The Basics of Data Mining in Healthcare
An Annotated Bibliography

Deirdre Crandall
INFO 522: Information Access & Resources
November 28 2012
Introduction and Scope

This bibliography provides an introduction to data mining and how it can be used in the healthcare setting. The majority of the articles focus on the different types of data mining and the challenges of data mining in healthcare. The remainder of the articles focus on diverse health care settings, such as a specialty provider’s office or the intensive care unit, to show how data mining can provide benefits to varying parts of the health care spectrum. Because many countries are ahead of the United States when it comes to data mining, there are articles included published in several different countries. All of the articles have been published between 2001 and 2012 because the technical nature of the topic it is important to look at more current articles.

Description

Data mining is a method of analyzing large datasets that has been used extensively in industries other than healthcare. In the past ten years, due to the explosion in the amount of health care data available, both electronic and other, and the increasingly powerful processing tools available, data mining has become a more common technique in the health care arena. It started in larger healthcare settings such as research facilities and tertiary care hospitals and is now moving into smaller and more specialized areas such as clinical practices and smaller hospital departments. It is likely that the use of data mining in health care will expand as more hospitals and clinical settings fully implement electronic health records (EHR) and data availability continues to expand.
Review of the Literature

There are a wide variety of applications for data mining in the healthcare arena. Although data mining started as a technique primarily used in other industries, it has more commonly been used in healthcare areas in the last ten years. There are significant differences in healthcare data, making it more challenging to collect, store and analyze. Healthcare can mean a clinical or research setting. Literature on data mining in healthcare runs from theoretical applications to current usage to critiques of current applications.

It is important to first generally define data mining. Since data mining is a highly technical topic, most of the literature includes at least a basic overview of how it has evolved and what it is. According to Bellazzi and Zupan (2008), data mining is “the process of selecting, exploring and modeling large amounts of data to discover unknown patterns or relationships which provide a clear and useful result to the data analyst” (p. 82). Data mining is sometimes used to mean the same thing as Knowledge Discovery in Databases (KDD) but is more generally defined as one step in the KDD process. The steps of KDD are most commonly defined as selecting the data, preparing the data, cleaning the data, data mining, interpreting the data and finally evaluating it (A. Berger & C. Berger, 2004, p. 124). These steps are not always done in the same order and sometimes are skipped or done more than once during the same analysis. Harrison (2008, p. 2) also goes on to classify data mining as the technical and statistical component of KDD.

In healthcare, data mining is generally used in one of two ways, either to “gain novel and deep insights and unprecedented understanding of large datasets (often accumulated for operational purposes) which can then be used to support decision
making” (Yoo et al., 2011, p. 2432) or to examine these large datasets for the purpose of finding new correlations between data elements. These two broad categories are generally defined in the literature as supervised or unsupervised learning. With supervised learning the analyst begins the process with a certain hypothesis and uses the results to prove or disprove the hypothesis. With unsupervised learning, the large datasets are analyzed with no prior hypotheses about correlating data to see what novel correlations might become evident.

Frequently mentioned in the literature is the unique nature of medical data and how it results in significant challenges for data mining in healthcare. Because of the continuing implementation of paperless electronic health records (EHR) to record patient data, there are increasingly more data available in a digital format. In some ways, this makes patient data ideal for data mining because the amount of data is too large for other methods of analysis, but it also presents challenges. (Ramon et al., 2007, p. 246). Clinical data is also largely heterogeneous and includes numerical data, image data, and natural language with no structured vocabulary (Cios & Moore, 2002, p.2). Spending time preparing data prior to mining can potentially improve the data mining results (Rajagopalan & Isken, 2001, p. 467).

There are numerous privacy and legal issues associated with healthcare data (Yoo et al., 2011, p. 2433). Because of HIPAA (The Health Insurance Portability and Accountability Act) privacy rules, data used for data mining must be anonymous to preserve patient confidentiality. There is also the possibility of discovering data correlations that have legal ramifications, such as fraud or medical malpractice.

Almost all articles on data mining include some discussion of specific algorithms used for data mining. Within the broad categories of supervised and unsupervised
learning techniques, there are a wide variety of methods used for data mining. A. Berger and C. Berger (2004, p. 127) give a good synopsis of the main algorithms used, primarily classification (frequently through decision trees), estimation and prediction for supervised data mining and clustering, association and feature extraction for unsupervised data. The common thought appears to be that no particular algorithm is better than another and that more complexity does not necessarily produce better results.

Because data mining is a relatively new technique for healthcare, the literature offers some thoughts on data mining versus other data analysis techniques. Obenshain (2004, p. 690) compares a standard database query to data mining by their level of complexity, where a database query would answer a relatively simple question and data mining is used for more complicated queries. There are also many comparisons to statistics in the literature. In most cases, data mining is presented as more flexible than statistics, more useful on extremely large datasets, and less dependent on a strictly mathematical model (Yoo et al., 2011, p. 2432).

Frequently, the articles about data mining are defined by the healthcare setting where they take place. In many cases, the data is from a large hospital database, but more specialized settings include the operating room (Harper, 2005), the smaller clinical setting of an allergy office (Dalan, 2010) and intensive care patients (Ramon et al., 2007).

The majority of the literature on data mining discusses actual scenarios where data mining has been used, but some offer ideas about theoretical future areas for data mining including more integration into smaller clinical practices and even consumer use of data mining to help choose a medical care provider (Reiner, 2011).
Data mining is not only a stand-alone data analysis technique. It can be combined with other things to produce an even better result. For example, many hospitals, especially larger ones, already extensively use simulation modeling for data intensive activities like patient flow and bed capacity. Data mining results can be fed into simulation programs to produce more timely and robust results from simulation software (Isken & Rajagopalan, 2002). Another area that is being explored is combining publicly available data from search engines like Google or Go Web with data mining to predict and define medical symptoms (Dalan, 2010, pp. 174-175).

As healthcare data becomes increasingly more available in an electronic format, the benefits of data mining will continue to increase. Improvements in software tools will aid in this process, as well as, further involvement by medical personnel who help define the requirements for the most useful results.

**Abstract:** “In this article we propose an approach for a decision support system (DSS) based on Knowledge Discovery from Databases (KDD). In such system, user must be involved throughout the decision-making process. In consequence we propose the integration of a Human–Computer Interaction (HCI) model into the development of DSS process based on KDD. The approach we propose is based on two systems development methods—the Unified Process (UP) from Software Engineering and the U model from HCI. In this article, we describe our combined approach (UP/U) and the way we used it to develop a DSS in a medical field.”

**Annotation:** This is a very long and complex article, primarily about Decision Support Systems (DSS), but it details how they would use data mining to feed the DSS. The article uses many abbreviations and acronyms without always defining them the first time they are used, which can be confusing. For example, they discuss nosocomial infections as “NI” before defining what it was. What is useful about this article is its emphasis on involving the user at all points of the process, from preparing the data for data mining to designing reports from the DSS.

**Search Strategy:** Keyword searching

**Database:** Web of Science (Web of Knowledge)

**Search String:** TS=(“data mining” OR kdd OR “knowledge discovery”) AND TS=(healthcare OR “health care”)

**Scholarly/Refereed Status:** Ulrichs lists this as an academic/scholarly journal and it is refereed.

**Abstract:**

“**Background**

The widespread availability of new computational methods and tools for data analysis and predictive modeling requires medical informatics researchers and practitioners to systematically select the most appropriate strategy to cope with clinical prediction problems. In particular, the collection of methods known as ’data mining’ offers methodological and technical solutions to deal with the analysis of medical data and construction of prediction models. A large variety of these methods requires general and simple guidelines that may help practitioners in the appropriate selection of data mining tools, construction and validation of predictive models, along with the dissemination of predictive models within clinical environments.

**Purpose**

The goal of this review is to discuss the extent and role of the research area of predictive data mining and to propose a framework to cope with the problems of constructing, assessing and exploiting data mining models in clinical medicine.

**Methods**

We review the recent relevant work published in the area of predictive data mining in clinical medicine, highlighting critical issues and summarizing the approaches in a set of learned lessons.

**Results**

The paper provides a comprehensive review of the state of the art of predictive data mining in clinical medicine and gives guidelines to carry out data mining studies in this field.

**Conclusions**
Predictive data mining is becoming an essential instrument for researchers and clinical practitioners in medicine. Understanding the main issues underlying these methods and the application of agreed and standardized procedures is mandatory for their deployment and the dissemination of results. Thanks to the integration of molecular and clinical data taking place within genomic medicine, the area has recently not only gained a fresh impulse but also a new set of complex problems it needs to address.”

Annotation: These authors seem to be more strongly involved in molecular biology than clinical applications, so accordingly their data mining perspective is more in-line with those arenas. This article is highly technical, but gives a good comparison of different data mining techniques. Especially useful is the example given comparing Decision Trees and Naïve Bayesian Classifiers because it uses comprehensible data to explain the modeling techniques.

Search Strategy: Author searching

Database: Web of Science (Web of Knowledge)

Search String: AU=BELLAZZI R
And then narrowed by refining to Web of Science category of Medical Informatics only

Scholarly/Refereed Status: Ulrichs lists this as an academic/scholarly journal and it is refereed.


Abstract: “Background: Medicine and biomedical sciences have become data-intensive fields, which, at the same time, enable the application of data-driven approaches and require sophisticated data analysis and data mining methods. Biomedical informatics
The Basics of Data Mining in Healthcare

provides a proper interdisciplinary context to integrate data and knowledge when processing available information, with the aim of giving effective decision-making support in clinics and translational research.

Objectives: To reflect on different perspectives related to the role of data analysis and data mining in biomedical informatics.

Methods: On the occasion of the 50th year of Methods of Information in Medicine a symposium was organized, which reflected on opportunities, challenges and priorities of organizing, representing and analysing data, information and knowledge in biomedicine and health care. The contributions of experts with a variety of backgrounds in the area of biomedical data analysis have been collected as one outcome of this symposium, in order to provide a broad, though coherent, overview of some of the most interesting aspects of the field.

Results: The paper presents sections on data accumulation and data-driven approaches in medical informatics, data and knowledge integration, statistical issues for the evaluation of data mining models, translational bioinformatics and bioinformatics aspects of: genetic epidemiology.

Conclusions: Biomedical informatics represents a natural framework to properly and effectively apply data analysis and data mining methods in a decision-making context. In the future, it will be necessary to preserve the inclusive nature of the field and to foster an increasing sharing of data and methods between researchers.”

Annotation: This article pulls together the opinions of a group of experts in the field of Biomedical Informatics. They each put forth their thoughts about where they think data analysis and data mining will be going in the near future. It is a highly technical and jargon-filled article, mostly focusing on the ongoing role of Biomedical Informatics, but also seems useful as a reference piece for someone interested in comparing different methods of data mining. It also discusses availability of and issues related to public data such as PubMed and the eMerge network, which collects electronic medical record data for genetic research.

Search Strategy: Keyword searching
Database: Web of Science (Web of Knowledge)

Search String: TS=(“data mining” OR kdd OR “knowledge discovery”) AND TS=(healthcare OR “health care”)

Scholarly/Refereed Status: Ulrichs lists this as an academic/scholarly journal and it is refereed.


Abstract: The ability to collect and store data has grown at a dramatic rate in all disciplines over the past two decades. Healthcare has been no exception. The shift toward evidence-based practice and outcomes research presents significant opportunities and challenges to extract meaningful information from massive amounts of clinical data to transform it into the best available knowledge to guide nursing practice. Data mining, a step in the process of Knowledge Discovery in Databases, is a method of unearthing information from large data sets. Built upon statistical analysis, artificial intelligence, and machine learning technologies, data mining can analyze massive amounts of data and provide useful and interesting information about patterns and relationships that exist within the data that might otherwise be missed. As domain experts, nurse researchers are in ideal positions to use this proven technology to transform the information that is available in existing data repositories into useful and understandable knowledge to guide nursing practice and for active interdisciplinary collaboration and research.”

Annotation: Although this article is geared toward the nursing professional, it really is a general overview of data mining techniques and how they are most frequently used in healthcare, either clinical or research settings. It describes data mining techniques without going into too much technical detail, which is useful for the less technically oriented reader. It gives some good examples of how data mining has been used in the
The Basics of Data Mining in Healthcare

healthcare arena and some thoughts about how it might be used more fully in the future.

Search Strategy:  Keyword searching

Database:  Web of Science (Web of Knowledge)

Search String:  
TS=(“data mining” OR kdd OR “knowledge discovery”) AND 
TS=(healthcare OR “health care”)

Scholarly/Refereed Status:  Ulrichs lists this as an academic/scholarly journal.


Abstract: “This article addresses the special features of data mining with medical data. Researchers in other fields may not be aware of the particular constraints and difficulties of the privacy-sensitive, heterogeneous, but voluminous data of medicine. Ethical and legal aspects of medical data mining are discussed, including data ownership, fear of lawsuits, expected benefits, and special administrative issues. The mathematical understanding of estimation and hypothesis formation in medical data may be fundamentally different than those from other data collection activities. Medicine is primarily directed at patient-care activity, and only secondarily as a research resource; almost the only justification for collecting medical data is to benefit the individual patient. Finally, medical data have a special status based upon their applicability to all people; their urgency (including life-or-death); and a moral obligation to be used for beneficial purposes.”

Annotation: This article focuses on what makes data mining of healthcare data unique. The authors describe the six steps of Knowledge Discovery in Databases (KDD), of which data mining is one step. They go on to discuss how the heterogeneous nature of healthcare data can make it more challenging to use as part of data mining. They also
The Basics of Data Mining in Healthcare

provide a detailed discussion of ethical and privacy issues surrounding the use of medical data.

**Search Strategy:** Footnote chasing

**Database:** N/A

**Search String:** Referenced in:

**Scholarly/Refereed Status:** Ulrichs lists this as an academic/scholarly journal and it is refereed.


**Abstract:** “Purpose of review: More data are anticipated from the expected increase in use of electronic health records (EHRs). Upcoming initiatives require reporting of quality measures, meaningful use of clinical decision support, alert systems, and pharmacovigilance – knowledge resulting through use of EHRs. Data mining is a new tool that will help us manage information and derive knowledge from these data, and is a part of evolving new disciplines of informatics and knowledge management.

Recent findings: Studies are reported from smaller clinic data marts to larger repositories and warehouses in various health systems, biomedical registries, and the medical literature on the Internet. Data mining technologies show promise and challenges. Outcome measures as structured data and narrative text can be mined with human assistance and newer automated natural language processing software. Despite advances, the growing diversity of clinic EHRs lack integration and interoperability with Internet-based biomedical databases.
The Basics of Data Mining in Healthcare

Summary: Allergists have the capability to mine clinic EHRs to discover new information, which may be hidden in charts. A central allergy computer can serve not just as a registry but also allows functionalities to enable EHRs' meaningful use. Harmonization of technological and organizational standards will allow seamless use of new natural language processing (NLP) tools and ontologies through a semantic web.”

Annotation: This article offers some unique strategies not seen in the other references. It discusses data mining in the smaller setting of an allergy practice. Because of limitations in natural language processing, the author describes a hybrid data mining/manual data analysis technique. Another interesting aspect was using Web resources in conjunction with data mining, primarily Google and another search engine called Go Web.

Search Strategy: Keyword searching

Database: Web of Science (Web of Knowledge)

Search String: TS=(“data mining” OR kdd OR “knowledge discovery”) AND TS=(healthcare OR “health care”)

Scholarly/Refereed Status: Ulrichs lists this as an academic/scholarly journal.


Abstract: “Variability and uncertainty are inherent characteristics of most health care processes. Patient pathways and dwelling times even within the same process typically vary from patient to patient, such as the flow of patients through a particular health care provider or patient progression through the natural history of a given disease. The challenge for the OR modeller is to adequately handle and capture the stochastic features within developed models. This paper will discuss the benefits of combining
patient classification tools (data mining techniques) with developed OR models, such as simulation tools, to more accurately capture patient outcomes, risks and resource needs. Illustrative applications will demonstrate the approach.”

**Annotation:** This article addresses a very specific application of data mining where the data derived from decision tree data mining of patient data is used to classify patients into homogeneous groups and is then fed into an existing simulation model to provide facility capacity information. It is useful because it addresses one particular type of data mining, decision trees, and how data mining can be used in conjunction with other tools that might be available, in this case a simulation modeling package.

**Search Strategy:** Keyword searching

**Database:** Web of Science (Web of Knowledge)

**Search String:**
TS=(“data mining” OR kdd OR “knowledge discovery”) AND
TS=(healthcare OR “health care”)

**Scholarly/Refereed Status:** Ulrichs lists this as an academic/scholarly journal and it is refereed.


**Abstract:** “The increasing volume of medical data online, including laboratory data, represents a substantial resource that can provide a foundation for improved understanding of disease presentation, response to therapy, and health care delivery processes. Data mining supports these goals by providing a set of techniques designed to discover similarities and relationships between data elements in large data sets. Currently, medical data have several characteristics that increase the difficulty of applying these techniques, although there have been notable medical data mining
successes. Future developments in integrated medical data repositories, standardized data representation, and guidelines for the appropriate research use of medical data will decrease the barriers to mining projects.”

**Annotation:** This article is an introductory article and gives a good general overview of data mining. The language is largely non-technical so it is a good choice for someone looking to understand data mining in the healthcare setting, but not requiring a lot of depth. It also gives references to other sources for deeper investigation into various topics related to data mining.

**Search Strategy:** Keyword searching

**Database:** Web of Science (Web of Knowledge)

**Search String:**

TS=“data mining” OR kdd OR “knowledge discovery”) AND TS=(healthcare OR “health care”)

**Scholarly/Refereed Status:** Ulrichs lists this as an academic/scholarly journal and it is refereed.


**Abstract:** “Spiraling health care costs in the United States are driving institutions to continually address the challenge of optimizing the use of scarce resources. One of the first steps towards optimizing resources is to utilize capacity, effectively. For hospital capacity planning problems such as allocation of inpatient beds, computer simulation is often the method of choice. One of the more difficult aspects of using simulation models for such studies is the creation of a manageable set of patient types to include in the model. The objective of this paper is to demonstrate the potential of using data mining
techniques, specifically, clustering techniques such as K-means, to help guide the
development of patient type definitions for purposes of building computer simulation or
analytical models of patient flows in hospitals. Using data from a hospital in the
Midwest this study brings forth several important issues that researchers need to
address when applying clustering techniques in general and specifically to hospital
data.”

**Annotation:** This article is co-authored and uses the same data as this article:
Rajagopalan, B. (2001). Exploiting data preparation to enhance mining and
knowledge discovery. *IEEE Transactions on Systems, Man and Cybernetics, Part C: 
Applications and Reviews*, 31(4), 460-467.
While the previous article explores the data preparation aspect, this article focuses on
using the data derived from data mining to run simulation models. Simulation modeling
has been widely used for healthcare applications, especially patient flow. The article
goes into depth about the particular data mining technique used called K-means cluster
analysis.

**Search Strategy:** Keyword searching

**Database:** Web of Science (Web of Knowledge)

**Search String:** TS=(“data mining” OR kdd OR “knowledge discovery”) AND
TS=(healthcare OR “health care”)

**Scholarly/Refereed Status:** Ulrichs lists this as an academic/scholarly journal
and it is refereed.

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Miller, W. G., … Knaus,
Abstract: “Clinical repositories containing large amounts of biological, clinical, and administrative data are increasingly becoming available as health care systems integrate patient information for research and utilization objectives. To investigate the potential value of searching these databases for novel insights, we applied a new data mining approach, HealthMiner®, to a large cohort of 667,000 inpatient and outpatient digital records from an academic medical system. HealthMiner® approaches knowledge discovery using three unsupervised methods: CliniMiner®, Predictive Analysis, and Pattern Discovery. The initial results from this study suggest that these approaches have the potential to expand research capabilities through identification of potentially novel clinical disease associations.”

Annotation: This article does evaluate a vendor system for data mining, which is not the intent of this bibliography, but it provides some good information about unsupervised data mining. Unsupervised data mining is where large volumes of data are analyzed with no a priori assumptions (assumptions that are formulated beforehand) to potentially find new relationships among the data. Many of the other articles focus more on supervised data mining.

Search Strategy: Keyword searching

Database: Library, Information Science & Technology Abstracts (Proquest)

Search String: (data min? OR kdd OR knowledge discovery) AND (healthcare OR health care OR health), narrowed by Scholarly (peer-reviewed) Journals.

Scholarly/Refereed Status: Ulrichs lists this as an academic/scholarly journal and it is refereed.

Abstract: “A high-level introduction to data mining as it relates to surveillance of healthcare data is presented. Data mining is compared with traditional statistics, some advantages of automated data systems are identified, and some data mining strategies and algorithms are described. A concrete example illustrates steps involved in the data mining process, and three successful data mining applications in the healthcare arena are described.”

Annotation: The author of this article is a consultant and it seems like she was promoting a product called SAS Enterprise Miner, to certain extent. In general though, I felt like this article provides a clear and concise overview of data mining in healthcare. It includes an excellent chart showing different modeling objectives and the different types of data mining that could be used to achieve the objectives and whether the data mining was supervised or unsupervised. It also gives a useful characterization of how data mining can be used to extract more complex information from healthcare data than a typical database query. The article also stresses the need for statisticians and data miners to work together to get the best results in a healthcare setting.

Search Strategy: Footnote chasing

Database: N/A


Scholarly/Refereed Status: Ulrichs lists this as an academic/scholarly journal and it is refereed.

Rajagopalan, B. (2001). Exploiting data preparation to enhance mining and knowledge

**Abstract:** “One of the major obstacles to using organizational data for mining and knowledge discovery is that, in most cases, it is not amenable for mining in its natural form. Using a data set from a large tertiary-care hospital, we provide strong empirical evidence that data enhancement by the introduction of new attributes, along with judicious aggregation of existing attributes, results in higher-quality knowledge discovery. Interestingly, we also found that there is a differential impact of data set enhancements on the performance of different data mining algorithms. We define and use several measures, including entropy, rule complexity and resonance, to evaluate the quality and usefulness of the knowledge discovered.”

**Annotation:** This article is highly technical, but clearly written and easy to understand. It focuses on how the data used for data mining can benefit from prior preparation. They offer three examples with increasing levels of preparation and how the highest level leads to the most useful results. The data they use is from a large tertiary care hospital, illustrating the use of data mining techniques on healthcare data. The article stresses the need for people knowledgeable about the processes being studied to be involved in data preparation and analysis to produce the best end results.

**Search Strategy:** Keyword searching

**Database:** Web of Science (Web of Knowledge)

**Search String:** TS=(“data mining” OR kdd OR “knowledge discovery”) AND TS=(healthcare OR “health care”)

**Scholarly/Refereed Status:** Ulrichs lists this as an academic/scholarly journal.
Abstract: “In this paper we describe the application of data mining methods for predicting the evolution of patients in an intensive care unit. We discuss the importance of such methods for health care and other application domains of engineering. We argue that this problem is an important but challenging one for the current state of the art data mining methods and explain what improvements on current methods would be useful. We present a promising study on a preliminary data set that demonstrates some of the possibilities in this area.”

Annotation: This article assumes a certain level of knowledge about data mining techniques. Its focus is on what is currently being done related to data mining for ICU (intensive care unit) patients and the direction that future data mining might take in that area. This article discusses, as do many of the other articles, the challenges presented by healthcare data. One very interesting facet was rather than using data collected and doing a retrospective study of it, the authors suggest using data collected on individual patients during their ICU stay to make clinical decisions during their time in the ICU.

Search Strategy: Keyword searching

Database: Web of Science (Web of Knowledge)

Search String: TS="data mining" OR kdd OR "knowledge discovery" AND TS=healthcare OR "health care"

Scholarly/Refereed Status: Ulrichs lists this as an academic/scholarly journal and it is refereed.

Abstract: “Commoditization pressures in medicine have risked transforming service provider selection from “survival of the fittest” to “survival of the cheapest.” Quality- and safety-oriented mandates by the Institute of Medicine have led to the creation of a number of data-driven quality-centric initiatives including Pay for Performance and Evidence-Based Medicine. A synergistic approach to creating quantitative accountability in medical service delivery is through the creation of consumer-oriented performance metrics which provide patients with objective data related to individual service provider quality, safety, cost-efficacy, efficiency, and customer service. These performance metrics could in turn be customized to the individual preferences and healthcare needs of each individual patient, thereby providing an objective methodology for service provider selection while empowering health care consumers.”

Annotation: This article is theoretical in nature, but it proposes an interesting application of data mining to help consumers choose a care provider. In this particular case, the author discusses the selection of a radiologist, but it could be applied to other general and specialty healthcare providers. There are few technical details, which make it easy to read, but since there are only ideas offered, it might not be feasible to implement.

Search Strategy: Keyword searching

Database: Web of Science (Web of Knowledge)

Search String: TS=(“data mining” OR kdd OR “knowledge discovery”) AND TS=(healthcare OR “health care”)

Scholarly/Refereed Status: Ulrichs lists this as an academic/scholarly journal and it is refereed.

associations between medications, laboratory results and problems. *Journal of Biomedical Informatics*, 43(6), 891-901.

**Abstract:** “Background The patient problem list is an important component of clinical medicine. The problem list enables decision support and quality measurement, and evidence suggests that patients with accurate and complete problem lists may have better outcomes. However, the problem list is often incomplete.

Objective To determine whether association rule mining, a data mining technique, has utility for identifying associations between medications, laboratory results and problems. Such associations may be useful for identifying probable gaps in the problem list.

Design. Association rule mining was performed on structured electronic health record data for a sample of 100,000 patients receiving care at the Brigham and Women's Hospital, Boston, MA. The dataset included 272,749 coded problems, 442,658 medications and 11,801,068 laboratory results.

Measurements. Candidate medication-problem and laboratory-problem associations were generated using support, confidence, chi square, Interest, and conviction statistics. High-scoring candidate pairs were compared to a gold standard, the Lexi-Comp drug reference database for medications and Mosby's Diagnostic and Laboratory Test Reference for laboratory results.

Results We were able to successfully identify a large number of clinically accurate associations. A high proportion of high-scoring associations were adjudged clinically accurate when evaluated against the gold standard (89.2% for medications with the best-performing statistic, chi square, and 55.6% for laboratory results using interest).

Conclusion. Association rule mining appears to be a useful tool for identifying clinically accurate associations between medications, laboratory results and problems and has several important advantages over alternative knowledge-based approaches.”
Annotation: This article focuses on two data mining techniques that seem to be less frequently used with healthcare data, frequent set mining and association rule mining. Interestingly, association rule mining is what Amazon uses to recommend books, based on purchase history. The article also illustrates a supervised learning technique where they already knew what they were looking for (an association between lab result problem lists and medication problem lists) and set up a data mining process to explore the relationship. There is also a discussion of data mining versus other techniques to get the same information.

Search Strategy: Citation

Database: Web of Science (Web of Knowledge)

Search String: Using this reference in Web of Science:

I clicked on “view all 33 citing references” and then I narrowed Document Type to “articles”.

Scholarly/Refereed Status: Ulrichs lists this as an academic/scholarly journal and it is refereed.


Abstract: “As a new concept that emerged in the middle of 1990’s, data mining can help researchers gain both novel and deep insights and can facilitate unprecedented understanding of large biomedical datasets. Data mining can uncover new biomedical
and healthcare knowledge for clinical and administrative decision making as well as generate scientific hypotheses from large experimental data, clinical databases, and/or biomedical literature. This review first introduces data mining in general (e.g., the background, definition, and process of data mining), discusses the major differences between statistics and data mining and then speaks to the uniqueness of data mining in the biomedical and healthcare fields. A brief summarization of various data mining algorithms used for classification, clustering, and association as well as their respective advantages and drawbacks is also presented. Suggested guidelines on how to use data mining algorithms in each area of classification, clustering, and association are offered along with three examples of how data mining has been used in the healthcare industry. Given the successful application of data mining by health related organizations that has helped to predict health insurance fraud and under-diagnosed patients, and identify and classify at-risk people in terms of health with the goal of reducing healthcare cost, we introduce how data mining technologies (in each area of classification, clustering, and association) have been used for a multitude of purposes, including research in the biomedical and healthcare fields. A discussion of the technologies available to enable the prediction of healthcare costs (including length of hospital stay), disease diagnosis and prognosis, and the discovery of hidden biomedical and healthcare patterns from related databases is offered along with a discussion of the use of data mining to discover such relationships as those between health conditions and a disease, relationships among diseases, and relationships among drugs. The article concludes with a discussion of the problems that hamper the clinical use of data mining by health professionals.”

Annotation: This is a very in-depth and current look at data mining. It examines every aspect of data mining in healthcare and would be a useful resource for any type of healthcare entity, from a small practice to a large research group because it discusses the specifics of different data mining techniques, limitations of each technique, and then how they can be used in different areas.

Search Strategy:  Keyword searching

Database:  Web of Science (Web of Knowledge)
Search String: TS=(“data mining” OR kdd OR “knowledge discovery”) AND TS=(healthcare OR “health care”)

Scholarly/Refereed Status: Ulrichs lists this as scholarly journal that is refereed. It is abstracted and indexed and is written for an audience of professionals in the medical field.

Personal Statement
I did not know much about data mining before I started this paper. I have a background in engineering and information technology, so I thought it would provide a nice meshing between things I already know about and what I am learning about in my library science courses. I discovered that it is a very technical subject. Many of the articles I reviewed either went into a level of depth that I did not understand, especially related to statistical analysis, or assumed a certain level of knowledge that I did not possess. The best articles for me either addressed a unique area of data mining or provided an overview, without going into too much depth. However, I know that in the real world, I would be preparing this for someone else, so I think it would be important to decide who my audience would be prior to choosing references. My guess is that librarians develop specialty areas depending on their client base.

When searching, I did not have trouble finding scholarly articles. I think the technical nature of this topic makes it unlikely that too many popular or non-scholarly articles would be written about it. In fact, as I stated previously, many of the articles were too scholarly for me to understand easily.

I found that I had one excellent search in Web of Science using keywords that yielded many good references, but for the purpose of this project I needed to try additional methods of searching. It felt a little artificial and when I did do my additional types of searching (author, footnote chasing, and citation), I found many repeats from my Web of Science keyword search.
I found it somewhat difficult to keep track of all of my articles on the computer and be able to continuously reference them, but I didn’t want to print them all out. I ended up making a handwritten list of all of my articles along with a few “tags” (e.g., data limitations, definition, history, algorithms) that I could quickly refer to while writing my literature review. Then I could more easily find the articles that addressed a specific topic.

I wasn’t sure if every reference had to be cited in the review of the literature. Much of the information was repeated in multiple articles and so I generally only cited one article that contained the information.

I learned that Tunisia has French-speaking residents.

I certify that:

- This assignment is entirely my own work.
- I have not quoted the words of any other person from a printed source or website without indicating what has been quoted and providing an appropriate citation.
- I have not submitted this assignment to satisfy the requirements of any other course.

Signature       Deirdre Crandall
Date            December 1 2012