Detection of Preanalytic Laboratory Testing Errors Using a Statistically Guided Protocol

Jason M. Baron, MD, Craig H. Mermel, MD, PhD, Kent B. Lewandrowski, MD, and Anand S. Dighe, MD, PhD

Key Words: Machine learning; Decision tree; Preanalytic error; Glucose; Informatics; Line draw

DOI: 10.1309/AJCPQIRIB3CT1EJV

A b s t r a c t

Preanalytic laboratory testing errors are often difficult to identify. We demonstrate how laboratories can integrate statistical models with clinical judgment to develop protocols for preanalytic error detection. Specifically, we developed a protocol to identify spuriously elevated glucose values resulting from improper “line draws” or related phlebotomy errors. Using a decision tree–generating algorithm and an annotated set of training data, we generated decision trees to classify critically elevated glucose results as “real” or “spurious” based on available laboratory data. Decision trees revealed that a 30-day patient-specific average glucose concentration lower than 186.3 mg/dL (10.3 mmol/L), a current glucose concentration higher than 663 mg/dL (37 mmol/L), and an anion gap lower than 16.5 mEq/L (16.5 mmol/L) suggested a spurious result. We then used the results from the decision tree analysis to inform the implementation of a clinical protocol that significantly improved the laboratory’s identification of spurious results. Similar approaches may be useful in developing protocols to identify other errors or to assist in clinical interpretation of results.

Phlebotomy errors and other preanalytic laboratory testing errors can lead to highly inaccurate test results, placing patients at risk for harm.1-3 While estimated to occur in a relatively small proportion of samples, preanalytic errors are significantly more common than analytic errors.1-4 Identifying results that are inaccurate because of preanalytic testing errors is an important but challenging responsibility for a clinical laboratory. Because many preanalytic errors occur outside the direct control of the central laboratory and may be subtle,2,3 laboratories need indirect methods to identify these errors. Examples of commonly used indirect preanalytic error identification methods include delta checks and refusal of specimens producing physiologically implausible or impossible results. However, preanalytic errors producing results that are plausible, yet inaccurate, would likely go undetected by most traditional error identification strategies. We thus sought to develop better methods for preanalytic error identification by leveraging concepts from statistics and computer science.

To develop better methods of preanalytic error identification, we selected as a proof of concept the identification of spurious glucose results arising from specimen contamination with intravenous (IV) fluid. These errors are relatively common but may be hard to identify because of the need for rapid turnaround and the lack of clinical information provided to the laboratory with most specimens. Specimens can become contaminated with IV fluid if improperly drawn from an intravascular catheter port without a sufficient discard volume or from a venipuncture site proximal to an active or recently discontinued IV line.5,6 Specimen contamination with small amounts of dextrose or total parenteral

CME/SAM

Upon completion of this activity you will be able to:

• describe ways that preanalytic laboratory testing errors impact the accuracy of test results and represent a challenge in quality management.
• discuss the utility of statistical and machine learning-based strategies in identifying patterns and making predictions based on laboratory data.
• list some limitations of machine learning and statistical classification methods with regard to laboratory data classification.

The ASCP is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education for physicians. The ASCP designates this journal-based CME activity for a maximum of 1 AMA PRA Category 1 Credit™ per article. Physicians should claim only the credit commensurate with the extent of their participation in the activity. This activity qualifies as an American Board of Pathology Maintenance of Certification Part II Self-Assessment Module.

The authors of this article and the planning committee members and staff have no relevant financial relationships with commercial interests to disclose.

Questions appear on p 480. Exam is located at www.ascp.org/ajcpcme.
nutrition–containing fluid may result in substantially elevated glucose results.\textsuperscript{5,6} For example, in a patient with a plasma glucose of 75 mg/dL (4 mmol/L), a specimen contaminated with 1 part 5% dextrose containing IV fluid to 10 parts blood would produce a measured plasma glucose level of nearly 800 mg/dL (44 mmol/L). This calculation assumes that all added glucose remains in the plasma, that no fluid shifts occur, and that the hematocrit is 40\% (0.40). Artificially elevated glucose results are often recognized or suspected to be spurious by clinicians. However, even when recognized, these spurious results provide no useful information, waste health care resources, and delay the acquisition of accurate results. Moreover, there exists the possibility that a spurious critical glucose value may not be identified as spurious, resulting in insulin administration and potential patient harm.

For our error identification strategy, we specifically sought an approach that would examine multiple laboratory parameters from each patient and classify the critically elevated glucose result (\textgtrsim 500 mg/dL [28 mmol/L]) as "real" or "spurious." Among the tools we used was a branch of artificial intelligence known as “machine learning.” One type of machine learning, known as supervised machine learning, uses an annotated training data set to generate a model for classification.\textsuperscript{7} The model can then be applied to nonannotated data to generate machine classifications. For example, in the case of chemistry error classification, the training data could consist of a list of tests manually classified (annotated) by retrospective medical record review as spurious or real. The annotated tests are then associated with predictor parameters. The predictor parameters can be other laboratory values, patient medications, patient location, or any other relevant variable. The classification model that is generated can then be prospectively applied to nonannotated data.

Herein, we describe the use of machine learning to distinguish spurious critical glucose results from true critical results. Among the hypotheses tested in this study is that critical glucose results from samples contaminated with IV fluid are statistically distinguishable from physiologic critical glucose results. We use machine learning to determine which chemistry parameters are predictive of spurious results. We then demonstrate the usefulness of the machine learning information to guide the laboratory in implementing a clinical algorithm to identify spurious glucose values.

**Materials and Methods**

**Setting and Patients**

This work was based on patients treated and samples collected and tested at the Massachusetts General Hospital (MGH), Boston, including affiliated outpatient clinics. MGH is a large (approximately 900 inpatient beds) academic tertiary care hospital. Most of this work was conducted under a clinical laboratory quality assurance program; remaining portions were conducted with approval of the hospital’s institutional review board.

**Training Data for Algorithm Development**

To compile data for algorithm development and decision tree training, a list of patients with 1 or more critical glucose values (\textgtrsim 500 mg/dL [28 mmol/L]) during a 6-month period was obtained from our laboratory information system. For each patient, all results for plasma sodium, potassium, chloride, bicarbonate, and glucose during a 1-year period were obtained. For each critical glucose measurement, the associated anion gap was calculated when the sodium, chloride, and bicarbonate values for that sample were available. In addition, a patient-specific mean glucose value during the 30 days preceding each critical glucose measurement was calculated.

**Clinical Data Annotation**

The clinical judgment of a physician based on retrospective review of the clinical record was used to determine whether a given glucose value was spurious or real. The most persuasive factor in making this determination was a rapid fall in subsequent glucose values without intervening insulin administration. Other factors considered included the presence or absence of diabetes, prior and subsequent glucose values, clinical condition, and the type of IV fluid. For example, infusion of 5\% dextrose (50 g/L) in water, 5\% dextrose in half-normal saline, or total parenteral nutrition would support the hypothesis of IV fluid contamination but would not be conclusive of a spurious designation. Most cases were straightforward and could be classified as real or spurious. For example, a patient with diabetic ketoacidosis treated with an insulin infusion and demonstrating progressively declining glucose values would have his or her values taken as real. Likewise, a surgical patient with no history of diabetes, receiving an infusion of 5\% dextrose, with multiple prior and subsequent normal glucose results, no insulin administration, and an isolated glucose of 900 mg/dL (50 mmol/L) would have the high value annotated as spurious. Because the compiled training dataset was large, we only annotated data from a subset of randomly selected patients.

**Machine Learning**

Decision trees were generated using the open source R statistical programming language (available at http://www.r-project.org/) and the “Rpart” package (available at http://cran.r-project.org/web/packages/rpart/index.html). Trees were built using the “rpart” function without specifying optional control parameters. Technical descriptions of
the tree building algorithms are available online. Briefly, the tree-generating function accepts as inputs predictor and response values. In this case, predictors are laboratory results and responses are the real or spurious annotations. As output, this function generates a decision tree using a recursive partitioning algorithm. Each partitioning step selects the tree-branching criterion based on the predictor variables that will split the parent data set into 2 daughter data sets, with the daughter sets being as “pure” (homogeneous) as possible as approximated according to a heuristic. Here, purity is measured with the Gini statistic with respect to the response variable, such that the more homogeneous a daughter set, the higher the purity. Partition steps are repeated, adding branches to the tree, until the data cannot be further purified by the partitioning algorithm and predictor values or other control conditions (none specified in our case) are met.

Algorithm Simplification and Implementation

As described in the Results, we combined the results from the decision tree analysis with clinical judgment and workflow considerations to develop an algorithm that could be applied in daily practice. Spurious results identified with this clinical algorithm were appended with the following comment: “Unexpected result. Possible causes include line draw with IV fluid interference. Clinical correlation required.” Performance, including the sensitivity and specificity of the applied algorithm, was determined by analyzing all inpatient critical (>500 mg/dL [28 mmol/L]) glucose values occurring during the first 4 weeks after the clinical algorithm was implemented. We also used this 4-week data set to validate the decision trees. Confidence intervals for sensitivity and specificity were calculated in R according to the Agresti-Coull method as implemented within the R “Binom” package.

Compiled Data Sets

An overview of the clinical data and sample sizes is provided in Figure 1. As shown, specimens missing any of the predictor variables used in generating a tree (eg, no sodium value) were disregarded in both training and testing the tree. Blood gas glucose results and outside hospital glucose results were included when calculating 30-day mean glucose values but were excluded for other purposes. Because of technical limitations in data extraction, plasma and blood-gas glucose values from the same accession had to be excluded from the training data set and calculations of 30-day mean glucose values. However, these values were sufficiently infrequent (estimated to represent <2% of the

![Figure 1](image)

Data used in the study. Sample sizes are shown for data used in each aspect of the analysis. Chemistry parameters of interest include sodium, potassium, chloride, bicarbonate, anion gap, and glucose.
average patient’s total glucose results) and random so as to not substantively alter or introduce bias into the analysis.

Results

Machine Learning Decision Trees

As described in the Materials and Methods, we used clinical record review to classify critically elevated glucose values as spurious or real. We then used machine learning techniques to create decision trees to determine the variables useful in predicting whether a given glucose value was spurious or real. We first used machine learning to construct a decision tree based on current chemistry results (sodium, potassium, chloride, bicarbonate, glucose, and calculated anion gap) as well as the patient’s mean glucose value during the previous 30 days. In this decision tree Figure 2A, the 30-day mean glucose was the most important predictor of whether a given result was spurious or real. In classifying the training data, this tree had a sensitivity (fraction of spurious results correctly classified) and specificity (fraction of real results correctly classified) of 93% and 89%, respectively Table 1. Sensitivity and specificity both remained greater than 80% when the tree was tested with an independent dataset not used in training (Table 1).

Implementing an algorithm incorporating values from prior laboratory specimens may be technically challenging in many laboratory information systems and unreliable in the setting of limited prior results. Therefore we also generated a decision tree based only on individual chemistry results available on the current sample. In this tree, the most significant predictor variables were the current glucose and current anion gap Figure 2B. The accuracy of this tree with respect to the training data was similar to the accuracy of the tree including mean glucose value, but this tree was less accurate in classifying the test data (Table 1).

Algorithm Application to Practice

We next sought to integrate the machine learning results with clinical judgment to develop an algorithm suitable for routine clinical practice. In particular, because of the constraints of laboratory workflow and the clinical need, the algorithm had to be specific and easy to use by laboratory staff. To aid in the development of a clinical algorithm, we plotted glucose results vs the anion gap for spurious and real values Figure 3. We selected anion gap and glucose values based on the output of the decision tree shown in Figure 2B. Using this data set, samples with nonelevated anion gaps (≤15 mEq/L [≤15 mmol/L]) and glucose results greater than 800

---

**Figure 2** Machine learning decision trees. **A**, Decision classification tree based on current chemistry results (sodium, potassium, chloride, bicarbonate, glucose, and calculated anion gap) and the 30-day mean glucose value. **B**, Decision tree based only on current chemistry results.
mg/dL (44 mmol/L) were always spurious and we used this finding as the basis for our clinical classification protocol. We decided to apply the algorithm only to inpatients because this is the primary setting for spurious glucose results. We were also concerned about the possibility of a patient presenting to the emergency department in a hyperosmotic, nonketotic state, despite the absence of such patients in our training data. With our simplified clinical protocol, glucose results with an associated nonelevated anion gap and with values greater than 800 mg/dL (44 mmol/L) were recorded with the comment, “Unexpected result. Possible causes include line draw with IV fluid interference. Clinical correlation required.” Results without sufficient data to calculate the anion gap were not flagged and were counted as though classified as real in our analysis. Laboratory staff could use their judgment to overrule this algorithm in either direction.

Algorithm Clinical Performance

We assessed the performance of the clinical algorithm during the first 4 weeks after implementation. During this period, 59 glucose values were found to be critical (>500 mg/dL [28 mmol/L]) and 20 were higher than 800 mg/dL (44 mmol/L). The algorithm identified 14 of 19 spurious inpatient glucose results higher than 800 mg/dL (44 mmol/L) for a sensitivity of 74% (95% confidence interval [CI], 51%-89%). The sensitivity of the algorithm was 33% (95% CI, 20%-48%, 14 of 43 values identified) with regard to the identification of all spurious critical glucose results higher than 500 mg/dL (28 mmol/L). The algorithm did not flag any of 16 true critical inpatient glucose results, for a specificity of 100% (95% CI, 77%-100%). Before implementation of our clinical algorithm, technologists were expected to use their judgment to manually review critical inpatient glucose results for accuracy. Analysis of a representative sampling (n = 57) of inpatient critical (>500 mg/dL [28 mmol/L]) glucose results in the preintervention period revealed that only 5 of 57 results had been flagged, corresponding to a sensitivity of 9% (95% CI, 3%-19%). In addition, only 1 of 24 spurious results higher than 800 mg/dL (44 mmol/L) was flagged before intervention. The sensitivity increase observed using the clinical algorithm was statistically significant (P < .01 for all critical glucose values and P < .001 for those over 800 mg/dL [44 mmol/L], Fisher exact test, 2-tailed).

Discussion

We demonstrate herein how an algorithm incorporating several items of data readily available to the laboratory can be useful in identifying preanalytic errors. Moreover, we show how machine learning can be paired with clinical judgment to develop useful data classification strategies. Although informative, the decision trees would not by themselves have been suitable for direct clinical application. Not only are the trees themselves complex, given workflow considerations, but they also lack specificity and have the potential to perform unpredictably...
in situations not observed in the training data. Likewise, with the vast quantity of laboratory data available for most patients, had we simply set out to develop a clinical algorithm without the assistance of machine learning, we likely would have been unable to optimally sort through the data. However, clinical judgment combined with machine learning produced an intuitive, easily implemented, and highly specific protocol.

The approach described herein differs from that of other studies that explored the use of machine learning and other related techniques in the clinical laboratory.11-21 The most significant of these differences is that we use machine learning to inform a clinical operating protocol, rather than directly applying machine learning as a clinical algorithm. We are not aware of any prior attempts to combine machine learning with medical knowledge to inform a clinical preanalytic error identification protocol. For example, Oosterhuis et al14 describe “LabRespond,” an innovative, statistics-based autovalidation system, and Doctor and Strelwicz20 present a novel Bayesian network method for identifying “wrong blood in tube errors.” Although these and other related reports demonstrate interesting and important findings, they differ from our approach in several fundamental respects. First, these error identification methods were published as prototypes and were validated in a simulation based on their ability to detect errors simulated by the investigators. We implemented our system and demonstrated its ability to detect real errors in real time. Moreover, none of these prior approaches used recursive partitioning to produce intuitive decision trees suited for informing a clinical protocol incorporating clinical judgment.

Our implemented algorithm is useful for several purposes, in addition to functioning as a proof of concept. Foremost, it can serve as a foundation for developing a more sensitive algorithm. In addition, this algorithm can be used to identify examples of improper phlebotomy practice. Because the ultimate goal is to eliminate these phlebotomy errors, we will be using our results to provide targeted feedback to nursing units to improve phlebotomy practice. We are also investigating opportunities to improve the sensitivity of the algorithm, particularly by incorporating additional parameters in a larger training set.

A potential limitation of our study is the lack of a “gold standard” for classifying glucose values as real or spurious. A gold standard classification is essential for the use of the machine learning approach as detailed here. Our surrogate gold standard, the clinical judgment of a physician based on record review, has some subjectivity. Nonetheless, we consider this to be appropriate and valid for several reasons. Because results were annotated by retrospective chart review, we were able to use information unavailable to clinicians evaluating the values in real time. Subsequent glucose results, with or without intervening insulin administration, were particularly helpful in this regard. In addition, we carefully reviewed multiple factors in each medical record as described in the methods, and most classifications were straightforward and indisputable.

Overfitting risk is an important consideration in the development of algorithms to classify data and represents a potential limitation of our approach. Overfitting occurs when a computer identifies random patterns in the tree-fitting data set, which occur by chance and are not truly predictive of the response variable.7,22 Because such random patterns do not generalize, they will be unhelpful in predicting subsequent testing sets. One technique to reduce overfitting is to use only predictor parameters that mechanistically or logically should be related to the outcome.22 For example, the anion gap has a mechanistic basis in distinguishing real from spurious glucose results whereas the total protein level does not. Another strategy is to limit the model’s complexity by limiting the number of predictor variables.7,22 This approach forces the algorithm to rely on only the most significant and presumably generalizable predictors. Increasing the size of the training data set relative to the number of predictor parameters also can help to reduce overfitting.

Testing models with data separate from those used in model training is important in assessing generalizability.7 Thus we tested our algorithms using a separate, prospectively collected data set to provide an unbiased assessment of each algorithm’s ability to classify glucose results. Although decision tree performance data are provided for both training and test data, the results on the prospectively collected test data provide a more realistic assessment of each tree’s classification ability. The decision tree incorporating 30-day mean glucose values performed comparably with both test and training data. In contrast, the tree with only current values performed considerably better on the training data than on the test data. This difference in performance between data sets for the tree without mean glucose values suggests some degree of overfitting. “Pruning” algorithms can be applied to assembled decision trees to remove selected branches starting from the bottom of the tree in an attempt to reduce overfitting. However, because the purpose of our decision trees was to inform a clinical algorithm, rather than serve as one, we did not apply pruning.

Decision trees built through recursive partitioning represent only 1 of many methods for machine classification of data. Other classification methods useful for predicting binary outcomes based on continuous predictor variables include support vector machines, artificial neural networks, and certain regression models.9,23 For example, logistic regression models estimate a probability of each outcome based on a linear combination of predictor variables.23 We evaluated a logistic regression model using the same training and test data used for the decision trees (data not shown). Although the logistic regression model performed
comparably in terms of its sensitivity and specificity, its output was less intuitive, and thus less well suited for clinical protocol development.

Machine learning approaches work well only to the extent that the training data set captures the various circumstances that might be seen subsequently. For example, because our training data did not include patients in hyperosmotic, nonketotic states, our decision trees may not perform well in classifying glucose results from such patients. Thus we decided to reduce the risk of misclassifying a previously unidentified hyperosmotic, nonketotic state by limiting the protocol to inpatients. One corollary is that larger training data sets will be more likely to capture less commonly occurring cases and may thus produce better trees.

We considered whether the predictive power of the decision trees could be improved by including a wider range of laboratory parameters, particularly hematocrit, serum urea nitrogen, or creatinine. However, we decided to limit the number of included parameters in our analysis to reduce overfitting risk, given the size of the training data set and because we were concerned about the reliability, consistency, and availability of certain parameters. For example, electrolytes frequently accompany glucose measurements, but hematocrit levels from the same specimen might be less often available, and when available, would not typically be included in the result simultaneously.

Other laboratories may wish to use strategies similar to those presented here. We consider it critical that any clinical protocol be internally validated with a laboratory’s own data before implementation because patient populations, assays, workflows, and the frequencies of various interferences differ among institutions. In addition, until a rule is well validated, laboratories should carefully consider the action taken in response to a value classified as potentially spurious. The strength of the response should account for an algorithm’s imperfect or not fully validated specificity.

The high proportion of total critical inpatient glucose results that are spurious is not indicative of the overall rate of phlebotomy errors at our institution. Although our analysis may have missed some IV fluid contamination errors, such as when the fluid is non–dextrose containing or when the contamination is not sufficient to raise glucose levels into the critical range, we estimate that spurious, critical glucose values represent fewer than 0.3% of the total inpatient plasma glucose results reported at our institution.

We anticipate that approaches similar to the one demonstrated here for glucose will prove useful for identifying other types of preanalytic errors. Furthermore, in addition to error identification, machine learning strategies using clinical and laboratory data and integrating clinical judgment may be useful in complex diagnosis and in the refinement of diagnostic algorithms. Substantial future work investigating these applications of machine learning is needed. We expect that the future development of better algorithms for laboratory error identification and diagnosis will lead to improved patient safety, diagnostic efficiency, and health care quality.

From the Department of Pathology, Massachusetts General Hospital, and Harvard Medical School, Boston, MA.

Address reprint requests to Dr Dighe: Bigelow Building, Room 510A, 55 Fruit St, Boston, MA 02114; asdighe@partners.org.

References


