# A Framework for Data and Mined Knowledge Interoperability in Clinical Decision Support Systems

Reza Sherafat Kazemzadeh, Kamran Sartipi and Priya Jayaratna

*Department of Computing and Software*
*McMaster University, Hamilton, ON, L8S-4K1, Canada*
*Email: {sherafr, sartipi, jayarasp}@mcmaster.ca*

**SUMMARY**

**Due to reliance on human knowledge the practice of medicine is subject to errors that endanger patients' health and cause substantial financial loss to both public and governmental health sectors. The computer-based decision making systems assist healthcare personnel to improve the quality of clinical practice. Currently, the decision making knowledge within most guideline modeling languages are represented by basic logical expressions. On the other hand, the results of data mining analysis on healthcare data can be employed as the source of knowledge to improve decision making. In this paper, we focus on encoding, sharing, and using the results of data mining analyses for clinical decision making at the point of care. For this purpose, a knowledge management framework is proposed that addresses the issues of data and knowledge interoperability by adopting healthcare and data mining modeling standards, HL7 and PMML respectively. In a further step data mining results are incorporated into guideline-based Clinical Decision Support Systems. A prototype tool has been developed as part of this research that provides an environment for clinical guideline authoring and execution capable of applying and interpreting data mining results. Also, three real world case studies have been presented, one of which is used as a running example throughout different phases of the proposed approach.**

KEY WORDS: Data Mining; Mined Knowledge; Clinical Decision Support System; Heterogeneous Healthcare Systems; Data Interoperability; Knowledge Interoperability; Clinical Guidelines; Guideline Interchange Format3; HL-7 v3; Clinical Document Architecture.

## 1. Introduction

Because of the paramount importance of the quality of public health services, these services consume a major portion of governmental spending in many countries and usually they are considered as significant measures of people's quality of life. In Canada the provincial

government of Ontario invested a total of $28.1 billion in healthcare services in 2003-2004 [31], and; the Canadian Institute for Health Information (CIHI) estimated that the total healthcare spending throughout Canada reached as high as $160 billion in 2007, up from $150.3 billion in 2006; this represents a forcasted annual increase of 6.6% [16]. However, the large volume of spending in healthcare does not necessarily translate to ideal and error-free health services. The inherent complexity and dynamic nature of the existing medical knowledge and overwhelming amount of diverse medical information adversely affect a practitioner's medical practice. According to HealthGrades' fifth annual Patient Safety in American Hospitals Study, [25] medical errors cost the federal Medicare program $8.8 billion and resulted in 238,337 potentially preventable deaths during 2004 through 2006 alone.

Clinical Decision Support Systems (CDSS) are computer applications that assist practitioners and healthcare providers in decision making through timely access to electronically stored medical knowledge [13] in order to improve the practitioners' medical practice. A CDSS interacts with practitioners and electronic medical records systems to receive the patient data as input and provides reminders, alerts, or recommendations for patient diagnosis, treatment, long-term care planning, and alike. A Clinical Decision Support System requires to access healthcare data and knowledge that are stored in data and knowledge bases. Since these repositories normally have diverse internal representations, data and knowledge interoperability are major issues. To achieve data interoperability two systems that participate in data communication should use the same vocabulary set, data model, and data interpretation mechanism. On the other hand, knowledge interoperability refers to the ability of healthcare information systems to incorporate and interpret the knowledge that is produced in other systems. *We are mainly interested in the knowledge that is generated by data mining algorithms and is represented using data mining specific data structures, called "data mining models".*

The most important issues in current healthcare industry that lead us to define the problem addressed in this paper are as follows: i) mining of healthcare data is a valuable source of knowledge; however the extracted knowledge is often used locally by the researchers in a proprietary system and is not made available to other interested users for easy and seamless integration and application [30]; ii) there is no specific approach or methodology for seamless integration of mined healthcare knowledge with the clinical decision making process, which is mainly due to the complexity of data mining results; iii) the healthcare industry suffers from the lack of standardization that is resulted from development of both ad-hoc systems and the inherited legacy systems; and iv) healthcare information systems are usually heterogeneous and are deployed in a distributed environment that necessitates careful handling of both data and knowledge for achieving interoperability. Based on the above observations we define the research problem in this paper as follows:

> *Devising methodologies, techniques, and tools to streamline the dissemination and application of data and mined knowledge for clinical decision making purposes in the heterogeneous clinical settings.*

However, in this paper we do not directly address the issues such as: patient data privacy, knowledge extraction from healthcare databases, and authenticity of the clinical best practice

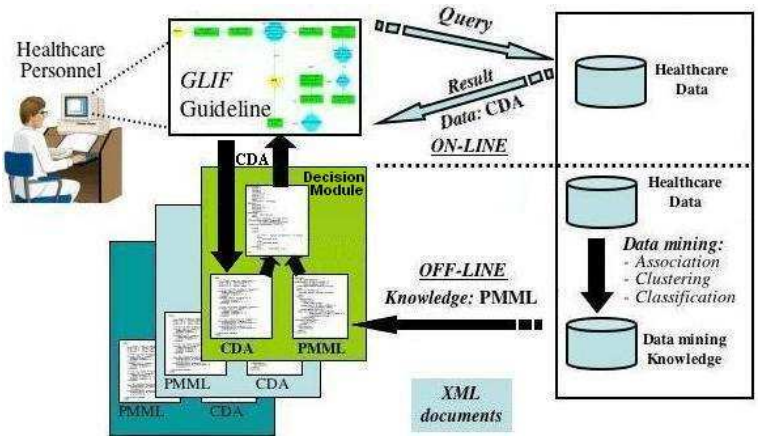


Figure 1. A guideline-based Clinical Decision Support System in a healthcare environment that allows data and knowledge interoperability among heterogeneous systems.

guidelines.

## Proposed approach

Figure 1 illustrates the proposed approach for clinical data and knowledge interoperability and interpretation within the context of a Clinical Decision Support System. The proposed approach relies on adoption of standards to encode healthcare data and knowledge. In an off-line operation, existing healthcare databases are mined using different mining techniques to extract and store clinical mined knowledge. In order to make this knowledge portable it is encoded as data mining models using a specialized XML-based standard, namely PMML (Predictive Model Markup Language) [19]. Also, the patient data that are stored in the Electrical Medical Record (EMR) systems are encoded using HL7 (Health Level 7)[2] version 3(v3) standard to be made portable between heterogeneous systems. At the point of care, a *decision module* accesses and operates on both data and knowledge in order to make patient-specific interpretations of the knowledge available to the healthcare practitioner. Within the CDSS we adopt a flow-oriented clinical guideline modeling language (GLIF3 [14]) to specify the overall decision making process. In this context, at different states of the flow-oriented guideline the CDSS accesses the patient data by making queries from the healthcare database. Moreover, to perform knowledge-based decision making, the CDSS supplies the patient data to the decision modules and receives the result of applying mined knowledge on the patient data. Finally, the healthcare personnel receive comments, recommendations, or alerts through

interaction with the CDSS system that allow them to make more knowledgeable decisions based on the system provided information.

The contributions of this paper are as follows: i) proposing a novel framework that supports data and (mined) knowledge dissemination and application in the context of Clinical Decision Support Systems; ii) extending clinical guideline modeling standard GLIF3 to use the HL7 and PMML standards respectively for data and knowledge interoperability; iii) developing a prototype tool for clinical guideline modeling and execution that supports integration of the data mining results and decision making based on the mined knowledge; and iv) applying the proposed approach on two real-world clinical data mining analysis from the literature.

The remainder of this paper is organized as follows: Section 2 provides an overview of the related work in the area of healthcare decision support systems research. Section 3 proposes a novel framework for data and knowledge interoperability that is composed of 4 phases which are elaborated in Sections 4 to 7. Section 4 addresses the knowledge extraction phase via data mining operations. Section 5 presents the standard-based data and knowledge interoperability approach used in this research. In Section 6, we elaborate on the application of the mined knowledge on the healthcare data at the point of use. Section 7 describes the proposed extensions to the GLIF3 flow-based guideline modeling language. Section 8 presents a guideline execution environment that is capable of interpreting the mined knowledge. Section 9 elaborates on the second real-world case study we used in this paper. Finally, Section 10 provides a short conclusion and sets the paths for future works.

## 2. Related work

In this section, we provide a short review of the related work on making healthcare data and clinical best practices available for use in a computer interpretable manner.

### GLIF3

Guideline Interchange Format 3 (GLIF3) [14] is a clinical guideline modeling language that represents the clinical best practices as flow charts. The GLIF3 specification was developed by the InterMed Collaboratory as a joint project of medical informatics laboratories at Harvard, Stanford, Columbia, and McGill universities. GLIF3 guidelines are modeled by expert medical researchers according to its specification and they are executed in the Clinical Decision Support System to provide decision making support and clinical best practice how-to for the healthcare professionals*.

Flow charts are defined in the *conceptual level* of GLIF3 that is meant to provide easy to encode and comprehend representation of the medical best practice. The control flow details including decision criteria, patient related data, and medical concepts are specified in the *computable level* which is integrated in the conceptual flow chart. At this level, GLIF3

*GLIF3 guidelines have been developed for a variety of purposes, including but not limited to heart failure, hypertension, thyroid screening, and many more.

uses simple logical predicates and conditional statements to direct the guideline flow. For GLIF3 guidelines to be eventually deployed at a healthcare institution, the patient data should be mapped from the institution-specific information systems to the guideline's internal data variables. In Section 7.1, we will describe extensions to the GLIF3 conceptual and computable levels in order to incorporate the data mining generated knowledge.

**COMPETE III**
Computerization Of Medical Practices for the Enhancement of Therapeutic Effectiveness (COMPETE) [11] is a Canadian project that intends to bring computer-based decision support facility for managing diabetes, hypertension, cholesterol, previous heart stroke, and chronic disease patients. The knowledge-base of COMPETE III consists of a set of guidelines each represented using a function table, where each row in the table specifies conditional statements in the form of logical expressions, and rows represent possible situations that can trigger an action, i.e., output a message. The rules encode output in a coloring scheme which is used to convey particular meanings to the user. Simplicity and understandability of the tables are considered as the strength of this approach, however the scope of knowledge that can be represented and managed in this approach is limited.

***e*-MS**
Electronic Medical Summary (*e*-MS ) [20] is not a Clinical Decision Support System, instead it addresses the data interoperability problem in healthcare. *e*-MS project intends to make subsets of patient data stored in one healthcare institution available to other stakeholders interested to access them. The *e*-MS project defines an XML-based Electronic Medical Summary (*e*-MS ) document format, and an Electronic-Medical Summary Exchange Protocol (e-MSEP). The documents encode different pieces of patient data into a structured textual format based on the HL7 CDA [24]. The document structure and semantics associated with different data items in the CDA structure are used to interpret the meaning of the whole document.

## 3. Framework for interoperability of data and mined knowledge

Figure 2 illustrates the overall view of the proposed framework for interoperability of data and knowledge[†] in a heterogeneous healthcare environment, where the shaded area illustrates the contributions of this paper. The framework consists of four phases *Preparation, Interoperation, Application*, and *Interpretation* as briefly described below.

*Phase 1 – Preparation.* Data is collected, pre-processed and mined to extract patterns and trends. The resulting mined-knowledge is stored in a knowledge repository to be incorporated in the CDSS process in the subsequent phases.

*Phase 2 – Interoperation.* Patient data and mined clinical knowledge are encoded into HL7 and PMML standards respectively, in order to be supplied to the CDSS. This ensures

[†]In this paper, the term "knowledge" refers to "mined knowledge".

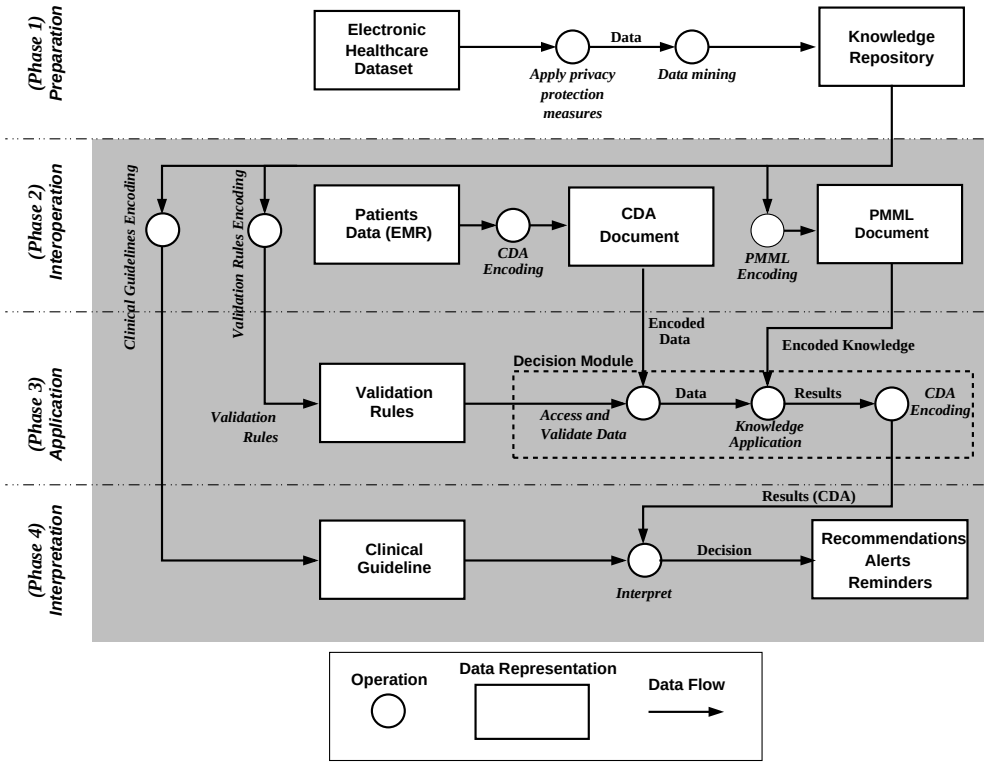


Figure 2. Healthcare framework for interoperability of data and mined knowledge. The shaded area designates the contribution of this paper.

interoperability amongst medical record repositories and knowledge repositories that support heterogeneous data formats and our CDSS.

*Phase 3 – Application.* This phase is performed at the point of care using a *decision module* (dashed line box) that processes the patient data and mined knowledge encoded in phase 2. The decision module parses and accesses the HL7 message containing patient data. The knowledge is then accessed from PMML documents and is applied on the patient data. The result of this application is encoded as an HL7 message to be used and interpreted by the CDSS.

*Phase 4 – Interpretation.* This phase is also performed at the point of care. The results of application of mined knowledge on patient data are accessed to influence the decision-making process of the CDSS.

| Activities | Performer |
|---|---|
| Definition of purpose | Healthcare researchers |
| Data collection | Healthcare researchers/staff |
| Data mining | Data mining and healthcare researchers |
| Data interoperability | Data modelers and managers |
| Knowledge interoperability | Computer scientists and data mining researchers |
| Clinical guidelines definition | Medical researchers |
| Clinical guidelines encoding | Healthcare informatics experts |

Table I. Different activities in the proposed framework and their corresponding performers.

The activities in the framework are carried out in a decentralized and collaborative healthcare environment in which different stakeholders (i.e., healthcare organizations and researchers) with diverse knowledge are involved. Hence a clear task separation and description of activities is required. Table I summarizes different tasks in the proposed framework and those who will perform them. In the remaining of this paper the phases of the framework are discussed in greater detail.

## 4. Knowledge preparation (phase 1)

In phase 1 of the framework (Figure 2) knowledge is extracted in an off-line process by application of different data mining techniques on a healthcare data collection. The extracted knowledge may represent interesting facts and trends in the data that can be used to describe different characteristics of the original dataset, or be used to carry out predictions on new cases in the future. Three major steps are carried out in this phase, i) data collection, ii) patient data protection, and iii) data mining operations. Figure 3 illustrates different steps that are carried out in the preparation phase. The description of these steps follows.

### 4.1. Data collection

The dataset that is used in the data mining operation is usually compiled from diverse sources (e.g., existing records in EMR systems, patient questionnaires, and surveys), and contains a set of data items relevant to the purpose of analysis (e.g., medical records of patients with specific conditions, selected clinical measurements, or the results of particular laboratory test). The collected data would then be anonymized as discussed below.

### 4.2. Patient data privacy

An extremely important concern in healthcare is to protect the patient's data privacy [35]. Privacy protection laws are currently in effect for this purpose in many countries including

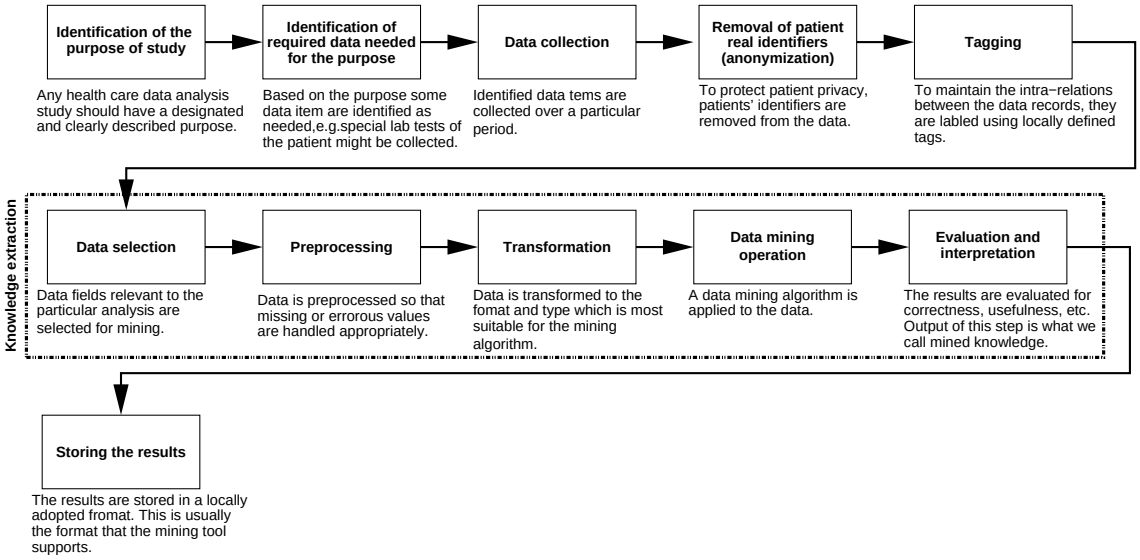


Figure 3. Different steps involved in phase 1 of the proposed framework (preparation).

Canada, United States, Europe, and Japan. They strictly regulate the collection, and use of the patients private data. Before any data collection takes place, the governmental and/or local healthcare authorities and the ethics board of the involved institutions must ratify the purpose of the data collection according to the laws that apply.

As a common privacy protection practice, patient identifiers such as name and Social Security Number are excluded from the collected dataset in an operation known as *anonymization*. However, in many cases a person may have several different records in the dataset and removal of identifiers results in loss of relationships between these data records. To reconnect such interrelated records, a fast solution is to use local identifiers to label the records prior to anonymization.

### 4.3. Knowledge extraction

The data mining step is carried out on the anonymized dataset and produces mined knowledge in the form of *data mining models*. This step contains several different activities including "data selection", "preprocessing", "transformation", "mining operation", "evaluation", and "interpretation" [23]. Data mining models are the data structures that represent the results of data mining operations. There are several categories of these operations, referred to as mining techniques. In the following subsections, we briefly describe the major techniques, (e.g., classification, clustering, and association-rules mining) where each category consists of a variety of algorithms.

| 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 II. Discovered association rules in mining of heart disease data [32].

In our framework, we don't differentiate between different implementations and algorithms of data mining techniques, as far as their results are representable by the general concepts of the corresponding data mining type. For instance, different association rules discovery algorithms take different approaches in extracting the frequent item sets and opt to choose different measures to exclude intermediary sets and hence prevent explosion in the results set. Some may refine the set based on standard constraints of support and confidence, others may apply additional constraints on the size of the rules' antecedent and consequent.

### 4.4. Data mining applications in healthcare

Healthcare data mining analysis produces valuable knowledge that can be used for decision making. Various types of mining models (e.g., clustering, classification, and association rules) can be used to represent interesting facts and hidden patterns and trends in clinical datasets with numerous applications in medical practice. In this subsection we briefly review some of the applications of data mining in healthcare from the literature.

Churilov et al. [18] describe a clustering method using an optimization approach to extract risk grouping rules for prostate cancer patients. The data record fields are the patient's age, tumor stage, Gleason score, and PSA level[‡]. The clustering algorithm generates 10 clusters that are then grouped to low, intermediate and high risk categories. Ordonez et al. [32] 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. Table II illustrates three association rules among all the generated rules that are of more importance.

A decision tree-based classification approach has been applied to mass spectral data to help diagnosis of ovarian cancer suspects [33]. While association rule classifiers have been applied to diagnose breast cancer using digital mammograms [37]; Land et al. use Neural Network based classification approach for the same purpose [26]. Li et al. [28] 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. Association rules mining is also applied over data of human sleep time [27]. Wilson et al. [36] discuss potential uses of data mining techniques in pharmacovigilance to detect adverse drug reactions. Duch et al. [9] compare various data mining methods supporting

[‡]In this paper the medical meaning of these fields are not of our interest.

diagnosis of Melanoma skin cancer. The last study mentioned above, serves as a running case study that is discussed throughout this paper. In the followings, we briefly discuss three data mining techniques.

**Classification models**
A classification algorithm assigns a patient's data record with specific attributes and attribute-values to a predefined class. The classification techniques in healthcare can be applied for diagnostic purposes. Suppose that certain symptoms or laboratory measurements are known to have a relation with a specific disease. A classification model is built using a set of relevant attribute-values (records) such as clinical observations or measurements that allows us to generate different classes representing different categories of records. Then, by comparison of a new patient's record with those of patients in different classes, we can determine to which class the new patient belongs, e.g., determining that "the patient has lung cancer".

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

$$\{People\ with\ a\ smoking\ habit\} \Rightarrow \{People\ having\ heart\ disease\}$$

with a high confidence signifies that: the number of people with heart disease is high among people with smoking habit.

**Clustering models**
Clustering is originated from mathematics, statistics, and numerical analysis [15]. In this technique the dataset is divided into groups (clusters) of similar records [15]. The clustering algorithms usually try to group records in a way to maximize a similarity metric between the members of the cluster. In many cases, closeness is the similarity metric and the aim is to maximize the cumulative closeness between data records in a cluster. The researchers then study the properties of the members of the generated clusters.

### 4.5. Running case study (knowledge extraction)

In this section, we describe a healthcare data mining analysis research from literature that we have chosen as our case study in this paper. This analysis is performed in order to study the relationship between the patients' skin marks and the possibility of presence of skin cancer, using a classification data mining technique [9]. The classifier is a decision tree that classifies patients according to their skin marks into four types: *Benign, Blue, Suspicious*, or *Malignant Melanoma*. The data have been collected in the Outpatient Center of Dermatology in Rzeszw, Poland containing 250 records.

| **Data item** | **Accepted values** |
|---|---|
| *Asymmetry* | Symmetric-spot = 0, 1-axial asymmetry = 1, 2-axial asymmetry = 2 |
| *Border* | Values from 0 to 8 |
| *Color* (Binary coded) | White, Blue, Black, Red, Light brown, Dark brown |
| *Diversity* (Binary coded) | Pigment globules, Pigment dots, Branched strikes, Structureless areas, Pigment network |
| $C - Blue$ | Absent = 0, Present = 1 |

Table III. Description of the different data items accessed by the decision tree classifier.

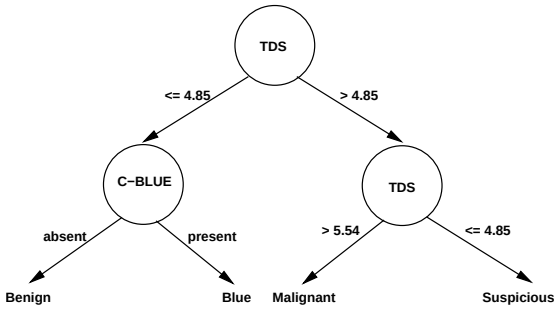


Figure 4. The decision tree classifier for Melanoma skin cancer [9].

The data selection for data mining contains five variables, indicating presence or absence of *C-Blue, asymmetry, border, color*, and *diversity* of the skin cancer's mark structure. All the variables except *C-Blue* are used to calculate an index, called *Total Dermatoscopy Score* (TDS) by the following formula:

$$TDS = 1.3 * Asymmetry + 0.1 * Border + 0.5 * \Sigma Colors + 0.5 * \Sigma Diversities \qquad (1)$$

The types of the variables have been casted to integer as indicated in Table III. The data mining operation has been carried out on the calculated TDS index and C-Blue variable to build the decision tree classifier that is illustrated in Figure 4. We will use this example decision tree in the following sections as a running case study to explain different activities in our framework in order to make this classifier available in Clinical Decision Support Systems.

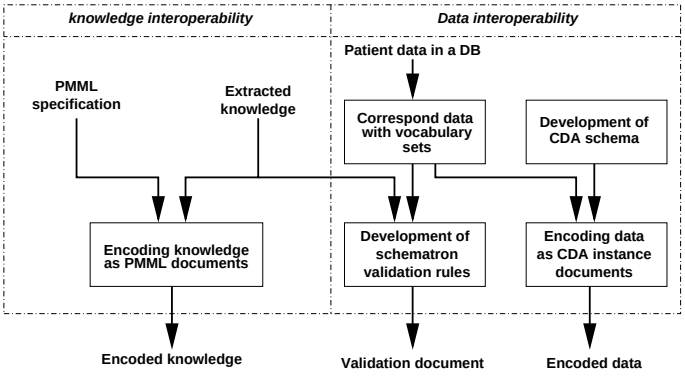


Figure 5. Different steps involved in phase 2 of the framework (interoperation).

## 5. Data and knowledge interoperation (phase 2)

Data and knowledge should be made available for access and use at heterogeneous healthcare institutions, and this phase ensures interoperability among such software systems.

### 5.1. Data interoperability

Data interoperability refers to the ability of information systems to interpret data from heterogeneous sources correctly. Our goal in this research is to make the CDSS interoperable with heterogeneous patient data sources so that multiple healthcare institutions can benefit from its use. Data interoperability consists of two different parts, *syntactic* data interoperability and *semantic* data interoperability.

*Syntactic* data interoperability is the ability of information systems to communicate using the same terms. This is achieved by adopting shared vocabulary sets uniquely identifying each term by its code. Several standard vocabulary sets exist that cover different portions of the medical terminology domain. LOINC (Logical Observation Identifiers Names and Codes) [6], UMLS (Unified Medical Language System) [10], SNOMED CT (Systematized Nomenclature of Medicine Clinical Terminology) [8], ICD (International Classification of Diseases) [1], and MeSH (Medical Subject Headings) [7] are among the most important standards. Alternatively, a new set can be developed for communication of terms that are not covered in these standards.

*Semantic* data interoperability refers to the joint ability of the sender and the receiver of data to share the exact context in which a term is used. In healthcare, there are multiple concepts that in many cases are referred to by identical terms. For instance suppose that two systems use the term *cough* to refer to the general concept of coughing (syntactic data interoperability in place). If the first system receives this term in a message, it should also be able to understand what is meant by that, as this term by itself does not convey any useful meanings. *Cough* can be sent as "a symptom of a disease"; "an observation in the patient's historical clinical

data"; or even "the cause of death of a person". It is important that this associated context be conveyed to the receiver in a proper way so that it can recognize the meaning of the term *cough* properly. The realization of semantic interoperability can be achieved using shared standard data models.

**HL7 Standard-based data interoperability**
HL7 is an international community of healthcare experts and information scientists collaborating to create standards for the exchange, management and integration of electronic healthcare information. HL7 v3 (version 3) specifies a hierarchy of static information models to describe information in the healthcare domain. Reference Information Model (RIM) is at the highest level in the hierarchy. The HL7 messaging process applies object-oriented development methodology on RIM and its extensions to create messages. Then these standard messages are used to transfer data between different healthcare systems.

**HL7 message refinement process**
HL7 methodology uses RIM, HL7-specified vocabulary domains, and HL7 v3 data type specification and establishes the rules for refining these base standards to specify Message Types and equivalent structures in v3. The strategy for development of these message types and their information structures is based upon the consistent application of constraints on HL7 RIM and HL7 Vocabulary Domains, to create representations that address a specific healthcare requirement. The different parts of the refinement process are are discussed below.

*Domain Message Information Model (D-MIM)* is a subset of the RIM that includes a fully expanded set of class clones, attributes and relationships that are used to create messages for any particular domain (e.g., accounting and billing, claims, and patient administration)

*Refined Message Information Model (R-MIM)* is used to express the information content for one or more messages within a domain. Each R-MIM is a subset of the D-MIM and only contains the classes, attributes and associations that are required to compose those messages.

*Hierarchical Message Description (HMD)* is a tabular representation of the sequence of elements (i.e., classes, attributes and associations) represented in an R-MIM. Each HMD produces a single base message template from which the specific message types are drawn.

*Message Type* represents a unique set of constraints on message identification that are presented in different forms such as: grid, table, or spreadsheet.

**Our approach**
We use a *document-based* communication approach by adopting the HL7 v3 CDA standard to encode input patient data and also the output of application of mined knowledge on patient data (described in detail in phase 3) to achieve semantic data interoperability in our framework. We shall justify our decision to use CDA as opposed to messages in the CDA vs messages topic later in this section. CDA documents need to be encapsulated as MIME packages in other

HL7 message types that support MIME. We chose messages from Medical Record domain to embed CDA documents carrying patient medical records for our purpose. These messages have a class Act.text of type ED (Encapsulated Data Type) that is suitable for embedding CDA. CDA instance is an XML document that conforms to a CDA schema and contains patients' clinical data. RIM has a large number of classes, relationships, data types, and coded values. CDA documents make wide use of coded values to present various semantics associated with clinical terms. For instance, it is possible for a clinical observation to represent "an event that has happened", "an objective and goal", "an intent or a plan", or "a request to perform" such an observation. On the other hand, there are various elements to encode different types of activities or information, e.g., a procedure, administration of medications, inclusion of medical images in CDA documents, or a patient encounter. To support syntactic data interoperability, the CDA schema provides mechanisms to specify clinical terms from a vocabulary system that is shared among the senders and receivers. In our case we used SNOMED.

**A message-based interoperability project in healthcare**

We are involved in a project with collaboration of an industrial partner and two medical research groups to integrate a clinical decision support system (CDSS) with a specialist application taking a message-based communication approach. This is one of the pioneer projects in integrating healthcare systems in compliance with new HL7 v3 and Canada Health Infoway standards [3]. COMPETE III Vascular Tracker (C3VT) is a decision support system that assists physicians to observe and ideally control patients risk factors within the domains of cardiovascular, diabetes, hypertension, and dyslipidemia diseases. C3VTs database contains a large body of knowledge gathered by rigorous study of the related literature in compliance with the internationally accepted methodology known as evidence-based medicine. As a further step in COMPETE III project, the research group would like to extend its scope by providing its services to other specialized applications. The expansion of the C3VT capabilities will take place within a pilot project that will allow C3VT to interoperate with a Cardiac Rehab Center (CRC) in a different city. In this integration a portion of patient data form CRC is sent to C3VT, such that the C3VT algorithms can be run on these data and corresponding recommendations and guidelines are returned to CRC.

**Project steps**. We have extracted and modeled different use cases and scenarios of the integrated system. Infoway specifies the set of standard transactions which are required to perform the selected scenario. The set of additional information to be transferred with each clinical concept such as dates, requester and performer are defined for high levels of data schema (such as vital signs, procedure and lab). In the next step, the data schema of the two systems are mapped to the standard format. In our project, this process resulted in four mapping files for SNOMED, LOINC, Infoway and HL7 which are finalized in an integrated extended version. The C3VTs responses (i.e., recommendations and guides) are expressed using HL7 suggestion for CDSS expression. Once these data schemas are mapped to HL7 standards, a framework is set up to transmit these data between two systems which completes the semantic interoperability.

**Data and Service mapping process**
In order to facilitate the mapping process, initially the C3VT data schema is completely mapped onto the SNOMED CT. By this mapping, the time required to trim irrelevant clinical terminology branches based on domain selection is reduced dramatically. Also the knowledge gained about SNOMED hierarchy through this mapping data schema to SNOMED is necessary to select the relevant domain from standard documents. The outcome of data mapping process is a mapping between legacy system data schema; HL7 messages; clinical domain; clinical terminology system hierarchy; and finally the extracted standard code for legacy data. In this sample, the Flu shot field from C3VT is mapped onto the ActProfessionalServiceCode domain and the appropriate message and the specific field of the message is identified. The complete service mapping is performed in two steps.

*First mapping* occurs among the standard documents. For example, in order to transfer the flu-shot information of a patient, we need to use the immunization storyboard in Infoway[3]. The first mapping, should find the proper transaction; interactions; trigger events and application roles, that are associated with the immunization storyboard. The mapping is performed by navigating the Scope and Package Tracking Framework document.

*Second mapping* is dedicated to selecting standard transactions that are required to perform a transaction of the legacy system. In our case study, two transactions are responsible for recording and retrieving the medical record of a patient. Considering the sections defined for a patient profile in our system (such as clinical observations and clinical procedures) we select the appropriate transactions to transfer these data. The Service Type of the Professional Service Procedure Recordmessage is the selected field to transfer flu-shot data. The message also contains additional information such as the performer of the procedure, the reason for procedure request,the person who should be notified of procedure completion, etc.

**Document-based (CDA) vs message-based communication**
The input to the CDSS is a subset of fields from a patient's medical record. CDA is a specialized domain in HL7 that specifies structure of Clinical Documents and is meant to be used precisely for such purposes as the transfer of medical records. While HL7 messages are attached to a control act (trigger event, sending application role, receiving application role etc.), the CDSS does not care about such context. It only cares about the values of a patient's clinical variables. Therefore, using a message-based approach would only add administrative overhead to the application. Furthermore, each HL7 version 3 message belongs to a domain or sub domain and serves a very specific purpose. For example, the POLB_MT004000UV message type of the Laboratory domain can only be used to communicate laboratory results data. Therefore, if we took a message-based approach, we would have had to employ multiple messages just to provide the input for a single guideline. Since each guideline requires many different inputs, message types would also vary depending on the guideline. This approach promised to make the implementation of parsers, message generators, logic modules and even data sources unnecessarily complicated and a maintenance nightmare, since adding new guidelines would mean adding a whole set of new messaging implementations. HL7 v3 fulfills all of our messaging

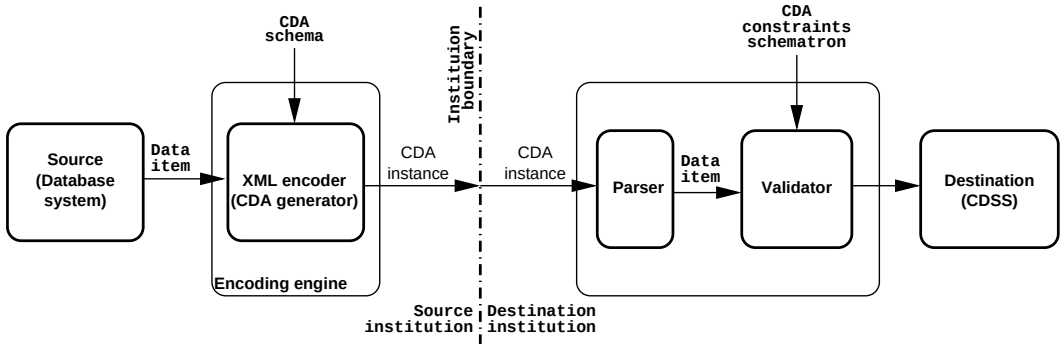


Figure 6. Data exchange using CDA: data source encodes the data items according to the CDA schema that is shared with the receiver. At the destination the CDA instance is parsed and validated to access the data items.

requirements with a minimum overhead and complexity making it an excellent alternative for our research project.

### 5.2. Knowledge interoperability

In addition to data, knowledge available in heterogeneous formats should be made available to our CDSS. For this purpose, knowledge that is deemed appropriate to be used by the CDSS is encoded in an XML-based standard called PMML (Predictive Model Markup Language) [19] in an off-line operation.

The PMML specification is developed by Data Management Group (DMG) [5] and is in the form of an XML schema. It provides a language to encode various types of data mining models, including clustering, regression, and association rules models. An apparent correspondence exists between the activities involved in data mining operation (described in Section 4.3) and different parts of PMML documents. The data items used in the "data selection" activity of the data mining operation are reflected in the PMML document's *data dictionary*. Similarly, the data processing and transformation are encoded mainly in the *transformation dictionary*. Input data transformations are done as part of the data mining process to make the data suitable for the mining algorithms. Several transformations may be applied in a sequence in the data mining operation. For correct interpretation and application of the data mining models, the input data should undergo the same sequence of transformations. For example, consider discritization of the age attribute to bins of *infant*, *youth*, *adult*, and *senior* according to the ranges of $[0, 3)$, $[3, 12)$, $[12, 50)$, and $[50, 120]$. If the age attribute of the mined data set is transformed by this transformation then it should also be performed during the application process.

Moreover the data mining models' data structures are also encoded along with their mining specific parameters. For example, an association rules mining model has frequent item sets,

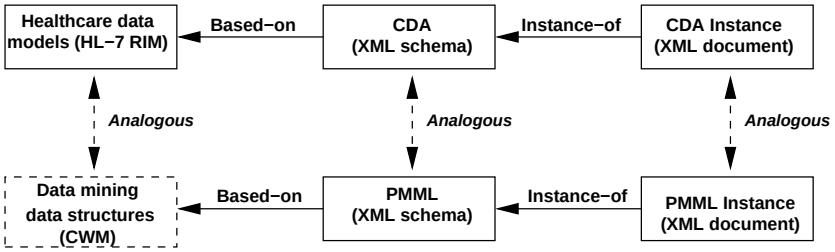


Figure 7. Analogy between different constituent artifacts for data and knowledge interoperability.

and association rules with support and confidence factors. The results of data mining technique are also encoded appropriately. The encoding process should preserve different properties associated with the data mining models discussed in Section 4.3.

Data mining models also require the input values to meet specific constraints, where these constraints are critical for correct application of the models. As an example the values of a numerical mining attribute should be within pre-specified valid ranges, or categorical attributes should not have values outside of the acceptable values set. Also, very often the data input to the data mining models are physical measurements and hence, careful attention should be paid in handling of units. We use *schematron validation documents* to validate CDA instances that contain the patient data. The validation documents encode the data mining model's input data constraints and hence are tightly bound to the data mining model's PMML document. These two documents should be ported to the usage site so that the data mining model can be applied on validated data items.

Even though data and knowledge interoperability address interoperability of different types of information, however as illustrated in Figure 7 knowledge interoperability artifacts are closely analogous to those of data interoperability. The PMML instance documents encode the mined knowledge according to the PMML specification's XML schema. Similarly, the CDA instance documents encode actual patients' data according to the CDA schema. PMML and CDA schemas are based on HL7 RIM and CWM [29] models, that define the data structures for healthcare-data and data mining results, respectively.

### 5.3. Running case study (data and knowledge interoperability)

In this section we continue our case study from phase 1 and provide the details for achieving data and knowledge interoperability. The data items for the decision tree model have been identified in Section 4.5 and presented in table III (page 10).

To validate the data items, a schematron document is developed with rules that perform the required checks according to the data specification which was provided in table III. The XML code snippet below shows two validation rules. The first rule (lines 3 to 11) defines an

assertion that checks whether the *C-Blue* data item (which is a required field for the classifier) exists in the document or not. The second rule (from line 13 to line 22) checks the validity of the terms that are used as the value of *C-Blue*[§]

```
<!-- Check for existance of 'C-Blue' -->
<pattern name="Check whether the required data elements exist.">
<rule context="/hl7:ClinicalDocument">
   <assert test="count(hl7:component/hl7:structuredBody/
    hl7:component/hl7:section/hl7:entry[@typeCode='COMP']/
    hl7:observation[@moodCode='EVN' and @classCode='OBS']/
    hl7:code[@code='1234-5' and
    @codeSystem='2.16.840.1.113883.6.2'])=1">
     The required data element 'C-Blue' does not exist in the input CDA document.
    </assert>
  </rule>
<!-- Check for validity of the values of 'C-Blue' -->
<rule context="/hl7:ClinicalDocument/hl7:component/hl7:structuredBody/
hl7:component/hl7:section/hl7:entry[@typeCode='COMP']/
hl7:observation[@moodCode='EVN' and @classCode='OBS']/
hl7:code[@code='1234-5' and @codeSystem='2.16.840.1.113883.6.2']">
   <assert test="../hl7:entryRelationship[@typeCode='COMP']/
   hl7:observation[(@classCode='OBS') and (@moodCode='EVN')]/
   hl7:code[(@code='1234-5-1' or @code='1234-5-2')]">
     Invalid value for 'C-Blue' data element.
    </assert>
  </rule></pattern>
```

The code below illustrates the structure of the decision tree classifier as encoded in the PMML file. Lines 3 to 7 defines the input and output of the classifier; and encoding of the tree structure is then followed (lines 8 to 24).

```
<TreeModel modelName="Decision Tree Model" splitCharacteristic="multiSplit"
algorithmName="decisionTree" functionName="classification">
<MiningSchema>
    <MiningField name="TDS" usageType="active"/>
    <MiningField name="C-BLUE" usageType="active"/>
    <MiningField name="DIAG" usageType="predicted"/>
</MiningSchema>
<Node score="UNKNOWN">
    <True/>
    <Node score="UNKNOWN">
        <SimplePredicate operator="lessOrEqual" value="4.85" field="TDS"/>
        <Node score="Benign-nevus">
            <SimplePredicate operator="equal" value="absent" field="C-BLUE"/></Node>
        <Node score="Blue-nevus">
            <SimplePredicate operator="equal" value="present" field="C-BLUE"/></Node>
    </Node>
    <Node score="UNKNOWN">
        <SimplePredicate operator="greaterThan"
```

[§]In our locally adopted vocabulary set, `1234-5`, `1234-5-1`, and `1234-5-2` are respectively the codes for *C-Blue*, *Present*, and *Absent*. The latter two are the values that *C-Blue* can have.

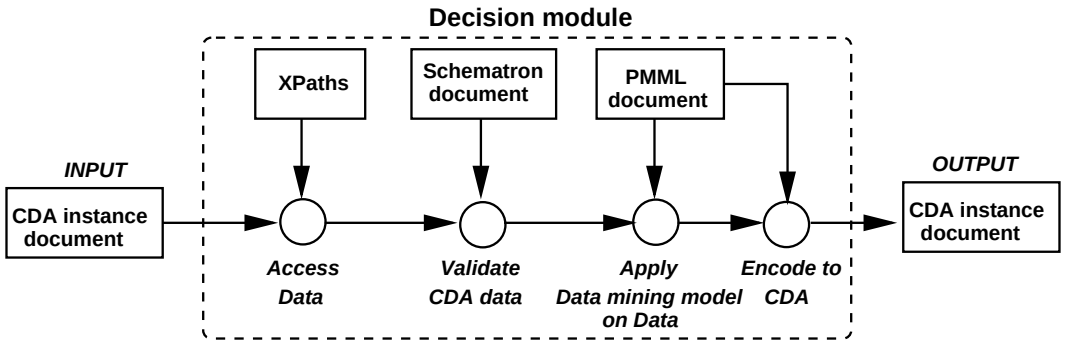


Figure 8. Different steps to apply the mined knowledge on the patient data. The input and output are encoded as CDA documents to further modularize the application of knowledge from the decision making process.

```
            value="4.85" field="TDS"/>
        <Node score="Malignant">
            <SimplePredicate operator="greaterThan" value="5.54" field="TDS"/></Node>
        <Node score="Suspicious">
            <SimplePredicate operator="lessOrEqual" value="5.54" field="TDS"/></Node>
        </Node> </Node>
</TreeModel>
```

## 6. Application (phase 3)

In this phase, the knowledge that is prepared and encoded in PMML documents (in the previous phases) is applied on the patient data encoded in CDA documents. The application is performed online at the usage site (i.e., point of care), by a program called decision module which is part of the CDSS. The decision module's functionality is provided through a simple interface, that receives a CDA and outputs a CDA. Figure 8 illustrates the steps of the decision module. These steps are described below in the order of execution.

1. *Accessing data items.* The data items in the CDA instance are accessed by an XML parser. We use *XPath expressions* [21] for this purpose. XPath is a language for referencing different locations within an XML document that is parsed by an XPath engine.

2. *Validating input data items.* The patient data obtained above are validated using the validation rules encoded in schematron documents to ensure data mining model specific constraints are met.

3. *Applying knowledge.* A parser component extracts information about the data mining model, in particular its type, the name of the algorithm used to build it, its data dictionary (input/output data types), the data transformations, and the structure of the data mining model from the PMML document. Based on this information, it builds an internal representation of this structure. Based on the type and technique of mining model received, the decision module then invokes an appropriate method to apply mining model on patient data. For example, a decision tree classification model is applied by invoking a method that traverses the decision tree based on the evaluation of the conditional expressions between its nodes. The PMML document specifies input and output data for the model it represents. The decision module uses XPath as specified in step 1 to extract required inputs from patient data CDA instance.

4. *Encoding the results.* The output of knowledge application is encoded in a CDA document to be used by the CDSS. The resulting CDA document in our running case study, for a patient with TDS<=4.85 and C-BLUE absent will contain a "diagnosis" node indicating that the patient has a "benign" Melanoma cancer.

### 6.1. Running case study (knowledge application)

A sample XPath expression that accesses the value of the *Skin Marks Border* data item in the XML file is provided below, where *Skin Marks Border* has been encoded as `1234-2` in our own vocabulary set.

```
/hl7:ClinicalDocument/hl7:component/hl7:structuredBody/
hl7:component/hl7:section/hl7:entry[@typeCode='COMP']/
hl7:observation[@moodCode='EVN' and @classCode='OBS']/
hl7:code[@code='1234-2' and
@codeSystem='2.16.840.1.113883.6.2']/../hl7:value/attribute::value
```

The validation document for our case study was illustrated in Section 5.3. After validation, the decision module calculates the *TDS* index according to equation 1. The calculated *TDS* index and the *C-Blue* data item (that is accessed directly from the CDA document) are then input to the data mining model (as decision tree). The following XML code shows part of the PMML code that encodes the output classes of the data mining classifier (line 3) as well as the custom messages that describe them (lines 4 to 6).

```
1 <DataField displayName="DIAGNOSIS" dataType="string"
2   name="DIAG" isCyclic="0" optype="categorical">
3 <Value displayValue="Benign-nevus" property="valid" value="Benign-nevus" >
4    <Extension extender="CAS" name="Description"
5    value="TDS is low and C-Blue is absent, so the
6       result of classification is Benign-nevus."></Extension>
7     </Value>
8      .....
9 </DataField>
```

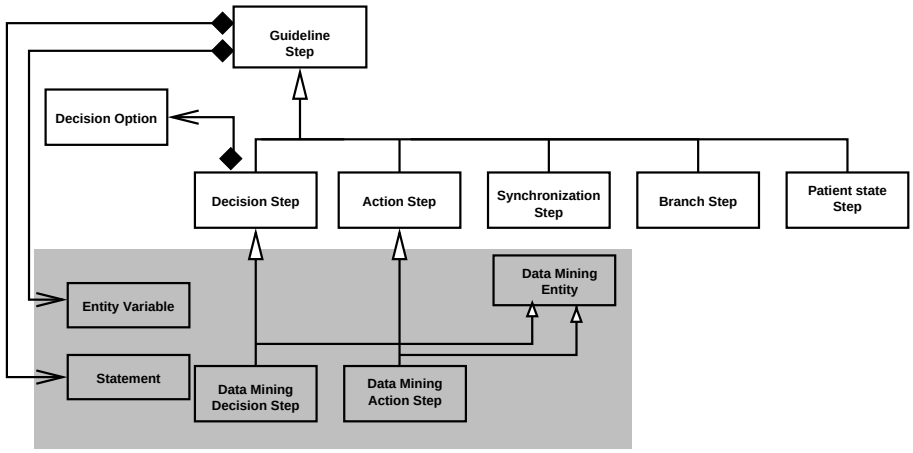


Figure 9. Top-level view of the modified GLIF3 model. The proposed extension classes are shown in the shaded area.

## 7. Interpretation (phase 4)

This phase involves using the outputs of the knowledge application (phase 3) in the context of Clinical Decision Support Systems. We use flow-based CDSS that uses the Guideline Interchange Format 3 (GLIF3) as its clinical guideline modeling language. In these systems the CDSS follows the flow of events encoded using a clinical guideline modeling language. GLIF3 is described briefly in Section 2. In the rest of this section we provide details on our extensions to enable incorporation of mined knowledge in clinical guidelines.

### 7.1. Incorporating mined knowledge in clinical guidelines

To incorporate data mining extracted knowledge for clinical decision making, we have adopted and extended GLIF3 specification. GLIF3 models are defined using five types of *steps*, which are nodes in the flow graph. They are as follows: i) *decision step* is a node in a guideline model's flow graph that determines the direction of the flow based on a decision criterion; an example is the age of the patient which is compared to a specific age as a decision criterion to direct the flow; ii) *action step* is a node that performs an action, such as a prompt to the user to prescribe medications, order tests, retrieve patient's medical records, or recommend treatments; iii) *synchronization step* is used to merge two or more concurrent decision flows into a single decision flow, such as receiving the lab test report and observing the effectiveness of the prescribed medication, before continuing to proceed to the next step; iv) *branch step* is used to fork and generate two or more concurrent decision making guideline-flows, such as, ordering a lab test and prescribing medication both at the same time; and v) *patient state step*

| Function | Description |
|---|---|
| *getDocument(String varName, String location, String docID, String patientID)* | Assigns the content of a CDA document from *location* to variable *varName*. *docID*, specifies the type of CDA document, and *patientID* specifies a patient identifier who the CDA document belongs to. |
| *getDataItem(String varName, String cdaDoc, String xpath)* | Assigns the value at location specified by *xpath* in the CDA document *cdaDoc*, to the variable specified by *varName*. |
| *setVariable(String varName, String value)* | Assigns the string value specified by *value* to variable by *varName*. |
| *evaluateExpression(String varName, String expr)* | Evaluates the logical expression specified by *expr* and assigns the resulting value to the variable *varName*. |
| *decisionModule(String cdaResult, String cdaInput)* | Invokes the decision module specified in the step's *decision_module* slot. *cdaInput*, and *cdaInput* are the module's input and output CDA documents. |
| *alert( String message )* | Outputs the message specified by *message* to the user. |

Table IV. The most important functions and corresponding arguments at the computable level of the extended GLIF3.

is a node in the flow graph that designates a specific patient's condition, such as presence of a symptom, previous treatments, or diagnoses.

Figure 9 illustrates the proposed extension to the GLIF3 model that represents the high-level clinical flow chart constructs. At this level, we defined a new abstract class, namely *Data Mining Entity* with an attribute that holds the name of the decision module used for interpreting the associated data mining model. Two new classes, called *Data Mining Decision Step* and *Data Mining Action Step* are then defined that extend the *Data Mining Entity* class, and the *Decision Step* and *Patient Step* classes, respectively. These classes add the functionality that is necessary to access and interpret a data mining model.

At the computable level of the flow-chart we provide a new mechanism to handle data and decision making logic which is based on the data and knowledge interoperability approaches that we discussed in phase 2 and 3 of the framework. The top level class, *Guideline Step* (and hence its sub-classes) is related to a *statement* class that is used to invoke pre-defined functions. These functions manipulate data and variables, apply the mined knowledge from the knowledge-base, or output the results to the user. The kinds of knowledge a particular step in the guideline expects to receive is pre-defined. Therefore it is equipped with the logic to handle that input. As the flow arrives at a step, the functions specified in that step are invoked and the related actions are performed. For example at a *decision step* different options (i.e., the following steps) are presented to the user. This enables the user to make the final decision and decide which path the flow should go to. Table IV summarizes the most important functions at the computable level of the extended GLIF3.

### 7.2. Running case study (interpretation of data and knowledge)

Continuing with our case study from Section 6.1 where we applied the mined knowledge on the patient data, in this section we demonstrate parts of a guideline model that was developed

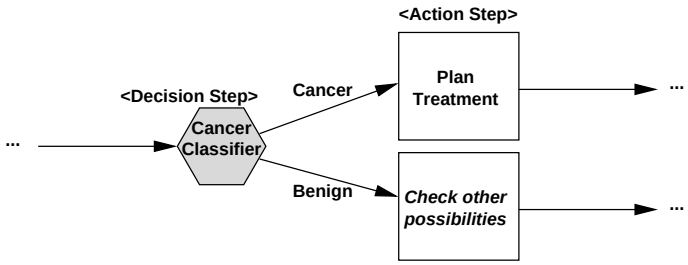


Figure 10. A portion of the guideline model for Melanoma skin cancer classifier at conceptual level.

for the Melanoma skin cancer decision tree classifier. The guideline model at its conceptual level is illustrated in figure 10 where a decision step and two action steps are used to decide whether or not the skin cancer exists.

At the *Data Mining Decision Step* the following functions are encoded to access and retrieve the required data items as well as running the *decision module* and providing the results and available options to the user. In the following code snippet belonging to the decision step, we retrieve the CDA document at line 2, and read its content in a variable. The variable contains the XML document and is passed to the decision module at line 4. The results are then stored in the *resultsCdaVar* variable. In lines 6 and 7 two variables contain two XPath expressions for accessing the "diagnosis", and diagnosis description from the CDA results document. Thus At lines 9 and 10, we retrieve the data values associated with the XPath expressions from the CDA results document stored in the *resultsCdaVar* variable. Finally, a message is provided to the user at line 12 detailing the results of the Melanoma skin cancer classifier.

## 8. The guideline execution environment

There are basically two approaches in executing a guideline model [34]. In the first approach, a new software is built for each individual guideline instance that implements the guideline flow as specified by the interconnection of the steps. This approach has many drawbacks including re-developing a large portion of the functionalities. Also, small changes in the model may require considerable recoding. Hence, necessary flexibility is obviously missing and therefore, this is not considered as a favorable approach.

On the other hand, we can think of an environment with an engine that receives a guideline model as input and interprets the model according to its specification. In this environment, we define a set of software modules that are responsible for performing the necessary actions as determined by the guideline model. The environment's execution engine is capable to follow the guideline model and invoke the corresponding modules. During the execution of the guideline,

```
1   //Read content of CDA document containing the required data items into cdaVar variable
2   getDocument(cdaVar, "/data-repository", "skin-tests", "Reza Sherafat")

3   //Execute the decision module and store the result CDA document in resultCdaVar
4   decisionModule(resultCdaVar, $cdaVar)

5   //Define variables to hold the XPath expressions referencing the result of the classification
6   setVariable(DiagnosisXpath, "/hl7:ClinicalDocument/hl7:component/hl7:structuredBody/
      hl7:component/hl7:section/code[@code='1292' and @codeSystem='2.16.840.1.113883.6.2']/
      ../hl7:entry[@typeCode='COMP']/observation[@classCode='ALRT' and @moodCode='INT']/
      code/attribute::code")
7   setVariable(DiagnosisDescriptionXpath,
      "/hl7:ClinicalDocument/hl7:component/hl7:structuredBody/hl7:component/hl7:section/
      code[@code='1292' and @codeSystem='2.16.840.1.113883.6.2']/../hl7:text")

8   //Retrieve the classification result from the results CDA document
9   getDataItem(resultVar1, $resultCdaVar, $DiagnosisXpath)
10   getDataItem(resultVar2, $resultCdaVar, $DiagnosisDescriptionXpath)

11   //Output a message to the user with the content of the classification
12   alert("Output of the Melanoma skin cancer classifier is $resultsVar1.
       Descriptions follow. Please select an option to continue: $resultsVar2")
```

the environment keeps track of the guideline's execution flow and provides the required data retrieval and knowledge interpretation facilities.

We adopted the second approach and implemented an environment and an execution engine to automatically interpret and execute a clinical guideline that has been defined according to the specification that we described in Section 7. In our implementation of the guidelines execution environment, we have developed a plug-in in the *Protégé* [4] *ontology editor tool*. GLIF3 modeling constructs are represented as ontology[¶] classes, where guideline authoring is carried out through graph widgets of *Protégé* and using the classes that we defined and were described in the previous section.

The environment allows multiple guideline models to be defined and the user can select a guideline for execution from a list. An instance of the engine is then instantiated to execute the selected guideline model. Execution is started from the initial step and continues along the links that connect different steps. The engine supports multiple flows of execution for each running guideline, since individual flows can fork at *branch step*s. Each flow points to a guideline step as its active step. Active steps are executed by the engine upon arrival of the flow to that step. After execution of a step and providing the user with the outputs, the engine waits for the user to signal continuation of the flow to the next step. At this point, the engine

[¶]An ontology in general refers to a collection of concepts and their relationships in different representations in a domain. In the context of this paper, the medical data and concepts of the GLIF3 guidelines are defined as an ontology.

retrieves the next step from the ontology model, updates the signaled flow, and executes the new active step.

The decision modules are implemented in Java and are run wherever the *decisionModule* function is invoked within a *Data Mining Decision Step* or *Data Mining Action Step* of a guideline model. They are implemented using the XELOPES library [12] and access locally stored PMML files that contain the corresponding data mining model. As described in Section 6 the output of the decision module is also encoded as a CDA document.

## 9. Other case studies

We have also worked on the application of our data and knowledge interoperability framework on other clinical data mining studies from the literature.

The following case study is an association rule mining analysis [32] that was briefly introduced in Section 4.4. This study was carried out on 665 data records each having 25 selected attributes including heart's perfusion measurements, heart vessels disease, and the heart disease risk factors (e.g., age, sex, smoking habit, and cholesterol level). The data mining model (i.e., a set of association rules) is extracted in the knowledge extraction phase and some of the rules are illustrated in Table II where LM, LAD, LCX, and RCA refer to heart vessels' narrowing measurements. In the interoperation phase, we adopt CDA schemas that can encode the data items defined in this model. Also, we encode the mined knowledge as a PMML document, and the constraints on its input data items as rules in a schematron document. A portion of the encoded PMML model is provided in the following code snippet[‖]

```
<AssociationRule id='1' support='0.18' confidence='0.80' antecedent='1' consequent='4'/>
<AssociationRule id='2' support='0.12' confidence='0.66' antecedent='2' consequent='5'/>
<AssociationRule id='3' support='0.20' confidence='0.53' antecedent='3' consequent='6'/>
```

In the knowledge application phase, the decision module accesses the data items from the CDA instances, validates them against the data mining model's data constraints, and applies the knowledge. Application is performed by matching the *antecedent* of the rules with the input data and providing the *consequent, support*, and *confidence* of the triggered rules as the result. This result along with the associated custom tags are encoded in a CDA document to be used in the guideline flow. The CDSS running the clinical guideline will then access the output CDA instance and provide the user with the results. These results help the practitioner to gain more insight into the clinical situation, and potentially improves her medical practice.

Similarly, the following case study is a data mining analysis found in the literature on association between pre-term births and antidepressants, drugs, alcohol and nicotine. The study has identified effects of taking these substances at various stages of pregnancy and

[‖]Values of the `'antecedent'` and `'consequent'` refer to identifiers associated with each discovered frequent item set. These are encoded in a separate part of the PMML document that is not shown.

| Association Rule | Support | Confidence |
|---|---|---|
| *Exposure to sedative medicine and tobacco before pregnancy* $\Rightarrow$ *(pre − term birth)* | 10% | 26% |
| *Exposure to sedative medicine before pregnancy* $\Rightarrow$ *(pre − term birth)* | 10% | 22% |
| *Exposure to sedative medicine and anti − depressant before or during pregnancy* $\Rightarrow$ *(pre − term birth)* | 5% | 21% |
| *Exposure to sedative medicine and tobacco during pregnancy* $\Rightarrow$ *(pre − term birth)* | 13% | 17% |

Table V. Discovered association rules in mining of data on substance, anti-depressant and sedative use during pregnancy [17].

| Data Mining Method | Clinical Variables | Use in CDSS |
|---|---|---|
| *Classification* | *Asymetry, Border, Color, Diversity, C − Blue* | *diagnosis of cancer* |
| *Association* | *heart region condition, heart valve condition* | *predict risk of heart disease* |
| *Association* | *exposure to anti − depressant, sedative, alcohol, drug, tobacco, pregnancy stage* | *predict risk of pre − term birth* |

Table VI. Discovered association rules in mining of heart disease data [32].

predicts risk percentage for pre-term birth. The study was carried out on 6231 records with 7 selected attributes and yielded 13 rules. Table V illustrates some of the prominent rules discovered. By incorporating the discovered rules into a CDSS, physicians can identify patients who are at higher risk for pre-term birth and take necessary steps accordingly.

The PMML encoding for some of the rules discovered is as follows:

```
<AssociationRule support=".10" confidence=".222" antecedent="4" consequent="14" />
<AssociationRule support=".10" confidence=".261" antecedent="5" consequent="14" />
<AssociationRule support=".05" confidence=".214" antecedent="6" consequent="14" />
<AssociationRule support=".13" confidence=".174" antecedent="8" consequent="14" />
```

Table VI offers a comparative analysis of the three case studies introduced in this paper with respect to the data mining method employed, clinical variables involved and potential practical use of incorporating the resulting rules into CDSS.

## 10. Conclusion

In this paper we described a novel framework for dissemination and application of the data and mined knowledge among the heterogeneous healthcare information systems. For data interoperability, we used HL7 CDA schema to define the required structure for encoding patients' health related data. We used schematron documents to define validation rules that perform consistency checks on CDA document instances. The healthcare researchers extract mined knowledge by mining existing healthcare data in an off-line operation and store in proprietary databases. In this context, we used the PMML specification to encode the produced mined knowledge to achieve knowledge interoperability between sources of knowledge and their users. To our best knowledge, there has been no methodology prior to this research to make this type of knowledge portable and available at the application sites. Further on, decision modules will access patient data from CDA documents and supply them into the data mining models from the PMML documents. The results of this operation are also provided as CDA documents to allow interoperability of the results. Moreover, we utilized the mined knowledge in our proposed extension to the GLIF3 clinical guideline modeling language that provides recommendations and warnings to the healthcare personnel based on the results of knowledge application. We demonstrated the application of our framework on case studies from the literature. We also described a prototype tool that has been implemented.

The approach taken in this research represents the mined knowledge as PMML models which are self-describing XML documents. The research should continue to examine the incorporation of the results of different types of data mining techniques. Fortunately, the database and data mining communities have been active in recent years to define and extend the PMML specification to a variety of different types of models, as well as developing libraries to implement different algorithms. Our research is focused on an application area of data mining results in the healthcare domain rather than the knowledge extraction process.

However, we note that there are still many obstacles that have to be taken care of. Before a Clinical Decision Support System can be deployed in a real world healthcare environment, it must undergo rigorous evaluations and tests to ensure high quality decision making and safety. The knowledge-base content, the system, and its processes need careful attention. Finally, we argue that this work provides the required methodology for research teams in data mining, healthcare standardization, and computer scientists to seek a collaborative healthcare environment where the data and knowledge are disseminated and utilized to improve the current state of healthcare services.

### REFERENCES

1. International Classification of Diseases (ICD). URL = `http://www3.who.int/icd/vol1htm2003/fr-icd.htm`. [Online; accessed 1-August-2008].
2. Health Level-7. URL = `http://www.hl7.org`. [Online; accessed 1-August-2008].
3. Canada health infoway website. URL = `http://www.infoway-inforoute.ca/en/home/home.aspx`. [Online; accessed 18-August-2008].
4. *Protégé* ontology editor tool. URL = `http://protege.stanford.edu/`. [Online; accessed 1-August-2008].
5. Data Management Group (DMG) website. URL = `http://www.dmg.org/`. [Online; accessed 1-August-2008].

6. Logical Observation Identifiers Names and Codes (LOINC). URL = http://www.regenstrief.org/loinc/. [Online; accessed 1-August-2008].
7. Medical Subject Headings (MeSH). URL = http://www.nlm.nih.gov/mesh/. [Online; accessed 1-August-2008].
8. Systematized Nomenclature of Medicine Clinical Terminology (SNOMED CT). URL = http://www.snomed.org/snomedct/index.html. [Online; accessed 1-August-2008].
9. Rules for Melanoma skin cancer diagnosis, URL = http://www.phys.uni.torun.pl/publications/kmk/. [Online; accessed 1-August-2008].
10. Unified Medical Language System (UMLS). URL = http://www.nlm.nih.gov/research/umls/. [Online; accessed 1-August-2008].
11. Computerization Of Medical Practices for the Enhancement of Therapeutic Effectiveness (COMPETE). URL = http://www.compete-study.com. [Online; accessed 1-August-2008].
12. P. AG. XELOPES library documentation - version 1.3.1. URL = http://www.prudsys.com/Service/Downloads/bin/1133983554/Xelopes1.3.1_Intro.pdf. [Online; accessed 1-August-2008].
13. Australia's National Electronic Decision Support Taskforce. Electronic decision support for Australia's health sector. URL = http://www.ahic.org.au/downloads/nedsrept.pdf. [Online; accessed 1-August-2008], January 2003.
14. Guideline Interchange Format (GLIF)3.5 - technical specification. URL = http://smi-web.stanford.edu/projects/intermed-web/guidelines/GLIF_TECH_SPEC_May_4_2004.pdf. [Online; accessed 1-August-2008], May 2004.
15. P. Berkhin. Survey of clustering data mining techniques. URL = http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf. [Online; accessed 1-August-2008].
16. Canadian Institute for Health Informatics (CIHI). Health expenditure in Canada. URL = http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=media_13nov2007_e. [Online; accessed 5-September-2008], November 2007.
17. Y. Chen, L. H. Pedersen, W. W. Chu, and J. Olsen. Drug exposure side effects from mining pregnancy data. *SIGKDD Explorations*, 9(1):22–29, June 2007.
18. L. Churilov, A. M. Bagirov, D. Schwartz, K. A. Smith, and M. Dally. Improving risk grouping rules for prostate cancer patients with optimization. In *Hawaii International Conference on System Sciences (HICSS)*, 2004.
19. Data Management Group (DMG). Predictive Model Markup Language (PMML) version 3.0 specification. URL = http://www.dmg.org/pmml-v3-0.html.
20. Electronic-Medical Summary (e-MS). The e-MS project page. URL = http://www.e-ms.ca/. [Online; accessed 1-August-2008].
21. The XML XPath Language - version 1.0. URL = http://www.w3.org/TR/xpath. [Online; accessed 27-September-2006], November 1999.
22. C. M. Farquhar, E. W. Kofa, and J. R. Slutsky. Clinicians' attitudes to clinical practice guidelines: a systematic review. *Medical Journal of Australia*, 177(9):502–506, 2002.
23. U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.
24. Health Level 7. The Clinical Document Architecture (CDA) standard specification. URL = http://www.hl7.org. [Online; accessed 1-August-2008].
25. HealthGrades. Fifth annual patient safety in american hospitals study, 2008. URL = http://www.healthgrades.com. [Online; accessed 18-August-2008].
26. W. H. L. Jr, T. Masters, J. Y. Lo, D. W. McKee, and F. R. Anderson. New results in breast cancer classification obtained from an evolutionary computation/adaptive boosting hybrid using mammogram and history data. In *2001 IEEE Mountain Workshop on Soft Computing in Industrial Applications*, pages 47–52, 2001.
27. P. Laxminarayan, C. Ruiz, S. A. Alvarez, and M. Moonis. Mining associations over human sleep time series. In *18th IEEE Symposium on Computer-Based Medical Systems (CBMS 2005)*, pages 323–328, 2005.
28. J. Li, A. W.-C. Fu, H. He, J. Chen, H. Jin, D. McAullay, G. Williams, R. Sparks, and C. Kelman. Mining risk patterns in medical data. In R. Grossman, R. Bayardo, and K. P. Bennett, editors, *KDD*, pages 770–775. ACM, 2005.
29. Object Mamanagement Group (OMG). The Common Warehouse Metamodel (CWM). URL = http://www.omg.org/technology/cwm/. [Online; accessed 30-August-2008].
30. Object Management Group (OMG). Healthcare Data Interpretation Facility (HDIF). URL = http://www.omg.org/docs/corbamed/98-03-07.pdf. [Online; accessed 1-August-2008], August 1998.

31. Ontario Ministry of Finance. The right choices: Investing in health care. URL = http://www.fin.gov.on.ca/english/budget/bud03/budhi1.html. [Online; accessed 1-August-2008], March 2003.
32. C. Ordonez, C. A. Santana, and L. 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.
33. A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman. Diagnosis of ovarian cancer using decision tree classification of mass spectral data. *Journal of Biomedicine and Biotechnology*, 5:308–314, 2003.
34. D. Wang, M. Peleg, S. W. Tu, A. A. Boxwala, O. Ogunyemi, Q. Zeng, R. A. Greenes, V. L. Patel, and E. H. Shortliffea. Design and implementation of the GLIF3 guideline execution engine. *Journal of biomedical informatics*, 2004 Oct;37(5):305-18.
35. Wikipedia. Health Insurance Portability and Accountability Act (HIPAA)— wikipedia, the free encyclopedia., 2006. URL = http://en.wikipedia.org/w/index.php?title=Health_Insurance_Portability_and_Accountability_Act&oldid=66219756. [Online; accessed 1-August-2008].
36. A. M. Wilson, L. Thabane, and A. Holbrook. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology*, 57(2):127–34, Feb 2004.
37. O. R. Zaïane, M.-L. Antonie, and A. Coman. Mammography classification by an association rule-based classifier. In *MDM/KDD*, pages 62–69, 2002.