34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Statistical Error Detection for Clinical Laboratory Tests

Todd K. Leen[a], Deniz Erdogmus[b] and Steven Kazmierczak[c]

Abstract— **Errors in clinical laboratory tests lead to increased costs and patient risks. Such errors are relatively rare, affecting $\sim 0.5\%$ of samples. Existing techniques for detecting errors have either far too low sensitivity or specificity to be useful. This preliminary study develops statistical sample selection criteria that capture faults upwards of fifty times more efficiently than expected from random sampling. Although this is only the first step towards an integrated discriminant system for reliable detection of laboratory errors, the statistical detection scheme demonstrated here outperforms existing methods.**

I. Introduction

The clinical laboratory is the major producer of information used to diagnose, treat, and monitor patients. Carraro and Plebani [1] estimate that 40% of all decisions concerning intensive care patient management are based on laboratory data. Errors in clinical laboratory tests lead to delays in treatment, or erroneous treatment and clinical investigation, and hence additional costs and increased patient risks.

The accuracy of data generated by the clinical laboratory is critical for optimum patient care, safety, and economy. The actual total costs associated with laboratory error have been difficult to quantify because estimates often do not account for consequences such as patient trauma, emotional and economic costs to the patient's family, and stress on vital organs contributing to co-morbidity and to premature death [2][1].

The ability to accurately identify true laboratory errors, and take the necessary corrective action when such errors are discovered is difficult in the clinical laboratory setting. The large number of samples that are submitted to the clinical laboratory daily, the multitude of analytes measured, and an emphasis on reporting tests results as quickly as possible are not conducive to detection of laboratory errors.

The current dominant technique for error identification uses measurement of quality control materials to identify instrument mis-calibration. These *quality control checks* are inadequate for several reasons. First, they are most sensitive to errors in instrument calibration. The technique is completely blind to errors made during the *pre-analytic phase* that includes sample collection, transport, storage and handling, all before samples reach the analytical instrument. Carraro and Plebani [1][4] estimate that *more than 60% of errors* occur in this pre-analytic phase. Thus, more than one-half of laboratory errors are *in principal* undetectable by the current most common assurance method. Second, quality control test materials are often animal-based with added stabilizers and surfactants and do not react with analytic reagents the same as human samples, leading to possible errors. Third, quality control checks cannot identify errors that are *specific to particular samples* and thus specific to an individual patient's test results [5][6]. Finally, quality control checks are performed infrequently — typically three times a day. If an instrument falls out of calibration between checks, hundreds of test results may be erroneous and those samples need to be re-analyzed, or even re-collected. This impairs critical patient care and increases costs.

A. Using Patient Data to Detect Errors

The limitations of quality control checks for error detection led to the development of techniques based on patient data [5][7]. In principle, such techniques can detect errors in the pre-analytic phase, and they can respond to analytic instrument faults quicker than quality control checks. However the existing methods — delta checks, absurd value checks, and anion gap analysis — have significant shortcomings.

Delta Checks use measurements from two consecutive samples produced within fairly short time intervals. The changes in concentration of the compounds measured (called the *analytes*) are recorded. If these changes exceed established limits (based on maximum expected physiological change between the sample collection times), then the analyte measurement is repeated on *both samples*. If the second measurement set also exceeds the change limit, one or both of the samples are at fault and new samples must be collected. Repeated measurement and collection are costly, and the false alarm rate is so high (low specificity) that detection flags are routinely ignored in practice [8]. Delta checks are inadequate.

Absurd Value Checks are univariate outlier detection tests. Measurement values generally considered to be incompatible with life are flagged as not likely to be correct [5]. These tests do not consider the dependencies between multiple analytes [5][9], and have extremely low sensitivity. Goldschmidt and Lent [10] estimate that up to 75% of laboratory errors produce measurements falling within acceptable univariate reference intervals, and so would remain unflagged by absurd value checks. Efforts to date to use interdependencies between analytes are *rule-based* and typically evaluate only two to four analytes at a time. Unlike probabilistic methods, they

This work was supported by NSF under grant IIS-1118061.

[a]Todd K. Leen (leent@ohsu.edu) is with the department of Biomedical Engineering, Oregon Health & Science University, Portland, OR.

[b]Deniz Erdogmus (erdogmus@ece.neu.edu) is with the department of Electrical and Computer Engineering, Northeastern University, Boston, MA.

[c]Steve Kazmierczak (kazmierc@ohsu.edu) is with the Clinical Chemistry & Toxicology department at Oregon Health & Science University, Portland, OR.

[1] One study estimates that errors in measured total calcium concentrations due to instrument mis-calibration alone cost \$60M to \$199M annually in the United States [3].

978-1-4577-1787-1/12/\$26.00 ©2012 IEEE

are intolerant of missing data values and cannot easily be extended to deal with many analytes, and disease-dependent or treatment-dependent contingencies.

Anion Gap Analysis is based on the concentrations of sodium, chlorine, and total carbon dioxide; if the quantity

$$C_{\mathrm{Na}} - (C_{\mathrm{Cl}} + C_{\mathrm{Total\ C0_2}})$$

is zero or negative, then one of these analytes was incorrectly measured. The test is only sensitive to errors in these analytes.

The study reported here shows that simple *multivariate statistical* criteria identify laboratory test errors that univariate outlier analysis *completely misses*, consist with Goldschmidt and Lent's estimates. The study also shows that statistical detection can outperform delta checking.

B. Effective Detection of Clinical Laboratory Errors

As discussed above, existing automated techniques are inadequate, suffering from low detection and high false alarm rates and long delays. Techniques based on statistical modeling and fault detection that can reliably detect true laboratory errors as they occur would represent a *significant advance* over current capabilities. Such a system would reduce costs due to sample re-collection and re-measurement, and reduce patient risk due to delays and unnecessary or erroneous clinical treatment.

The scarcity of faults and the cost of labeling samples by a human expert collude to create a significant challenge. Roughly 0.5% of all samples carry faults leading to measurement error [4][11][12]. To develop and evaluate statistical fault-detection algorithms requires (at least) hundreds of examples of *faulty* as well as non-faulty sample measurements. At the anticipated 0.5% fault rate, random selection from the population would require tens of thousands of samples to be labeled to provide sufficient faulty samples. Labeling, carried out by clinical laboratory test experts, uses direct examination of laboratory test data together with review of instrument logs and patient charts. This costs upwards of US $10 dollars per sample. Clearly, random sampling is an inadequate strategy to capture sufficient examples of faulty samples; model-based selection of samples is required just to proceed with labeling.

The goal of this preliminary study was to demonstrate that multivariate statistical techniques can be used to efficiently pre-screen samples for expert labeling. As a by-product, we show that even simple multivariate statistical techniques are significantly more sensitive error detectors than quality control, delta-checking, absurd value, and anion gap analysis. Furthermore they do not incur the time lags or multi-sample costs associated and delta checking.

This study is the first step towards development and implementation of a comprehensive fault detection system for the clinical laboratory. We envision a system that uses discriminant algorithms trained iteratively with data economically selected for labeling using active learning. A comprehensive system will also use patient measurement trajectories during treatment [13] to refine detection.

II. Statistical Detection of Clinical Laboratory Errors

A. Data Description

This study used data collected at the OHSU clinical laboratory. The data contains tests on 25,596 specimens collected from inpatients across various hospital units. Measurements from two separate (but identical) instruments were used in the study.

Appropriate for our future focus on kidney disease, this study used analytes from the Centers for Medicare and Medicaid Services Renal Function Panel (RFP) which includes measurements of: albumin, blood urea nitrogen (BUN), total calcium, carbon dioxide, chloride, creatinine, glucose, phosphate, potassium, and sodium. (The database contains up to 30 analyte measurements for each sample, but most specimens have fewer than the full array measured.)

We excluded the glucose measurements due to its large variability across and within subjects. In order to avoid modeling complications arising from missing data values, only specimen records that have *all nine* of the remaining RFP analyte measurements were retained for this study[2]. This left 3,524 specimen records.

We partitioned the specimens in the database into two groups: (1) those with creatinine ≥ 2.0 (indicative of renal insufficiency), and (2) those with creatinine < 2.0 (indicative of normal renal function). This is a very rough division of the population. A future full study will use diagnosis from a standard estimate of glomerular filtration rate to identify renal insufficiency. Having identified the samples belonging to each group, the data were normalized to zero-mean and unit-variance in each of the nine analyte values.

B. Outlier Detection

We evaluated two *multivariate criteria functions* applied to the task of pre-screening samples to select those with high probability of being faults, and so worth labeling. Our goal was to capture a much higher percentage of faults than would result from random sampling from the population ($\sim 0.5\%$). The first is the *likelihood* of the specimen's measured analyte values under the probability distribution of the full population sample (containing individuals from both renal groups). This criterion evaluates a sample based on its consistency with the distribution of the general population. The second criteria uses both a *likelihood ratio* and a likelihood. The ratio used is the likelihood of the analyte values under the distribution for low-creatinine samples divided by the probability of the analyte values under the distribution for high-creatinine samples. This ratio compares a sample's consistency with both the high and low creatinine groups. (Although this has the flavor of a Bayesian discriminant function, we are *not* attempting to classify individuals into the two groups from their analyte measurements, but rather identify measurements that appear inconsistent with the actual *known* group membership.)

[2] Well-established methods for accommodating missing data values will be applied in follow-on work.

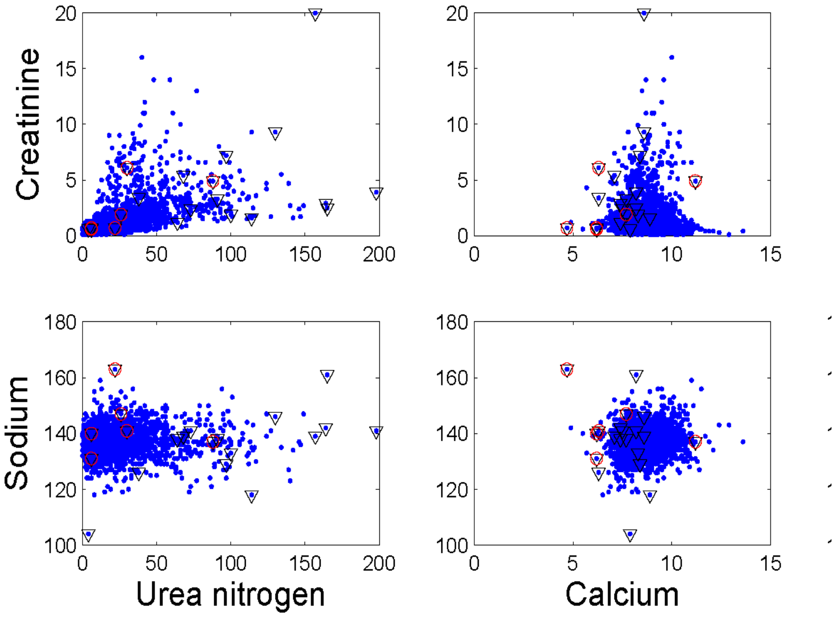

Fig. 1. A subset of the pairwise analyte scatter-plots. The 20 least-likely measurement vectors are indicated with a superimposed ∇. Confirmed errors (from expert review) are marked in red.

1) Likelihood-Based Outlier Detection: We pooled the data from all $N = 3,524$ specimens and constructed a model probability density function $p_{\text{all}}(x)$ using a kernel density estimate (KDE) [14][15]. The model density is

$$p_{\text{all}}(x) = \frac{1}{N}\sum_{k=1}^{N} G_\sigma(x - x_i) \qquad (1)$$

where x_i is the vector of nine analyte measurements from specimen i, and $G_\sigma(z)$ is a zero mean, isotropic Gaussian kernel with variance σ^2. We determined the kernel width σ by maximizing the jackknife estimate [15][16] of the average log-likelihood

$$L(\sigma) = \frac{1}{N}\sum_{j=1}^{N} \log\left(\frac{1}{N-1}\sum_{i \neq j}^{N} G_\sigma(x_j - x_i)\right) \quad . \qquad (2)$$

We used a Fibonacci line search to maximize L with respect to σ.

We evaluated the likelihood of each of the specimens x_i under this model. (When evaluating equation (1) at one of the dataset points x_j, the term $k = j$ is left out of the sum, and the normalization factor N replaced by $N-1$.) We sorted the samples sorted by $p_{\text{all}}(x)$. The 20 lowest likelihood specimens were selected as candidate faults, and are indicated in scatter plots of the data shown in Figure 1. The full 9×9 array of scatter-plots confirms the view of the 2×2 subset in Fig. 1 which shows that while obvious outliers with respect to univariate and bivariate distributions are captured by the likelihood criteria, a *significant* portion of the outliers identified by the full multivariate distribution *would not* have been detected by univariate or bi-variate analysis because the measurements reside in mid-range intervals for the analytes. Multivariate dependencies between the analyte values are *critical* for locating faults.

One of us (S. Kazmierczak) reviewed the data, patient charts, and instrument logs associated with the 20 candidate faults in order to identify which were true errors. Six of the 20 candidates (marked red in Figure 1) were confirmed as errors. None of these would have been detected by absurd value checks or *univariate* outlier methods. (Bivariate checks of BUN vs. calcium would identify *some* of them, depending on threshold.) None of the errors would have been detected by anion gap calculations. Out of the six confirmed errors, two would have been caught by delta-check comparison with the previous value. Three would have been caught by delta-check after the subsequent specimen was taken *several hours later*, and one had no previous or subsequent specimen for comparison by delta-check. For the three errors caught by delta-check after subsequent sample collection, a multivariate statistical method would raise an alarm *instantly*, thus a new specimen could have been ordered quickly.

2) Likelihood-Ratio-Based Outlier Detection: The second approach used two multivariate probability density models calculated on measurements of eight analytes (glucose, urea, creatinine, sodium, potassium, calcium, albumin, and hemolysis level). We constructed two model densities, $p_{\text{H}}(x)$ for the high-creatinine (≥ 2) group, and $p_{\text{L}}(x)$ for the low-creatinine (< 2) group. The models were estimated using the KDE described above. The sample likelihood under each model was computed, and the ratio $l(x) \equiv p_{\text{L}}(x)/p_{\text{H}}(x)$ formed. Specimens in the *high-creatinine group* with both high $l(x)$ *and* high $p_L(x)$ values are inconsistent with the bulk of the high-creatinine samples, and were selected as potential laboratory errors. This can be viewed as simple discriminant test using the pair of features $(l(x), p_L(x))$.

The 19 error candidates (largely different from the first set) thus determined were reviewed to identify which are true errors. This investigation yielded *seven confirmed errors* out of the 19 candidates. None of these errors would have been detected by absurd value checks on single analytes, and they would not have been detected by delta-check since there was no available previous (or subsequent) data. Neither would they have been found by anion gap calculations.

Some of the candidate samples were shown to be without error by comparing the measured values with those from previous and subsequent samples from the *same patient* (but separated by too many hours to qualify for delta-check analysis). This suggests that using patient-specific measurement histories, or analyte trajectories over time will be useful to reduce the false alarm rate.

III. Discussion

Our results show that simple outlier detection techniques using a kernel density estimate (KDE) are efficient at identifying potential faults for labeling of laboratory test data. The sensitivity demonstrated is *vital* for constructing a human-labeled database with sufficient number of faulty samples. The overall frequency of errors in clinical laboratory tests is estimated in the review by Bonini et al. [12] (and the references therein [4][11]) to fall in the range of 0.47-0.61% of samples. Given that a database for discriminative training will require hundreds of confirmed errors, the low prior for errors in the population coupled with the high cost of expert labeling precludes random sampling from the population for labeling. (At 0.5% error rate and a target of 300 expert-labeled faulty samples — to insure sufficient accuracy in the measured sensitivity — drawing at random from the population would require labeling 60,000 samples at a cost of approximately $10 per sample.)

Thus selective pre-screening for faults is essential. We have demonstrated that capability. The frequency of expert-confirmed errors among the tests flagged by our statistical criteria was 30-37%, or 50-79 times the overall error rate expected on the basis of Bonini's review. Thus our multivariate outlier detection is *far more efficient than random sampling* for pre-screening laboratory tests for expert labeling.

In fact, the sensitivity of our methods are already far better than the best current *detection methods*. Our simple multivariate statistical methods identify errors that *would not* have been identified by univariate absurd value checks (or other univariate criteria), by anion gap calculations, or by checks using quality control materials. Less than half of the confirmed errors would have been caught by delta-checks; and those that were may well have been ignored due to the history of high false alarm rates associated with that technique. Furthermore, those that could have been found by delta-checks would have been identified and responded to far more quickly with a statistical test.

Finally, this study used only very basic information (the analyte tests) and simple algorithmic methods (outlier tagging on unlabeled data) for modeling. A future full system will integrate information from sample hemolysis (which is predictive of erroneous excess serum potassium), patient demographics (age, race, and gender), and the facility from which the samples were collected (e.g. intensive care vs. hemo-dialysis). It will use detectors developed by *discriminative training*, which one expects to increase sensitivity and specificity beyond that obtained with outlier methods. Lastly, moving beyond static distributions to models incorporating *patient measurement histories* (i.e. *trajectories* in the space of analyte values) will further enhance detection performance [13]. The resulting detectors will improve patient care by reducing costs and patient risks associated with laboratory test errors.

References

[1] P. Carraro and M. Plebani, "Errors in a stat laboratory: Types and frequencies 10 years later," *Clin Chem*, vol. 53, pp. 1338–42, 2007.

[2] E. Cavenaugh, "A method for determining costs associated with laboratory error," *Am J Public Health*, vol. 71, pp. 831–834, 1981.

[3] K. Downer, "How much does calibration error cost? NIST report suggests $60-199M for calcium testing alone," *Clin Lab News*, vol. 30, August 2004.

[4] M. Plebani and P. Carraro, "Mistakes in a stat laboratory: Types and frequency," *Clin Chem*, vol. 43, pp. 1348–1351, 1997.

[5] S. Kazmierczak, "Laboratory quality control: Using patient data to assess analytical performance," *Clin Chem Lab Med*, vol. 41, pp. 617–627, 2003.

[6] E. Pearlman, L. Bilello, J. Stauffer, A. Kamarinos, R. Miele, and M. Wolfert, "Implications for autoverification for the clinical laboratory," *Clin Lead Manage Rev*, vol. 16, pp. 237–239, 2002.

[7] I. Crolla and J. Westgard, "Evaluation of rule-based auto-verification protocols," *Clin Lead Manage Rev*, vol. 17, pp. 268–272, 2003.

[8] P. Sher, "An evaluation of the detection capacity of a computer-assisted real-time delta check system," *Clin Chem*, vol. 25, pp. 870–2, 1970.

[9] F. Smith and S. Kroft, "Optimal procedure for detecting analytical bias using patient samples," *Am J Clin Pathol*, vol. 108, pp. 254–268, 1997.

[10] H. Goldschmidt and R. Lent, "Gross errors and work flow analysis in the clinical laboratory," *Clin Biochem Metab*, vol. 112, pp. 14–21, 1995.

[11] M. Stahl, E. Lund, and I. Brandlund, "Reasons for a laboratory's inability to report results for requested analytical tests," *Clin Chem*, vol. 44, pp. 2197–7, 1998.

[12] P. Bonini, M. Plebani, F. Ceriotti, and F. Rubboli, "Errors in laboratory medicine," *Clin Chem*, vol. 48, pp. 691–698, 2002.

[13] S. Kazmierczak, T. Leen, D. Erdogmus, and M. Carreira-Perpinan, "Reduction of multi-dimensional laboratory data to a two-dimensional plot: A nobel technique for the identification of laboratory error," *Clin. Chem. Lab. Med*, vol. 45, no. 6, pp. 749–752, 2007.

[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition, 2nd edition*. Academic Press, Inc., 1990.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2001.

[16] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM, 1982.