10

Data and Mined-Knowledge Interoperability in eHealth Systems

Kamran Sartipi, Mehran Najafi and Reza S. Kazemzadeh
McMaster University
Canada

1. Introduction

The advancement of software development through component technology and system integration techniques has resulted in a new generation of very large software systems. This new paradigm has intensified challenges including interoperability of heterogeneous systems, sharing and reusing services, management of complexity, and security and privacy aspects. Examples of such globalized systems include: communication systems, banking systems, air traffic systems, transportation systems, and healthcare systems. These challenges require new development and management technologies and processes that fulfill the emerging demands in the networked systems.

Modern healthcare is experiencing major changes and as a result traditional conceptions are evolving: from health provider-centric to patient and family-centric; from solitary decision making to collaborative and evidence-based decision making; from decentralized and generalized care to centralized and specialized care. The need for better quality of service, unique identification of health records, and efficient monitoring and administration requires a uniform and nation-wide organization for service and data access.

Among other requirements, these changes require extensive assistance from modern software and information technology domains. Until recently there has been little attention to the IT infrastructure of healthcare systems. However, global trends are shifting healthcare towards computerization by developing electronic health record systems, IT-standardization through HL7 (Health Level 7) initiatives and nation-wide infrastructure specifications (Canada Health Infoway) and utilization of current evidences in decision-making processes. Most current systems are monolithic, isolated, paper-based, error-prone legacy systems which cause huge costs for the governments and healthcare organizations. The new systems need to take advantage of modern information and distributed systems to meet the emerging demands in healthcare environments.

As a result governments and private sectors are investing on healthcare information technology infrastructures to reduce huge costs of the existing systems and improve the quality of public care.

Health informatics (eHealth) is a new field which embodies a variety of techniques in information and knowledge management, data mining, decision support systems, web services, and security and privacy. Therefore, researchers with multi-disciplinary research interests from these fields need to collaborate in order to advance the state of the art in eHealth. In this chapter, we attempt to cover the core technologies that need to work

Source: Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou,
ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria

seamlessly to allow healthcare professionals and administration to use the available services effectively and efficiently. Also, we pay particular attention to the current research activities and issues regarding to the interoperbility of information and knowledge extracted from data mining operations. We propose a new architecture for interoperability of data and mined-knowledge (knowledge extracted from data mining algorithms). Finally, we propose new research avenues on the combination of data mining, eHealth, and service oriented architecture and discuss their characteristics.

The structure of this chapter is as follows: Section 2 presents the application of data mining in healthcare. Section 3 introduces different forms of knowledge representations in medical domain. Section 4, describes messaging standards in this domain. After these introductions to knowledge and messaging standards, we propose our framework in Section 5. In Section 6 an architecture for the framework is discussed and finally in Section 7, some research avenues are elaborated for applying our framework on the architecture that is explained in Section 6. We conclude the discussion of the chapter in Section 8.

2. Application of data mining in healthcare

Data mining or Knowledge Discovery from Databases (KDD) refers to extracting interesting and non-trivial relations and patterns from data in large databases. The results would be used for different purposes and domains stretching from marketing, administration, research, diagnosis, security, and decision making that are characterized by large amount of data with various relations. Healthcare is an important source for generating large and dynamic data repositories that can not be analyzed by professionals without help from computers. Typical healthcare databases containing information about patients, hospitals, bed costs, claims, clinical trials, electronic patient records, and computer supported disease management, are ideal sources to apply data mining operations for discovery purposes.

Data mining solutions have been used in healthcare to overcome a wide range of business issues and problems. Some of these problems include:

- Segmenting patients accurately into groups with similar health patterns.
- Evidence based medicine, where the information extracted from the medical literature and the corresponding medical decisions are key information to leverage the decision made by the professional. Therefore, the system acts as an assistant for the healthcare professionals and provides recommendations to the healthcare professionals.
- Planning for effective information systems management.
- Predicting medical diagnosis, treatment costs, and length of stay in a hospital.

After identifying an appropriate problem in a domain for applying data mining, we need a methodology. Different steps for applying data mining on a large database are as follows.

Data Selection: extract the data fields of interest. The selection is a subset of the data attributes that were collected in the data collection activity. For instance, in data collection the research team might choose to monitor or collect data from all of the patient's physical examinations, while in this step, particular fields, e.g., weight or height measurements are selected to be used for the data mining operation.

Data preprocessing: this step involves checking data records for erroneous values, e.g., invalid values for categorical data items, and out of range values for numerical attributes. In the real world practice, many records may have missing values. The researchers may decide to exclude these records from the data set, or substitute missing attributes with default or calculated values, e.g., the average of the values in other records for a missing numerical attribute.

Data transformations: several different types of transformations may be applied to the data items to make them more appropriate for the particular purpose of mining. The transformations can be considered as changing the basis of the space in which data records reside as points in this space. For example the patient's height in millimeter has probably too much precision; hence a conversion to centimeter or meter may be considered. Additionally, the data mining expert may choose to transform the weight value into discretized bins to further simplify things for the mining process. Also, there might be some fields that are derived from other data attributes, e.g., the duration of an infection can be derived by subtracting the initial diagnosis date from the date that the treatment was completed.

Data modeling: in this step, a data mining algorithm is applied to the data. The choice of the algorithm is decided by the researchers and depends on the particular type of analysis that is being carried out. There are wide ranges of algorithms available, but we can group them into two categories: those that describe the data and those that predict on future cases (Prudsys, 2006). The algorithms can also be grouped based on the type of mining they perform, e.g., clustering, classification, and association rules mining.

Evaluation and interpretation of results: it is essential that the results be evaluated in terms of meaningfulness, correctness, and usefulness. Based on the evaluation of results, the researchers may choose to go some steps back and perform them again differently. This makes the knowledge discovery process an iterative process. After completion of the discovery process, we refer to the extracted results as mined knowledge. These results are eventually stored in some application (data miner tool) specific format for future access and use.

Data mining models

Data mining models are data structures that represent the results of data mining analysis. There are many types of data mining models. In this section we briefly describe some major types: classification, clustering, and association-rules models. There are numerous algorithms in each category that typically differ in terms of their data or application specific fine tunings, their performance and approach in building the models, or the case or domain-specific heuristics they apply to increase the efficiency and performance of the mining process.

Classification models:

A classification algorithm (e.g., neural network or decision tree) assigns a class to a group of data records having specific attributes and attribute-values. 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 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, the classes can identify "whether a patient has been diagnosed with a particular cancer or not", and the classifier model assigns each patient's case to one of these classes. Some classification techniques that are applied on healthcare include: i) Neural network which is modeled as a large number of inter-connected data processors (known as neurons) that possess a small amount of local memory. These neurons learn based on algorithms that are inspired by biological neurons. A neural network (Haykin, 1998) can be used as a tool in the data mining modeling step; and ii) Bayesian (statistical) modeling (Jensen, 1998) is another modeling alternative that uses conditional probability to form a model.

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 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 the 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 mined data set. For example, the rule:

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

with a high confidence; might signify a cause-effect relationship between smoking and the 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.

Clustering models:

Clustering is originated from mathematics, statistics, and numerical analysis (Berkhin, 2006). In this technique the data set is divided into groups of similar objects. The algorithms usually try to group elements in clusters in a way to minimize the overall distance measure (e.g., the Cartesian distance) among the cluster's elements. Data items are then assigned to the clusters based on a specific similarity measure, and the researchers study the other properties of the generated clusters.

Applications of the above data mining techniques in healthcare that have been reported in literature are as following:

- By applying the k-NN classifier (i.e., an instance based method (AhaD, 1989)), Burroni et al. developed a decision support system to assist clinicians with distinguishing early melanoma from benign skin lesions, based on the analysis of digitized images obtained by epiluminescence microscopy (Burroni et al., 2004).
- Neural networks have been used in the computer-aided diagnosis of solid breast nodules. In one study, ultrasonographic features were extracted from 300 benign and 284 malignant biopsy-confirmed breast nodules (Joo et al., 2004).
- Another application of neural networks is detection of the disposition in children presenting to the emergency room with bronchiolitis (inflammation of small airways) (Walsh et al., 2004).
- In (Perou et al., 2000), k-means clustering is used to explore breast cancer classification using genomic data.
- The usefulness of Bayesian networks in capturing the knowledge involved in the management of a medical emergency service have been studied by Acid et al. (Acid et al., 2004).
- Segal et al. have shown that by learning Bayesian networks from data it is possible to obtain insight into the way genes are regulated (Segal et al., 2003).

As far as we are concerned in our framework, we don't differentiate between different implementations and algorithms of any of the data mining categories, if their results can be represented by the general constructs of the corresponding data mining type. For instance, different association rules mining algorithms take different approaches in extracting the *frequent-itemsets* and opt to choose different measures to exclude intermediary sets and hence prevent explosion in the results set. Based on standard constraints of support and confidence, others may apply additional constraints on the size of the rules' antecedent and consequent.

As we discussed, data mining has been used widely in healthcare for extracting knowledge in medical data. Finally this knowledge helps physicians to make a better decision. In the following sections, we will introduce a framework for transferring the generated mined-knowledge along with the electronic health records (EHR) of patients from a source organization to the point of use in another organization.

3. Knowledge representations in medical domain

In this section, major international methodologies and standards for representing medical and healthcare body of knowledge will be discussed. Best practice clinical workflows known as "clinical guidelines" are developed by medical researchers and represent human-based medical knowledge through rule-based or flow-based guideline techniques. On the other hand mined-knowledge can be automatically extracted through data mining techniques to be incorporated into human-generated knowledge in order to enhance their decision-making processes. We will elaborate on different computer based knowledge representations (i.e., GLIF3 and PMML) that are used to represent and transfer extracted knowledge. Finally, at the point of care the medical experts need clinical decision support systems to use these knowledge.

Clinical Guidelines

In the healthcare domain, a medical guideline (Grimshaw & Russell, 1993) (or a clinical guideline) is a document that assists healthcare personnel in making decision with respect to diagnosis, management or treatment of disease. There are two major types of medical guidelines as rule based and flow based guidelines. Rule based guidelines (Quaglini et al., 2003) are built with decision tree or association rules mining. There is one corresponding decision tree for each rule based guideline. In other words, we can convert each decision tree to its corresponding rule based guideline and vice versa. Figure 1 (left) illustrates a decision tree representation of a rule-based guideline.

A flow-based guideline (Panzarasa & Stefanelli, 2006) consists of different steps that form a treatment or diagnostic process for a patient. In every implementation of a flow-based guideline, there is an engine that runs the workflow. At each step, the workflow either changes the state of the system or transfers the control to the next step based on the condition and constraints on the working information.

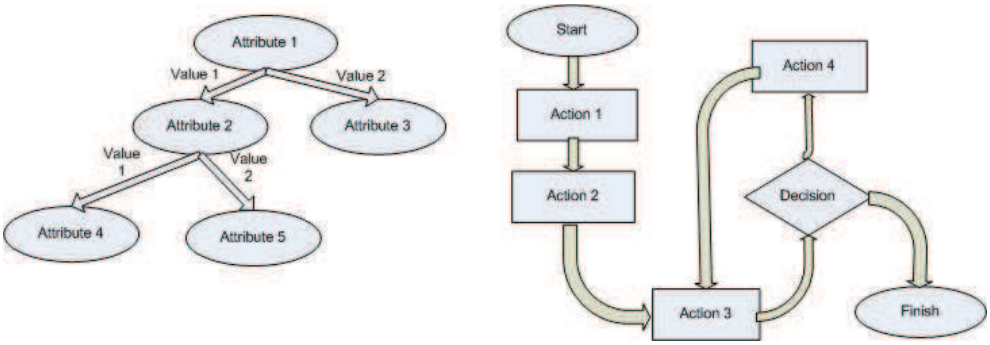

Fig. 1. A rule based guideline (left) and a workflow guideline (right).

GLIF (Guide Line Interchange Format)

Guideline Interchange Format 3 (GLIF3) (*Guideline Interchange Format)*, is a guideline modeling language that represents the clinical best practices as flowcharts. Expert medical researchers execute medical guidelines in GLIF3 format and in their Clinical Decision Support Systems to provide decision making support and clinical best practice. GLIF3 guidelines have been developed for a variety of purposes, including but not limited to heart failure, hypertension, thyroid screening, and many more. GLIF3 guidelines are defined in three levels of abstraction:

- **Conceptual level**: the first level is a flow chart that represents different states and actions in a structured graph. This level provides an easy to comprehend conceptualization of the guideline. At this level, the details of decision making are not provided and hence the guideline models are not computable. Different types of nodes in GLIF3 are as follows:
  - *Decision step* determines the direction of the flow based on a decision criterion specified in an expression language. For example, the age of the patient might be compared to a specific age as a decision criterion to direct the flow.
  - *Activity step* is a node that performs an action, e.g., prompts to prescribe medications; order tests; retrieve patient's medical records; or recommends treatments.
  - *Patient state step* is a node in the flow graph that designates a specific patient's condition, e.g., presence of a symptom, previous treatments, or diagnoses. Also, guideline models start with a patient state step.
  - *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.
  - *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.
- **Computable level**: to allow a guideline flow to be computed, the author has to specify the control flow, decision criteria, medical concepts, and relevant patient data. These are specified in the computable level.
- **Implementation level**: for GLIF3 guidelines to be actually deployed at an institution site, the patient data and actions should be mapped to institution specific information systems. The required mappings are specified in this level.

Predictive Model Markup Language

Predictive Model Markup Language (PMML) (Data Management Group) is an XML-based language to describe data mining models (clustering, associations, etc.). Also it represents the required constructs to precisely describe different elements, input parameters, model specific parameters, transformations, and results of a variety of types of data mining models. PMML is meant to support the exchange of data mining models between different applications and visualization tools. PMML provides independence from application, platform, and operating system, and simplifies the use of data mining models by other applications (consumers of data mining models).

Clinical Decision Support System

Clinical (or diagnostic) Decision Support Systems (CDSS) (Spiegelhalter & Knill-Jones, 1984) are interactive computer programs which are designed to assist physicians and other health professionals with decision-making tasks. The basic components of a CDSS include a *dynamic* (medical) knowledge base and an *inference mechanism* (usually a set of rules derived from the experts and evidence-based medicine) and implemented through medical logic modules based on a language such as Arden syntax (Hripcsak, 1991). It could also be based on expert systems or neural network.

Some of the most common forms of decision support systems include: *drug-dosing calculators* which are computer-based programs that calculate appropriate doses of medications based

on clinician's input key data (e.g., patient weight, indication for drug, serum creatinine). These calculators are especially useful in managing the administration of medications with a narrow therapeutic index. More complex systems include computerized diagnostic tools that, although labor intensive and requiring extensive patient-specific data entry, may be useful as an adjunctive measure when a patient presents with a combination of symptoms and an unclear diagnosis.

Both simple and complex systems may be integrated into the point-of-care and provide accessible reminders to clinicians based on previously entered data. These systems may be most practical when coupled with computerized physician order entry and electronic medical records. Finally, through their integration with practice guidelines and critical pathways, decision support systems may provide clinicians with suggestions for appropriate care, thus decreasing the likelihood of medical errors. For example, a guideline for the management of community-acquired pneumonia may include a clinical tool that, after the input of patient-specific data, would provide a recommendation regarding the appropriateness of inpatient or outpatient therapy.

An example of a clinical decision support system using decision trees can be found in a study by Gerald et al (Gerald et al., 2002). The authors developed a decision tree that assisted health workers in predicting which contacts of tuberculosis patients were most likely to have positive tuberculin skin tests. Also, a clinical decision support system for predicting inpatient length of stay is proposed in (Zheng et al., 2005).

In our proposed framework in Section 5, we use PMML to encode the result of data mining of the healthcare data and transfer this information to the point of care to be used by a clinical decision support system.

4. Communication standards for electronic health records

Standard-based interoperability between heterogeneous legacy systems is one of the main concerns in healthcare domain. Health Level 7 (HL7) is the most important standard messaging model that is widely adopted by the new healthcare systems. In addition to messaging standards, mapping clinical concepts and terms among different healthcare systems is an essential requirement for interorability provision. SNOMED CT is a comprehensive clinical terminology system that will be introduced in this section. Electronic Health Record (EHR) is a major component of the health informatics domain that is defined as: *digitally stored healthcare information about an individual lifetime with the purpose of supporting continuity of care, education and research*. It allows an information system engineer to represent data about observations, laboratory tests, diagnostic imaging reports, treatments, therapies, administrated drug, patient identifying information, legal permissions, etc. There are three main organizations that create standards related to EHR, including: HL7 (Dolin et al., 2005) in United States which is widely adopted, CEN TC 215 (De Moor et al., 2004) operates in most European countries, and ASTM E31 (Hripcsak et al., 1993) is specialized for commercial laboratory vendors in United States.

Health Level 7 (HL7)

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 version 3 (V3) has defined Reference Information Model (RIM) which is a large class diagram representation of the clinical data and identifies the life cycle of events that a message will carry. 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 EHR 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. Figure 2 illustrates the refinement process specified in HL7 methodology, where the different parts are discussed below.

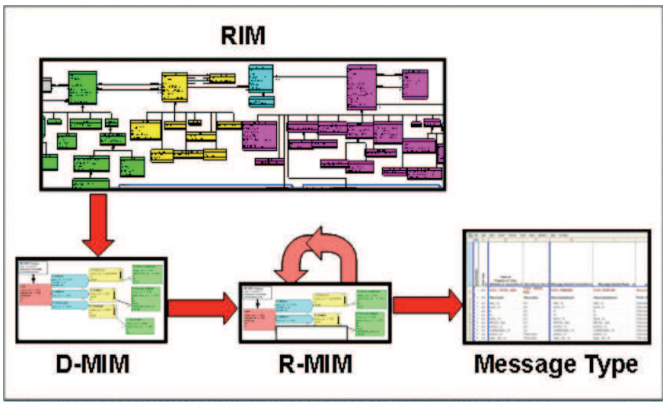

Fig. 2. Refinement process specified in HL7 methodology (Dolin R., Alschuler L., Beebe C., et al. The HL7 clinical document architecture.)

- *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.

Clinical Document Architecture (CDA)

CDA is an XML-based markup standard intended to specify the encoding, structure and semantics of clinical documents to be exchanged. The content of a CDA document consists of a mandatory textual part (which ensures human interpretation of the document contents) and optional structured parts (for software processing). The structured part relies on coding systems (e.g., from SNOMED (Andrews et al., 2007) and LOINC (McDonald et al., 2003)) to

represent concepts. The CDA standard doesn't specify how the documents should be transported. CDA documents can be transported using both HL7 v2 and HL7 v3 messages. CDA contributes in simplifying the healthcare message composition and transportation between non HL7 standard legacy systems, as well as to communicate more complex information.

Clinical terminologies (SNOMED CT)

A clinical terminology system facilitates identifying and accessing information pertaining to the healthcare process and links together terms with identical clinical meanings; hence it leverages the provision of healthcare services by the care providers.

SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms), is a systematically organized computerized collection of medical terminology that covers most areas of clinical information such as diseases, findings, procedures, microorganisms and pharmaceuticals. The terminology is comprised of concepts, terms and relationships with the objective of precisely representing clinical information across the scope of healthcare. It allows a consistent way to index, store, retrieve, and aggregate clinical data across specialties and sites of care. It also helps organizing the content of medical records, reducing the variability in the way data is captured, encoded and used for clinical care of patients and research. Concepts are clinical meanings and identified by a unique and human-readable numeric identifier (ConceptID) that never changes (e.g., 25064002 for "Headache").

In a clinical terminology system, *concepts* represent various levels of clinical detail, from very general to very specific with finer granularity. Multiple levels of granularity improve the capability to code clinical data at the appropriate level of detail. (e.g., *Dental Headache* is a kind of *Headache* in general). Concept descriptions are the terms or names assigned to a SNOMED CT concept. Multiple descriptions might be associated with a concept identified by its ConceptID. Every description has a unique DescriptionID.

So far, we introduced: i) standards for representing healthcare information and extracted mined-knowledge, ii) technology for a standard-based messaging mechanism, and iii) required healthcare terminology systems to provide unique medical concepts and terms. These together will allow the healthcare professionals to access to medical information and communicate their medical knowledge and use the computerized services available at other healthcare organizations. In the next section, we elaborate on a framework that allows both data and knowledge to be communicated and used.

5. A conceptual framework for data and mined knowledge interoperability

We have proposed a framework for interoperability of data and mined knowledge in clinical decision support systems which is based on HL7 v3 messaging, web services for communication, and flow-based clinical guidelines.

Figure 3 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.

Knowledge preparation

In this phase, the data mining knowledge is extracted from healthcare 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. Different data mining techniques have been presented in Section 2. This phase starts by removing the healthcare data attributes that can identify a patient or reveal their private data. Some studies (Mielikainen, 2003) have also shown that the privacy breaches can occur even when the data is anonymized. After anonymization the knowledge extraction process starts through: 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.

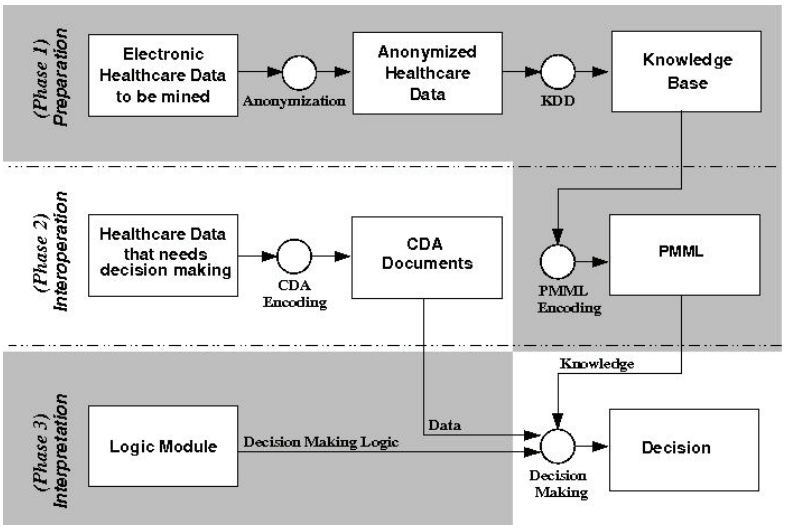

Fig. 3. Healthcare knowledge management framework. The shaded areas designate off-line parts.

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., healthcare data that needs decision making) in a source institution's internal data representation (e.g., EMR systems) is encoded into HL7 v3 messages or 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 of the proprietary data format of the involved institutions

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.

The proposed framework is at the conceptual level and requires an infrastructure to be applied. In the next section, we discuss some of the existing architectures in this domain which allow our conceptual framework to be implemented.

6. Existing architectures for interoperability support

Healthcare systems are large and complex information systems that require huge investments from a government to provide a nation-wide infrastructure. Such an infrastructure implements an electronic health record (EHR) system that spans different jurisdictional regions and connect a large number of distributed healthcare systems. In this context, service oriented architecture (SOA) has been widely adopted to solve the interoperability of the involving heterogeneous distributed systems.

In the rest of this section, first we introduce SOA, then we describe web services as a implementation technology for SOA and finally, we elaborate on Canada Health Infoway as an example of a service based architecture that connects different healthcare systems.

Service Oriented Architecture

Service Oriented Architecture (SOA) (Krafzig et al., 2004) plays a key role in the integration of heterogeneous systems by the means of services that represent different system functionality independent from the underlying platforms or programming languages. SOA contributes in relaxing the complexity, leveraging the usability, and improving the agility of the business services. On the other hand, new services may need to be adopted by the SOA community. Service is a program that interacts with users or other programs via message

exchanges. An (SOA) consists of the following concepts: *application frontend, service, service repository*, and *service bus*; each summarized as follows. Application frontends use the business processes and services within the system. A service consists of implementation, service contract, functionality and constraint specification, and service interface. A service repository stores service contracts. A service bus connects frontends to the services. A service-oriented architecture is a style of design that guides all aspects of creating and using business services throughout their lifecycle (from conception to retirement). An SOA is also a way to define and provide an IT infrastructure to allow different applications to exchange data and participate in business processes, regardless of the operating systems or programming languages underlying those applications.

Web Services

In more technical terms, a service is a program that interacts with users or other programs via message exchanges, and is defined by the messages not by the method signatures. Web services technology is defined as a systematic and extensible framework for application-to-application interaction built on top of existing web protocols. These protocols are based on XML and include: Web Services Description Language (WSDL) to describe the service interfaces, Simple Object Access Protocol (SOAP) for communication between web services and client applications, and Universal Description, Discovery, and Integration (UDDI) to facilitate locating and using web services on a network. These protocols are briefly defined below.

SOAP is an XML based protocol for messaging and remote procedure call using HTTP and SMTP. It defines how typed values can be transported between SOAP representation (XML) and application's representation by using XML schema definition. It also defines where various parts of Remote Procedure Call (RPC) are defined, including object identity, operation name, and parameters.

WSDL has an XML format that describes web services as a collection of communication end-points that can exchange certain messages. A complete WSDL service description has two parts: i) web service description (abstract interface), and ii) protocol-dependent details (concrete binding) that users must follow to access service at a service end- point.

UDDI is an XML based standard that provides a unified and systematic way to find service providers through centralized registry of services.

BPEL is a language for specifying business process behavior based on web services. These processes export and import functionality by using web service interfaces.

Web services are widely adopted as standard technology for implementation of service oriented architecture (SOA).

Infoway EHRi

Canada Health Infoway (EHRS Blueprint) is an organization that provides specifications for a standard and nationwide healthcare infrastructure. Infoway defines specifications and recommendations for development of an interoperable Electronic Health Record (EHR) system which is compatible with HL7 standards and communications technologies.

Infoway aims at integrating information systems from different health providers and administrations (e.g., hospitals, laboratories, pharmacies, physicians, and government agencies) within different provinces, and then connect them to form a nationwide healthcare network with standard data formats, communication protocols, and a unique health history file for each patient. In such a large infrastructure the individual's health information is accessible using common services according to different access privileges for patients and providers. Infoway provides EHRi (Electronic Health Record Infostructure) which is

designed based on service oriented architecture technology and consists of several components, as illustrated in Figure 5. The SOA main components within Infoway are discussed below.

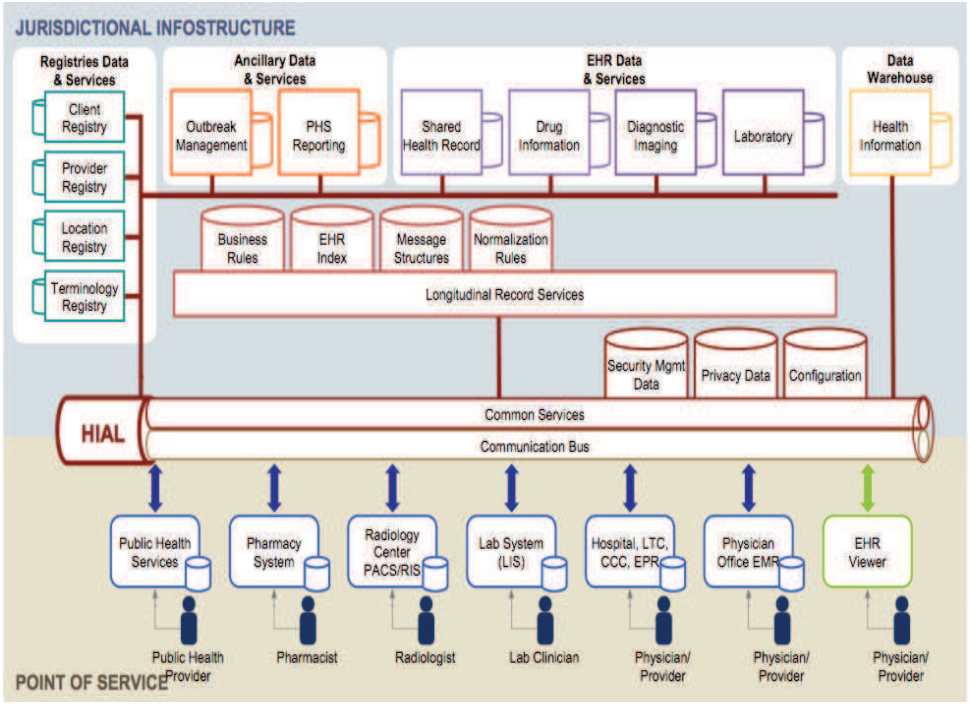

Fig. 5. Infoway EHRi (Source: Infoway (Canada Health Infoway))

As mentioned earlier, a typical SOA architecture consists of four main parts: frontend applications, service repository, services, and service bus. Application frontends are point-of-service components for physician, pharmacy, patient, and EHR viewer. Services are provided by different components; for an extended list of these services and their types, refer to (EHRS Blueprint). Service repositories consist of three groups "registries data & services", "ancillary data & services", and "EHR data & services". HIAL (Health Information Access Layer) is responsible for the functionality of the service bus. Considering the heterogeneity and age of the connected healthcare systems in a network, the integration process should be performed through a multi-technology facility, provided by "intermediary services" of SOA architecture. Infoway's Infostructure is mainly intended for transporting clinical documents through a communication framework. However, this architecture can also be used for other purposes such as tele-medicine, where performance guarantees are required. In such cases the performance of the service bus can be configured using technologies such as Message Oriented Middleware. Data warehouse represents a separate capability to compile, aggregate and consolidate EHR data for reporting and statistical or research analysis, as well as health prevention initiatives. After this introduction to Infoway Infostructure, in the next section we will discuss how to incorporate data mining services in such infrastructures.

7. Potential research avenues in electronic healthcare

In Section 5, we proposed a conceptual framework for data and mined-knowledge interoperabilty where the emphasize was put on the extraction of knowledge from healthcare data and transporting it to the point of care to be used by decision support systems. We discussed information and knowledge representation technologies (CDA and PMML) in an abstract manner.

In this section, we shift our focus towards research guidelines for defining new services for data and knowledge transportation based on SOA with emphasize on healthcare applications. We propose the enhancement of SOA services to allow domain knowledge to be available for users throughout the whole system as well as to provide monitoring and supervisory facilities for the administrative personnel. Other services include decision support facilities and mined-knowledge provisions that are crucial in administrative or technical decision making processes. The application domains include: financial analysis, tourism, insurance, healthcare, and transportation. These services will be discussed in the followings.

Mined-knowledge services

Current SOA services are either "data-centric", i.e., they transport data between two systems, or "logic-centric", i.e., they encapsulate business rules. However, there has been less attention given to knowledge-based services where the embodied knowledge in a specific application domain could be available through standard services. The main reason is the difficulty of precisely encoding different aspects of knowledge so that they can be correctly interpreted at the point of use. Researchers in the healthcare domain are actively working on encoding terminology semantics through terminology systems and clinical guidelines. This has resulted in encoding important business rules and best practice work flows into guidelines to be used by the community. Another source of valuable knowledge is the knowledge that is extracted from a large data set by applying data mining algorithms. This knowledge includes non-trivial patterns and trends among data that are not easily visible without computing assistance. We discussed different data mining algorithms and their applications in healthcare in Section 2. The application of the provided mined-knowledge at the point of use would boost the accuracy and convenience of decision making by the administrative personnel. In order to provide such mined-knowledge interoperability for large systems, enterprise service bus technology (Chappell, 2004) provides the required facilities.

Decision support system (DSS)

The proposed DSS component provides two types of services. I) Standard workflow services that reflect the best practices in a domain. In the healthcare domain, best clinical practices are developed by researchers and practitioners as clinical guidelines to help healthcare professionals and patients make decisions about screening, prevention, or treatment of a specific health condition. The DSS component offers these workflows to the users all over the system. COMPETE (COMPETE) is a pioneering Canadian healthcare project in electronic health research that is specialized in developing clinical guidelines based on different available resources. II) Customizable workflow generation services that allow users to define workflows for their enterprises. Such components may provide services for generating new guidelines by professionals as a part of standardization of

workflows. This service for generating clinical guidelines would complement the service for accessing a predefined clinical guideline.

Supervisory and visualization services

We propose services that allow administrative personnel or government agencies to secure effective control and supervision over the quality of service of the networked systems through activity visualization and identification of distribution patterns of services and bottlenecks. The visualization of activities in a large network of systems is crucial for administrative personnel to obtain updated and comprehensive insight into the active or passive relationships among different systems within the network. Examples of such networks include: bus and train transportation systems, air traffic control systems, transactions between financial institutions, and healthcare systems. As an application, consider the integrated network of healthcare systems as a large graph with different types of nodes representing clinical and administrative institutions, and types of edges representing categories of interactions among these institutions. Specific software analysis techniques from software reverse engineering can be applied to visualize the static or dynamic architecture of a large region of healthcare infrastructure (e.g., a province) from different view points. In this approach, data mining techniques are used to discover complex patterns of interactions among the nodes of the network and to provide means for self-management of the network.

Case study for Healthcare Network Visualization

A domain model is required to specify the graph node-types (i.e., different kinds of healthcare institutions such as: hospitals, physicians, pharmacies, laboratories, and government agencies) and graph edge-types as abstractions of health related communications between any two nodes (e.g., lab referral, drug prescription, billing statements, and patient electronic record acquisition). In this context, each edge-type will have several sub-types; for example, an edge-type "drug prescription" can have sub-types that represent a specific category of drugs that can be prescribed. This service will be implemented as follows. A common service in each node of the network is needed so that it enables the "supervisory component" to inquire about the network transaction activities of every node in the network within a certain period of time. On a daily basis, the interactions among the healthcare institutions will be logged and at the end of the day these logs will be sent (through service invocation) to a supervisory component to be analyzed. Consequently, the supervisory component can provide different views of the graph of the healthcare network, where the nodes and edges are color-coded according to the type of institutions and the types of interactions. The application of association rules mining algorithms on the generated graph would identify groups of maximally associated graph nodes according to specific graph edge types. In this context, maximal association refers to a group of nodes that all share the same services from a maximal group of service providers (e.g., specific categories of medications in pharmacies; and blood tests, X-rays, and ultrasounds in laboratories). A variety of data mining applications can be used to explore nontrivial properties in this network such as: spread of epidemics; distribution patterns of patients in particular regions; or distribution patterns of specific health services. The discovery of such patterns would enable the healthcare administrations and government agencies to restructure the service locations in order to reduce the cost of services and to increase the

accessibility of services to a larger population. The results would be available through a set of "monitoring and supervisory services" to the healthcare administrative personnel for analysis and policy decisions.

8. Conclusion

Current healthcare infrastructures in the advanced societies can not fulfil the demands for quality public health services which are characterized by patient-centric, seamless interoperation of heterogeneous healthcare systems, and nation-wide electronic health record services. Consequently, the governments and healthcare institutions are embracing new information and communication technologies to provide the necessary infrastructures for healthcare and medical services. In this chapter, we attempted to cover background preparation, advanced technology, architectural considerations, and research avenues within the new and critical domain of electronic health to address these emerging demands and presented the state-of-the-art solutions. The emphasize has been on the exploration power of data mining techniques to extract patterns and trends from large healthcare databases, and the means to deliver these knowledge along with the healthcare data to the point of use for enhanced decision making by professionals. Also, we discussed the trends towards raising the level of abstraction of services to the users which resulted in adopting service oriented architecture by the nation-wide healthcare infrastructures. Such high-level abstraction of healthcare services provides ease of use, vendor-independence, and seamless integration of the legacy systems with new systems. Healthcare domain is pioneer in systematically tackling the semantic interoperability by providing large and comprehensive terminology systems that allow common understanding of the medical terms and concepts. Furthermore, healthcare domain provides a well-defined process for representing and refining the whole body of medical information to develop standard HL7 messages for communication. As a result, the healthcare domain has acquired the necessary means to evolve towards a nation-wide and fully interoperated network of healthcare systems. Such a healthcare network is characterized by collaborating service providers and service users and enhanced techniques for more accurate clinical decision making.

9. References:

Prudsys A. *XELOPES,* library documentation - version 1.3.1. URL= http://www.prudsys.com/Service/Downloads/bin/1133983554/Xelopes1.3.1_Intro.pdf, 2006.

Haykin S., *Neural Networks: A Comprehensive Foundation*, book, Prentice Hall PTR, 1998.

Jensen F., *Introduction to Bayesian Networks*, book, Springer, 1998.

Berkhin P., *Survey of clustering data mining techniques*. URL = http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf, 2006.

AhaD W. ,*Incremental, instance-based learning of independent and graded concept descriptions*. International Workshop on Machine Learning, pp. 387–391, 1989.

Burroni M., Corona R., Dell'Eva G., et al., *Melanoma computer-aided diagnosis: reliability and feasibility study,* Clin Cancer Res, pp.1881–1886,2004.

Joo S.,Yang Y., Moon W., Kim H., *Computer-aided diagnosis of solid breast nodules: use of an artificial neural network based on multiple sonographic features*. IEEE Transact Med Imaging, pp 1292–1300, 2004.

Walsh P., Cunningham P., Rothenberg S., O'Doherty S., Hoey H., Healy R. *An artificial neural network ensemble to predict disposition and length of stay in children presenting with bronchiolitis*. Eur Journal Emerg Med. pp 259–564, 2004.

Perou C., Sorlie T., Eisen M., et al. *Molecular portraits of human breast tumours*, Nature, pp. 747–752, 2000.

Acid S., De Campos LM., Fernandez-Luna J., Rodrguise S., Rodrguise J., Salcedo J., *A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service*. Artificial Intelligence in Medicine; pp 215-232, 2004.

Segal E., Shapira M., Regev A., Pe'er D., Botstein D., Koller D., Friedman N., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.*, Nat Genet; pp 16-176, 2003.

Grimshaw J., Russell I., *Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations.* Lancet. pp. 1317–1322, 1993.

Quaglini S., Stefanelli M., Cavallini A., Micieli G., Fassino C., Mossa C., *Guideline-based careflow systems*, Artificial Intelligence in Medicine , pp. 5 – 22, 2003.

Panzarasa S., Stefanelli1 M., *Workflow management systems for guideline implementation*, Neurological Sciences, pp. 245-249, 2006.

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, 2004.

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

David J. Spiegelhalter and Robin P. Knill-Jones, *Statistical and Knowledge-Based Approaches to Clinical Decision-Support Systems, with an Application in Gastroenterology*, Journal of the Royal Statistical Society. Series A (General), pp. 35-77, 1984.

Hripcsak G., *Arden Syntax for Medical Logic Modules,* MD Compution, 1991

Gerald L.,Tang S., Bruce F., et al. *A decision tree for tuberculosis contact investigation*. Am J Respir Crit Care Med, pp. 1122–1127, 2002.

Zheng Y., Peng L., Lei J., *R-C4.5 decision tree model and its applications to health care dataset,* Services Systems and Services Management conference, pp. 1099- 1103, 2005

Dolin R., Alschuler L., Beebe C., et al. *The HL7 clinical document architecture*. J Am Med Inform Assoc, pp.552–569, 2005.

De Moor g., Claerhout B., van Maele G., Dupont D., *e-Health Standardization in Europe: Lessons Learned*, E-Health: Current Situation and Examples of Implemented and Beneficial E-Health Applications, book, Publisher: IOS Press, Pages233-237, 2004.

Hripcsak G., Wigertz O., Kahn M., Clayton P. , *ASTM E31.15 on health knowledge representation: the arden syntax* , book, progress on standardization in health care informatics, IOS press, 1993.

Andrews J., Richesson R., and Krischer J. ,*Variation of SNOMED CT Coding of Clinical Research Concepts among Coding Experts*, J. Am. Med. Inform. Assoc., pp. 497 – 506, 2007.

McDonald C., Huff S., Suico J., Hill G., Leavelle D., Aller R., Forrey A., Mercer K., DeMoor G., Hook J., Williams W., Case J., Maloney P. , *LOINC, a universal standard for identifying laboratory observations: a 5-year update*, MEDLINE, pp.624-33, 2003.

Mielikainen T., On inverse frequent set mining. *Workshop on Privacy Preserving Data Mining (PPDM)*, pp 18–23, 2003.

Krafzig D., Banke K., Slama D.,Enterprise *SOA: Service-Oriented Architecture Best Practices*, Book, Prentice Hall PTR, 2004.

Canada Health Infoway, http://www.infoway-inforoute.ca/.

EHRS Blueprint - Infoway Architecture Update. http://www.infoway-inforoute.ca/.

Chappell D., *Enterprise Service Bus: Theory in Practice*, book, O'Reilly, 2004.

COMPETE. Computerization of Medical Practice for the Enhancement of Therapeutic Effectiveness. http://www.compete-study.com/index.htm.

Data Mining in Medical and Biological Research

Edited by Eugenia G. Giannopoulou

ISBN 978-953-7619-30-5
Hard cover, 320 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

This book intends to bring together the most recent advances and applications of data mining research in the promising areas of medicine and biology from around the world. It consists of seventeen chapters, twelve related to medical research and five focused on the biological domain, which describe interesting applications, motivating progress and worthwhile results. We hope that the readers will benefit from this book and consider it as an excellent way to keep pace with the vast and diverse advances of new research efforts.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Kamran Sartipi, Mehran Najafi and Reza S. Kazemzadeh (2008). Data and Mined-Knowledge Interoperability in eHealth Systems, Data Mining in Medical and Biological Research, Eugenia G. Giannopoulou (Ed.), ISBN: 978-953-7619-30-5, InTech, Available from: http://www.intechopen.com/books/data_mining_in_medical_and_biological_research/data_and_mined-knowledge_interoperability_in_ehealth_systems

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821