Data Mining and Knowledge Discovery: Applications, Techniques, Challenges and Process Models in Healthcare

Shaker H. El-Sappagh[1], Samir El-Masri[2], A. M. Riad[3], Mohammed Elmogy[4]

[1]Faculty of Computers & Information, Mansoura University, Egypt
[2]College of Computer and Information Sciences, King Saud University, Saudi Arabia
[3,4]Faculty of Computers & Information, Mansoura University, Egypt

ABSTRACT

Many healthcare leaders find themselves overwhelmed with data, but lack the information they need to make right decisions. Knowledge Discovery in Databases (KDD) can help organizations turn their data into information. Organizations that take advantage of KDD techniques will find that they can lower the healthcare costs while improving healthcare quality by using fast and better clinical decision making. In this paper, a review study is done on existing data mining and knowledge discovery techniques, applications and process models that are applicable to healthcare environments. The challenges for applying data mining techniques in healthcare environment will also be discussed.

Key Words: *Data mining, knowledge discovery, healthcare, Electronic health record, health informatics.*

I. INTRODUCTION

The advances in health informatics as Electronic Health Record (HER) make healthcare organizations overwhelmed with data.There is a wealth of data available within the healthcare industry that would benefit from the application of KDD tools and techniques. These techniques transform the huge mounds of data into useful information for decision making [3]. A proper medical database created with intention mining can provide a useful resource for data mining and knowledge discovery [1]. The tools and techniques of KDD have achieved impressiveresults in other industries, and healthcare needs to take advantage of advances in this exciting field. A hospital typically has detailed data about every charge entered on a patient's bill, which easily can reach hundreds of charges for only a few days' stay. Each lab test, radiology procedure, medication, and so on is recorded, whether in a clinical information system or in a financial billing system.

There is an enormous volume of data generated [8], but few tools exist in the healthcare setting to analyze the data fully to determine the best practices and the most effective treatments. In general, the healthcare industry lags far behind other industries in terms of information technology expenditures. As healthcare continues to become more complex, the industry needs to find an effective means of evaluating its large volume of clinical, financial, demographic and socioeconomic data.

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Integration of KDD tools with EHR could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome.

EHR is only a first step in capturing and utilizing health-related data – the problem is turning that data into useful information. Models produced via data mining and predictive analysis can form the backbone of Clinical Decision Support Systems (CDSS).

KDD(also known as knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, and business intelligence) is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) meaningful patterns or knowledge from huge amount of data stored in multiple data sources such as file systems, databases, data warehouses and etc by automatic or semi-automatic means [10].

KDD has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing.

More traditional query tools require the user to make many assumptions. For example, a user may use a query tool to ask the question, "What is the link between cholesterol and heart disease?" By running this query, the analyst assumes that there is a relationship between cholesterol and heart disease. The power of KDD is that it will search the dataset for all relationships, including those that may not have occurred to the analyst. With large datasets, there may be many variables interacting with one another in very subtle ways. KDD can help find the hidden relationships and patterns within data.TheKDD is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. This process must have a model to control its execution steps.

KDD has many applications in healthcare such as patient diagnosis, patient treatment, management of chronic diseases, prediction of patients at risk for specific diseases, and in public health.Because of the complexity of healthcare environment, there are many challenges that face the application and complete benefit from KDD. This paper is organized as follows: section 2 discusses data mining applications. Section 3 discusses KDD process models. Section 4 discusses data mining techniques. Section 5 is the data mining challenges, and conclusion is discusses in section 6.

II. DATA MINING APPLICATIONS IN HEALTHCARE

There is vast potential for data mining applications in healthcare. Generally, these can be grouped as the evaluation of treatment effectiveness; management of healthcare; andCustomer Relationship Management (CRM). More specialized medical data mining, such as analysis of DNA micro-arrays, lies outside the scope of this paper.

- *Evaluation of treatment effectiveness:*Data mining applications can be developed to evaluate the effectiveness of medical treatments (evidence based medicine). By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective such as predict optimum medication dosage.
- *Healthcare management:* Data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims. It can search for patterns that might indicate an attack by bio-terrorists.Moreover, this system can be used for hospital infection control, or as an automated early-warning system in the event of epidemics. Accurate prognosis and risk assessment as survival analysis for AIDS patients, predict pre-term birth risk, determine cardiac surgical risk, predict ambulation following spinal cord injury, and breast cancer prognosis.
- *CRM:* Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings. It determines the preferences, usage patterns, and current andfuture needs of individuals to improve their level of satisfaction.

Other applications of KDD in healthcare are:

- Many providers are migrating toward the use HER [12]. EHR store a large quantity of patient data on test results, medications, prior diagnoses, and other medical history. This is a valuable source of information that could be better used by employing KDD techniques. Several examples include identifying patients who should receive flu shots, enroll in a disease management program and are not in compliance with a treatment plan. Moreover, historical data in EHR can help in management of chronic diseases and anticipating patient's future behavior on the given history. EHR stores spatial and demographic data which can help in public health management and planning.
- Finding themselves in an increasingly competitive market, many healthcare organizations are now employing sophisticated marketing efforts. KDD can help in this arena in ways similar to other industries. Organizations can use their data to identify those most likely to use their facilities and the most effective marketing activities to reach those individuals.
- KDD has proved as successful at identifying fraud in the healthcare industry as it has in other industries. Several insurance companies use KDD techniques to sift through their claims, seeking to identify fraudulent providers.
- KDD is used to predict patient problems based on his medical history.
- When researchers investigate the records of patients with a particular disease to determine if there are any risk factors in their histories that could help predict the occurrence of the disease in other patients; they may identify a new risk factor that could help detect the disease sooner in other patients and allow for more timely intervention.
- Data mining techniques could be successfully applied for detection and diagnosis with a reasonably high performance of many medical diseases as cancer [4], diabetes [6, 24], kidney [7] heart diseases [2], andsleep problems [5].
- Hospital management for optimize allocation of resources and assist in future planning for improved services. For example, forecasting patient volume, ambulance run volume, etc and predicting length-of-stay for incoming patients.

III. DATA MINING PROCESS MODELS IN HEALTHCARE

Knowledge discovery is a process, and not a one-time response of the KDD system to a user's action. As any other process, it has its environment, its phases, and runs under certain assumptions and constraints. Figure 1 shows a typical decision making environment. In the healthcare environment, the source data in clinical databases and/or EHRs can be queried directly using SQL. A Data Warehouse (DW) can be created to integrate data from many sources and enhance data quality. Analysts can apply OLAP tools on the DW. Data warehouse is not enough for data analysis. Data mining is required to discover hidden patterns in either EHR or DW.

KDD is an iterative or cyclic process that involves a number of stages. Although the specific techniques may vary from project to project, the basic process is the same for all KDD problems.

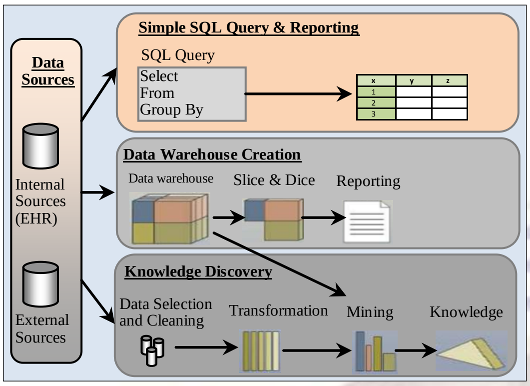

Figure 1: Decision Making Environment

Following standard process in KDD help analysts by guiding them through analysis process exposing those aspects that could otherwise be forgotten or neglected. Figure 2shows the basic phases of the KDD process [11].

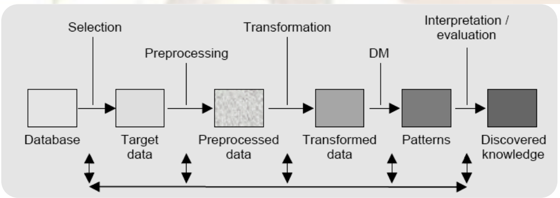

Figure 2: Phases in the KDD process

Selection phase generates the target data set from the database. *Preprocessing*solve issues aboutnoise, incomplete and inconsistent data.The next phase is *transformation* of the preprocessed data into a suitable form for performing the desired DM task. In the *DM* phase, a procedure is run that executes the desired DM task and generates a set of patterns. However, not all of the patterns are useful. The goal of *interpreting and evaluating* all the patterns discovered is to keep only those patterns that are interesting and useful to the user and discard the rest. Those patterns that remain represent the discovered knowledge.

This process is a time-consuming, incremental, and iterative process by its very nature, hence many repetition and feedback loops exists in Figure 2. Individual phases can be repeated alone, and the entire process is usually repeated for different data sets.

Many data mining process methodologies are available. However, the various steps do not differ much from methodology to methodology. The major KDD process models will be discussed next:The first reported KDD model consists of nine steps and was developed by Fayyad et al. in the mid-1990s [10]. The next model, by Cabena et al., consists of five steps and was introducedin 1998 [14]. The third model, which consists of eight steps, was developed byAnand& Buchner at about the same time [15]. TheCRISP-DM (CRoss-Industry Standard Process for DM) process model that includes six steps wasfirst proposed in early 1996 [17] by a consortium of four companies: SPSS (a provider of commercialDM solutions), NCR (a database provider), Daimler Chrysler, and OHRA (an insurancecompany). The last two companies served as sources of data and case studies. Next is the six-step process model of Cios et al. [16] by adopting the CRISP-DM model to the needs ofacademic research community. The main extensions of the latter model include providing a moregeneral, research-oriented description of the steps, introduction of several explicit feedbackmechanisms and a modification of the description of the last step, which emphasizes that knowledgediscovered for a particular domain may be applied in other domains. Another model is the six-step KDD process model by Adriaans&Zantinge (1996), which consists of DataSelection, Cleaning, Enrichment, Coding, DM, and Reporting. Another model is the four-step model by Berry &Linoff (1997), which consists of Identifying the Problem, Analyzingthe Problem, Taking Action, and Measuring the Outcome. Another model is the five-step SEMMA model by the SAS Institute Inc. (1997), which consists of steps namedSample, Explore, Modify, Model, and Assess. This model was incorporated into commercial KDsoftware platform SAS Enterprise Miner. Another model is the seven-step model by Han &Kamber (2001), which consists of Learning the Applicationdomain, Creating a Target Data Set, Data Cleaning and Preprocessing, Data Reduction andTransformation, Choosing Functions of DM, Choosing the Mining Algorithm(s), DM, PatternEvaluation and Knowledge Presentation, and Use of Discovered Knowledge. Another model is the five-step model by Edelstein (2001), which consists of Identifying the Problem, Preparing theData, Building the Model, Using the Model, and Monitoring the Model. Another model is the seven-step model by Klosgen&Zytkow (2002), which consist of Definition and Analysis ofBusiness Problems, Understanding and Preparation of Data, Setup of the Search for Knowledge,Search for Knowledge, Knowledge Refinement, Application of Knowledge in Solving theBusiness Problems, and Deployment and Practical Evaluation of the Solutions. Finally is the seven-step model by Haglin et al. (2005), which consists of Goal Identification, Target DataCreation, Data Preprocessing, Data Transformation, DM, Evaluation and Interpretation, andTake Action steps.

A. NEW TRENDS IN KDD PROCESS MODELS

The future of KDD process models is in achieving overall integration and distribution of the entire process through the use of other popular industrial standards. Another currently very important issue is to provide interoperability and compatibility between different software systems and platforms, which also concerns KDD models. Such systems would serve end-users in automating, or more realistically semi-automating the KDD systems.

A current goal is to enable users to carry out KDD projects without possessing extensive background knowledge, without manual data manipulation, and without manual procedures to exchange data and knowledge between different DM methods. This requires the ability to store and exchange not only the data, but most importantly knowledge that is expressed in terms of data models generated by the KDD process and domain expert, and meta-data that describes data and domain knowledge used in the process.The technologies, which can help in achieving these goals, are XML and Predictive Model Markup Language (PMML). From the KDD point of view, XML is the key technology to [18]:

- Standardize communication between diverse data mining (DM) tools and databases.
- Build standard data repositories sharing data between different DM tools that work on different software platforms.
- Implement communication protocols between the DM tools.
- Provide a framework for integration and communication between different DMKD steps. The information collected during domain and data understanding steps can be stored as XML documents. Then, the information can be used for data preparation and data mining steps, as a source of already-accessible information, cross-platforms, and cross-tools.

IV. DATA MINING TECHNIQUES IN HEALTHCARE

There are a variety of data mining techniques available, all with pros and cons, depending on the business problem at hand and the data available for analysis. One of the most difficult tasks is to choose the right data mining technique which requires more and more expertise. The final choice depends basically on:

- The main goal of the problem to be solved.
- The structure of the available data.

Conceptual map ofmining technique helps in selection for non-expert miners [26]. Discovered model (pattern, knowledge) from data mining algorithms can be *predictive* if they predict the value of an attribute (referred to as the target attribute or target), by making use of the remaining attributes (referred to as predictive attributes). This is also called *supervised learning*. The target attribute is the focus of this process, and the data usually includes examples where the values of the target attribute have been observed. The goal is to build a model that can predict the value of the target attribute for new unseen examples. If the target value is nominal then the prediction task is known as *classification* and each possible value for the target variable is referred to as its class-label or class. For real-valued target attributes, the prediction task is known as *regression*. *Unsupervised learning* refers to modeling with an unknown target variable. In that case, models are solely descriptive. The goal of the process is to build a model that describes interesting regularities in the data. *Clustering* is an example of a descriptive algorithm that is concerned with partitioning the examples in similar subgroups.

A. A GLANCE AT DATA MINING TECHNIQUES

Data mining techniques can be classified based on the database, the knowledge to be discovered and the techniques to be utilized.

Classification: if the target attribute is categorical then the applicable techniques are decision tree induction, Bayesian classification, backpropagation (neural network), based on concepts from association rule mining, k-nearest neighbor, case based reasoning, genetic algorithms, rough set theory, support vector machine and fuzzy set. If the target attribute is continuous then we use linear, multiple and non-linear regression.

Clustering: many methods are applicable as partitioning methods (e.g. k-means, k-medoids), hierarchical methods (e.g. chameleon, CURE), density-based methods (e.g. DBSCAN, OPTIC), grid-based methods (e.g. STING, CLIQUE) and model-based methods (e.g. statistical and neural network approaches).

Association rules:There are many classifications of these algorithms such as according to the types of values handled (e.g. Boolean and quantitative rules), according to the number of dimensions (e.g. single and multi dimensional) and according to the level of abstraction (e.g. single and multi level rules).

Graph mining, Social Network Analysis (SNA), link mining and multi-relational data mining: these techniques become increasingly important in modeling sophisticated structures and their interactions and relationships.

Data mining algorithms are needed in almost every step in KDD process ranging from domain understanding to knowledge evaluation. For example, in data preprocessing, algorithms (e.g. genetic algorithms) are used in data integration, transformation and cleaning and in knowledge discovery. Because of space restriction, we will not give examples of applicable techniques in healthcare real problems.

The most used software tools that provide these techniques are SPSS/ SPSS Clementine, Salford

Systems CART/MARS/TreeNet/RF, Yale (open source), SAS / SAS Enterprise Miner, Angoss Knowledge Studio / Knowledge Seeker, KXEN, Weka (open source), R (open source), Microsoft SQL Server, MATLAB and Oracle Data Miner.

V. DATA MINING CHALLENGES IN HEALTHCARE

Application of KDD in healthcare faces many challenges as [19, 20]:

1- Need for algorithms with very high accuracy because it is an issue of life or death. The problems of missing, noisy and inconsistent data complicate this objective. Moreover, unlike other (arguably simpler) domains, the medical discipline itself is diverse, complex and, to an outsider, relatively opaque.

2- Active data mining: It needs two types of triggers. First type used to fire data mining technique to analyze the data automatically after some time or making some updates. The second type is to enforce the discovered knowledge by embedding the discovered knowledge within clinical information systems.

3- For the same problem, we need to apply many data mining techniques, compare their results and select the most interesting one.

4- Automated or semi-automated KDD system: this challenge is critical since EHR data and medical knowledge are changed continuously. KDD process needs to start automatically or with a minor help from analyst.

5- Previous (background) knowledge as concept hierarchy, domain expert knowledge, previously mined knowledge and rule template must be taken in accountin data preparation, modeling and model evaluation phases [25]. Background knowledge can be expressed in different formats: examples may be found in the areas of decision rules, Bayesian models, fuzzy sets and concept hierarchies. Figure 3 shows that background knowledge must be taken in to account in the KDD process.

5- The result of data mining system must be appended in the existing knowledge base (knowledge fusion). The new knowledge may update, remove, or add constraints to the existing knowledge.

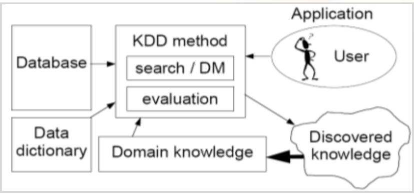

Figure 3: Background knowledge role in KDD process

6- Longitudinal, temporal and spatial support: There is a need for advancement in data mining techniques to deal with EHR environment [27]. EHR contains clinical data collected and summarized from many healthcare systems, temporal data forming patient history, social network of patients (Patient ID, Mother ID, Father ID, and Family ID), and spatial data as patient addresses. EHR may contain data with different formats as audio, video, image and text data. These formats also need advanced data mining techniques. The use of static techniques thus oversimplifies (or may hide) possible relationships and thus support for longitudinal, temporal and spatial semantics within the mining process is highly desirable. Episodic data is often the key to good data mining. Techniques as anomaly detection [21], difference detection [22], longitudinal X analysis [22] and temporal/spatial X analysis [23] are used.

7- Mining complex knowledge from complex data: Using multi-relational data mining, mining knowledge in the form of graphs and mining non-relational data as text and image are future issues.Using text mining is challenging in analysis of physician free text describing patient diagnoses and free text prescription. It can be used to summarize patient conditions.

8- Utilizing data mining in CDSS creates a system of real-time data-driven clinical decision support, or "adaptive decision support." It can "learn" over time, and can adapt to the variation seen in the actual real-world population. The approach is two pronged – developing new knowledge about effective clinical practices as well as modifying existing knowledge and evidence-based models to fit real-world settings [9].

9- Most KDD systems today work with relational databases. KDD and data mining need to be extended to object-oriented and multimedia databases.

10- Various security and privacy aspects: how to ensure the users' privacy while their data are being mined? One solution is*Anonymousness and identification transformation.* Besides data preprocessing, in order to separate the relative between patients and their records which may referenced some private information, anonym-oussness and identification transformation are also needed.

11- Distributed data mining, mining heterogeneous and multi-agent data: many schemas can be utilized as collecting data from all sites and then mine it, or mine the local data and distribute the discovered knowledge. In distributed mining, one problem is how to mine across multiple heterogeneous data sources such as multi-database and multi-relational mining.

12- Large databases: clinical databases are very large and massive with hundreds of tables and fields, millions of records. Using more efficient algorithms, sampling, approximation, feature selection, and parallel processing can mitigate the problem.

13- High dimensionality: clinical environment has a very large number of fields. It can be solved by using prior knowledge to know irrelevant variables.

14- Changing data and knowledge: Rapidly changing data make previously discovered knowledge invalid. Solving the problem of automatic KDD system can solve this problem.
15- Integration with other systems: A standalone discovery system may not be very useful. Integration issues include integration with a DBMS (e.g. via a query interface), CPOE (Computer-based Physician Order Entry system), spread sheets and visualization tools.
16- Distributed data mining: As EHR and clinical databases are mostly distributed. This situation require enhancement to data mining algorithms and KDD process to work with distributed data.
17- Developing a unified framework of data mining that can deal with the healthcare environment.
18- Mining sequence data and time series data: A particularly challenging problem is the noise in time series data. It is an important open issue to tackle
19- Encoding problem: Healthcare data comprises both numeric and textual information for medications diagnostic tests, demographics, problem lists, staff notes and images for radiology. Standard coding, such as International Classification of Disease (ICD-9-CM) codes for diagnoses and National Drug Codes (NDC) for drugs, SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms), Logical Observation Identifiers Names and Codes (LOINC) and Classification of Interventions and Procedures (OPCS-4) was implemented when possible rather than having free text dataentry to facilitate research and analyses.
20- Collecting the default knowledge is a challenge. Default knowledge as "Only women can pregnancy" is used by domain expert or physician when making decision. This knowledge needs to be collected and added in the data mining system's knowledge base. As a result, if a person is pregnant, then the system, by default, knows that she is a woman.

VI. CONCLUSION

In this paper, we tried to make a review of all KDD applications in healthcare. As the healthcare is a complex environment, there are many useful and proved applications of data mining such as diagnoses, prognoses, treatment and others. The existing process models for data mining life cycle are also reviewed. These models are critical for any KDD project because knowledge discovery contains many phases. Existing data mining algorithms are also reviewed, and all problems, challenges that face KDD in healthcare environment are collected.

REFERENCES

[1] S. K.Wasan, V. Bhatnagar, H. Kaur, The impactofdataminingtechniquesonmedical diagnostics,in the proceeding of Data Science Journal, Vol. 5, 2006, 119-126.

[2] K.Srinivas, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, International Journal on Computer Science and Engineering (IJCSE), 02(02), 2010, 250-255.

[3] M. Khajehei, F. Etemady, Data Mining and Medical Research Studies, IEEE Second Int. Conference on Computational Intelligence, Modeling and Simulation (CIMSiM),2010, 119 – 122.

[4] L. Li, et al., Data mining techniques for cancer detection using serum proteomic profiling, Artificial Intelligence in Medicine, vol. 32(2), 2004, 71-83.

[5] O. Shmiel, et al., Data mining techniques for detection of sleep arousals, Journal of Neuroscience Methods, vol. 179(2), 2009, 331-337.

[6] J. L. Breault,Data mining a diabetic data warehouse, Journal of Artificial Intelligence in Medicine, 26 (1-2), 2002, Pages 37-54.

[7] A. Kusiak, et al., Predicting survival time for kidney dialysis patients: a data mining approach, Computers in Biology and Medicine, vol. 35(4), 2005, 311-327.

[8] N. Aggarwal, A. Kumar, H. Khatter, V. Aggarwal, Analysis the effect of data mining techniques on database, Advances in Engineering Software 47(1), 2012, 164–169.

[9] C. Bennett, T. Doub; Data Mining and Electronic Health Records: Selecting Optimal Clinical Treatments in Practice, Proceedings of the 6th Int. Conference on Data Mining, pp. 313-318, 2011.

[10] Fayyad, U., Piatetsky-Shapiro G. and Smyth, P., from data mining to knowledge discovery: an Overview, In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R (eds) Advances in Knowledge Discovery and Data Mining. AAAI Press, 1996, 1-34.

[11] U. Fayyad, P. Smyth, G. Piatetsky, knowledge discovery and data mining towards a unifying framework, KDD-96, American Association for Artificial Intelligence (AAAI), 1996.

[12] LI Yun, LI Xiang-sheng,The data mining and knowledge discovery in biomedicine,IEEE 5th Int. Conf. onComputer Science & EducationHefei (ICCSE), China. 2010, 1050, 1052.

[13] J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.

[14] Cabena, P, Hadjinian, P, Stadler, R, Verhees, J and Zanasi, A, Discovering Data Mining: From Concepts to Implementation, Prentice Hall. 1998.

[15] Anand, S, Buchner A; Decision Support Using Data Mining, Financial Time Management, London. 1998.

[16] Cios, K, Kurgan L.,Trends in data mining and knowledge discovery," In Pal, N and Jain, L (eds) Advanced Techniques in Knowledge Discovery and Data Mining. Springer, pp. 1–26. 2005.

[17] CRISP-DM, 2003, CRoss Industry Standard Process for Data Mining, http://www.crisp-dm.org.

[18] K. J. Cios, G. W. Moore, Uniqueness of Medical Data Mining, Artificial Intelligence in Medicine journal, 26(1-2), 2002, 1-24.

[19] Q. YANG, 10 Challenging problems in data mining research, Int. Journal of Information Technology & Decision Making, 5(4)2006, 597–604.

[20] Ruben D. Canlas Jr., Data mining in healthcare: Current applications and issues, MSc. of Science in Information Technology, Carnegie Mellon University, Australia 2009.

[21] Y. Li, N. Wu, S. Jajodia, X.S. Wang,Enhancing profiles for anomaly detection using time granularities," J. of Computer Security, 10(1,2), 2002, 137–157.

[22] Anthony K. H. Tung, Hongjun Lu, Jiawei Han, Ling Feng, Efficient mining of intertransaction association rules, IEEE Transactions on Knowledge and Data Engineering, 15(1), 2003, 43–56.

[23] C. Freksa, Temporal reasoning based on semi-intervals, Artificial Intelligence, 54, 1992, 199–227.

[24] Noah Lee, A. F. Laine, J. Hu, F. Wang, J. Sun, Sh.Ebadollahi, Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients, First IEEE Int. Conf. on Healthcare Informatics, Imaging and Systems Biology (HISB), on page(s): 250 – 257, 2011.

[25] F. Alonso, L. Martínez, A. Pérez, J. P. Valente,Cooperation between expert knowledge and data mining discovered knowledge: Lessons learned, ACM An International Journal on Expert Systems with Applications, 39(8), 2012, 7524–7535.

[26] K. Gibert, M. Sànchez-Marrè, V. Codina, Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation Techniques, International Environmental Modeling and Software Society (iEMSs), www.iemss.org, 2010.

[27] G. Hripcsak, C. Knirsch, L. Zhou, A. Wilcox, G. B. Melton, Bias Associated with Mining Electronic Health Records, Journal of Biomedical Discovery and Collaboration, 6, 2011, 48-52.