]. Available at: http://www.R-project.org. [Accessed 01 04 2015].

Ramon, J. et al., 2007. Mining data from intensive care patients. *Advanced Engineering Informatics,* 21(3), pp. 243–256.

Ren, X., Lange, R. & Balentine, J., 2014. *Aortic Stenosis.* s.l.:s.n.

Ridler, T. & Calvard, S., 1978. Picture thresholding using an iterative selection method. *IEEE transactions on Systems, Man and Cybernetics,* 8(8), pp. 630–632.

Robson, M. E. et al., 2004. A combined analysis of outcome following breast cancer: differences in survival based on BRCA1/BRCA2 mutation status and administration of adjuvant treatment. *Breast Cancer Res,* 6(1), pp. R8--R17.

Rudnick, A., 2004. An introductory course in philosophy of medicine. *Medical humanities,* 30(1), pp. 54–56.

Sacha, J. P., Cios, K. J. & Goodenday, L. S., 2000. Issues in automating cardiac SPECT diagnosis. *Engineering in Medicine and Biology Magazine, IEEE,* 19(4), pp. 78–88.

Sani, Z. A., Shalbaf, A., Behnam, H. & Shalbaf, R., 2014. Automatic Computation of Left Ventricular Volume Changes Over a Cardiac Cycle from Echocardiography Images by Nonlinear Dimensionality Reduction. *Journal of digital imaging,* pp. 1–8.

Schloeffel, P. et al., 2006. The relationship between CEN 13606, HL7, and openEHR. *HIC 2006 and HINZ 2006: Proceedings,* p. 24.

Schneider, C. A. et al., 2012. 671 nih image to imageJ: 25 years of image analysis. *Nature methods,* 9(7).

Shalbaf, A., Behnam, H., Alizade-Sani, Z. & Shojaifard, M., 2013. Automatic classification of left ventricular regional wall motion abnormalities in echocardiography images using nonrigid image registration. *Journal of digital imaging,* 26(5), pp. 909–919.

Silver, M. et al., 2001. Case study: how to apply data mining techniques in a healthcare data warehouse. *Journal of healthcare information management,* 15(2), pp. 155–164.

Skjaerpe, T., Hegrenaes, L. & Hatle, L., 1985. Noninvasive estimation of valve area in patients with aortic stenosis by Doppler ultrasound and two-dimensional echocardiography.. *Circulation,* 72(4), pp. 810–818.

Smith, K. A. & Gupta, J. N., 2000. Neural networks in business: techniques and applications for the operations researcher. *Computers \& Operations Research,* 27(11), pp. 1023–1044.

Soille, P., 2013. *Morphological image analysis: principles and applications.* s.l.:Springer Science & Business Media.

Spečkauskienė, V. & Lukoševičius, A., 2009. A data mining methodology with preprocessing steps. *Information Technology and Control,* 38(4), pp. 319–324.

Špečkauskienė, V. & Lukoševičius, A., 2009. Methodology of adaptation of data mining methods for medical decision support: Case study. *Electronics and Electrical Engineering,* 2(90), pp. 25–28.

Stacey, M. & McGregor, C., 2007. Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial intelligence in medicine,* 39(1), pp. 1–24.

Stroetmann, K. A., Artmann, J., Stroetmann, V. N. & Whitehouse, D., 2011. European countries on their journey towards national eHealth infrastructures. *Final European progress report,* pp. 1–47.

Tai, K.C., 1979. The tree-to-tree correction problem.. *Journal of the ACM (JACM),* Volume 26.3, pp. 422–433.

Tak, T., Mathews, S. & Chandraratna, P., 1996. Severity of aortic regurgitation assessed by digital image processing of Doppler spectral recordings. *Echocardiography,* 13(3), pp. 259–263.

Tanwani, A. K., Afridi, J., Shafiq, M. Z. & Farooq, M., 2009. Guidelines to select machine learning scheme for classification of biomedical datasets. In: *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics.* s.l.:Springer, pp. 128–139.

Thomson Reuters Web of Science, 2014. *Thomson Reuters Web of Science.* [Online]. Available at: http://thomsonreuters.com/thomson-reuters-web-of-science/. [Accessed 01 04 2015].

Tibco, Inc., 2010. *TIBCO Spotfire Miner™ 8.2 User's Guide.* [Accessed 01 04 2015]

Van Laer, W. & De Raedt, L., 2001. How to upgrade propositional learners to first order logic: A case study. In: *Machine Learning and Its Applications.* s.l.:Springer, pp. 102–126.

Waljee, A. K. H. P. D. &. S. A. G., 2013. A primer on predictive models. *Clinical and translational gastroenterology,* 5(1), p. e44.

Wehlou, M., 2014. *Rethinking the Electronic Healthcare Record.* s.l.:MITM - Man In The Middle AB.

Wilson, A., Thabane, L. & Holbrook, A., 2004. Application of data mining techniques in pharmacovigilance. *British journal of clinical pharmacology,* 57(2), pp. 127–134.

Wu, X. et al., 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems,* 14(1), pp. 1–37.

Yang, Q. & Wu, X., 2006. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making,* 5(04), pp. 597–604.

Yeh, J.Y., Wu, T.-H. & Tsao, C.-W., 2011. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems,* 50(2), pp. 439–448.

Yin, X., Han, J. & Yu, P. S., 2005. *Cross-relational clustering with user's guidance.* s.l., s.n., pp. 344–353.

Yin, X., Han, J. & Yu, P. S., 2006. *LinkClus: efficient clustering via heterogeneous semantic links.* s.l., s.n., pp. 427–438.

Yoo, I. et al., 2012. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems,* 36(4), pp. 2431–2448.

Ziou, D., Tabbone, S. & others, 1998. Edge detection techniques-an overview. *Pattern Recognition And Image Analysis C/C Of. Raspoznavaniye Obrazov I Analiz Izobrazhenii,* Volume 8, pp. 537–559.

ANNEXES

Annex A. Aortic Valve Stenosis Predictive Model in PMML Format

```
<PMML version="2.0">
    <DataDictionary numberOfFields="9">
        <DataField name="class_id" optype="categorical">
            <Value value="0"/>
            <Value value="1"/>
            <Value value="2"/>
            <Value value="3"/>
        </DataField>
        <DataField name="pt_sex" optype="categorical">
            <Value value="M"/>
            <Value value="V"/>
        </DataField>
        <DataField name="pt_age" optype="continuous"/>
        <DataField name="xAV_2" optype="continuous"/>
        <DataField name="xAV" optype="continuous"/>
        <DataField name="cAV" optype="continuous"/>
        <DataField name="xLV_2" optype="continuous"/>
        <DataField name="xLV" optype="continuous"/>
        <DataField name="cLV" optype="continuous"/>
    </DataDictionary>
    <TreeModel modelName="class_id" functionName="classification" splitCharacteristic="binarySplit">
        <Extension extender="Insightful" name="X-IMML-XTProps">
            <X-IMML-XTProps>
                <X-IMML-Property name="criterion" value="entropy"/>
            </X-IMML-XTProps>
        </Extension>
        <MiningSchema>
            <MiningField name="class_id" usageType="predicted"/>
            <MiningField name="pt_sex"/>
            <MiningField name="pt_age"/>
            <MiningField name="xAV_2"/>
            <MiningField name="xAV"/>
            <MiningField name="cAV"/>
            <MiningField name="xLV_2"/>
            <MiningField name="xLV"/>
            <MiningField name="cLV"/>
        </MiningSchema>
        <Node score="0" recordCount="270">
            <Extension extender="Insightful" name="X-IMML-XTProps">
                <X-IMML-XTProps>
                    <X-IMML-Property name="id" value="1"/>
                    <X-IMML-Property name="group" value="0"/>
                    <X-IMML-Property name="deviance" value="728.507509882887"/>
                    <X-IMML-Property name="entropy" value="728.507509882887"/>
                    <X-IMML-Property name="gini" value="0.730864197530864"/>
                    <X-IMML-Property name="risk" value="174"/>
                    <X-IMML-Property name="yprob" value="0.3556 0.2667 0.1778 0.2"/>
                    <X-IMML-Property name="improvement" value="179.00075"/>
                </X-IMML-XTProps>
            </Extension>
            <SimplePredicate field="cAV" operator="lessThan" value="2.80272083859873"/>
            <Node score="0" recordCount="168">
                <Extension extender="Insightful" name="X-IMML-XTProps">
```

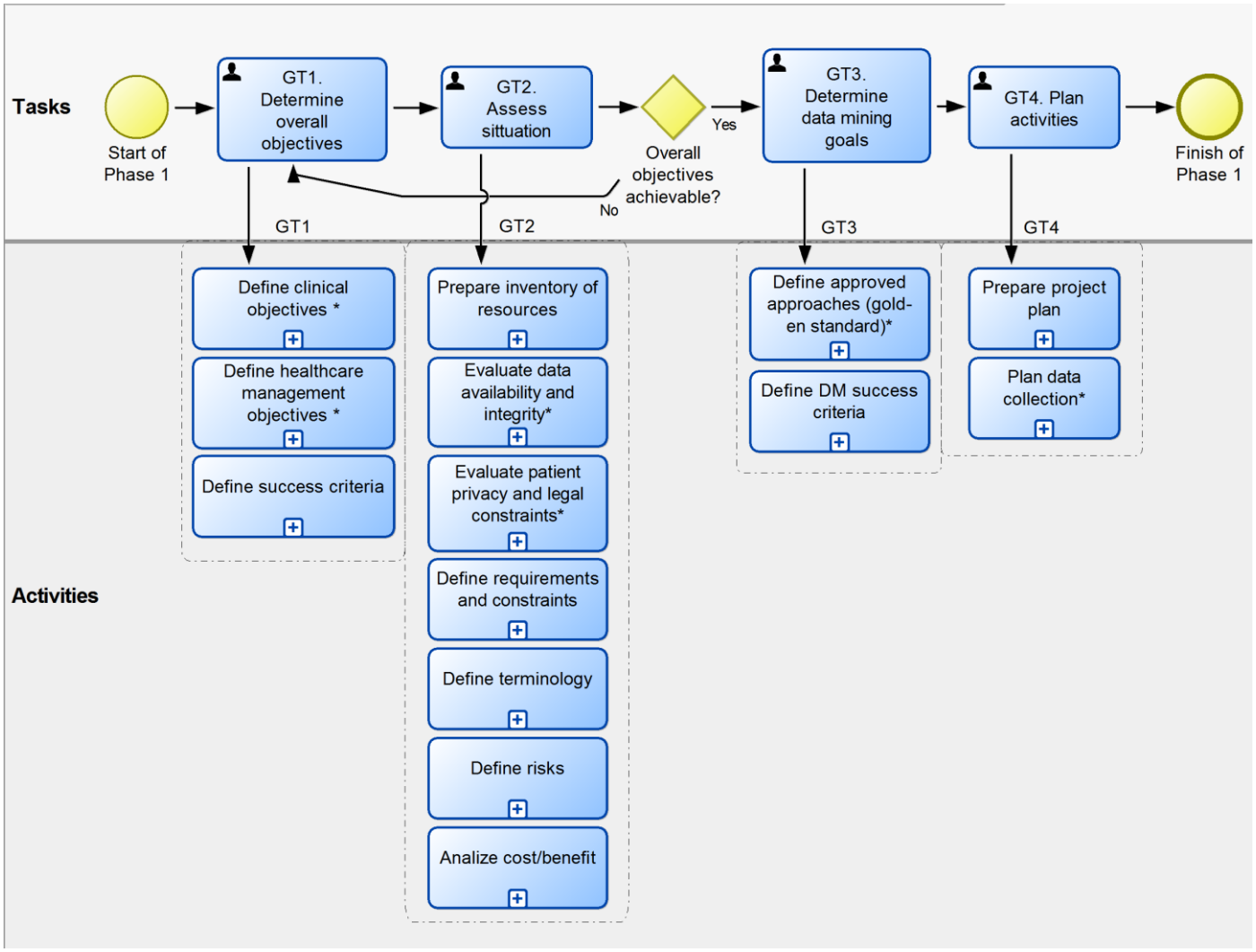
```
                    <X-IMML-XTProps>
                        <X-IMML-Property name="id" value="2"/>
                        <X-IMML-Property name="group" value="1"/>
                        <X-IMML-Property name="deviance" value="229.457"/>
                        <X-IMML-Property name="entropy" value="229.457"/>
                        <X-IMML-Property name="gini" value="0.48979"/>
                        <X-IMML-Property name="risk" value="72"/>
                        <X-IMML-Property name="yprob" value="0.57 0.429 0 0"/>
                        <X-IMML-Property name="improvement" value="99.059"/>
                    </X-IMML-XTProps>
                </Extension>
                <SimplePredicate field="cAV" operator="lessThan" value="1.46935040364626"/>
                <Node score="0" recordCount="92">
                    <Extension extender="Insightful" name="X-IMML-XTProps">
                        <X-IMML-XTProps>
                            <X-IMML-Property name="id" value="4"/>
                            <X-IMML-Property name="group" value="1"/>
                            <X-IMML-Property name="deviance" value="0"/>
                            <X-IMML-Property name="entropy" value="0"/>
                            <X-IMML-Property name="gini" value="0"/>
                            <X-IMML-Property name="risk" value="0"/>
                            <X-IMML-Property name="yprob" value="1 0 0 0"/>
                            <X-IMML-Property name="improvement" value=""/>
                        </X-IMML-XTProps>
                    </Extension>
                    <True/>
                </Node>
                <Node score="1" recordCount="76">
                    <Extension extender="Insightful" name="X-IMML-XTProps">
                        <X-IMML-XTProps>
                            <X-IMML-Property name="id" value="5"/>
                            <X-IMML-Property name="group" value="2"/>
                            <X-IMML-Property name="deviance" value="31.3411916962512"/>
                            <X-IMML-Property name="entropy" value="31.3411916962512"/>
                            <X-IMML-Property name="gini" value="0.0997229916897509"/>
                            <X-IMML-Property name="risk" value="4"/>
                            <X-IMML-Property name="yprob" value="0.05263 0.94736842 0 0"/>
                            <X-IMML-Property name="improvement" value=""/>
                        </X-IMML-XTProps>
                    </Extension>
                    <True/>
                </Node>
            </Node>
            <Node score="3" recordCount="102">
                <Extension extender="Insightful" name="X-IMML-XTProps">
                    <X-IMML-XTProps>
                        <X-IMML-Property name="id" value="3"/>
                        <X-IMML-Property name="group" value="2"/>
                        <X-IMML-Property name="deviance" value="141.048879833892"/>
                        <X-IMML-Property name="entropy" value="141.048879833892"/>
                        <X-IMML-Property name="gini" value="0.498269896193772"/>
                        <X-IMML-Property name="risk" value="48"/>
                        <X-IMML-Property name="yprob" value="0 0 0.47059 0.5294"/>
                        <X-IMML-Property name="improvement" value="51.6875067414223"/>
                    </X-IMML-XTProps>
                </Extension>
                <SimplePredicate field="cAV" operator="lessThan" value="4.14862756695093"/>
                <Node score="2" recordCount="54">
                    <Extension extender="Insightful" name="X-IMML-XTProps">
                        <X-IMML-XTProps>
                            <X-IMML-Property name="id" value="6"/>
                            <X-IMML-Property name="group" value="1"/>
                            <X-IMML-Property name="deviance" value="37.6738663510475"/>
                            <X-IMML-Property name="entropy" value="37.6738663510475"/>
                            <X-IMML-Property name="gini" value="0.197530864197531"/>
```

```
                                <X-IMML-Property name="risk" value="6"/>
                                <X-IMML-Property name="yprob" value="0 0 0.888888889 0.11111"/>
                                <X-IMML-Property name="improvement" value="18.8369331755237"/>
                            </X-IMML-XTProps>
                        </Extension>
                        <SimplePredicate field="cLV" operator="lessThan" value="1.24886031775343"/>
                        <Node score="2" recordCount="48">
                            <Extension extender="Insightful" name="X-IMML-XTProps">
                                <X-IMML-XTProps>
                                    <X-IMML-Property name="id" value="12"/>
                                    <X-IMML-Property name="group" value="1"/>
                                    <X-IMML-Property name="deviance" value="0"/>
                                    <X-IMML-Property name="entropy" value="0"/>
                                    <X-IMML-Property name="gini" value="0"/>
                                    <X-IMML-Property name="risk" value="0"/>
                                    <X-IMML-Property name="yprob" value="0 0 1 0"/>
                                    <X-IMML-Property name="improvement" value=""/>
                                </X-IMML-XTProps>
                            </Extension>
                            <True/>
                        </Node>
                        <Node score="3" recordCount="6">
                            <Extension extender="Insightful" name="X-IMML-XTProps">
                                <X-IMML-XTProps>
                                    <X-IMML-Property name="id" value="13"/>
                                    <X-IMML-Property name="group" value="2"/>
                                    <X-IMML-Property name="deviance" value="0"/>
                                    <X-IMML-Property name="entropy" value="0"/>
                                    <X-IMML-Property name="gini" value="0"/>
                                    <X-IMML-Property name="risk" value="0"/>
                                    <X-IMML-Property name="yprob" value="0 0 0 1"/>
                                    <X-IMML-Property name="improvement" value=""/>
                                </X-IMML-XTProps>
                            </Extension>
                            <True/>
                        </Node>
                    </Node>
                    <Node score="3" recordCount="48">
                        <Extension extender="Insightful" name="X-IMML-XTProps">
                            <X-IMML-XTProps>
                                <X-IMML-Property name="id" value="7"/>
                                <X-IMML-Property name="group" value="2"/>
                                <X-IMML-Property name="deviance" value="0"/>
                                <X-IMML-Property name="entropy" value="0"/>
                                <X-IMML-Property name="gini" value="0"/>
                                <X-IMML-Property name="risk" value="0"/>
                                <X-IMML-Property name="yprob" value="0 0 0 1"/>
                                <X-IMML-Property name="improvement" value=""/>
                            </X-IMML-XTProps>
                        </Extension>
                        <True/>
                    </Node>
                </Node>
            </Node>
        </TreeModel>
</PMML>
```

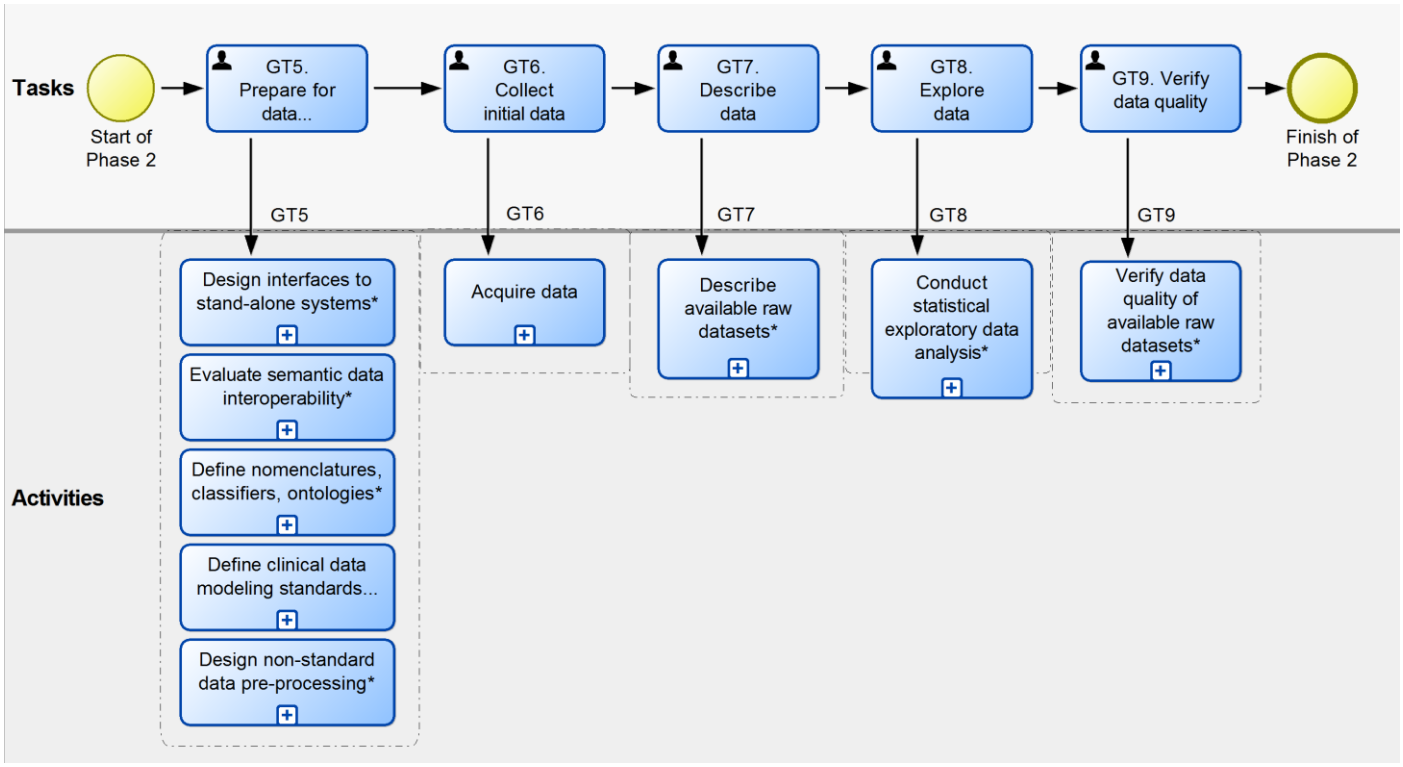
Supplementary Materials

Supplementary Figures. CRISP-MED-DM Process Flow Diagrams

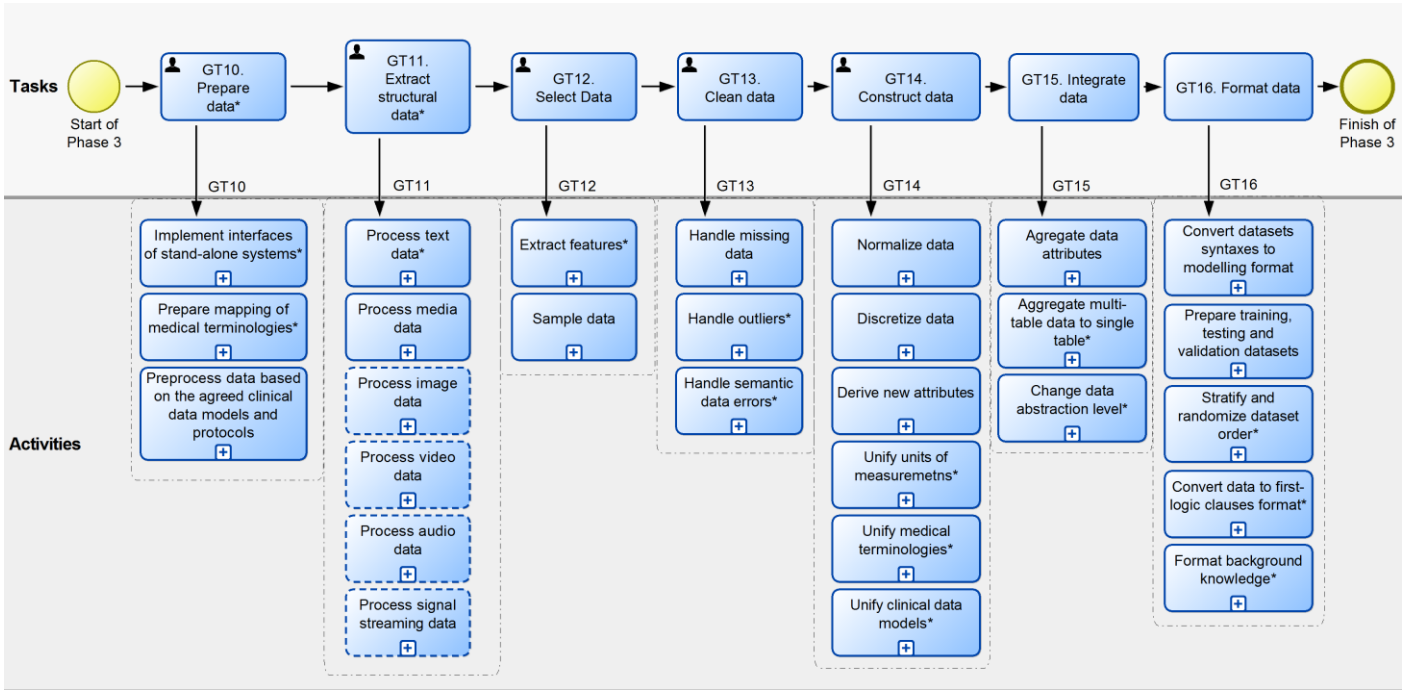
Phase 1 – Problem Understanding



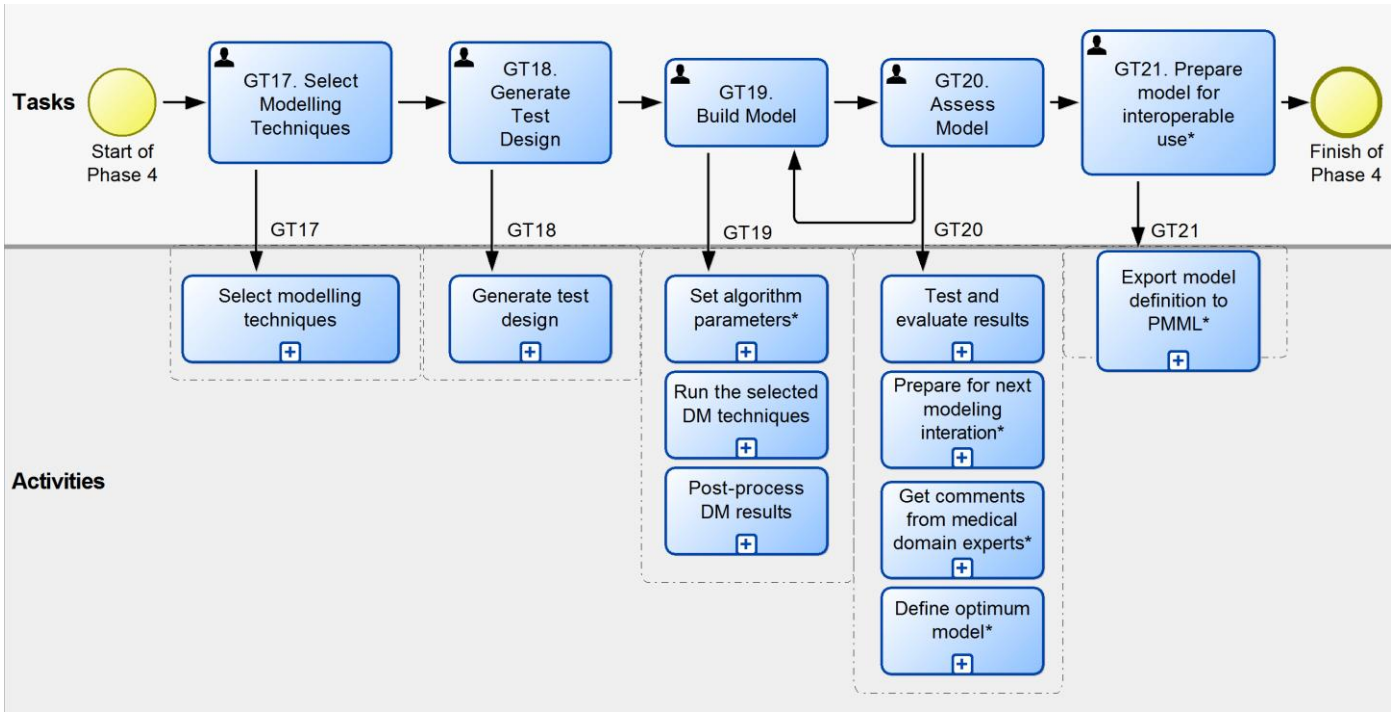
Phase 2 – Data Understanding



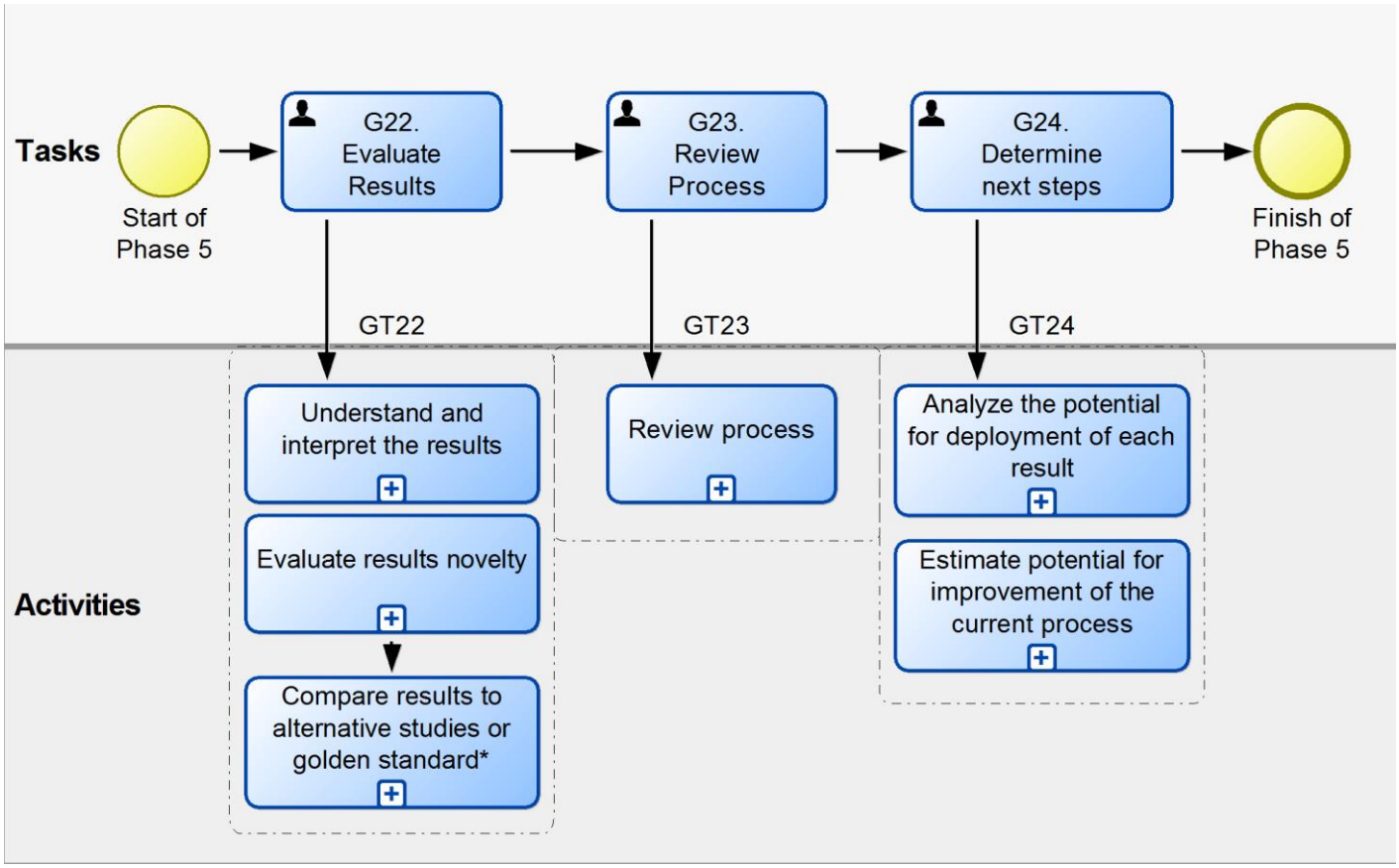
Phase 3 – Data Preparation



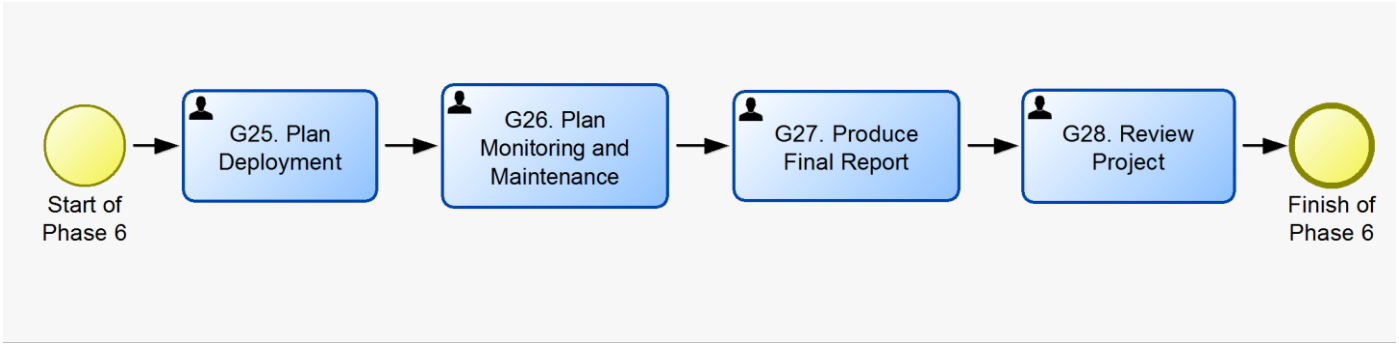
Phase 4 – Modelling



Phase 5 – Evaluation



Phase 6 – Deployment



Olegas NIAKŠU

DEVELOPMENT AND APPLICATION OF DATA MINING METHODS IN MEDICAL DIAGNOSTICS AND HEALTHCARE MANAGEMENT

Doctoral Dissertation

Technological Sciences,
Informatics Engineering (07 T)

Editor Alison Koczanski