

Introduction to the Mining of Clinical Data

James H. Harrison, Jr, MD, PhD

*Division of Clinical Informatics, Departments of Public Health Sciences and Pathology,
University of Virginia, Suite 3181 West Complex, 1335 Hospital Drive,
Charlottesville, VA 22908-0717, USA*

The progressive increase in the amount of clinical data stored in electronic form is—for the first time in history—making it possible to carry out large-scale studies that focus on the interaction between genotype, phenotype, and disease at a population level. Such studies have extraordinary potential to determine the effectiveness of treatment and monitoring strategies, identify subpopulations at risk for disease, define the real variability in the natural history of disease and comorbidities, discover rational bases for targeting therapies to particular patients, and determine the incidence and contexts of unwanted health care outcomes. Matching patient responses (phenotype) with gene expression and known metabolic pathway relationships across large numbers of individuals may be the best hope for understanding the complex interplay between multiple genes and the environment that underlies some of the most common and debilitating health problems [1]. Although serious issues remain to be resolved before the large-scale secondary use of health data for research can become routine, this topic has been recognized and identified as a national priority in Canada [2] and the United States [3]. Clinical laboratory databases contain perhaps the largest available collection of structured medical data representing human phenotypes of disease progression and response to therapy. Alone and especially in combination with other clinical and environmental data, laboratory databases have substantial value for translational research, including correlative studies linking gene expression with phenotype, and for identifying groups of patients with similar characteristics for follow-up analysis or inclusion in clinical studies.

Large-scale clinical databases permit targeted observational and correlative studies that complement randomized clinical trials [4,5]. These databases also hold the promise of more comprehensive analyses to reveal

E-mail address: james.harrison@virginia.edu

unknown, useful real-world relationships among clinical data. The data volume and comprehensiveness that make these data sets useful, however, also make them difficult or impossible to analyze by manual or traditional statistical methods. Analogous challenges have occurred previously in other domains, including the need to identify purchasing associations among billions of retail transactions [6], the need to identify similarities in patterns among terabytes of geologic data for oil exploration [7], and the need to identify patterns in planetary mapping data [8], among many other examples. These needs have been addressed using a set of techniques from the machine learning and pattern recognition fields collectively referred to as “data mining.” In recent years, biomedical science has also begun to apply these techniques to large-scale data analysis, as evidenced by the dramatic increase in biomedical publications referring to data mining over the past 10 years (Fig. 1).

Brief overview of data mining

Data mining has been described as the “extraction of implicit, previously unknown and potentially useful information” [9], such as associations and correlations between data elements, from large repositories of data. It is the technical and statistical component of the process of “knowledge discovery in databases” (KDD, Refs. [10,11]), which has a primary goal of identifying useful new information and is sometimes used synonymously with KDD. Although the data mining label is sometimes also applied to techniques designed to determine whether and to what extent prespecified patterns exist in data sets, those primarily data querying methods are distinct

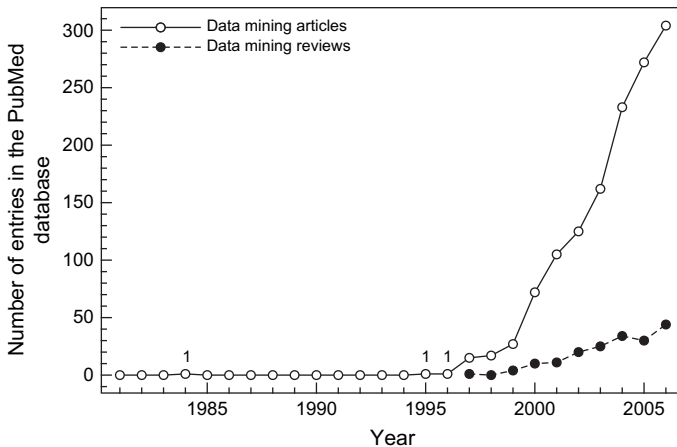


Fig. 1. Dramatic increase in articles and reviews in biomedical science mentioning data mining since 1981. The first article appeared in 1984, and single articles also appeared in 1995 and 1996. The first review appeared in 1997. Articles and reviews include clinical research topics and studies using bioinformatics/high-throughput analytic techniques. In 2006 there were 304 articles and 44 reviews mentioning data mining.

from data mining. The latter, in contrast, relies on unsupervised machine learning and statistical techniques [12] to create a catalog of patterns found in data, without prespecification, and to assess their potential usefulness.

Data mining techniques include methods to group data elements that have similar features (pattern discovery, eg, association rules and clustering) and methods to determine predictive and other relationships between groups of data elements (model building, eg, regression techniques, classification, neural networks, and support vector machines, among others) [12]. For best results, these techniques require consistent, accurate data that are appropriately structured, and thus they are often applied in the setting of integrated data repositories or warehouses designed for mining. Mining typically yields a large number of patterns and relationships, some of which are previously known or trivial. Ultimately, evaluation of the usefulness of these patterns requires human expertise, although statistical and heuristic methods are important for prioritizing, or “pruning,” found pattern sets to yield a subset most likely to be of interest that is small enough to evaluate manually. Data mining is often performed in settings in which the number of features or dimensions (eg, over-expressed genes) contributing to a comparison between items of interest (eg, tumor tissue specimens) is relatively large with respect to the population of items. Because traditional statistical methods are not well-suited to evaluating the probability of coincidental patterns in these high-dimensional data sets, methods specialized for data mining applications, such as false discovery rate calculations, have been developed [12].

This issue presents an initial set of foundational articles in data mining (techniques, tools, databases, and time sequence mining) followed by a set of articles addressing data mining applications in health care. Basic data mining techniques for discovering data patterns and their relationships are introduced in the article by Brown, elsewhere in this issue. A substantial number of high-quality software tools for applying these data mining techniques are available. An overview of open-source data mining tools is provided in the article by Zupan and Demsar, elsewhere in this issue and detailed descriptions of commercial tools are available from vendors, such as SAS [13], SPSS [14], Cognos [15], Insightful [16], and Oracle [17]. Key topics related to aggregation of data for mining are discussed in an article by Lyman and colleagues, elsewhere in this issue and topics related to data handling for specific applications are addressed in later articles. Special considerations related to the developing field of time series mining are discussed in the article by Post and Harrison, elsewhere in this issue. The articles by Siadaty and Harrison, Harrison and Aller, Brossette and Hymel, Klee, and Lee and colleagues, elsewhere in this issue examine data mining applications relevant to laboratory medicine, including pattern detection in laboratory databases in conjunction with knowledge bases, the characteristics and usefulness of several types of regional databases that may incorporate laboratory data, mining microbiology data in infection control, data mining for biomarker discovery, and mining applied to high throughput

genomics data sets. Some foundational topics are also addressed in the later articles in the setting of specific applications, for example pattern pruning and false discovery rate evaluation in the articles by Siadat and Harrison, and Lee and colleagues, elsewhere in this issue.

Although topics in data mining can often be approached in an intuitive manner, data mining methodology is based on mathematic and statistical principles. Effective application of data mining techniques, including effective use of data mining software, requires a reasonable understanding of these principles. A full introduction to the mathematics of data mining is beyond the scope of this brief review issue, but several references are generally available for interested readers. Those who find the mathematic discussions in the articles by Brown, Klee, and Lee and colleagues, elsewhere in this issue too advanced may wish to start their study of data mining by reviewing Tan's *Introduction to Data Mining* [18], which presents data mining topics intuitively using visualizations, or the initial sections of Dunham's *Data Mining: Introductory and Advanced Topics* [19], which takes a bit more mathematic approach. General reviews introducing data mining topics are also available from Fayyad and colleagues [10], Hand and colleagues [20], and Cios and Moore [21].

Special characteristics of medical data

Medical data have characteristics that make them uniquely difficult to analyze in an automated fashion by traditional techniques or by data mining [21]. Some of these characteristics appear in data sets from other domains, but medical data seem to combine more problematic and challenging features at once than almost any other type of data.

High dimensionality

High dimensionality means that many different data elements, each representing a dimension that can vary in value, characterize an item of interest, such as a patient, disease, or specimen. It is not unusual for a patient's medical record to contain 50 or even 100 different types of data elements. With so many variables across a limited number of comparisons, the likelihood of patients sharing coincidental patterns is high and appropriate techniques must be used to minimize identification of spurious patterns.

Heterogeneity

Medical data may include textual descriptions, various types of images, and discrete values using multiple scales. Values may be obtained from multiple methods, some of which may produce incompatible results for the same observation. Data mining requires consistent data, which means that large volumes of clinical data may need to be transformed to compatible representations. Some of these challenges are further addressed in the articles by Siadat and Harrison, and Harrison and Aller, elsewhere in this issue.

Imprecision

Unlike retail transactions, which directly reflect a particular purchase act, medical observations commonly indicate a probability that a condition exists based on their sensitivity and specificity as indicators for that condition. A given feature thus may be consistent with more than one condition, or the condition may exist without the feature. Furthermore, sensitivities and specificities for most features are not known precisely for all data sets and at best may be estimated based on values obtained in other data sets. For these reasons, linking data elements to true characteristics of patients is not straightforward.

Interpretations

Diagnoses and other summary data in medical records are generally human interpretations of aggregates of observations and objective data values. Interpretations by different individuals may differ or even conflict, and are often expressed as text that must be further interpreted to a form that may be processed during data mining.

No canonical form

Although substantial progress continues to be made in developing standard medical terminologies, in the absence of a generally accepted representation for important medical concepts many clinical data are still expressed in idiosyncratic ways.

Incomplete and inconsistent data

Patients who have the same conditions may have substantially different types and timing of observations, unlike retail transactions, surveys, or technical data, which generally have comparable complements of data elements obtained at similar times. Clinical data may be inconsistent or conflicting for various reasons. These qualities add noise and spurious patterns that increase the difficulty of identifying real patterns of interest.

Difficult mathematic characterization

Medicine deals with concepts, such as inflammation, comorbidities, and disease severity, that strongly influence clinical outcome but are difficult to quantitate and incorporate into mathematic relationships with diagnoses and disease progression models.

Temporal patterns

Data elements in clinical records may not be meaningful outside of a particular temporal context. This phenomenon is particularly true for laboratory databases, which are largely composed of time sequences. For

example, a time sequence of viral antigen and antibody values associated with hepatitis might be interpreted differently if it were randomly shuffled. The important information that is implicit in time sequences presents a problem because it is not accessible to data mining methods that catalog associations between individual data elements. The article by Post and Harrison, elsewhere in this issue discusses methods for abstracting temporal data sequences to meaningful new data elements that can be used by data mining tools.

Need for generalization

The need to understand and generalize information about the mechanism and causation of relationships in medical data makes data mining techniques that do not reveal meaningful information about discovered relationships (eg, neural networks) less desirable than those that do (eg, classification trees), thus limiting applicable techniques.

Ethical, legal, and social issues

Important questions remain about the ownership and appropriate use of clinical information, the need to ensure an appropriate balance of risk to patients with benefit to society, and the need to protect patient privacy and confidentiality. Controversy and confusion exist related to these issues that complicate the aggregation and analysis of clinical databases for reasons other than individual patient care. The call for a national framework for the secondary use of health information referenced above [3] is largely a recommendation that society, health care providers, and government resolve these issues so that the community—and ultimately patients—can benefit from techniques such as data mining.

These typical characteristics of medical data do not preclude data mining, but they may necessitate extra data processing and compensatory techniques that substantially complicate mining, increasing effort and costs.

Summary

Mining medical data, including laboratory data, is currently a challenging exercise in data acquisition, aggregation, and reconciliation. Significant political and legal challenges may also exist for medical data mining projects. Partly for these reasons, data mining techniques have not been widely used in laboratory medicine (with a few notable exceptions). Substantial potential exists, however. The last five articles in this issue show that data mining can be applied successfully to clinical care, public health, and research problems. As the volume of data online increases, standard data representations become more widespread, and issues related to secondary analysis of health data are resolved, the cost and effort barriers to data mining projects will decrease. Laboratory data represent a substantial volume of objective,

relatively well-standardized and well-characterized data that directly express patient phenotype in disease presentation and response to therapy. Because of the value of these data, laboratories have the opportunity to contribute to mining projects and to lead projects that particularly target laboratory data. This is an opportunity that should not be missed and for which laboratories can begin to prepare.

References

- [1] Rees J. Complex disease and the new clinical sciences. *Science* 2002;296:698–700.
- [2] Canadian Institute for Health Research. Secondary use of personal information in health research: case studies. 2002. Available at: <http://www.cihr-irsc.gc.ca/e/1475.html>. Accessed August 26, 2007.
- [3] Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American medical informatics association white paper. *J Am Med Inform Assoc* 2007;14(1):1–9.
- [4] Grossman J, Mackenzie FJ. The randomized controlled trial: gold standard, or merely standard? *Perspect Biol Med* 2005;48(4):516–34.
- [5] Jager K, Stel V, Wanner C, et al. The valuable contribution of observational studies to nephrology. *Kidney Int* 2007;72(5):539–42.
- [6] Babcock C. Parallel processing mines retail data. *Computerworld* 1994;28(39):6.
- [7] Harrison D. Backing up 100 terabytes. *Network Computing* 1993;413:98–104.
- [8] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: an overview. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, editors. *Advances in knowledge discovery and data mining*. Menlo Park (CA): AAAI Press; 1996. p. 1–34.
- [9] Lee S, Siau K. A review of data mining techniques. *Industrial Management & Data Systems* 2001;100(1):41–6.
- [10] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Magazine* 1996;17(3):37–54.
- [11] Hipp J, Güntzer U, Nakhaeizadeh G. Data mining of association rules and the process of knowledge discovery in databases. In: Perner P, editor. *Advances in data mining: applications in e-commerce, medicine, and knowledge management*. Berlin (Germany): Springer; 2002. p. 207–26.
- [12] Hand DJ. Principles of data mining. *Drug Saf* 2007;30(7):621–2.
- [13] SAS Institute Inc. SAS Enterprise Miner®. Available at: <http://www.sas.com/technologies/analytics/datamining/miner/index.html>. Accessed August 26, 2007.
- [14] SPSS Inc. Clementine®. Available at: <http://www.spss.com/clementine/>. Accessed August 26, 2007.
- [15] Cognos Inc. Data mining. Available at: <http://www.cognos.com/data-mining.html>. Accessed August 26, 2007.
- [16] Insightful Corp. Insightful Miner®. Available at: <http://www.insightful.com/products/iminer/default.asp>. Accessed August 26, 2007.
- [17] Oracle Corp. Oracle data mining. Available at: <http://www.oracle.com/technology/products/bi/odm/index.html>. Accessed August 26, 2007.
- [18] Tan P, Steinbach M, Kumar V. *Introduction to data mining*. Boston: Addison-Wesley Longman Publishing Co., Inc.; 2005.
- [19] Dunham MH. *Data mining: introductory and advanced topics*. Upper Saddle River, NJ: Prentice Hall; 2002.
- [20] Hand D, Blunt G, Kelly M, et al. Data mining for fun and profit. *Stat Sci* 2000;15(2):111–26.
- [21] Cios KJ, Moore GW. Uniqueness of medical data mining. *Artif Intell Med* 2002;26(1–2): 1–24.