Interoperability of Data and Knowledge in Distributed Health care Systems

Reza Sherafat Kazemzadeh and Kamran Sartipi
Department of Computing and Software
McMaster University
Hamilton, L8S 4K1, Canada.
{*sherafr, sartipi*}*@mcmaster.ca*

Abstract

In this paper we propose a knowledge management framework for distributed health care systems consisting of data- and knowledge-bases that contain patient data and mined knowledge from health care institutions. The framework takes advantage of data mining techniques, enabling technologies and standards to provide decision making support for the health care personnel. The application areas of the new framework range from clinical care to administrative decision support. With the guidance of the health care researchers the available patient data is mined off-line to extract meaningful knowledge from medical data which can be shared with other institutions through XML-based documents (known as PMML) to achieve knowledge interoperability among different heterogeneous health care systems. Data interoperability is achieved through an XML-based clinical data representation standards (HL7 CDA) that is used to encode patient data. A clinical guideline and a logic module will receive inputs form both PMML and CDA documents to enable decision making at a higher level based on patient data and the mined knowledge. We also applied the proposed framework on three clinical case studies.

Keywords: Interoperability; Data Mining; Knowledge Management; Health care Informatics; Clinical Decision Support Systems; Clinical Guideline.

1 Introduction

Due to the paramount importance of the quality of public health care services, these services represent a major portion of the government spending in most countries. In Canada the Provincial Government of Ontario invested a total of $28.1 billion in health care services in 2003-4 [19]; it is expected that the total health care spending throughout Canada in 2005 reaches as high as $142 billion [3]. On the other hand, today's health care professionals are overwhelmed with information and preventable medical errors are estimated to be the main cause of 44,000-98,000 deaths and costs upto $29 billion per year in the United States alone. Studies show that computerized Clinical Decision Support Systems are helpful and have positive effects in improving clinical practices [12]. Any improvement in the health care sector will result in savings that benefit all the stakeholders, from patients to health care providers and government agencies in both financial aspects and the quality of care.

Currently, Information Systems (IS) have been deployed by many health care organizations for a wide range of different purposes, such as telemedicine, patient care, Electronic Health Record (EHR) systems, decision support and many more. A major obstacle to the widespread use of IT in health care settings is the high degree of heterogeneity between health care Information Systems. These systems are interconnected and the need for proper communication and collaboration is apparent. Data should be semantically understandable by the receiver who is not necessarily using the same data format as the source. The data interoperability problem has been tackled by some health care data standards some of which are based on *Health Level 7 Reference Information Model* (HL-7 RIM).

In many cases, the health care personnel lack the necessary experience and need decision making assistance from the IS that they are working with. While some decision support approaches focus on providing plain clinical knowledge in the form of clinical guidelines GLIF [6, 1], our approach focuses on the knowledge in the form of patterns and trends extracted through a data mining process. Examples of such patterns include: likelihood of coincidence of particular diseases; outbreak of a communicable disease; adverse drug usage event; and diagnosis based on previous cases.

Widespread applications of data mining techniques in health care produce valuable knowledge that should be made available to users other than those who extracted the knowledge, in order to improve administrative or clinical

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE

decision making. It is usually the case that the knowledge extracted by data mining techniques is only accessible to the institutions that perform the studies and it rarely happens that the results become available to others in a form that can be easily integrated with their information systems.

In this paper, we study the interoperability of the data and knowledge within the distributed health care environment and will try to promote the decision support functionality of the health care systems by incorporating data mining models. In this context, the term *mined knowledge* refers to the knowledge extracted from stored patient data through some offline data mining techniques such as, association, classification, and clustering. We will describe a knowledge management framework that enhances dissemination of the mined knowledge to a usage point in a distributed health care environment. Also, since the extracted knowledge is meant to be shared among health care organizations we pay special attention to the interoperability issues that arise among heterogeneous systems. Hence, we establish our framework on standards and technologies that enable interoperability in the health care. We also explain the integration of this framework with the current state of the art decision support approaches based on clinical guidelines (GLIF) and provide examples of different types of mining methods.

The contributions of this paper are as follows: i) proposing a novel guideline-based decision-support approach using data mining techniques to extract health care knowledge from patient data; ii) incorporating knowledge interoperability with data interoperability in the distributed health care systems; iii) applying the proposed framework on three real clinical data mining cases from the literature.

The outline of the remaining sections of the paper follows. Section 2 provides a literature review on applications of data mining techniques in health care as well as the clinical decision support systems. Section 3 describes the distributed health care systems to provide enhanced decision making based on the mined knowledge. Section 4 presents the current state of the data and knowledge interoperability techniques including the discussion of relevant data mining techniques and clinical guidelines. Section 5 describes our proposed knowledge management framework. Section 6 demonstrates three case studies of materializing the proposed approach. Finally, the paper concludes in section 7.

2 Related work

Because of multi-disciplinary nature of our approach, we have carried out a literature review in both data mining application in health care and Clinical Decision Support Systems (CDSS).

Numerous applications of data mining techniques over medical data are carried out by researchers. Churilov et al. [7] describe a clustering method using an optimization approach to extract risk grouping rules for prostate cancer patients. Ordonez et al. [20] propose a new algorithm to mine association rules in medical data with additional constraints on the extracted rules and applies the method for predicting heart disease. A decision tree based classification approach has been applied to mass spectral data to help diagnosis of ovarian cancer suspects [24]. Grzymala-Busse and Hippe [14] compare various data mining methods supporting diagnosis of melanoma skin cancer. While association rule classifiers has been applied to diagnose breast cancer using digital mammograms [25]; Land et al. use Neural Network based classification approach for the same purpose [11]. Li et al. [16] discuss the problem of mining risk patterns in medical data using statistical metrics in the context of an optimal rule discovery problem and apply the method to find patterns associated with an allergic event for ACE inhibitors. Mining associations is also applied over human sleep times series data [15]. Wilson et al. [21] discuss potential uses of data mining techniques in pharmacovigilance to detect adverse drug reactions. In our approach, we use the result of three data mining techniques in our case studies and will provide the means for interoperability of the results for different data mining techniques. However, generation of the mining results is out of the scope of our discussion and this paper.

Many Clinical Decision Support Systems have also been developed. CHICA [23] is a CDSS, developed to improve preventive paediatric primary care. It uses a knowledge-base of 290 *if-then* rules for guideline modeling. The rules are encoded in a procedural language called Arden Syntax [1] and are grouped into modules called Medical Logic Modules (MLMs). Dynamic forms are generated and tailored to patients' needs based on the MLMs. PRESGUID [22] is a decision support system that integrates clinical practice guidelines with a drug database and supports prescribing in a primary care settings. Clinical guidelines defined as decision trees are coded in XML format and the system provides recommendations through a web based interface. Evidence-Based Guidelines And Decision Support System (EGADSS) [9] is an open source standards-based tool that provides an extensible clinical decision support framework. EGADSS integrates clinical guidelines coded in separate modules with Electronic Medical Record (EMR) systems to provide alerts and reminders in primary care. Modules contain the decision making logic and access patients' data from the EMR system encoded in structured clinical documents (CDA). To enable interoperability with heterogeneous health care systems, each data item in the CDA document should be mapped to the corresponding data items at the deployment site. In contrast to the above approaches, our approach is based on mined knowledge and uses a clinical guideline standard in conjunction with a log-

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE

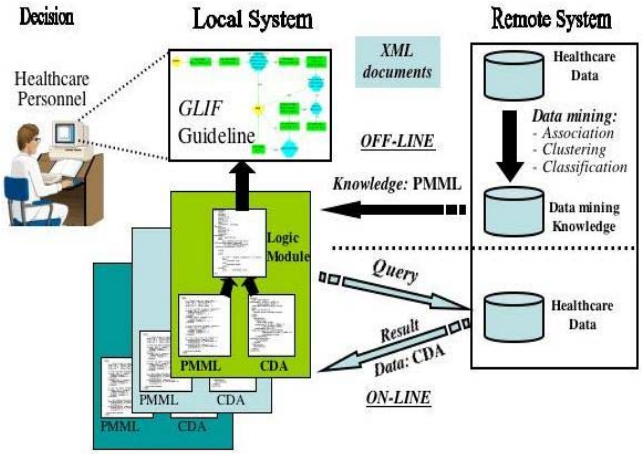


Figure 1. Decision support system in distributed health care environment.

ical module that incorporate data from CDA documents and knowledge from PMML documents (to be discussed later). Therefore, both data and knowledge would be employed to provide an enhanced decision making environment.

3 Proposed Health care environment

Figure 1 illustrates the operating environment for the proposed data and knowledge interoperability framework within the context of a guideline-based decision support system in a distributed health care setting. The steps for conducting a scenario (health care event) are described below. The system provides a form-based user interface for the health care personnel (physician, nurse, pharmacist, medical researcher) according to a GLIF guideline for a specific clinical encounter. The GLIF guidelines are basically flow-charts that provide decision making support at the point of care. The flow of actions in the guideline is mainly controlled at decision steps which consult with the knowledge-base to make appropriate decisions. The steps consist of operations, such as: acquisition of patient information; actions according to the patient information; sub-decision making; and making a final decision that concludes the guideline. The accuracy and effectiveness of the GLIF guided decisions are highly dependent on the accuracy of the information (such as patient's clinical history and the test results) and the degree of user's knowledge to direct the flow of information to a final decision about the case at hand. In this context, any additional knowledge that enhances the effectiveness of the decision making process would be highly valuable. In a distributed health care system the patient data is usually stored at a remote site. In this environment, the data required in a guideline step is queried and retrieved *on-line* from health care databases and the result of the query is encoded into an XML-based document as a CDA (Clinical Document Architecture) document to be used in the guideline. The proposed environment takes advantage of the data mining operations to explore the non-trivial relations, patterns, and trends in the health care data. Since data mining operations are known to be time consuming and require dedicated hardware and software, usually they are not handled locally. Instead, a specialized institution is equipped to perform all the data mining tasks on the health care data in an *off-line* operation. The relevant data mining operations include: association rules, clustering, and classification.

The off-line generated data mining results are shared among heterogeneous institutions; hence the interoperability of the data mining results (as the extracted knowledge) is crucial. We adopt a standard XML-based data model, namely PMML (Predictive Model Markup Language) to achieve the interoperability of the mined knowledge. PMML documents carry the relevant concepts and results of the different data mining techniques to be used for educated clinical or administrative decision making. In Figure 1 each Logic Module along with its corresponding CDA (for health care data) and PMML (for health care knowledge) is used to provide the required logic for a GLIF guideline.

4 Data and knowledge interoperability

The heterogeneous nature of the health care systems has been the focus of a lot of research activities. In this paper, we address both data and knowledge interoperability; however the knowledge interoperability is the main focus of the paper. In the following subsections, these two complementary aspects will be discussed in details.

4.1 Data interoperability

The underlying technology for data interoperability consists of 4 layers: i) at the lowest layer a communication channel is shared between acting information systems; ii) on top of the communication channel, services are defined that provide basic data interchange functionality, such as secure or reliable data transmission; these two layers lay the foundation for data exchange between information systems; iii) for information systems to actually be able to understand each other, a common syntax must be established to which all the messages that flow between them should conform to; this is referred to as syntax interoperability which can be considered as a third layer located on top of the service layer; and finally, iv) a common semantic framework is necessary to interpret every message with the right meaning.

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE

An established common vocabulary and a unified representation are required elements for syntax interoperability. For example, one party in the communication might be talking in English while others know French. Also, it might encode the messages in plain text while others use XML encoding.

Semantic interoperability, on the other hand, needs more delicate attention since it is by nature harder to achieve. It is not usually enough to use the same words, but also to interpret the words with the right meaning. Here is an example; a doctor working at his clinic might need to access his patient's test results on the laboratory database. On the other hand he might also be interested in comparing this document to a normal test document stored on the hospital database. While these two pieces of data can both be referred to as laboratory report that have the same data elements and an identical structure, they represent totally different things and the system should be able to distinguish these two.

The HL7 data modeling methodology with its core information model and the associated vocabularies and data types provide the syntax and semantic interoperability among heterogeneous health care information systems. Based on this methodology data is encoded from a proprietary data model into HL7 messages (or vice versa) as it crosses the organizations' boundaries.

Following an Object-Oriented approach, the standard defines a Reference Information Model (RIM) in its core which models health care data domain and provides a unified view of the health care domain. To define a message a subset of the RIM classes relevant to the message context is selected (Message Information Model), refined into R-MIM (refined MIM) model and flattened by tracing the diagram from a root class. Then using different vocabularies for different parts; LOINC terminology for observations; SNOMED for procedures and UMLS for medication treatments, the messages are encoded in an Implementation Specific Technology (IST), e.g. XML, and transmitted between information systems.

The Clinical Document Architecture (CDA) is another health care standard that enables semantically interoperable exchange of clinical documents such as history information, discharge summary, and progress note. CDA is a document markup standard that can be used to define the structure of XML-based clinical documents through specifying an XML schema that is assigned a unique identifier for different types of documents. CDA documents may be stored in databases [17], and transmitted as part of an HL7 message or any other messaging framework.

4.2 Knowledge interoperability

A definition for decision support systems for health care has been proposed in the literature [5] as "*systems that provide access to knowledge which is stored electronically to aid patients, carers and service providers in making decisions on health care*". These systems aim to disseminate clinical knowledge in the right time, for the right person and in the right way [10], to improve quality of care. Below, we describe the enabling techniques for sharing clinical knowledge and data mining results.

Clinical guidelines

We are interested in approaches that try to capture the clinical knowledge in the form of best practice computer-readable clinical guidelines. One such approach, Arden Syntax [1] defines individual decision making modules that code the decision criteria for a specific situation. Medical knowledge is encoded in a set of discrete modules called Medical Logic Modules (MLMs) that enable the exchange of knowledge in many institutions. When the rules conclude as $true$ then an action specified within the module is executed which is normally in the form of a reminder or an alert. To realize interoperability, each module defines its own data slot used to read data items into a set of internal variables. This section of the module is the only part that is affected when the module is deployed in different health care settings. The other important part of an MLM is the logic slot, which contains the decision making rules in the form of *if-then* statements. These rules use the variables initialized in the data slot. An action slot defines the actions that should be triggered when certain rules are satisfied.

Guideline Interchange Format (GLIF) [6] specifies another format for sharable representation of computer interpretable clinical guidelines in the form of flow diagrams. The main objectives in design of GLIF include precision, non-ambiguity, human-readability, computability, and platform independence. GLIF defines three abstraction levels as follows. In the first level, a conceptual flowchart model that is easy to write and comprehend represents the flow of events, states, and actions. Nodes in the flowchart represent *Decision Steps*, *Branch Steps*, *Synchronization Steps* and *Patient State Steps*. In the second level, namely computable level, the details of decision criteria, relevant patient data, and iteration information are provided to make the guidelines computable; in this level, the guidelines reference standard vocabularies and standard medical data models that are to be institution independent. In the third level, the implementation details and necessary mappings into a specific institution's information system are handled.

Data Mining

Data mining aims at building and fitting a *model* to the data.

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE

This model can be used to describe the data or to do predictions on new cases [4]. Standards have been developed to make the models that are built by one party to be available and understandable by other parties. Predictive Model Markup Language (PMML) provides a way to share and exchange statistical and data mining models where users can develop a model with one vendor's tool and then visualize, analyze, evaluate, and use the model using other applications. The specification comes in the form of an XML DTD or XML schema, and the current version 3.0 supports a wide range of models including association rules, clustering, naive bayes, neural network, regression, sequences, and trees.

According to the specification [8], in each PMML document a *data dictionary* defines the data attributes that are used in the document and a *transformation dictionary* specifies the required data transformations to be performed prior to the application of the model. Each document may contain one or more data mining models. As an example, an association rules mining task and its results can be specified in a PMML document by the following information: data attributes and their types and value ranges (numeric, categorical), data transformations; mining specific parameter (minimum support and minimum confidence); items; itemsets; and finally the rules that were discovered along with the corresponding supports and confidences.

5 Proposed framework

Figure 2 illustrates the overall view of the proposed distributed knowledge management framework. The framework consists of three phases, as: preparation, interoperation, and interpretation. The description of each phase follows.

5.1 Phase 1 - knowledge preparation

In this phase, the data mining knowledge is extracted from health care data in an off-line operation. For this purpose data is mined and a data mining model is fit to the data. This model might describe the data or be used to carry out future predictions on new data. Examples of such applications are: classifying a disease based on its symptoms to help diagnosis; clustering the patients based on relevant risk factors; verifying known medical facts; and expressing useful hidden patterns in data as in association rules mining.

This phase starts by removing the health care data attributes that can identify a patient or reveal their private data[1]. After anonymization the knowledge extraction process begins which is a complex and time consuming activity

[1]Based on the definition of privacy, some studies [18] have also shown that the privacy breaches can occur even when the data is anonymized. The investigation of the privacy breaches is not in the scope of this paper.

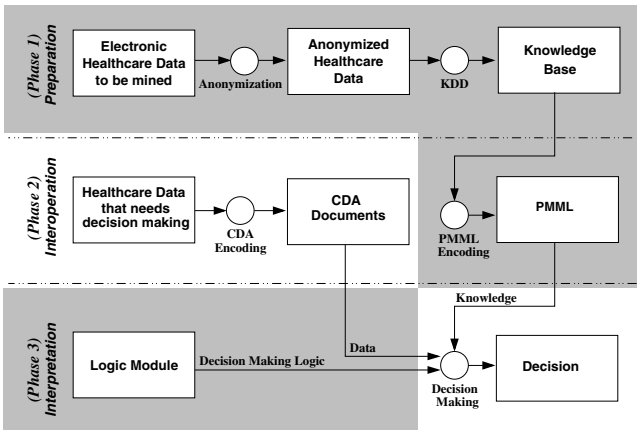


Figure 2. Health care Knowledge management framework. The shaded areas designate off-line parts.

that has many steps and iterations. The discovery starts by data selection, data cleaning, and data transformation which are followed by the actual data mining operation. Finally, the results are assessed in terms of usefulness, validity, and understandability [13]. In this paper we are interested in encoding the results of the different data mining techniques not the extraction process.

5.2 Phase 2 - knowledge interoperation

In this phase, two separate flows of data and knowledge are properly encoded to be used at the point of care. This phase ensures the interoperability among the institutions with different data and knowledge representations. In an off-line operation, the extracted knowledge in phase 1 should be ported to the parties that will use it for decision making. This is performed by employing PMML specification to encode the mined results into XML based documents. The XML schema for each data mining result describes the input data items, data mining algorithm specific parameters, and the final mining results. In an on-line operation, the subject data (i.e., health care data that needs decision making) in a source institution's internal data representation (e.g., EMR systems) is encoded into a CDA document to be interpreted for decision making in a destination institution. The encoded PMML knowledge can also be stored and used locally by health care institutions.

The PMML and CDA documents provide the interoperability of knowledge and data in our framework in the sense that the Decision Support System (DSS) will be independent to the proprietary data format of the involved institutions.

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE

5.3 Phase 3 - knowledge interpretation

In this phase, a final decision is made based on the results of applying the mined models to the subject data. The logic of decision making is programmed into the logic modules that access, query, and interpret the data and knowledge that flow from the previous phase. The final decision might be to issue an alert or to remind a fact.

For the mined knowledge to be actually used at the point of care, the data mining models should be interpreted for the subject cases (patient data). Three different documents are involved in this phase. The first document is the CDA document from phase 2 that contains a case data for a patient (e.g., a particular laboratory report) and is accessed on-line; the second document is the PMML document, containing the knowledge extracted in an off-line data mining process in phase 1 that was made portable by proper encoding in phase 2; and the third document is a program (logic module) that contains the necessary logic to interpret the data and knowledge.

The logic modules are independent units, each responsible for making a single decision based on the facts extracted in phase 1. In principle, they are similar to the idea of Arden Syntax Medical Logic Modules (MLM) with the exception that they can access and query mined knowledge bases. Each logic module contains the core decision making operation for a specific application and is bound to a specified data mining model in a PMML file. The overall structure of a logic module is described below.

The decision making is carried out in 3 main steps, retrieving the right data fields from the data source; applying the mined models to the data; and eventually taking an action or a set of actions. To do this, first the local variables in each logic module are populated by accessing the corresponding data fields in the CDA document. Before the model is applied to the data that was read, the required transformations are performed on the data. These transformations are specified in the transformation dictionary section of the PMML document. Based on the results of this application, the module takes an action. For example, if the module was invoked at a *Decision Step* in a guideline, it may branch to a specific path; or it may simply display the results in the form of a reminder or an alert.

6 Case studies

In this section, we describe three different data mining techniques along with a real-world clinical application for each case and explain how these techniques fit in our proposed framework.

6.1 Classification

A classification algorithm (e.g., neural network or decision tree) assigns a class to a group of data records having specific attributes.The classification techniques in health care can be applied for diagnostic purposes or recommendation for treatments. Suppose that certain symptoms or laboratory measurements are known to have a relation with a specific disease. A classification model is built that receives a set of relevant attribute-values, such as clinical observations or measurements, and outputs the class to which the data record belongs. As an example, a class can identify "whether a patient has been diagnosed with a specific disease". Such a classification model is built at phase 1 in the framework on a possibly large dataset by a different party; the classification model is then encoded into a PMML document in phase 2 to be shared and exchanged with the other health care institutions who will eventually use the model; and in phase 3, the model will be applied to a patient's medical case to yield a result.

To incorporate classifications into the GLIF guidelines, we extend the decision step nodes in the guideline flowchart. In these nodes, the classification model is run over the subject data; and based on the classification results the guideline flows to a particular path. As described in phase 3, a logic module is defined that takes care of accessing data from CDA documents, running the model and taking the right action.

We chose a decision tree classifier that was built to classify historical data to diagnose melanoma skin cancer [2]. The inputs to the classifier (i.e., TDS and C-BLUE) are of Numerical, Categorical, and Boolean types. Figure 3.a illustrates the decision tree classifier. Figure 3.c, shows part of the PMML encoding for the same decision tree and Figure 3.b illustrates the guideline that uses this classifier.

6.2 Association rules

An association rule $X \Rightarrow Y$ is defined over a set of transactions T where X and Y are sets of items. In a health care setting, the set T can be the patients' clinical records and items can be symptoms, measurements, observations, or diagnosis. Given S as a set of items, $support(S)$ is defined as the number of transactions in T that contain all members of set S. The *confidence* of a rule is defined as $support(X \cup Y)/support(X)$ and the support of the rule itself, is $support(X \cup Y)$. The discovered association rules can show hidden patterns in the dataset that was mined. For example, the rule:

$\{$*Patient has smoking habit*$\}$
$\Rightarrow \{$*Patient has coronary heart disease*$\}$

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE

Association Rule	Support	Confidence
$SeptoAnterior \Rightarrow (LAD \geq 50\%)$	18%	80%
$InferoSeptal \Rightarrow (RCA \geq 50\%)$	12%	65%
$InferoLateral \Rightarrow (LCX \geq 50\%)$	20%	53%

Table 1. Most significant discovered association rules.

with a high confidence; might signify a cause-effect relationship between smoking and diagnosis of heart disease. Although, this specific rule is a known fact that is expected to be valid, there are potentially many more rules that are not known or documented.

To make this type of knowledge available at the point of care to the practitioners, we consult with the knowledge-base to retrieve the rules that are applicable to the current patient's data. The valid rules can then be displayed to the user with their corresponding support and confidence factors. In order to consult with the knowledge-base, certain data items that are relevant to the current guideline flow should be retrieved, (e.g., whether the patient smokes, when he complains from the chest pain) and the data should be applied to the set of association rules to find the relevant rules. Again the process of data access and knowledge-base consultation is encoded in a logical module with the same overall structure that was described in the case of the classifier models.

The case study that we present in this section attempts to discover rules that associate risk factors and perfusion measurements to disease measurements [20]. Three of the most significant rules are shown in Table 1. The mining algorithm used is a variation of Apriori and the mining parameters, i.e., support and confidence are selected appropriately. The left hand side of the rules are perfusion measurements on specific regions of the heart; while the right hand side corresponds to the heart disease measurements. The detailed meaning of these attributes is out of the context of this paper.

Figure 4.b illustrates part of a clinical guideline that branches to consult the association rules in the knowledge-base which will eventually yield applicable rules. The rules can then be displayed to the user. Figure 4.a illustrates part of the PMML document that contains the association rules shown in Table 1. The data attributes illustrated in Figure 4.a are derived data fields that were subject to certain transformations. These transformations are not shown in the figure. It is worth mentioning that in some situations, numerous rules might apply. These rules should be further constrained based on an appropriate measure. A simple solution is to filter out some rules and just leave the ones with higher confidence measure.

6.3 Clustering

The last group of data mining techniques that we describe in this section is clustering. To demonstrate our approach, we used the results of applying a clustering method on prostate cancer patients [7]. The data record fields are the patient's age, tumour stage, Gleason score, and PSA level (in this paper the medical meaning of these fields are not of our interest). The clustering algorithm generates 10 clusters. Figure 5.a illustrates parts of the PMML encoding for the results. As the original paper claims, the clusters can further be used to classify patients into risk groups of low, intermediate, and high risks based on appropriate metrics. Based on this method, clusters 10 and 5 belong to low risk groups, whereas clusters 1, 3, 4 and 6 include high risk patients and the other clusters represent intermediate risk groups. Figure 5.b illustrates a decision step in a clinical guideline that takes different paths after applying the data mining encoded PMML model to the patient data to identify the cluster and hence the risk group associated with the patient.

7 Conclusion

In this paper, we proposed a novel knowledge management framework for distributed health care systems that incorporate the knowledge extracted by data mining techniques into health care information systems for decision making. We also described a health care environment at the point of care that takes advantage the knowledge provided by our framework within the context of clinical guidelines to improve clinical decision making. We note that the framework can be potentially used for administrative decision making purposes as well. At the end, three examples demonstrating different data mining techniques (i.e., decision tree based classification; association rules mining; and clustering) that were applied to clinical data in the real-world research studies were selected and using our proposed framework we showed how they can be adopted by guideline-based decision support systems. We demonstrated how a variety of standards (CDA, PMML), guidelines (GLIF), and algorithms (data mining techniques above) can be integrated into a unified framework to work seamlessly towards an enhanced health care decision making environment. The proposed framework is a part of reference architecture for health care systems that the authors are involved in. The proposed approach is an on-going research work and a prototype system is being implemented.

References

[1] The Arden Syntax for Medical Logic Systems, URL = http://cslxinfmtcs.csmc.edu/hl7/arden/.

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE

[2] Rules for melanoma skin cancer diagnosis, URL = http://www.phys.uni.torun.pl/publications/kmk/.

[3] Canadian institute for health informatics (cihi) project page. URL = http://www.cihi.ca/.

[4] Prudsys AG. Xelopes library documentation - version 1.3.1. URL = http://www.prudsys.com/Service/Downloads/bin/1133983554/Xelopes1.3.1_Intro.pdf.

[5] Electronic decision support for australia's health sector, report to health ministers by the national electronic decision support taskforce. URL = http://www.ahic.org.au/downloads/nedsrept.pdf, November 2002.

[6] Guideline interchange format 3.5 - technical specification. URL = http://smi-web.stanford.edu/projects/intermed-web/guidelines/GLIF_TECH_SPEC_May_4_2004.pdf, May 2004.

[7] Leonid Churilov, Adil M. Bagirov, D. Schwartz, Kate A. Smith, and M. Dally. Improving risk grouping rules for prostate cancer patients with optimization. In *Hawaii International Conference on System Sciences (HICSS)*, 2004.

[8] Data Management Group (DMG). Predictive model markup language (pmml) version 3.0 specification. URL = http://www.dmg.org/pmml-v3-0.html.

[9] EGADSS.org. Evidence-based guidelines and decision support system (egads) project page. URL = http://egads.org.

[10] J. Eisenberg and E. Power. Transforming insurance coverage into quality health care: voltage drops from potential to delivered quality. *Journal of the American Medical Association*, 284:2100–2107, 2000.

[11] W.H.Jr Land et al. New results in breast cancer classification obtained from an evolutionary computation/adaptive boosting hybrid using mammogram and history data. In *Proceedings of the 2001 IEEE Mountain Workshop on Soft Computing in Industrial Applications*, pages 47–52, 2001.

[12] Cynthia M. Farquhar, Emma W. Kofa, and Jean R. Slutsky. Clinicians' attitudes to clinical practice guidelines: a systematic review. *Medical Journal of Australia*, 2002 177 (9): 502-506.

[13] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[14] Jerzy W. Grzymala-Busse and Zdzislaw S. Hippe. Data mining methods supporting diagnosis of melanoma. In *18th IEEE Symposium on Computer-Based Medical Systems (CBMS 2005)*, pages 371–373, 2005.

[15] Parameshvyas Laxminarayan, Carolina Ruiz, Sergio A. Alvarez, and Majaz Moonis. Mining associations over human sleep time series. In *18th IEEE Symposium on Computer-Based Medical Systems (CBMS 2005)*, pages 323–328, 2005.

[16] Jiuyong Li, Ada Wai-Chee Fu, Hongxing He, Jie Chen, Huidong Jin, Damien McAullay, Graham Williams, Ross Sparks, and Chris Kelman. Mining risk patterns in medical data. In Robert Grossman, Roberto Bayardo, and Kristin P. Bennett, editors, *KDD*, pages 770–775. ACM, 2005.

[17] Zheng Liang, Peter Bodorik, and Michael Shepher. Storage model for cda documents. In *Hawaii International Conference on System Sciences (HICSS)*, page 159, 2003.

[18] Taneli Mielikäinen. On inverse frequent set mining. In Wenliang Du and Christopher W. Clifton, editors, *Proceedings of the 2nd Workshop on Privacy Preserving Data Mining (PPDM), November 19, 2003, Melbourne, Florida, USA*, pages 18–23. IEEE Computer Society, 2003.

[19] Ontario Ministry of Finance. The right choices: Investing in health care. URL = http://www.fin.gov.on.ca/english/budget/bud03/budhi1.html, March 2003.

[20] Carlos Ordonez, Cesar A. Santana, and Levien de Braal. Discovering interesting association rules in medical data. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 78–85, 2000.

[21] Br. J. Clin Pharmacol. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology*, 57(2):119–20, Feb 2004.

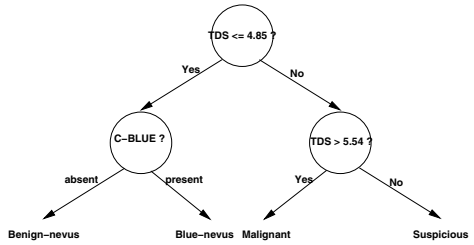
[22] Prescription and guidelines (presguid) project page. URL = http://cybertim.timone.univ-mrs.fr/CybErtim/LERTIM/Recherche/PRESGUID/contenu.htm.

[23] Anand V., Biondich PG., Liu G., and Rosenman M.and Downs SM. Child health improvement through computer automation: The chica system. *Medinfo*, 2004;2004:187-91.

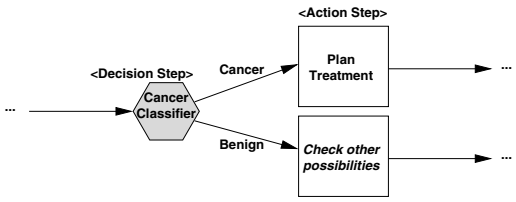
[24] Antonia Vlahou, John O. Schorge, Betsy W. Gregory, and Robert L. Coleman. Diagnosis of ovarian cancer using decision tree classification of mass spectral data. *Journal of Biomedicine and Biotechnology*, 5:308–314, 2003.

[25] Osmar R. Zaïane, Maria-Luiza Antonie, and Alexandru Coman. Mammography classification by an association rule-based classifier. In *MDM/KDD*, pages 62–69, 2002.

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE



Part a. The decision tree for diagnosis of Melanoma skin cancer.



Part b. A decision step in a clinical guideline that takes different paths based on the case classification results.

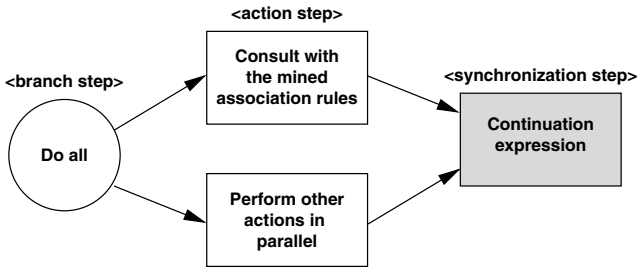
```
<PMML version="3.0" >
<Header copyright="NA" description="A decision tree encoded in PMML
and used to classify Melanoma skin cancer"/>
<DataDictionary numberOfFields="2" >
    <DataField name="TDS" optype="continuous"/>
    <DataField name="C−BLUE" optype="categorical" >
<Value value="absent"/>
<Value value="present"/>
    </DataField>
    <DataField name="diagnosis" optype="categorical" >
<Value value="Benign−nevus"/>
<Value value="Malignant"/>
<Value value="Suspicious"/>
    </DataField>
</DataDictionary>

<TreeModel modelName="MelanomaSkinCancer" functionName="classification">
    <MiningSchema>
<MiningField name="TDS"/>
<MiningField name="C−BLUE"/>
    <MiningField name="whatIdo" usageType="predicted"/>
    </MiningSchema>
    <Node score="Benign−nevus">
     <True/>
     <Node score="Benign−nevus">
      <SimplePredicate field="TDS" operator="lessThanOrEqual" value="4.85"/>
      <Node score="Benign−nevus">
         <SimplePredicate field="C−BLUE" operator="equal" value="absent" />
      </Node>
      <Node score="Blue−nevus" >
         <SimplePredicate field="C−BLUE" operator="equal" value="present" />
      </Node>
     </Node>

     <Node score="Benign−nevus">
      <SimplePredicate field="TDS" operator="greaterThan" value="4.85"/>
      <Node score="Malignant">
         <SimplePredicate field="TDS" operator="greaterThan" value="5.54" />
      </Node>
      <Node score="Suspicious" >
         <SimplePredicate field="TDS" operator="lessThanOrEqual" value="5.54" />
      </Node>
     </Node>

    </Node>
  </TreeModel>
</PMML>
```

Part c. The PMML representation of the Melanom skin cancer classifier.

Figure 3. The classification example.

```
<PMML version="3.0">
(header)
(data dictionary)
(transformation dictionary)

<AssociationModel ... ...

<Item id="1" value="SeptoAnterior"/>
<Item id="2" value="InferoSeptal"/>
<Item id="3" value="InferoLateral"/>
<Item id="4" value="LAD_greater_than_50_percent"/>
<Item id="5" value="RCA_greater_than_50_percent"/>
<Item id="6" value="LCX_greater_than_50_percent"/>

<Itemset id="1" numberOfItems="1" support="..."> ...
<ItemRef itemRef="1"/> </Itemset>
<Itemset id="2" numberOfItems="1" support="..."> ...
<ItemRef itemRef="2"/> </Itemset>
<Itemset id="3" numberOfItems="1" support="..."> ...
<ItemRef itemRef="3"/></Itemset>
<Itemset id="4" numberOfItems="1" support="..."> ...
<ItemRef itemRef="4"/></Itemset>
<Itemset id="5" numberOfItems="1" support="..."> ...
<ItemRef itemRef="5"/></Itemset>
<Itemset id="6" numberOfItems="1" support="..."> ...
<ItemRef itemRef="6"/></Itemset>

<AssociationRule id="1" support="0.18" confidence="0.80"
antecedent="1" consequent="4"/>
<AssociationRule id="2" support="0.12" confidence="0.65"
antecedent="2" consequent="5"/>
<AssociationRule id="3" support="0.20" confidence="0.53"
antecedent="3" consequent="6"/>

</AssociationModel>
</PMML>
```

Part a. Part of the PMML encoding representing the association rules for predicting patients heart disease



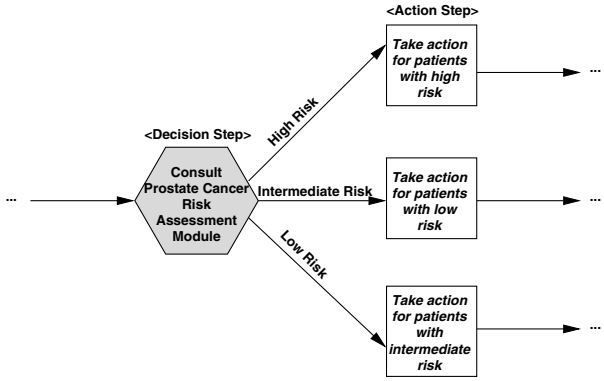
Part b. An action step in a clinical guideline that consults with the knowledge base containing association rules.

Figure 4. The association rule mining example.

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE

```
<PMML version="3.0">
  <Header copyright=""/>
  <DataDictionary numberOfFields="4">
    <DataField name="Gleason score" optype="categorical">
      <Value value="1"/> ... <Value value="10"/>
    </DataField>
    <DataField name="PSA" optype="continues"/>
    <DataField name="Age" optype="numerical"/>
    <DataField name="Tumor stage" optype="categorical">
      <Value value="1a"/>...<Value value="4"/>
    </DataField>
  </DataDictionary>

  <ClusteringModel modelName="Prostate Cancer Clustering"
      functionName="clustering"   modelClass="centerBased"
      numberOfClusters="10">
   <MiningSchema>
        <MiningField name="Tumor stage"/>
        <MiningField name="Gleason score"/>
        <MiningField name="PSA"/>
        <MiningField name="Age"/>
   </MiningSchema>

   <ComparisonMeasure kind="distance">  <squaredEuclidean/>
   </ComparisonMeasure>

   <ClusteringField field="Tumor stage" compareFunction="absDiff"/>
   <ClusteringField field="Gleason score" compareFunction="absDiff"/>
   <ClusteringField field="PSA" compareFunction="absDiff"/>
   <ClusteringField field="Age" compareFunction="absDiff"/>

   <CenterFields>
   (derived and transformed Tumor stage)
   (derived and transformed Gleason score)
   (derived and transformed PSA)
   (derived and transformed Age)
   </CenterFields>
   ...
   <Cluster name="cluster1">
    <Array n="4" type="real"> 7.2 7 9.3 68</Array> </Cluster>
   ...(other clusters)
   <Cluster name="cluster10">
    <Array n="4" type="real">3.2 5.7 15 67</Array> </Cluster>
  </ClusteringModel>
</PMML>
```

Part a. Part of a PMML document representing the clusters for prostate cancer patients.



Part b. A decision step in a clinical guideline that takes a differe path based on the risk assessment of the patient.

Figure 5. Clustering example.

Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)
0-7695-2639-X/05 $20.00 © 2005 IEEE