Variability of Laboratory Test Results

Shahram Shahangian, PhD,1 and Richard D. Cohn, PhD2

Key Words: Variability of laboratory results; Laboratory test imprecision; Serum total cholesterol; Serum potassium; Audit sample; Split specimen; Hospital laboratories; Physician office laboratories

Abstract

Variabilities of serum total cholesterol and potassium results provided to 11 medical clinics were assessed using an audit sample–split specimen design. This involved collection of 3 tubes of blood from each of 302 patients, with 1 split specimen divided into 3 audit samples: 1 was sent to the original participating laboratory, another to a commercial referral laboratory, and the third to an academic referee laboratory. Two methods were used to assess variability of test results. Method 1 was based on result pairs corresponding to the split specimen and its corresponding audit sample. Method 2 was based on audit sample results only. The 2 methods provided comparable results for total cholesterol; the estimated coefficient of variation was 1.0% to 3.7%. However, method 1 consistently provided higher estimates of variability for potassium; the estimated SD was 0.096 to 0.168 mmol/L for method 1, while it was 0.035 to 0.090 mmol/L for method 2. Method 1 is more practical, but method 2 can provide a more accurate assessment of analytic variability.

The quality of clinical laboratory testing is vital for promoting and maintaining the public’s health. In 1988, the US Congress enacted the Clinical Laboratory Improvement Amendments of 1988 (CLIA) in response to the concern that laboratory errors were important public health issues. During the debate preceding passage of CLIA, little scientific evidence was available to document the frequency and type of laboratory errors affecting, among others, the accuracy and reliability of laboratory test results.1 Provisions of CLIA subsequently directed the US Department of Health and Human Services to assess the nature and extent of laboratory-related problems. Initially, the Centers for Disease Control and Prevention (CDC) published a review of the scientific literature relating to CLIA.2 In 1990, the CDC solicited a study proposal to address the 5 study areas outlined in CLIA.3 Insufficient resources, however, prevented implementation of the proposed design.4,5 Therefore, the CDC began conducting focused independent projects to address issues that would have been covered by the more comprehensive design.

In 1994, the CDC began developing and evaluating a prototype process4,5 that used a split-specimen (SS) and an audit-sample (AS) design to determine the frequency and type of problems that occurred in certain portions of the total testing process for specific laboratory tests performed in physician offices and hospital laboratories. A 3-month study was conducted to assess the feasibility of such a system to monitor problems in the testing process.6 Based on these findings, a full-scale evaluation7 was initiated involving 1,378 patients from 11 medical clinics with their own office laboratories (n = 8) or affiliated with different hospital laboratories (n = 3). Two of the 3 analytes evaluated in the feasibility study were selected for the full-scale study: serum total cholesterol (TC) and serum potassium (K). Results of the full-scale evaluation were useful for assessing the potential for this design to characterize problems occurring in a portion of the total testing process.7
The objective of the present study was to use and compare 2 methods for assessing test result variability using an AS design alone or in combination with an SS design; it is, to our knowledge, the first account of the use of such controlled experimental schemes to do so in different medical and laboratory environments. The methods described herein can be used to evaluate reproducibility of laboratory test results. Audit samples (and sometimes also split specimens) already are collected in many laboratories for possible retesting (and parallel testing) to assess the reliability of laboratory results. Laboratory results were obtained using blood specimens collected from 302 patients randomly selected from each clinic population who participated in the 1,378-patient full-scale evaluation study. For 1 of the 3 specimens collected from each of these 302 patients, audit samples were prepared and sent to the original participating laboratory, to the referral laboratory, and to an academic referee laboratory. Subsequently, during clinic site visits, the result from S1 specimen (S1) was abstracted from the participating patient’s medical record.

Participating facilities included 8 physician office and 3 hospital laboratories, a commercial referral laboratory that analyzed split specimens within 18 hours of blood collection, and an academic referee laboratory for conducting AS analysis. Selection criteria and characteristics of the testing facilities have been described. Four of the physician office laboratories served 1 to 9 clinicians (average, 6.0 clinicians) in different multispecialty practices, 2 served 1 and 8 clinicians in internal medicine practices, and 2 served 3 and 48 clinicians in family practice settings. Hospital laboratories were in facilities that ranged in size from approximately 100 to 1,000 beds. One served a referral multispecialty practice, the second provided laboratory results to 13 clinicians in an internal medicine practice, and the last served 2 clinicians in a family practice setting. The referee laboratory was certified through the Lipid Standardization Program of the CDC to perform TC testing using standardized testing procedures referenced to the Abell-Kendall reference method and participated satisfactorily in an accredited College of American Pathologists proficiency testing (PT) program for TC and K. Serum samples obtained from S3 specimens were sent to the contractor for this study (Analytical Sciences Inc, Durham, NC), where integrity of each sample was examined visually and noted if compromised, recentrifuged if necessary, and stored at –20 C, and later retrieved and divided into 3 audit samples.

Serum TC and potassium (K) tests were selected because of their clinical relevance, specimen stability, availability of a standard reference method, and common use for ambulatory and hospitalized patients. Total cholesterol
was measured using cholesterol oxidase methods (analyzers were Abbott Spectrum, Beckman Synchron CX-7, Kodak Ektachem DT-60, Kodak Ektachem Vitros 250, Vitros 750 and Vitros 950, Olympus AU5200, and Roche Cobas MIRA), and K was assayed using ion-selective electrodes (analyzers were Abbott Spectrum Series 1, Beckman Synchron CX-7 Delta, Electronucleonics AVL, Electronucleonics Starlyte, Boehringer-Mannheim Hitachi 747-100, Kodak Ektachem Vitros 250, Olympus AU5200, and Roche Cobas MIRA and Cobas MIRA Plus). There were 150 patients with K results and 152 patients with TC results; 14 to 30 patients (average, 28 patients) per clinic participated in the study.

Methods for Assessing Test Result Variability

Two methods were used to assess variability of test results. Method 1 was based on the variability exhibited by results of duplicate tests as defined by the SS and AS pairs for participating laboratories (S,k, A,k) and for the referral laboratory (S,2, A,2). Variability of K results was reported as SD, while TC variability was reported as coefficient of variation (CV; see “Additional Computational Procedures”). Throughout this report, Sk represents the result for SS k, and (CV; see “Additional Computational Procedures”).

The following equation was used to measure variance of Ak results [var (Ak)] for k = 1 or 2:

\[ \text{var} \left( A_k \right) = \text{var} \left( A_k - A_3 \right) - \text{var} \left( A_3 \right) \]

where var (A3) is an independently estimated variance of the A3 result. We obtained independent estimates of A3 variability from the CDC Lipid Standardization Program specimen pool results for TC and from replicate K test results of selected patient specimens (6 measurements on each of 9 specimens). We estimated var (A,k) directly from the paired differences and subtracted the known var (A3) from the result to obtain an estimate of var (A,k).

The noncomputational differences between the 2 methods used to assess laboratory result variabilities are shown in Table I.

Additional Computational Procedures

To define appropriate measures of variability, we assessed any relationship between the SD of duplicates and the analyte concentration by linearly regressing SD against the average of the duplicate (S,k, A,k) values. For K, SD was independent of concentration (slope of the linear regression line not significantly \[ P < .05 \] different from zero). Therefore, variability of K results was reported as SD, computed as (pooled variance)1/2. For TC, SD was related linearly to concentration (positive slope significantly different from zero and intercept not significantly different from zero). So, TC variability was reported as CV. We calculated this by averaging CV2 for duplicate (S,k, A,k) results for each facility. For method 2, Ak and A3 were adjusted as follows to account for the relationship between result magnitude and variability:

\[ A_k' = A_k \left[ \frac{1}{2} (A_k + A_3) \right] \]

and

\[ A_3' = A_3 \left[ \frac{1}{2} (A_k + A_3) \right] \]

and these adjusted values were used in calculating variabilities. Again for TC, variance was set to CV2, and the paired differences were represented by relative differences \( (A_k - A_3) \left[ \frac{1}{2} (A_k + A_3) \right] \).

For both methods, we excluded outliers before all estimations. Removal of outliers was based on 1 studentized residual sweep by testing whether any observation was outside of the average ± 3 SD of the remaining observations. If so, such an observation was called an outlier and excluded from future analyses. For K, either the SD of the duplicate SS-AS results was used, when using method 1, or the difference with the referee laboratory’s AS result (i.e., A3) was used, when using method 2. Removing outliers for TC was done similarly except that the CV for the duplicate SS-AS results was used instead of the SD for method 1, and the relative difference in AS results was used instead of the actual difference for method 2. Between 0.0% and 3.4% of all the results were outliers.

By using the Bartlett test, we verified homogeneity of the variance of the paired A2 – A3 differences for audit samples obtained from different participating laboratories. Since

Table I
Noncomputational Differences Between Methods Used to Assess Laboratory Result Variabilities

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Method 1</th>
<th>Method 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of nonanalytic processes on result variability</td>
<td>More than method 2</td>
<td>Less than method 1</td>
</tr>
<tr>
<td>Type of test material*</td>
<td>Both AS and SS</td>
<td>Only AS</td>
</tr>
<tr>
<td>Use of referee laboratory results</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

* Type of test material is audit sample (AS) or split specimen (SS).
method 2 was based on calculating the difference between 2 independently estimated variances (ie, those of \(A_k - A_3\) [where \(k = 1\) or \(2\)] and \(A_3\)) as opposed to a traditional variance estimation. Monte Carlo simulation was used to obtain confidence intervals for the result variabilities. Specifically, 1,000 simulations were performed, introducing uncertainty from 2 sources (the estimated referee laboratory’s result variability, \(\text{var}(A_3)\), and the estimated variability of paired AS result differences, \(\text{var}(A_k - A_3)\)) under the assumption that each of these acted as variables with chi squared distributions.

**Additional Statistical Considerations**

For both analytes, 2 assumptions were validated for method 1. The first was that the audit process itself did not produce a bias in laboratory results. This was done by testing (and failing to reject) the null hypothesis that mean \((S_k) = \text{mean}(A_3)\) using the paired \(t\) test. The second assumption was that variabilities of paired SS and AS results were not significantly different. This was done by testing (and again failing to reject) the null hypothesis that pooled variance \((S_1^2 - S_2^2) = \text{pooled variance}(A_1 - A_2)\) using the \(F\) test. For TC, we justified equivalence of pooling CV\(^2\) to pooling variance, based on a proof using Taylor’s series expansion (not described herein) and the observation that for this analyte, measurement variability is small relative to the magnitude of the measurement.

Note that the required relationship of method 2 does not hold exactly under 2 complicating circumstances. The first complication results if there is a nonconstant bias between \(A_k\) and \(A_3\), for example if systematic differences between the measurement. One was based on the maximum allowable variability by using the mid range of the abscissa of the Westgard operation process specification charts quality control lines\(^9\) using CLIA PT standards at bias = 0 (CLIA PT-based variability limits). For K, this was an SD of 0.114 mmol/L, while for TC it was a CV of 2.25%. The second criterion was derived from the maximum allowable imprecision based on biologic variation data,\(^10\) setting maximum CV to one half of published intraindividual CV (biologically based variability limits). For TC, this was 3.05%. The biologically based criterion for K was not used because this analyte had a constant variance structure.

**Results**

**Variability of Total Cholesterol Results**

With method 1, the variability (CV) in TC results for the referral laboratory was 2.2%, and it ranged from 1.0% to 3.5% for the 6 participating facilities that performed TC testing. With method 2, variability in TC results for the referral laboratory was 1.9%, and it ranged from 1.6% to 3.7% for the participating facilities. These may be compared with maximum variability (CV) limits of 2.25% and 3.05% based on CLIA PT-based and biologically based criteria, respectively (see “Materials and Methods”). The estimated variabilities along with 95% confidence intervals are illustrated in **Figure 2**.

**Variability of Potassium Results**

With method 1, variability (SD) in K results for the referral laboratory was 0.098 mmol/L, and it ranged from 0.096 mmol/L to 0.168 mmol/L for the 5 participating facilities that performed K testing. With method 2, variability in K results for the referral laboratory was 0.089 mmol/L, and it ranged from 0.035 mmol/L to 0.090 mmol/L for the participating facilities. These may be compared with maximum variability (SD) limit of 0.114 mmol/L based on CLIA PT-based performance criterion (see “Materials and Methods”). The estimated variabilities along with 95% confidence intervals are illustrated in **Figure 3**.

**Discussion**

Laboratory result variability (related to test imprecision) results from random errors in the testing process and, along with result bias that results from systematic errors in the testing process, contributes to result inaccuracy (ie, total analytic error). Although both contribute to inaccurate laboratory results, in the context of monitoring patients’ conditions in the same facility, variability is more relevant...
provided that bias does not change systematically with time. Any assessment of laboratory result variability requires some form of repeated test measurements; however, the data collected in the present study are likely to provide different assessments of result variability than usual laboratory method evaluations that include determination of testing imprecision by repeatedly analyzing the same specimen pools over time. Such imprecision studies may underestimate the true test result variability under somewhat controlled conditions (eg, same and generally more experienced operators, knowledge of being involved in the imprecision study, resulting in estimated test result variabilities that may be lower than actual field experience).

In the present study, 2 different methods were studied to assess test result variability for 2 common laboratory tests, TC and K, in 11 participating laboratories and a referral laboratory. One method (method 2) was based on the strict use of an AS scheme, using a referee laboratory with known test result variabilities. This is probably a more accurate method for assessing analytic variability because all audit
Shahangian and Cohn / Variability of Laboratory Test Results

samples arise from the same SS (S3). For example, for K if S3 is hemolyzed, results for A1, A2, and A3 will be artificially elevated. This is an advantage for this method if one is concerned primarily with analytic sources of result variability, but it is a limitation if one desires to identify certain nonanalytic factors affecting laboratory results. However, a major limitation of this method, aside from its involved computational procedure, is that a referee laboratory with known test result variabilities needs to participate.

Our findings demonstrate that sometimes different conclusions may be drawn on the adequacy of laboratory test result variability depending on the analytic approach used for data analysis. Differences between the 2 methods were more pronounced for K result variability compared with TC result variability. For K, method 1 consistently provided greater SDs (Figure 3). Methods 1 and 2, however, provided fairly comparable results for TC (Figure 2). One explanation for these observations may be that among the 2 analytes evaluated in the present study, K is more susceptible to relevant nonanalytic factors, such as hemolysis and other effects of serum-clot contact time,11 that can affect its measured concentration. Another source potentially contributing to the difference between the 2 analytes may lie in the different assessment of variability of the referee laboratory results, since this laboratory’s TC result variability was based on the CDC Lipid Standardization Program’s database, while its K result variability was assessed using an imprecision study.

Since method 1 did not use a referee laboratory with known and established result variability, no prior assessment of variability of a referee laboratory results is required. Therefore, although method 2 is better to assess only variability of the analytic stage of the testing process, method 1 is more practical. We used method 2 as a comparison method to find out how laboratory result variability obtained using the practical method 1 approach, based on AS-SS pairs, compared with an impractical or theoretical assessment of result variability, ie, method 2. TC results exhibited relatively small differences between the 2 methods. For K, although it was expected that method 1 would provide greater variability, the extent of increased variability (ie, some 2-fold) using method 1 over method 2 could be determined only by conducting the present study. Finally, the present study is the only one of its kind in which such a well-controlled empiric scheme was used to determine test result variability. We are not aware of method 2 ever being used to assess test result variability (related to testing imprecision). One should note that to assess laboratory test result variability, an alternative experimental design may be used with the practical proviso of the participating and referral/referee laboratories being the same. Such experiments are already part of the quality assurance programs of many clinical laboratories.

The present study, unlike our previous ones, does not evaluate the feasibility6 or implementation7 of an AS-SS experimental design; rather, it uses the design used in the implementation study7 along with 2 data analytic approaches to assess variability of laboratory test results. In the implementation study,7 to evaluate an AS-SS design, a result discrepancy rate was used that is related to an additive function of both result variability and bias. The major emphasis of the present study was to use data analytic schemes using an evaluated AS-SS design7 to assess only test result variability. Our goal was not to actually determine variability of TC and K test results with any generalizable certainty, considering the small number of patient samples and specimens as well as facilities used. Moreover, other factors limiting the generalizability of our results include the laboratory selection bias owing to the small sample (11 facilities) and oversampling certain types of facilities in a limited geographic region.

Method 1, based on the evaluation of the results obtained from a specimen and an AS obtained from the same specimen, can become part of the quality assurance programs of any laboratory as audit samples are collected routinely in many laboratories for possible retesting to verify accuracy of a questionable result or to assess the reliability (ie, reproducibility) of laboratory results. Although method 2, as described in the present article, is computationally and logically more involved, a modification of it may be used by making the participating and referee laboratories the same. Still an assessment of the test result variability of the laboratory in question needs to be made, but the computational challenge of a nonconstant bias and the logistic issue involving participation of a second referee laboratory can be obviated, since the same laboratory would be analyzing both audit samples.

Conclusions

Variability of laboratory test results may be assessed using an AS-SS design. For some analytes (such as K), analytic variability may be better assessed using a less practical, purely AS design. To assess laboratory test result variability, an alternative AS-SS design may be used with the practical proviso of the participating and referral/referee laboratories being the same. Such AS-SS designs are being used in some clinical laboratories as part of their overall quality assurance programs.

From the 1Laboratory Practice Assessment Branch, Division of Laboratory Systems, Public Health Practice Program Office, Centers for Disease Control and Prevention, Atlanta, GA; and the 2Statistics and Public Health Research Division, Analytical Sciences Inc, Durham, NC.

An account of this work was presented at the national meeting of the American Association for Clinical Chemistry, New York, New York, in January 1999.

Address reprint requests to Dr Shahangian: Laboratory Practice Assessment Branch, Division of Laboratory Systems, Public Health Practice Program Office, CDC, 4770 Buford Highway, NE, Mail Stop G-23, Atlanta, GA 30341-3724.

Acknowledgments: We are indebted to D. Joe Boone, PhD, John S. Hancock, BS, and Harvey B. Lipman, PhD, Division of Laboratory Systems, Public Health Practice Program Office, CDC, for their contributions to the design, development, implementation, and completion of this study.

References