Healthcare Informatics: Data Mining

Wullianallur Raghupathi
Fordham University, New York, New York, U.S.A.

Abstract
With the development and maintenance of large health data repositories of structured and unstructured data, health organizations are increasingly using data analytics, including data mining, to analyze and utilize the patterns and relationships found in the data to make improved clinical and other health-related decisions. This entry discusses the potential of data mining in healthcare and describes the various applications of data mining methods and techniques. A brief review of examples of data mining in healthcare is also offered. An ongoing project in the mining of the unstructured information in cancer blogs is also described. Conclusions are then offered.

INTRODUCTION
Health data, including general patient profiles, clinical data, insurance data, and other medical data, are being created for various purposes, including regulatory compliance, public health policy analysis and research, and diagnosis and treatment. The data include both structured data (e.g., patient histories as records in a database) and unstructured data (e.g., audio/video clips, textual information such as in blogs or physician’s notes). Data mining methods can be applied to search and analyze these large repositories to shed light on a wide range of health issues, including drug reactions, side effects, and other issues. For example, data mining techniques revealed the association between Vioxx, the arthritis drug, and increased risk of heart attack and stroke. The drug was withdrawn from the market (http://www.informationweek.com/news/business_intelligence/mining/showArticle.jhtml?articleID = 207300005).

In another example, IBM has been working with the Mayo Clinic for mining the data of millions of patient records to “analyze the information, look for similarities from one patient and another, and identify patterns” (http://www.healthcareitnews.com/news/data-mining-key-phase-2-ibm-mayo-partnership).

Healthcare organizations, including hospitals, health maintenance organizations (HMOs), and government entities such as the Centers for Disease Control, are establishing numerous health data repositories. These are typically large, relational databases that store different types of clinical and administrative data from primary electronic health sources such as hospital admission records. These repositories collect comprehensive data on large patient groups in longitudinal fashion, thereby permitting the examination and analysis of patterns and trends over time. Tasks include utilization statistics and outcomes. The data can be used for quality assurance and clinical management queries. Although the breadth and depth of the repositories include a variety of health and medical data, including genetic data, biomedical data, and data for general health issues such as quality control (e.g., medical error patterns), data mining applications are relatively new. Additionally, challenges are also foreseen. For example, large repositories may lead to a combinatorial explosion of alternatives. On the other hand, the multiple dimensions of the data for very complex relationships are typically rarely available because the relationships are spread thinly across the several dimensions. Fortunately, the developing large medical and health repositories can alleviate these challenges to an extent. These are providing integrated views of the patient encounters. Data mining of these quantitative and qualitative data has great potential for improving the quality of healthcare and reducing the costs of healthcare delivery.

In this entry we discuss the potential of data mining in healthcare. An outline and discussion of the steps involved is also provided. Our ongoing research in the data mining of health-related blogs using the Unstructured Information Management Architecture is then described. Finally, conclusions are offered.

DATA MINING IN HEALTHCARE
The Value of Data Mining
Data mining is defined as “the nontrivial extraction of implicit previously unknown and potentially useful information from data.” The value for healthcare delivery is enhanced when the data mining has specific purposes and health/medical questions to answer. Typically, the healthcare process being data rich, many potential patterns can be discovered by the use of different types of algorithms. However, the patterns have
value for enhancing the quality of the healthcare delivery process only when specifically addressing a particular issue or question. For example, using a data mining method involving clustering, a user can automatically discover distinct patient sets classified by one or more variables. It is not necessary to hypothesize a solution or delve into the details of the clustering. An application with a well-defined user interface has the potential to make the mining process transparent and seamless. The application with the underlying algorithm works on the data repository to enable the user to find solutions (e.g., categorizing patients, grouping patients by drug reactions, profile of emergency visit patients) in the most promising way. The data themselves become an active part of the solution. To this end, data mining is data driven. As Mullins et al. suggest, it is pertinent as a strategy to “discover” patterns already known to be true in the preliminary stages of the data mining task. It is important to confirm the tool, build confidence in the approach, and often, serendipitously, revelations may occur.

In healthcare, therefore, pattern-discovering algorithms in the data mining process can transform raw data into useful decision-making information with minimal intervention by the user, be it a physician or a hospital administrator. The data repositories created by health delivery organizations and health insurance companies are not in vain as the role of the data is enhanced. These organizations can tap into the discovery role of data mining, just as the financial services industry has done, and provide higher-quality healthcare as participants in the healthcare delivery process are empowered with useful information.

In the healthcare domain encompassing bioinformatics, medical informatics, and health informatics, data mining offers many new opportunities for practitioners and researchers. Some of the more significant ones include:

- Discovery of previously unknown facts (e.g., correlation between a drug and side effect). In this situation the application learns associations and flags the user, or the application facilitates the health data to identify value (e.g., potential drug discovery).
- Organization of large repositories of health and medical data for very complex problems (e.g., pandemic patterns and clusters). In this regard, the application can provide real-time alerts as to particular situations that require immediate attention, as well as provide insight into what might occur next.
- Prediction of the future in various situations and scenarios (e.g., what-if analysis in clinical trials, consequences of certain actions on public health policy). Data mining can help forecast trends (as in epidemics) and threats as well as opportunities, thereby enabling the organization, be it for profit or nonprofit, to deal with the future effectively with knowledge.

Data Mining as a Process in Healthcare

The typical types of healthcare questions that are solvable by data mining techniques can be divided into two main categories: those that are solved by discovery techniques and those solved by predictive techniques. If the healthcare problem requires the researcher to find useful patterns and relationships in the data (e.g., relationships between a particular diet and blood pressure), that problem will lead the researcher to a discovery method. On the other hand, if the healthcare problem requires the researcher to predict some type of value (e.g., the radiation dosage for a particular profile of cancer patients), that problem would obviously lead the researcher to a predictive method. In the pharmaceutical industry, for example, a range of methods, including associations, sequences, and predictive methods for clinical disease management, associations and prediction for cost/quality management, and segmentation and clustering for patient groupings in clinical trials, have tremendous potential.

Typical questions include:

- What do my patients look like?
- What is the drug dosage–patient profile association?
- With which other drugs does the new drug interact negatively?
- What effect does use of a particular drug for a disease have on other conditions?
- Does the drug cause side effects, and if so, what?

Many of the typical problems and questions can be resolved by one of a few data mining techniques. They include three discovery techniques (clustering, associations, and sequences) and two predictive techniques (classification and regression).

Discovery techniques

The data mining techniques based on this method find health or medical patterns that preexist in the data, but with no a priori knowledge of what those patterns may be. One could think of these patterns as serendipitously discovered, although the goals are inherently present in the data themselves. Three of the popular discovery techniques include:

1. The clustering technique groups health/medical records into segments by how similar they are based on the characteristics under study. Clustering could be used, for example, to find distinct symptoms of
diseases with similar characteristics to create a disease/patient segmentation model.

2. Association is a type of relationship analysis that finds relationships or associations among the health/medical records of single transactions. A potential use of the association method is for health group analysis, that is, to find out what diseases tend together to form a group, such as viral or bacterial (or patient groups), which is quite useful in epidemic/pandemic surveillance and identifying cause-treatment protocols for particular diseases.

3. The sequential pattern discovers associations among health/medical records, but across sequential transactions. A hospital could use sequential patterns (longitudinal studies) to analyze admissions over time, and to provide customized patient care.\[10\]

Predictive techniques

The predictive techniques of classification and regression are data mining techniques that can help forecast some type of categorical or numerical value (e.g., optimal dosage of a drug, drug pricing, etc.).

1. The technique of classification can be used to forecast the value that would fall into predefined grouping or categories. For example, it can predict whether a particular treatment will cure, harm, or have no effect on a particular patient.

2. On the other hand, the technique of regression is used to predict a numerical value on a continuous scale, for example, predicting the expected number of admissions each hospital will make in a year. In contrast, if the range of values is between 0 and 1, then this becomes a probability of an event occurring, such as the likelihood of a patient dying (repeat visits) or getting well, for example.\[10\]

In many instances a combination of data mining techniques is necessary (e.g., first perform patient segmentation using the clustering technique to identify a target group of patients); this is followed by a grouping analysis using the associations technique with the transactions (data) only for the target group to find drug affinities on which to base treatments.\[10\]

Mining and scoring

The five mining techniques outlined above are used against current health/medical data to create a data mining model. The process of applying an existing mining model against new data is called scoring.\[10\] Each of the five techniques has an associated scoring method that is used to apply against new data. Cluster scoring can be used, for example, to assign a new patient to the appropriate clinical trial based on the existing cluster model (or a drug for treatment).

In order to select the initial mining technique, one may develop a short list of typical health questions that the most common mining method or combination of methods may help answer. For example:

1. What do my patients look like? Clustering
2. Which patients should be targeted for drug (treatment) promotion (trial)? Clustering
3. Which drugs should I use for the trial (treatment)? Association or sequential patterns
4. Which drugs should I replenish in anticipation of an epidemic? Associations
5. Which of my patients are most likely to get well (based on a protocol)? Classification or regression
6. How can I identify high-risk patients? Clustering
7. When one drug fails, which others are most likely to fail too? Sequential patterns
8. Who is most likely to have another heart attack? Classification or Regression
9. How can I improve quality of care (or patient satisfaction)? Clustering plus associations

Build and deploy data mining application

The process of building and deploying a data mining application is highly iterative.\[10\] This process may include three specific steps:

1. Health/medical data preparation
2. Creation and verification of the particular mining model
3. Deployment of the model in some way

The process of data preparation involves finding and organizing the health/medical data for the chosen mining technique. Once the data are ready for use, the mining technique can be involved in the development of the mining model, which is then confirmed by the developer. It is possible that the process goes through several iterations until one obtains a refined data model. After confirmation, the model is ready to be deployed for use. Generally speaking, the data preparation step comprises the identification of the specific data requirements, the appropriate location of the data, and the extraction and transformation of the data into the appropriate format for the chosen mining technique.\[10\]

The data mining model is created once the data are transformed and ready for use. The particular application/tool is used on the data set after choosing the technique and providing the parameters. Multiple algorithms may be used with the input parameters. In predictive techniques additional steps may be involved, including a training phase and a testing phase. The
resulting model can be stored and possibly viewed using an appropriate visualization tool in the application.\textsuperscript{10} The visualization process plays a critical role in presenting information about model quality, specific results such as associations, rules, or clusters, and other information about the data and results pertinent to the particular model. This information enables the data mining analyst to evaluate the model quality and determine whether the model fulfills its healthcare purpose. If need be, improvements to the input data, model parameters, and modeling technique can then be made to obtain a good model that reflects the healthcare objective.\textsuperscript{10} In the final step, the data mining results are deployed in the healthcare organization as part of a business intelligence (data analytics) solution. Data mining results can be deployed by several means:

1. \textit{Ad hoc} decision support: Use data mining on an \textit{ad hoc} basis to address a specific nonrecurring question. For example, a pharmaceutical researcher may use data mining techniques to discover a relationship between gene counts and disease state for a cancer research project.

2. Interactive decision support: Incorporate data mining into a larger health intelligence application for ongoing interactive analysis.

3. Scoring: Apply a data mining model to generate some sort of prediction for each health/medical record, depending on model type. For example, for a clustering model, the score is the best-fit cluster for each patient. For the association model, the score is the highest-affinity item (variable), given other items (variables). For a sequence model, the score is the most likely action to occur next. For a typical predictive model, the score is the predicted value or response.\textsuperscript{10}

**EXAMPLES**

Mullins et al.\textsuperscript{1} report on the application of Health Miner to a large group of 667,000 inpatient and outpatient digital records from an academic medical system. They used three unsupervised methods: Clici mines predictive analysis, and pattern discovery. The initial results from their study suggested that these approaches had the potential to expand research capabilities through identification of potentially novel clinical disease associations. In other examples, the prior analyses using large clinical data sets have typically focused on specific treatment or disease objects.\textsuperscript{14} Most have examined specific treatment procedures, for example, cesarean delivery rate,\textsuperscript{11} coronary artery bypass graft (CABG) surgery volume,\textsuperscript{12} routine chemistry panel testing,\textsuperscript{13} patient care, cancer risk for nonaspirin nonsteroidal/anti-inflammatory drugs users,\textsuperscript{14} preoperative beta-blocker use and mortality and morbidity following CABG surgery,\textsuperscript{15} and incidence and mortality rate of acute (adult) respiratory distress syndrome.\textsuperscript{16} to name a few. These studies have several factors in common: large sample size, clinical information source, and they support or build upon preestablished hypotheses or defined research paradigms that use specific procedures or disease data. Clinical outcome algorithms have also been applied to harness large health information databases in order to generate models directly applicable to clinical treatment. These models have been used successfully to create mortality risk assessments for adults\textsuperscript{17–19} and pediatric intensive care units.\textsuperscript{20}

In other studies, Uramoto et al. describe the application of IBM Text Analysis and Knowledge Mining (TAKMI) for biomedical documents to facilitate knowledge discovery from the very large text databases characteristic of life science and healthcare applications. MedTAKMI dynamically and interactively mines a collection of documents to obtain characteristic features within them.\textsuperscript{21} By using multifaceted mining of these documents together with biomedically motivated categories for term extraction and a series of drill-down queries, users can obtain knowledge about a specific topic after seeing only a few key documents.

Inokuchi et al.\textsuperscript{22,23} and Wang et al.\textsuperscript{23} describe MedTAKMI-clinical decision intelligence (CDI), an online analytical processing system that enables the interactive discovery of knowledge for CDI. CDI supports decision making by providing in-depth analysis of clinical data from multiple sources.\textsuperscript{22,23} These and other examples indicate the potential and promise of data mining in healthcare.

**MINING OF CANCER BLOGS WITH THE UNSTRUCTURED INFORMATION MANAGEMENT ARCHITECTURE (UIMA)**

In this section we describe our ongoing research project in the use of the UIMA in mining textual information in cancer blogs. Health organizations and individuals such as patients are using information in blogs for various purposes. Medical blogs are rich in information for decision making. Current software such as web crawlers and blog analysis are good at generating statistics about the number of blogs, top ten, etc., but they are not advanced/useful computationally to help with analysis and understanding of the social networks that form in healthcare and medical blogs, the process of diffusion of ideas (e.g., the commonality of symptoms and disease management), and the sharing of ideas and feelings (support and treatment options, what worked). Therefore, there is a critical need for sophisticated tools to fill
this gap. Furthermore, there are hardly any studies or applications in the content analysis of blogs.

There has been an exponential increase in the number of blogs in the healthcare area, as patients find them useful in disease management and developing support groups. Alternatively, healthcare providers such as physicians have started to use blogs to communicate and discuss medical information. Examples of useful information include alternative medicine and treatment, health condition management, diagnosis-treatment information, and support group resources. This rapid proliferation in health- and medical-related blogs has resulted in huge amounts of unstructured yet potentially valuable information being available for analysis and use.\textsuperscript{[2]} Statistics indicate health-related bloggers are very consistent at posting to blogs.

The analysis and interpretation of health-related blogs are not trivial tasks. Unlike many of the blogs in various corporate domains, health blogs are far more complex and unstructured. The postings reflect two important facets of the bloggers: the feeling and the mind of the patient (e.g., an individual suffering from breast cancer but managing it). How does one parse and extract the deep semantic meanings in this environment? Mere syntactic analysis would not do.

The UIMA defines a framework for implementing systems for the analysis of unstructured data.\textsuperscript{[2,24–26]} In contrast to structured information, whose meaning is expressed by the structure or the format of the data, the meaning of unstructured information cannot be so inferred.\textsuperscript{[2]} Examples of data that carry unstructured information include natural language text and data from audio or video sources. More specifically, an audio stream has a well-defined syntax and semantics for rendering the stream on an audio device, but its music score is not directly represented.\textsuperscript{[27]} The UIMA is sufficiently advanced and sophisticated computationally to aid in the analysis and understanding of the content of the health-related blogs. At the individual level (document-level analysis) one can perform analysis and gain insight into the patient in longitudinal studies. At the group level (collection-level analysis) one can gain insight into the patterns of the groups (network behavior, e.g., assessing the influence within the social group), for example, in a particular disease group, the community of participants in an HMO or hospital setting, or even in the global community of patients (ethnic stratification). The results of these analyses can be generalized. While the blogs enable the formation of social networks of patients and providers, the uniqueness of the health/medical terminology comingle with the subjective vocabulary of the patient compounds the challenge of interpretation. Taking the discussion to a more general level, while blogs have emerged as contemporary modes of communication within a social network context, hardly any research or insight exists in the content analysis of blogs. The blog world is characterized by a lack of particular rules on format, how to post, and the structure of the content itself. Questions arise: How do we make sense of the aggregate content? How does one interpret and generalize? In health blogs in particular, what patterns of diagnosis, treatment, management, and support might emerge from a meta-analysis of a large pool of blog postings? The overall goal, then, is to enhance the quality of health by reducing errors and assisting in clinical decision making.

Additionally, one can reduce the cost of healthcare delivery by the use of these types of advanced health information technology.

Therefore, the objectives of our project include:

1. To use UIMA to mine a set of cancer blog postings from http://www.thecancerblog.com
2. To develop a parsing algorithm and clustering technique for the analysis of cancer blogs
3. To develop a vocabulary and taxonomy of keywords (based on existing medical nomenclature)
4. To build a prototype interface with Eclipse (based on our existing work in the use of Eclipse in the development of an electronic health record system)
5. To contribute to social networks in the semantic web by generalizing the models from cancer blogs

The following levels of development are envisaged:

- First level: Patterns of symptoms, management (diagnosis/treatment)
- Second level: Glean insight into disease management at individual/group levels
- Third level: Clinical decision support (e.g., generalization of patterns, syntactic to semantic)

Typically, the unstructured information in blogs comprises:

- Blog topic (posting)—What issue or question does the blogger (and comments) discuss?
  - Disease and treatment (not limited to)—What cancer type and treatment (other issues) are identified and discussed?
- Other information—What other related topics are discussed? What links are provided?

What Can We Learn from Blog Postings?

Unstructured information related to blog postings (bloggers), including responses/comments, can provide insight into “diseases” (cancer), “treatment” (e.g., alternative medicine, therapy), support links, etc.
1. What are the most common issues patients have (bloggers/responses)?
2. What are the cancer types (conditions) most discussed? Why?
3. What therapies and treatments are being discussed? What medical and nonmedical information is provided?
4. Which blogs and bloggers are doing a good job of providing relevant and correct information?
5. What are the major motivations for the postings (comments)? Profession (e.g., doctor) or patient?
6. What are the emerging trends in disease (symptoms), treatment and therapy (e.g., alternative medicine), support systems, and information sources (links, clinical trials)?

What Are the Phases and Milestones?

This project envisions the use of UIMA and supporting plug-ins to develop an application tool to analyze health-related blogs. The project is scoped to content analysis of the domain of cancer blogs at http://www.thecancerblog.com. Additional open-source plug-ins and an Eclipse development environment with Java/Extensible Markup Language plug-ins, limited AJAX capability, and a social network analysis tool such as Apache Agora would provide the desired capabilities. In a typical scenario, the cancer blogs can be stored in an open-source Derby database application.

- Phase 1 involved the collection of blog postings from http://www.thecancerblog.com into a Derby application.
- Phase 2 consisted of the development and configuration of the architecture—and taxonomy, clustering, correlations, keywords.
- Phase 3 entailed the analysis and integration of extracted information in the cancer blogs; preliminary results of initial analysis (e.g., patterns that are identified).
- Phase 4 involved the development of taxonomy.
- Phase 5 proposes to test the mining model and develop the user interface for deployment.

We propose to develop a comprehensive text mining system that integrates several mining techniques, including association and clustering, to effectively organize the blog information and provide decision support in terms of search by keywords.

CONCLUSIONS

The development and application of large repositories of patient-specific clinical, medical, and health data generated during patient encounters in the routine delivery of healthcare was, until recently, limited to static uses of utilization management, quality assurance, and cost management. However, with the focus on reducing medical errors through evidence-based health management, these repositories are being subjected to more sophisticated analyses using data mining techniques. These techniques offer numerous opportunities to perform in-depth analysis of the data to gain new insights into the healthcare process with the resultant decision support for a range of tasks. In the future, we will see not only an increased use of data mining techniques in healthcare but also their integration with health intelligence and health organization strategy. The overall goals include the delivery of quality care with a simultaneous decrease in costs.

REFERENCES


