The Performance of Delta Check Methods

Lewis B. Sheiner, Lawrence A. Wheeler,¹ and John K. Moore

The percentage of mislabeled specimens detected (true-positive rate) and the percentage of correctly labeled specimens misidentified (false-positive rate) were computed for three previously proposed delta check methods and two linear discriminant functions. The true-positive rate was computed from a set of pairs of specimens, each having one member replaced by a member from another pair chosen at random. The relationship between true-positive and false-positive rates was similar among the delta check methods tested, indicating equal performance for all of them over the range of false-positive rates of interest. At a practical false-positive operating rate of about 5%, delta check methods detect only about 50% of mislabeled specimens; even if the actual mislabeling rate is moderate (e.g., 1%), only about 10% of specimens flagged by a delta check will actually have been mislabeled.

Additional Keyphrases: quality control · computers · statistics

Delta check methods have been proposed for detecting erroneous laboratory test results (1–3). Because of the importance of accurate laboratory test results for patient care and the increasing pressure on hospitals and laboratories by government and accrediting agencies to improve quality-control techniques, it is appropriate to evaluate the effectiveness of delta check methods.

For a given patient, delta check methods compare the differences (deltas) between today's test values and corresponding previous test values with given thresholds. If a delta exceeds its threshold, the value for "today" fails the delta check and is suspected of being erroneous. Any source of laboratory error may cause one or more of a set of test values to fail a delta check. The method is particularly interesting as a method for detecting two important types of error, specimen mislabeling (i.e., assigning "today's" values to a patient different from the one from whom the sample was actually taken) and error in reporting test results, because these types of errors are not detectable by any other post hoc means, such as checking control specimens against quality-control limits.

This study reports an evaluation of the ability of three previously suggested delta check methods (1–3) and two linear discriminant functions to detect mislabeled specimens. The evaluation is based on simulated specimen identification errors with almost 3000 pairs of actual continuous-flow (SMA 6) determinations.

The essential finding of the study is that the previously suggested delta check methods for the usual SMA 6 tests are approximately equivalent when performance is judged by their relative abilities to correctly identify mislabeled specimens (true-positive rate) while operating at similar false-positive rates. However, the highest true-positive rate attainable by these delta check methods operating at acceptable false-positive rates (about 5%) is modest, only about 50%. The delta check is therefore no panacea for lax specimen-identification procedures.

Materials and Methods

The test results used in this study were collected in the clinical laboratory of the University of California Medical Center, San Francisco, with use of its Community Health Computing Laboratory computer system. The system's patient history file, as of autumn 1977 (containing all test results for the preceding 60 days), was searched to identify all pairs of SMA 6 (Technicon Instruments Corp., Tarrytown, NY 10591) results (electrolytes, serum urea nitrogen, creatinine) for which the second determination was made between 0.9 and 2.5 days after the first. No test result was used more than once. All patients in the history file were considered for inclusion in this study. Therefore, the data arise from an unselected sample of patients who vary with respect to age, sex, inpatient or outpatient status, and clinical state. The only basis for selection was that the physicians caring for these patients had ordered two or more SMA 6 tests within the specified period.

A total of 2988 pairs of SMA 6 results, representing data from 749 patients, was available. An examination of the pairs revealed that in 19 pairs one or more obviously nonphysiological results existed (e.g., the test value was zero or impossibly large). In each such case that pair of SMA 6 results was excluded from the study because we sought to evaluate only detection of specimen mislabeling.

The three delta check methods investigated differ in which of the six SMA 6 tests they check and in the magnitude of their thresholds. The method of Wheeler and Sheiner (3) checks all six SMA 6 values (Na⁺, K⁺, Cl⁻, HCO₃⁻, blood urea nitrogen, and creatinine), plus the serum urea nitrogen/creatinine ratio and the "anion gap" (Na⁺ + K⁺ − Cl⁻ − HCO₃⁻); this method will hereafter be designated method A. The method of Whitehurst et al. (2), which checks only the six SMA 6 values, will be designated method B. Our adaptation of the Ladenson method (1), method C, checks only Na⁺, K⁺, serum urea nitrogen, and creatinine.

The threshold values of methods B and C, given in the appropriate references, were selected intuitively by the methods' authors. The threshold values of method A are based on the empirically observed distribution of deltas in an unselected group of hospital patients. Method A defines three thresholds for each test value, the most stringent being so high that only 1% of (presumably) correctly identified specimen pairs will exceed it as a result of physiological shifts in test values; the

¹Present address: Indiana University, Department of Clinical Pathology, 1100 W. Michigan St., Indianapolis, IN 46202.

Received May 10, 1979; accepted Sept. 10, 1979.
second will be exceeded by physiological shifts 2% of the time; and the least stringent, 5% of the time. These three sets of thresholds will be designated as methods A01, A02, and A05, respectively.

Because each method checks a number of deltas simultaneously, one may define a hierarchy of rules for deciding that a specimen has failed the delta check. For example, one may fail the specimen only if one particular delta (e.g., Na⁺) exceeds its threshold, if at least one of all the deltas checked exceeds its threshold, if at least two deltas exceed their thresholds, etc. Each such distinct rule defines a submethod for each parent method.

We computed the false-positive rate for each of the submethods of each of the parent delta check methods by simply applying each submethod to the 2969 pairs of SMA 6 results we studied. All pairs failing by a submethod are, by definition, “false positives” because we assume that the original data include at most very few mislabeled specimens.

The true-positive rates—the proportion of mislabeled specimens detected by the various submethods—are computed from a “mislabeled” data set of 2969 pairs, each including a “mislabeled” specimen. Each of these pairs consists of the first SMA 6 value from one of the original 2969 pairs and the second SMA 6 value from another pair, selected randomly from among the pairs for other patients.

Obviously any submethod that has a higher false-positive rate and lower true-positive rate than some other submethod can be ignored in assessing the performance of a method. The remaining submethods of the method, whose performance cannot be bettered by any other submethod of that method at a given false-positive rate, are called “optimal.” The Receiver Operating Characteristic (ROC) curve of a method depicts its performance. An ROC curve is constructed by plotting the true-positive rates of optimal submethods on the vertical axis against their false-positive rates on the horizontal axis and connecting these points to form a curve (see Figure 1).

ROC curves make one kind of comparison of methods especially simple: one method is in all circumstances superior to another if its ROC curve everywhere lies to the left of and above the ROC curve of the latter, for then the former method has a higher true-positive rate than the latter at any false-positive rate.

Fig. 1. ROC curves (true-positive rate as a percentage vs false-positive rate as a percentage) for the various methods discussed in the text
A, the composite curve for the methods A01, A02, and A05 of Wheeler and Sheiner (3); B, the method of Whitehurst et al. (2); C, the method of Ladenson (1); D30, the 30-variable discriminant function; D7, the 7-variable discriminant function. The ROC curves of methods A, B, and C connect the ROC points of their optimal submethods (see text)

Linear discriminant functions, a familiar statistical technique for classifying multivariate observations into two groups (4), have not been used previously to our knowledge for delta checks. Our linear discriminant function is a formula that is linear in a set of variables derived from one SMA 6 pair. (One of the linear discriminant functions for which results are reported is given by equation 1, below.) Pairs are assigned to the “correctly labeled” and “mislabeled” groups according to whether their discriminant function value exceeds a certain threshold. Each different discriminant function can be considered a method. Adjustment of the threshold yields submethods with specified false-positive rates.

Given a set of variables derived from an SMA 6 pair, an “index set” of pairs assumed to be correctly labeled, and an index set of pairs known to be mislabeled, the “best” linear discriminant function of these variables has coefficients that minimize the total probability of misclassification of the two index sets into the two groups. Choice of the appropriate sets of variables to include in the discriminant function is aided by computer algorithms that determine which set of variables of a given size best differentiates the two groups (5). If many variables are available, there usually exist numerous discriminant functions with almost identical properties; the choice between them is essentially arbitrary.

A discriminant function is always influenced by the idiosyncracies of the data set from which it is computed. The more variables that are included in the function, the stronger is this influence. Hence, the discriminant functions reported here probably would not perform so well on other independent data sets.

Altogether, 56 variables computed from each SMA 6 pair were considered for use in the linear discriminant functions. Seven variables were computed from each of the eight pairs of values tested by delta check method A: the value of the analyte in the first SMA 6 set, the delta, the absolute value of the delta, the delta divided by the analyte value in the first SMA 6, the delta divided by the time elapsed between specimens, and the absolute values of the two quotients.

Results for two discriminant functions are reported in Results: the best function involving only seven variables, denoted D7, and the best function involving 30 variables, denoted D30. The D7 discriminant function is:

\[ F = -0.552|\Delta Na| -0.471|\Delta K| -0.0471|\Delta Cl| -0.0775|\Delta HCO_3| -0.0247|\text{urea}-N_I| -0.0959|\Delta Cr| +0.159 \text{ urea}-N_I +1.29 \]  

where \( |\Delta x| \) is the absolute value of the delta for test x, and where urea-N_I is the value of the first blood urea nitrogen of a pair. Given a cutoff value, \( F_c \) (see Results), pairs with \( F \) greater than \( F_c \) are classified as properly labeled, and pairs with \( F \) less than or equal to \( F_c \) fail the delta check, and are classified as mislabeled.

Results

Figure 1 presents the portions of the ROC curves corresponding to false-positive rates less than 20% for all the methods of this paper. There are only minor differences among the curves of methods A, B, and C. These differences become even less interesting when we note that the 95%-confidence limits of the estimated rates are on the order of ±2%, even for a data set as large as ours. To get an idea of the variability of these rates, we created a second “mislabeled” data set in exactly the same way as the first (see Methods) and re-estimated all the true-positive rates from it. Indeed, the standard deviation of the differences between corresponding true-positive estimates was 0.9, which yields an approximate 95%-confidence interval of ±2%.

At a false-positive rate of about 2% the D30 method does have about a 10% higher true-positive rate than any previously suggested method, but recall that the performance of the discriminant function on the index sets is likely to be better
than can be expected on any other data set. No such bias exists for methods A, B, and C.

Figure 1 reminds us of a universal characteristic of ROC curves: increased sensitivity (increased true-positive rate) is only possible if one is willing to pay the price of decreased specificity (increased false-positive rate). At a reasonable operating point on the ROC curve of 5% false-positive rate, Figure 1 indicates that one may expect to detect only 50-60% of truly mislabeled specimens.

Table 1 presents the performances of certain of the discriminant function submethods and compares them with the performances of comparable submethods of Methods A, B, and C. The above findings are again obvious.

Discussion

Laboratory error may be defined as reporting a value that differs from that which would have been obtained if the test were correctly performed on a specimen from the correct patient. Therefore, mistakes in labeling of specimens, performance of tests, and reporting of results all contribute to laboratory error. In this study, we evaluate the effectiveness of three previously described delta check methods for detecting “mislabeled” specimens. Although this is not the only function of delta checks, it is an important one, and our findings therefore merit some consideration.

Our results lead us to two conclusions. First, with respect to identifying mislabeled specimens, the particular method used for the delta check is relatively unimportant; all currently suggested methods are capable of approximately the same performance. Second, no currently available delta check method offers a truly substantial guarantee of detecting mislabeled specimens; a practical expectation is correct identification of only half the truly mislabeled specimens.

The first conclusion is supported by inspection of Figure 1: the eight tests used for method A (and the D7 method) add little to the performance possible with the four tests used for Method C. This is undoubtedly because of the high correlation between the various tests, especially among the electrolyte values. Although we did not specifically test for the possibility, we speculate that deleting serum urea nitrogen or creatinine (but not both) from the four Method C tests would have little detrimental effect, because these two tests are also highly correlated.

As mentioned in Results, only the 30-variable discriminant function appears to improve upon the performance of the other methods, and this result is highly suspect. Indeed, we deliberately have not given the equation for the 30-variable discriminant function because we have so little faith that its performance could be duplicated on a data set gathered at another institution. This is because the 30-variable function involves so many coefficients that it very likely achieves its performance by exploiting idiosyncracies of the data set. If, then, we regard the 30-variable discriminant function more as an overly optimistic upper bound on the performance of any (linear) delta check system, we are led directly to our second conclusion: one must not expect too much from the delta check.

Any practical system must take costs as well as benefits into account. The actual computation of the delta and its comparison to a threshold by a laboratory computer system has a negligible cost. The investigation of a specimen that fails the check is another matter. If mislabeling is suspected, simply rerunning the specimen does not constitute adequate investigation. Failure of a delta check should logically require drawing and running a new specimen. Operating this way at a false-positive rate of 5% will raise costs at least 5% and delay reporting of 3% of results. Although we think it unlikely that laboratory directors would permit operation at a false-positive rate higher than 5%, the exact rate at which to operate is, of course, a matter of individual choice, involving a trade-off between operating costs and the benefits of identifying some fraction of mislabeled specimens.

When one considers the actual fraction of all tests failing the delta check that actually will have been mislabeled, the true- and false-positive figures of Table 1 are seen in another light. This fraction depends, of course, on the actual mislabeling rate. For example, if the mislabeling rate is 1% (one of each 100 specimens is incorrectly identified), operating at a 5% false-positive rate means that (a) 5.5% of all specimens will fail the check, (b) almost half of the truly mislabeled specimens will not be detected, and (c) about 90% of the 5.5% of specimens identified as possibly mislabeled will, on subsequent investigation, prove to have been false alarms.

Of course, because the test values of the mislabeled specimens that are not detected by the delta check differ little from the first day's values, they will likely differ little from the second day's values for the patient; this lowers the "cost" of not detecting mislabeled specimens by an unknown amount. Further, although we feel that the major conclusions discussed above are warranted, we caution our readers not to place too much reliance on the exact values of our ROC figures, or to use them as more than a guideline for the choice of the point on
the ROC curve at which to operate their own delta check system.

There are three reasons for our caution.

First, the false-positive rates we find are upper bounds on the true values because our data set probably contains a few mislabeled specimens.

Second, the true-positive rates we find may be lower bounds because they refer only to detection of specimen mislabeling. Both of these considerations affect all delta check methods equally, and do not, therefore, weaken our first conclusion. We further doubt their being of sufficient quantitative importance to seriously weaken our second conclusion.

A third reason for some caution is that the particular pattern of clinical characteristics in our group of patients may not be representative of the characteristics of patients at other institutions. This could change the ROC curves for the various delta check methods. All delta check methods depend on the truth of two propositions: (a) serial specimens from a single patient are apt to be more “alike” than specimens selected at random from two different patients, and (b) “large” short-term changes in the test results of a patient are unlikely. A homogeneous population of patients (e.g., “normals” coming to a screening clinic) may violate the first proposition, and a population of desperately ill patients, undergoing abrupt and calamitous shifts in clinical status (e.g., patients in an intensive care unit) may violate the second. We cannot predict the delta check ROC curves for such patient groups. However, it seems reasonable that they would still display the same relative positions as in our study. If so, our first conclusion would be insensitive to differences in groups of patients. Either circumstance would likely shift the ROC curves to the right (poorer discrimination) so that our second conclusion still appears conservative.

Thus, despite some cautions concerning quantitative rather than qualitative matters, we have shown that all currently suggested delta check methods perform equivalently in detecting mislabeled specimens, but none perform so well that proper laboratory vigilance concerning specimen identification can be relaxed.

This work was supported in part by NIH Grant GM 16496, GM 00001.

References