Clin Chem Lab Med 2007;45(6):749–752 © 2007 by Walter de Gruyter • Berlin • New York. DOI 10.1515/CCLM.2007.177

Review

Reduction of multi-dimensional laboratory data to a two-dimensional plot: a novel technique for the identification of laboratory error

Steven C. Kazmierczak[1,*], Todd K. Leen[2], Deniz Erdogmus[2] and Miguel A. Carreira-Perpinan[2]

[1] Department of Pathology, Oregon Health & Science University, Portland, OR, USA
[2] Department of Computer Science and Engineering, OGI School of Science and Engineering, Portland, OR, USA

Abstract

Background: The clinical laboratory generates large amounts of patient-specific data. Detection of errors that arise during pre-analytical, analytical, and post-analytical processes is difficult. We performed a pilot study, utilizing a multidimensional data reduction technique, to assess the utility of this method for identifying errors in laboratory data.
Methods: We evaluated 13,670 individual patient records collected over a 2-month period from hospital inpatients and outpatients. We utilized those patient records that contained a complete set of 14 different biochemical analytes. We used two-dimensional generative topographic mapping to project the 14-dimensional record to a two-dimensional space.
Results and conclusions: The use of a two-dimensional generative topographic mapping technique to plot multi-analyte patient data as a two-dimensional graph allows for the rapid identification of potentially anomalous data. Although we performed a retrospective analysis, this technique has the benefit of being able to assess laboratory-generated data in real time, allowing for the rapid identification and correction of anomalous data before they are released to the physician. In addition, serial laboratory multi-analyte data for an individual patient can also be plotted as a two-dimensional plot. This tool might also be useful for assessing patient wellbeing and prognosis.
Clin Chem Lab Med 2007;45:749–52.

Keywords: data reduction techniques; error detection; laboratory error; serial data analysis.

Introduction

The clinical laboratory generates vast amounts of patient-specific biochemical data that represent an invaluable resource for error detection. Merging these biochemical data with other patient-specific information, such as demographic data and clinical information, and subsequent evaluation may lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management, and help in evaluating the likelihood that the data generated are accurate.

Data mining, sometimes referred to as Knowledge Discovery in Databases, is the search for relationships and global patterns that exist in large databases, but are hidden and not obvious due to the vast amount of data (1). The application of non-hypothesis-driven data-mining approaches to high-dimensional medical information assumes no prior knowledge about relationships among the data (2). This technique therefore has the potential for uncovering previously unknown relationships between the data.

The data mining process typically involves collecting data into a data warehouse, filtering the data to remove errors and checking for consistency of format, and then searching the data using statistical queries, neural networks, or other machine learning methods (3).

At present, the major applications of data mining have been in the commercial sector for use by companies with a strong consumer focus, such as retail, financial, and marketing organizations. Data mining enables companies to determine relationships among factors such as sales, costs, and inventory in order to forecast future sales and industry trends. These associations or relationships provide valuable information about historical patterns and may help in predicting future trends.

The utilization of large repositories of patient-specific biological and clinical data generated during the routine delivery of medical care has historically been limited to utilization management and for quality assurance purposes (2). Data mining techniques have rarely been applied to laboratory medicine. However, during the past 25 years, laboratory information systems have evolved into systems capable of collecting, organizing, and storing vast quantities of data. Electronic capture of a wide range of clinical and laboratory data will allow for the creation of large-scale data repositories with the potential to generate new insights into relationships between data. Associations among a wide repository of clinical and biological data with no prior assumptions can facilitate the generation and investigation of new hypotheses.

In the current pilot study, we screened a large database of biochemical data. Our goals were to develop and apply multivariate statistical data modeling and

*Corresponding author: Dr. Steven Kazmierczak, Department of Pathology, Oregon Health & Science University, Mailcode L-471, Portland, OR 97239, USA
Phone: +1-503-494-4208, Fax: +1-503-494-8148,
E-mail: kazmierc@ohsu.edu

outlier detection techniques to help identify errors in clinical laboratory test results. In this approach, historical data are used to build models of the data corresponding to valid or correct measurements. New data points with high probability under the model are deemed valid, while data points with low probability under the model are flagged as outliers. This modeling approach allows instantaneous identification of statistically plausible and statistically unlikely data. Thus, this approach has several distinct advantages compared to rule-based error detection systems. Multivariate modeling has significantly more power to detect anomalies in data compared with univariate techniques.

Materials and methods

All data used in this study were obtained from a proprietary data warehouse located at the Oregon Health & Science University Hospital, a 500-bed tertiary-care academic medical center in Portland, Oregon. Approximately 4.2 million tests are performed annually in the clinical chemistry and hematology laboratories. The current database includes information for up to 30 discrete analytes measured with two Beckman LX20 analyzers (Beckman Instruments, Brea, CA, USA) and five discrete hematological analytes measured with two Coulter GenS instruments (Beckman Instruments). The current database for biochemical analytes contains data from January 2004, while the database of hematological analytes includes data from September 2005. Information collected in the data warehouse includes patient demographics, date and time at which patient specimens were collected, analyzed, and reported, laboratory test results, and clinical findings. The study complied with all Health Insurance, Portability, and Accountability (HIPAA) regulations, which were designed to protect the privacy of personal health information (4).

For the purposes of this pilot study, we restricted our analysis to a 2-month period (April–May 2006). Data for both inpatients and outpatients were collected and compiled into a spreadsheet format. One row of 30 columns was generated for each unique patient sample that was collected and analyzed in the laboratory. Date and time of collection captured for each patient specimen allowed us to evaluate serial data for individual patients.

While it is possible to train the data reduction model using records with missing variables, for this study we selected a subset of data with no missing data. To do this, we selected 14 biochemical analytes (out of the total of 20+) so that the resulting dataset would contain as many complete-data records as possible. The resulting dataset contained 13,670 records, each a 14-dimensional vector. In order to give equal importance to each analyte, each analyte was normalized by its standard deviation.

Table 1 Analytes evaluated in the current study.

Glucose	Sodium
Potassium	Chloride
Total CO_2	Urea nitrogen
Creatinine	Calcium
Total protein	Albumin
Total bilirubin	ALP
AST	ALT

The entire dataset was modeled using a two-dimensional generative topographic mapping (GTM) using MatLab software (The MathWorks Inc., Natick, MA, USA) (5). In GTM, the 2D latent space is mapped by a nonlinear function into the 14-dimensional space, yielding a curved 2-dimensional submanifold in the 14-dimensional space. To project a 14-dimensional record to 2-dimensions, we used the posterior mean in latent space provided by GTM.

Results

We evaluated 13,670 individual data records collected from hospital inpatients and outpatients, collected over a 2-month period. While it is possible to train the models using data records with missing data, for this pilot study the data we evaluated contained no missing records. The 14 analytes contained in each individual patient record are listed in Table 1.

Figure 1 shows two-dimensional projections of various subsets of the whole data set. Although the two-dimensional space corresponds to an abstract representation of the data intended to mimic its distribution in 14-dimensional space, the method does show significant visual structure of the data. For example, Figure 1A, corresponding to 1000 patient data sets with normal values for all 14 analytes, occupies a distinct crescent-shaped region in the two-dimensional graph. Likewise, the majority of outpatient samples (Figure 1B) also occupy the same region as that for samples with normal results for all 14 analytes, consistent with the fact that outpatients tend to be healthier than inpatients. The data plotted for hospital inpatients (Figure 1C) occupy the largest region of the graph, consistent with the finding that biochemical data for inpatients generally show a much greater range of variability.

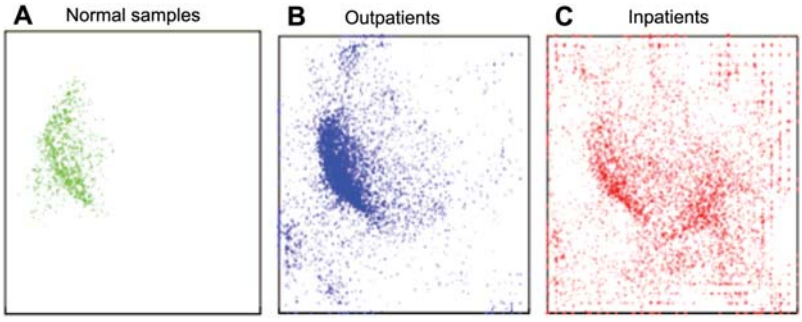

Figure 1 Two-dimensional projection of 14 different biochemical analytes.
(A) Data for 1000 patients with all results within the normal reference interval. (B) Data for hospital outpatients. (C) Data for hospital inpatients.

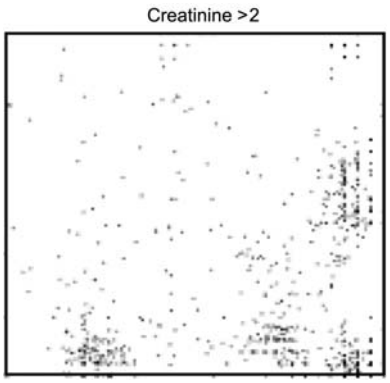

Figure 2 Two-dimensional projection of all patients with serum creatinine >150 μmol/L (2.0 mg/dL).

The inpatient data also reveal other distinct data groupings associated with various disease states. For example, Figure 2 shows the two-dimensional graph for only those patients with renal insufficiency, defined as serum creatinine >150 μmol/L (2.0 mg/dL). Four distinct data groupings emerge when the data for the 14 analytes are plotted as a two-dimensional graph. While patients whose data fall into one of these four distinct regions may indeed have renal insufficiency, of greater interest is whether the individual patient samples falling far from these regions represents error, or some other disease process. The significance of these ''outliers'' is further revealed when data for only those patients with urea nitrogen >18 mmol/L (50 mg/dL) are plotted (Figure 3). This plot mimics the plot for patient samples with creatinine >150 μmol/L (2.0 mg/dL) with two notable exceptions. First, the data grouping in the lower right corner of Figure 3 is not as pronounced as that in Figure 2. Second, and more notable, there are many fewer ''outliers'' in the plot for urea nitrogen compared with the plot for creatinine. This suggests that the creatinine outlier results in Figure 2 may represent faulty data and necessitate retesting to verify the initial result.

Similar two-dimensional projections with generative topographic mapping for patients with glucose >11.1 mmol/L (200 mg/dL) (Figure 4) or those with bilirubin >86 μmol/L (5.0 mg/dL) (Figure 5) also reveal distinct regions in the two-dimensional map

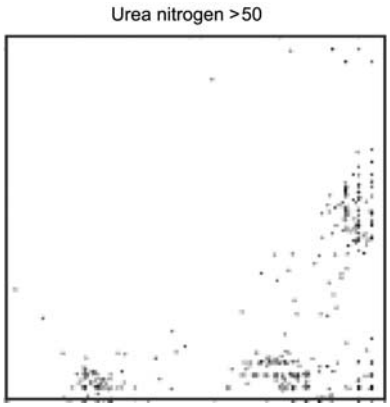

Figure 3 Two-dimensional projection of all patients with serum urea nitrogen >18 mmol/L (50 mg/dL).

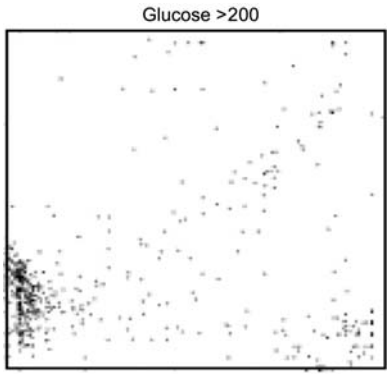

Figure 4 Two-dimensional projection of all patients with serum glucose >11.1 mmol/L (200 mg/dL).

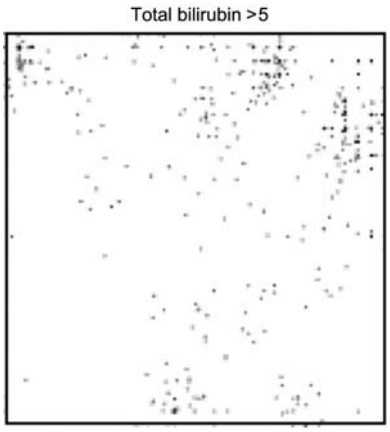

Figure 5 Two-dimensional projection of all patients with serum bilirubin >86 μmol/L (5.0 mg/dL).

that may be used for determining the accuracy of the generated data.

In addition to plotting data for all patients as a means to detect outliers or anomalous data, we also attempted to track serial patient data over time. Figure 6 shows serial results for a single outpatient who had eight sequential samples collected during the 2-month study period. The initial sample collected from this individual (circled point), when plotted on the two-dimensional graph, falls in a region slightly outside that described by samples with all normal data. The clinical course of this individual improved over

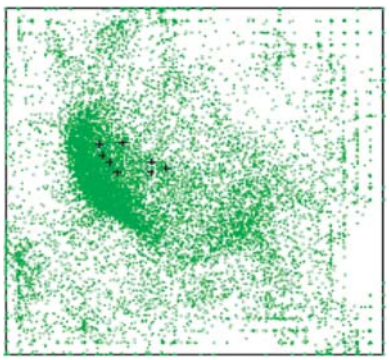

Figure 6 Two-dimensional projection showing serial trajectory of an outpatient superimposed on data generated for all outpatients.
The circled point is the earliest sample.

the subsequent interval, consistent with the trajectory towards the area defined by normal data.

Discussion

Improving the safe care of patients has gained national attention with the publication in 2000 of ''To err is human: building a safer health system'' (6). This report highlighted four major aspects of patient safety: 1) the problem of accidental injury is serious; 2) the cause of injury is usually not careless people, but faulty systems; 3) we need to redesign our systems to prevent patient injury; and 4) patient safety must become a national priority. The importance of clinical laboratories to patient safety is clearly demonstrated by the establishment of the Institute for Quality in Laboratory Medicine in 2003, with the assistance of the Centers for Disease Control and Prevention.

The practice of medicine consists largely of information management, and the clinical laboratory is a major producer of information used to diagnose, treat, and monitor patients. In the intensive care setting, it has been estimated that 40% of all decisions concerning patient management rely on laboratory data (7). The accuracy of patient-derived data generated by the clinical laboratory is a critical component of optimum patient care and patient safety. Surveys on errors in laboratory medicine generally agree that errors occur more frequently in the pre- and post-analytical phases of the testing process, rather than in the analytical testing phase itself (8). The variety of errors and mistakes that can occur in the pre-analytical, analytical, and post-analytical components of testing are almost limitless. Studies designed to evaluate laboratory error typically assess one specific type of error that might occur in the total testing process. This type of focused approach does enable laboratories to assess the extent and possible causes of certain specific types of error, but does little to address other types of errors that may occur.

Use of patient-generated data can help to detect analytical errors and can detect certain types of pre- and post-analytical errors (9). Examples of patient-derived data for error detection include use of delta checks, calculation of average-of-normals, and anion gap calculations for assessment of errors in electrolyte measurements. While all of these procedures have value in error detection, they suffer from several shortcomings. Complex algorithms for error detection are often difficult to implement and are therefore infrequently used. Current rule-based systems tend to be overly simplistic and typically catch only the most blatant of errors. In addition, these systems typically evaluate, at most, the interdependence of only three or four analytes at a single time, even though individuals typically have up to 20 or more discrete biochemical and hematological data values generated from a single phlebotomy procedure that could be evaluated for interdependence (9–11).

There are numerous statistical approaches that can be used for outlier (i.e., error) detection in multivariate data sets. In each method, the technique basically determines a probability density model for the valid measurements and identifies analyte values that have a probability lower than some threshold, and are therefore likely to be in error. The technique of dimensionality reduction is traditionally referred to in geometric terms in which data in multi-dimensional space are reduced to a lower-dimensional visual context. In our study, we reduced 14 different analyte values to a two-dimensional graphical format. The results of our pilot study suggest that this data reduction technique can be a very powerful tool for assessing the validity of laboratory data generated. Abnormalities in specific analytes representative of various pathophysiological processes can be readily identified with this technique. Data that do not fit a specific pathophysiological process, such as the outlier data in the plot of increased creatinine in Figure 2, are readily apparent and may indicate the need for further testing or follow-up. In addition to the identification of potential errors in individual samples, plotting the serial trajectory for an individual patient over time, such as that shown in Figure 6, could be useful for tracking therapeutic responses over time. Comparing an individual's trajectory with known examples indicative of favorable outcomes may provide a useful tool for assessing patient wellbeing and prognosis. Further studies are under way to more fully assess the potential of this data reduction technique for identifying anomalous laboratory data.

References

1. Hand D, Mannila H, Smyth P. Principles of data mining. Cambridge, MA: MIT Press, 2001.
2. Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG, et al. Data mining and clinical data repositories: insights from a 667,000 patient data set. Comp Biol Med 2006;36:1351–77.
3. Krivda CD. Data-mining dynamite. Byte 1995;95:97–102.
4. http://www.os.dhhs.gov/ocr/hipaa/. Accessed April 4, 2007.
5. Bishop CM, Svensen M, Williams CK. GTM: the generative topographic mapping. Neural Comput 1998;10: 215–34.
6. Kohn LT, Corrigan JM, Donaldson MS, editors. To err is human: building a safer health system. Washington, DC: National Academy Press, 2000.
7. Nykanen P, Boran H, Pince H, Clarke K, Yearworth M, Williams JL, et al. Interpretative reporting and alarming based on laboratory data. Clin Chim Acta 1993;222:37–48.
8. Plebani M. Errors in clinical laboratories or errors in laboratory medicine? Clin Chem Lab Med 2006;44:750–9.
9. Kazmierczak SC. Laboratory quality control: using patient data to assess analytical performance [review]. Clin Chem Lab Med 2003;41:617–27.
10. Valdiguie PM, Rogari E, Phillippe H. VALAB: expert system for validation of biochemical data. Clin Chem 1992; 38:83–7.
11. Oosterhuis WP, Ulenkate HJ, Goldschmidt HM. Evaluation of labrespond: a new automated validation system for clinical laboratory test results. Clin Chem 2000;46: 1811–7.