ELSEVIER
SAUNDERS

Clin Lab Med 28 (2008) xi–xiii

CLINICS IN
LABORATORY
MEDICINE

Preface

James H. Harrison, Jr, MD, PhD
Guest Editor

Clinical laboratory data are among the most detailed, objective, reliable, and useful measures of patient characteristics contained in the medical record. Numerous studies over the past 30 years based on laboratory data alone and in aggregate with other clinical and experimental data have revealed correlative and predictive patterns in laboratory data that have improved our understanding of disease, therapeutic response, and health care delivery processes. Additional useful patterns undoubtedly remain hidden in the data, awaiting discovery by creative, prepared minds using effective analysis techniques.

Some pathologists have recognized this opportunity; over the past 10 years there have been periodic reports in the literature that have used automated pattern recognition and modeling techniques collectively termed "data mining" to identify patterns in laboratory data for various purposes. Unfortunately, these efforts have been relatively few, whereas the use of data mining techniques in other medical domains has increased dramatically (see article by Dr. Harrison). There are several reasons that the use of data mining techniques has been inhibited in the laboratory. Data mining is a set of statistical approaches to data analysis that are relatively technical and that need to be correctly matched to an analysis task. This specialized knowledge is generally outside the scope of laboratorians' training. Software tools for data mining by non-experts have been very expensive and were often a poor fit for laboratory databases. Most were designed to discover associations

doi:10.1016/j.cll.2007.10.011

between discrete events and data elements in business analyses. As such, these tools are generally better suited to analyzing relationships between diagnosis or procedure codes in other medical domains than recognizing important patterns in the time sequences of data elements that make up much of laboratory databases. Various political and legal forces have prevented laboratories and other care providers from collaboratively building the large data sets that optimally support data mining. Finally, dedication of effort and resources to a pattern discovery project can be difficult when the outcome is (by definition) unpredictable at the time of the investment decision.

This situation is changing. Data mining techniques are becoming more widely known, particularly those that are associated with high throughput genomics and proteomics analyses (see articles by Drs. Klee and Lee). Medical informaticians, who can be familiar (if not expert) with data mining are also more generally available within pathology practices or as local collaborators. High-quality open source software for data mining is available that is appropriate for use by non-statisticians (see article by Drs. Zupan and Demsar). Techniques to incorporate time series databases into data mining analyses are being developed (see article by Drs. Post and Harrison). Efforts are underway to create a societal and governmental consensus for the secondary analysis of health care information (see article by Dr. Harrison). One of the most important developments promoting data mining in the mainstream, however, is the coming convergence and correlation of genomic and proteomic data with data representing patient phenotype (see article by Dr. Harrison). These studies will use data mining techniques and will make data mining approaches and tools broadly available for application to clinical data. Because laboratory data present a high-quality, reliable representation of patient phenotype, it will be of substantial interest for aggregation with high throughput genomics and proteomics data. Laboratorians will have opportunities to contribute to, or lead, parts of this work.

Our intent in assembling this issue is to provide an introduction to standard techniques for managing and mining clinical data and to illustrate these techniques with several applications related to laboratory medicine and associated research. The issue is divided into a foundations section, which provides a discussion of data mining techniques and tools, data warehousing, and time series analysis, and an applications section that presents a set of projects that illustrate data aggregation, detection of interesting and unusual patterns in laboratory data, infectious disease surveillance, and discovery of patterns indicating new biomarkers and gene expression profiles. Balancing the level of complexity in introductory material is always challenging; we present the statistical discussions at a moderate level so as to be useful to readers who have some familiarity with statistical concepts, and provide references to additional materials appropriate for novices (see article by Dr. Harrison) or experts. We hope that this issue is useful in raising the interest in data mining in the laboratory community and providing

a guide to the types of clinical and research opportunities that will become available over the next several years.

James H. Harrison, Jr, MD, PhD
Departments of Public Health Sciences and Pathology
University of Virginia
Hospital West Complex 3181
PO Box 800717
Charlottesville, VA 22908-0717, USA

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