System to Monitor a Portion of the Total Testing Process in Medical Clinics and Laboratories: Evaluation of a Split-Specimen Design

Shahram Shahangian,1* Richard D. Cohn,2 Edward E. Gaunt,2 and John M. Krolak1

To evaluate a split-specimen design to identify problems in the testing process in hospital and physician office laboratories, we examined the testing for serum total cholesterol (n=646) and potassium (n=732) at 11 medical clinics evaluating 30–199 patients (mean, 125). Clinic personnel collected three tubes of blood from each patient. One specimen was processed routinely, the second was sent to a referral laboratory (RL), and the third specimen was sent to a holding facility for storage. The corresponding stored sample was retrieved and divided into three audit samples randomly and when result difference for the first two specimens exceeded critical values; one audit sample was sent to the original participant, the second to the RL, and the third to a referee laboratory. When three criteria were used, the result discrepancy rates were 2.5–8.7% for potassium and 1.5–4.6% for cholesterol. The split-specimen design could be implemented and evaluated as a monitoring system for a portion of the testing process.

The quality of clinical laboratory testing is important in promoting and maintaining the public’s health. In 1988, the US Congress passed CLIA in response to concern that laboratory problems were an important public health problem. During the CLIA debate, little scientific evidence was available to document the frequency and type of laboratory problems or their impact on patient care (1). Provisions of the CLIA, therefore, directed the US Department of Health and Human Services to study and determine the nature and extent of such problems. The CDC reviewed the scientific literature relating to CLIA (2). This review and subsequent reports suggested that laboratory problems, at least in some areas of clinical laboratory testing, occurred 1–5% of the time (3–10). The rate of clinically significant errors affecting patient outcome was even smaller, 0.1% (9–11). These rates, however, are based on self-reports by medical clinics, clinical laboratories, and other sites involved in the total testing process (TTP). The true problem rate is greater, and is a function of the observed problem rate, the effectiveness of existing monitoring and reporting systems, and how a problem is defined (4, 5).

In 1990, CDC began conducting the five studies listed in Section 4(a) of the CLIA, including, “... study of the effect on laboratory test accuracy of errors in each of the components of the clinical testing process ... “. In late 1990, CDC solicited a study proposal to address the five study areas outlined in CLIA. Insufficient resources, however, prevented implementation of the proposed design (12, 13). Therefore, CDC began implementing focused, independent projects to address issues that would have been covered by the more comprehensive design. In 1994, CDC began developing and evaluating a prototype process that used a split-specimen (SS) design to determine the frequency and type of problems that occurred in certain portions of the TTP for specific laboratory tests performed in both hospital laboratories (HLS) and physician office laboratories (POLS). A 3-month feasibility study was conducted to evaluate the use of an SS process to determine the frequency and type of problems that
occurred in a portion of the TTP (14). On the basis of these findings (14), the present full-scale evaluation study was initiated to evaluate the SS design and its logistics, to assess the usefulness of the SS methodology for measuring test result discrepancies, and to assess the feasibility and effectiveness of the SS process as a quality-assessment system for laboratory testing.

**Materials and Methods**

**SS DESIGN**
Clinic personnel collected three tubes of blood from each patient. The first specimen (S1) was processed routinely, the second (S2) was sent to a referral laboratory (RL) for testing, and the third (S3) was sent to a central specimen holding facility where the serum was stored at −20 °C (Fig. 1). When the percentage of the difference between the result obtained from the S1 specimen (S1) and that from the S2 specimen (S2) exceeded set values [± 6.4% for total cholesterol (TC) and ± 8.5% for potassium (K)], the corresponding S3 sample was divided into three audit samples (A1, A2, and A3). Sample A1 was sent to the original participating laboratory, A2 was sent to the RL, and A3 was sent to a referee laboratory. Furthermore, a random audit sample analysis was initiated for up to 30 (mean, 28) patients per clinic. Subsequently, during clinic site visits, the S1 result was abstracted from the participating patient’s medical record.

**PARTICIPATING FACILITIES**
All 11 facilities (8 POLs and 3 HLs) were located within 240 kilometers (150 miles) of Durham, NC. Selection criteria for POLs, in addition to geographic location, were an annual volume of ≥5000 tests and performing each of the two (TC and K) tests. Four of the POLs served 1–9 (mean, 6.0) clinicians in different multispecialty practices, two POLs served 1 and 8 clinicians in internal medicine practices, and two POLs served 3 and 48 clinicians in family practice settings. Laboratories in three hospitals ranging in size from ≥100 to ≤1000 beds each served a referral multispecialty practice: 13 clinicians in an internal medicine practice, and 2 clinicians in a family practice setting.

**REFERRAL LABORATORY**
A commercial RL central to the study participants was selected. The RL analyzed specimens within 18 h of blood

---

**SPLIT-SPECIMEN DESIGN**

![Fig. 1. SS design used in this study.](image-url)
collection, with results then electronically transmitted and imported into the study database.

REFEREE LABORATORY
An academic referee laboratory was selected to adjudicate testing discrepancies and to conduct audit-sample analysis. This 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 (15) and participated satisfactorily in an accredited College of American Pathologists proficiency testing (PT) program for TC and K.

HOLDING FACILITY
Sera obtained from S3 specimens were sent to Analytical Sciences, Inc., the contractor for this study, where the integrity of each sample was examined visually and noted if compromised, recentrifuged if necessary, and stored at −20 °C for possible future retrieval. Randomly selected samples (up to 30 per clinic) were retrieved and divided into three audit samples. Samples were also divided when the percentage of the difference between S1 and S2 results exceeded set values (see the discussion of SS design earlier in this section).

PATIENT SELECTION
Any patient from 18 to 80 years of age for whom the laboratory test was ordered for a specific clinical reason and not as part of a laboratory test profile was eligible for inclusion. Excluded patients were those who were unable or unwilling to provide informed consent, those who had given ≥100 mL of blood, or those whose blood was collected by fingerstick. The six facilities involved in TC testing recruited 30–191 patients (mean, 108), whereas the five facilities involved in K testing recruited 75–199 patients (mean, 146) into the study. The number of patients recruited by each facility was positively related, although not linearly proportional, to its typical test volume.

LABORATORY TESTS USED
Serum TC and K were selected because of their frequent use for both ambulatory and hospitalized patients, their clinical relevance, specimen stability, and the availability of a standard reference method. There were 646 patients with TC specimens (four POLs and two Hls) and 732 patients with K specimens (four POLs and one HL). Laboratory measurements were made using cholesterol oxidase methods (for TC) and ion-selective electrodes (for K).

SS RESULT DIFFERENCES
Two different types of discrepancy criteria were used in this study. One was based on predefined standards (standard-based), whereas the other was based on the actual data collected from each facility, taking into account each center’s estimated measurement variability and its systematic result differences compared with the RL (data-driven). For assessing SS result discrepancies, two methods used predefined standards to determine whether SS result differences exceeded set thresholds. Standard-based methodologies are comprehensive because they may identify result discrepancies that are associated with any causes and are more widely accepted than their data-driven counterpart because they are based on preexisting and validated standards, allowing identification of discrepancies as data are received. The data-driven discrepancy method, on the other hand, tends to detect result discrepancies that are “out-of-context” on the basis of each facility’s results. The data-driven discrepancy method is, therefore, designed to identify isolated problems within the TTP by adjusting for systematic result differences between laboratories. Therefore, result discrepancy identification using data-driven criteria requires prior analysis of laboratory results from each facility.

For standard-based discrepancy criteria, denoting maximum allowable variance by $s^2$:

$$s^2(S_1 - S_2) = s^2(S_1) + s^2(S_2) = 2s^2$$

so that

$$Maximal\ allowable\ s(S_1 - S_2) = (2s^2)^{1/2} = \sqrt{2s}$$

The critical difference for the standard-based discrepancy criteria was obtained as follows:

$$Critical\ difference = 3\sqrt{2}s$$

where we used a multiplier of 3 for $s$ in conformity with a commonly accepted outlier criterion of 3 standard deviations. The maximum allowable standard deviation ($s$) was obtained as either the mid-range of the abscissa of the Westgard operation process specification charts quality-control lines using CLIA PT standards at bias = 0 [Westgard CLIA PT-based discrepancy method (16)], or was set to the maximum allowable imprecision on the basis of published biological variation data for each analyte, 1/2(intraindividual CV) [biologically-based discrepancy method (17, 18)]. The critical limits to define the SS result differences as discrepant were ± 9.6% ($s = 2.25\%)$ and ± 0.48 mmol/L ($s = 0.114$ mmol/L) for TC and K, respectively, when the CLIA-based discrepancy criteria were used, whereas they were ± 12.9% [$s = 1/2(6.1\%)$] and ± 10.4% [$s = 1/2(4.9\%)$] for TC and K, respectively, when the biologically-based discrepancy criteria were used.

Data-driven result discrepancies were identified by reexamining the $S_1$ and $S_2$ differences after accounting for systematic result differences exhibited by the participating facility and the RL and for measurement variability associated with the testing process. Generally, if the systematic result difference between $S_1$ and $S_2$ ($S_1 - S_2$) is denoted by $d$ and the standard deviation of the $S_1 - S_2$ difference is denoted by $s_d$, then the criterion for declaring a data-driven result discrepancy is $|S_1 - S_2| \geq 3s_d$. In practice, the systematic difference between $S_1$ and $S_2$ was characterized by a linear relationship rather than a
constant \( d \). Measurement error modeling procedures, with exclusion of outliers, were used in achieving this characterization (19). In the scatter plots showing the relationship between \( S_1 \) and \( S_2 \) using data-driven discrepancy criteria, \( S_1 \) was adjusted to eliminate any systematic difference between \( S_1 \) and \( S_2 \) results.

For the assessment of testing bias of the six laboratories whose TC testing was examined, an adjustment for \( S_2 \) results was made as follows: First, the referee laboratory’s \( A_3 \) results were adjusted to reflect the bias exhibited by the referee laboratory relative to its peer group in the CDC Lipid Standardization Program by linear regression, examining nearly 2 years of the referee laboratory’s Lipid Standardization Program TC performance data. Next, the \( S_2 \) result from the RL was adjusted by linear regression to account for the bias exhibited by the RL relative to the referee laboratory. Finally, to assess the participating laboratories’ bias, the linear relationship between \( A_1 \) and the adjusted \( A_3 \) was examined using measurement error modeling procedures while excluding outliers as described above. Bias for results from laboratories performing \( K \) testing was not computed because a laboratory standardization program for this analyte did not exist.

DECISION TREE ALGORITHM TO ASSESS EFFECTIVENESS OF THE SS DESIGN

We were interested to note what proportion of the time an \( S_1 \) result was consistent with its corresponding \( S_2 \) and the three audit sample results when the \( S_1 \) and \( S_2 \) results were called not discrepant by either standard-based discrepancy criterion. This helps in assessing the effectiveness of not calling an SS discrepancy in predicting a consistent (and probably “correct”) \( S_1 \) result. Alternatively, we wanted to know what proportion of the time an \( S_1 \) result was inconsistent with its corresponding \( S_2 \) and the three audit sample results when the \( S_1 \) and \( S_2 \) results were called discrepant by either standard-based discrepancy criterion. This would help in assessing the effectiveness of calling an SS discrepancy in predicting an inconsistent (and probably “incorrect”) \( S_1 \) result. A decision tree algorithm was used when audit sample results were available to assess if the \( S_1 \) and \( S_2 \) results that were discrepant or not discrepant by the standard-based discrepancy criteria were discrepant or not discrepant when the same discrepancy criteria were used and the \( S_1 - A_3 \), \( S_1 - A_2 \), and \( S_1 - A_1 \) differences were examined (Fig. 2). When the \( S_1 - S_2 \) difference exceeded the critical limits set by each discrepancy criterion for each analyte, using either one of the two standard-based discrepancy criteria, the \( S_1 \) result was called “in error” by the SS design. The following parameters were used to assess the effectiveness of the SS design in calling the \( S_1 \) result in error (discrepant) or not in error (not discrepant):

- sensitivity, the proportion of \( S_1 \) results called in error by the decision tree algorithm that was also called in error by the SS design;
- specificity, the proportion of \( S_1 \) results called not in error by the decision tree algorithm that was also called not in error by the SS design;

![Fig. 2. Decision tree algorithm used to call \( S_1 \) results in error.](image-url)
vs S2 and the estimation of intralaboratory variability as the coefficient of variation (CV, %) rather than as standard deviation. We excluded outliers from TC concentration, we chose to express intralaboratory variability based on A1 – A2 and A2 – A3 differences, by determining the studentized residual r and calling results as outliers if \( |r| > 3.0 \).

Regression lines were fitted using measurement error regression procedures, rather than simple least squares (19). This was necessary because the value represented by the independent variable was measured with some error, and simple least-squares analysis would have underestimated the slope of linear regression lines.

Audit sample results were used to characterize measurement variabilities associated with the participating laboratories and with the RL. Differences in the audit sample results between the participating and referee laboratories and between the RL and referee laboratory were used to estimate these variabilities because we accounted for the measurement variability of the referee laboratory by using data obtained from the specimen pool (TC) of the Lipid Standardization Program and from six measurements on each of nine selected specimens (K). The variability between the measurements by the participating laboratories and the RL was estimated using these estimates of the measurement variability at the referee laboratory together with the estimated variability associated with the A1 – A3 and A2 – A3 differences for randomly selected audit samples. Outliers were excluded before computing these estimates.

Because of considerable variation in result discrepancy rates across laboratories, standard statistical approaches were not appropriate for some statistical analyses. Specialized analysis techniques were used to account for the variation in result discrepancy and problem rates across facilities. In particular, variance estimation for the odds ratio analysis was based on first-order Taylor series approximations (20) to accommodate the complex sampling design; these were analyzed using the SUDAAN software (Research Triangle Institute) package (21).

**Characterization of Standard-Based Result Discrepancies**

Result discrepancies identified using either of the two standard-based result discrepancy criteria (CLIA PT-based and biologically-based) were attributed to three possible causes: isolated (classified as discrepancies by the data-driven criterion), measurement variability, or systematic result differences (biases) between laboratories. The following algorithm was used in classifying standard-based result discrepancies:

(a) standard-based discrepancies that were also identified as data-driven discrepancies were classified as attributable to isolated problems in the TTP;

(b) of the remaining result discrepancies, those that fell outside of either of the two standard-based discrepancy bounds, after S1 was adjusted for any bias against S2, were classified as attributable to measurement variability; and

(c) any remaining discrepancies were ascribed to systematic result differences between participating and referral laboratories.

**Documented Problems**

Clinic and laboratory records corresponding to individual patients and their corresponding specimens were reviewed. All problems were grouped into either a major stage of the TTP (preanalytical, analytical, and postanalytical) or as study-induced when the facility failed to follow the study protocol.

**Data Management**

A customized computer application was used for double-blind, 100% rekey data entry. Data entry inconsistencies were identified and remedied by a senior data manager.

**Results**

Scatter plots of S1 vs S2 results for all but two facilities are shown in Fig. 3. Each of the nine panels (A–E, G, H, J, and K) correspond to one of nine facilities. Each panel shows
the $y = x$ (dotted) and the linear regression (dashed) lines as well as the lines defining result discrepancies by the biologically-based discrepancy method. Scatter plots of $S_1$ vs $S_2$ for facilities F and I are shown in Fig. 4. The three panels for each facility, in addition to the $y = x$ (dotted) and the linear regression (dashed) lines, show the lines defining result discrepancies by the CLIA PT-based (left panel), biologically-based (center panel), and data-driven
Discrepancy methods. In the right panels, $S_1$ actually represents an adjusted $S_1$ result; therefore, the regression line is forced to correspond to the $y = x$ line.

| Table 1 lists both the observed and adjusted discrepancy rates for TC and K, using each of the three discrepancy criteria. The observed data-driven discrepancy rates were less than the observed CLIA PT-based discrepancy rates for both TC (1.5% vs 4.6%) and K (2.5% vs 8.6%). Although the observed data-driven discrepancy rate was less than the observed biologically-based discrepancy rate for K (2.5% vs 8.7%), it was nearly the same as the observed biologically-based discrepancy rate for TC (1.5% vs 1.4%).

Result discrepancy rates derived using the two standard-based (i.e., CLIA PT-based and biologically-based) discrepancy criteria for TC were also determined after accounting for the bias of the $S_2$ results from the RL relative to the $A_3$ results from the referee laboratory after the latter was adjusted for its bias relative to the group mean of the Lipid Standardization Program (see Materials and Methods). This caused the TC discrepancy rates to be reduced from 4.6% (Table 1) to 2.6% (not shown in Table 1) when the CLIA PT-based discrepancy criterion was used and from 1.4% (Table 1) to 0.9% (not shown in Table 1) when the biologically-based discrepancy criterion was used. Such an accounting for bias does not affect data-driven discrepancy rates because this discrepancy criterion is designed to account for systematic result differences between laboratories.

EVALUATION OF THE SS DESIGN FOR ASSESSMENT OF RESULT DISCREPANCIES

When the decision tree algorithm (Fig. 2) was used, the efficiency of the SS design to call a result discrepant was 93–98% for TC and 79–84% for K. Table 2 lists the efficiency, predictive values, sensitivity, and specificity of...
CHARACTERIZATION OF STANDARD-BASED RESULT DISCREPANCIES

Result discrepancies, as determined by either the CLIA PT-based or biologically-based discrepancy criterion, were classified as resulting from isolated problems (data-driven discrepancy), from excessive measurement variability, or from systematic result differences between participating and referral laboratory results (see Materials and Methods). The causes of the two standard-based result discrepancies for both analytes are listed in Table 3.

For TC, no discrepancies were attributed to measurement variability. However, for K, more discrepancies were attributed to measurement variability (4%) than to either systematic result difference (bias) between laboratories (2.7%) or isolated problems implied by the data-driven discrepancy method (1.9%). When either of the two standard-based criteria for K was used, ~50% of discrepancies could be ascribed to excessive measurement variability, whereas ~30% of discrepancies could be attributed to systematic result differences between the participating laboratory and the RL. For both analytes, 20–70% of standard-based discrepancies were a result of systematic result differences.

PARTICIPANTS’ PROBLEM-MONITORING SYSTEMS

Except for the usual quality-assurance procedures cited by the participating hospital laboratories, we found no evidence of existing monitoring systems specifically designed to detect problems throughout the TTP. All participating laboratories had facility-specific preanalytical procedures for specimen collection and processing for testing, analytical procedures for conducting requested tests on the provided specimens, and postanalytical procedures for processing and posting test results. All participating laboratories performed daily quality-control testing and participated in interlaboratory PT programs as required by CLIA for laboratory accreditation and licensure. We observed only two instances (0.1%) where actions had been taken because of problems identified by existing quality-assurance systems. In the first instance, the laboratory recommended that a specimen be re-collected because of hemolysis. In the second, quality-control testing identified instrument calibration problems that resulted in the re-analysis of the study specimen.

DOCUMENTED PROBLEMS

For the 1378 patients included in this study, 40 problems were identified by review of medical and laboratory records; all of these occurred in 5 of the 11 participating facilities. Fourteen (35%) of the problems detected during review of records were study-induced (the study protocol was not adhered to or the study itself caused a facility to commit an error they would not have committed routinely). This included eight patients, all from the same clinic, who were inappropriately recruited because thyroid tests were ordered instead of TC tests and four patients from another clinic for whom TC was assayed twice, once to report the result to Analytical Sciences, Inc. (the study contractor), and a second time the next day to incorporate

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Discrepancy criterion</th>
<th>Observed rate</th>
<th>Adjusted rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>CLIA PT-based</td>
<td>4.6%</td>
<td>3.5%</td>
</tr>
<tr>
<td></td>
<td>(0.0–13.3%)b</td>
<td>(0.0–7.2%)</td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>Biologically based</td>
<td>1.4%</td>
<td>1.3%</td>
</tr>
<tr>
<td></td>
<td>(0.0–3.2%)</td>
<td>(0.0–2.5%)</td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>Data-driven</td>
<td>1.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0–4.3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>CLIA PT-based</td>
<td>8.6%</td>
<td>8.6%</td>
</tr>
<tr>
<td></td>
<td>(4.4–12.6%)</td>
<td>(0.6–17.7%)</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Biologically based</td>
<td>8.7%</td>
<td>10.6%</td>
</tr>
<tr>
<td></td>
<td>(4.4–13.3%)</td>
<td>(2.6–18.6%)</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Data-driven</td>
<td>2.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.6–8.7%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Adjusted discrepancy rates were determined on the basis of the discrepancy assessment using: (a) the decision tree algorithm (Fig. 2) in cases where audit-sample results were available; or (b) the predictive values for positive or negative results for the split-specimen design in cases where audit-sample results were not available.

b The ranges of the discrepancy rate (in parentheses) refer to the range for the six laboratories performing total cholesterol tests and the five laboratories involved in potassium testing.

Table 2. Efficiency, predictive values, sensitivity, and specificity of the split-specimen design.

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Criterion</th>
<th>Efficiency</th>
<th>PV(+)</th>
<th>PV(−)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>CLIA PT</td>
<td>93%</td>
<td>50%</td>
<td>99%</td>
<td>83%</td>
<td>94%</td>
</tr>
<tr>
<td></td>
<td>(81–100%)</td>
<td>(0–86%)</td>
<td>(96–100%)</td>
<td>(75–100%)</td>
<td>(81–100%)</td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>Biological</td>
<td>98%</td>
<td>67%</td>
<td>100%</td>
<td>86%</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td>(97–100%)</td>
<td>(0–100%)</td>
<td>(99–100%)</td>
<td>(50–100%)</td>
<td>(97–100%)</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>CLIA PT</td>
<td>84%</td>
<td>52%</td>
<td>95%</td>
<td>80%</td>
<td>85%</td>
</tr>
<tr>
<td></td>
<td>(76–97%)</td>
<td>(13–86%)</td>
<td>(90–100%)</td>
<td>(60–100%)</td>
<td>(77–97%)</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Biological</td>
<td>79%</td>
<td>43%</td>
<td>93%</td>
<td>68%</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>(71–97%)</td>
<td>(0–88%)</td>
<td>(86–100%)</td>
<td>(0–100%)</td>
<td>(73–96%)</td>
<td></td>
</tr>
</tbody>
</table>

* The ranges in parentheses correspond to those obtained for the six laboratories involved in TC testing and the five laboratories performing K tests.

PV(+) and PV(−) are the predictive values for positive and negative results, respectively.
the result into each patient’s lipid panel. Two study-induced problems occurred in a third clinic. One problem occurred when the clinician had not specifically ordered a K test on the S1 specimen although the patient had been recruited in the study. The other study-induced problem occurred when the clinic inadvertently processed an S2 specimen and sent it for testing by a laboratory that was not participating in this study. Of the remaining 26 problems, two (8%) were preanalytical. In one case, the actual specimen collection date and the collection date reported on the test result differed by 1 day. In the other case, involving a different facility, when a worker indicated that the S1 specimen for K testing was hemolyzed, another specimen was requested by the clinician. The patient was recalled on another date, and the S1 result from a second specimen was reported for comparison with the S2 result from the original specimen. Twenty-two of the 26 (85%) problems were postanalytical (2 in facility A, 1 in facility F, 1 in facility G, and 18 in facility J). Fourteen of these problems related to the untimely posting of the patient’s result on the medical record (delays exceeding 2 weeks). Eleven of these 14 problems were associated with facility J. The remaining eight postanalytical problems were all related to transcription of laboratory results, seven of which were also associated with facility J. This was the only facility in the study that did not use a computer for managing clinical or laboratory data. All results were transcribed by hand from the analyzer printout to an accession log and then onto the medical record. Because of understaffing and the high workload in facility J, results were not always timely and transcribed accurately. Because of three discrepancies with potential clinical impact involving three patients (TC S1 = 4.55 mmol/L vs S2 = 6.34 mmol/L, TC S1 = 4.78 mmol/L vs S2 = 8.87 mmol/L, and K S1 = 5.3 mmol/L vs S2 = 4.8 mmol/L), we reviewed all available medical and laboratory records for these patients extensively. Although we suspected that the preanalytical specimen for one of these three patients was switched with another study patient, we could not document it. In this case, which involved TC testing, two patients from the same clinic whose blood was collected only 5 min apart provided what we believe is strong, albeit circumstantial, evidence for preanalytical specimen switches. For TC testing, one-half of the documented problems were study-induced, and problem rates were 20–40% of the rates for K (Table 4). For K tests, 90% of problems were related to the routine testing process, and only 10% were study-induced; the 24% problem rate refers to facility J.

Of the 24 problems identified during document review that were not study-induced and for which complete results were available, 5 result discrepancies were identified. Of the remaining 1354 cases with complete data (among which no document review problems were found), 100 were associated with result discrepancies. Thus, we were 3.3 times more likely to observe result discrepancies when problems could be documented and vice versa. The odds ratio of 3.30 was significantly ($P = 0.0125$) different from 1.

**CLINICAL IMPACT OF RESULT DISCREPANCIES**

Laboratory data may be used to assess the nature and extent of a potential impact resulting from result discrepancies. Our only attempt to assess clinical impact was TC result classification by the National Cholesterol Education Program screening guidelines of ≥5.17 mmol/L (≥200 mg/dL) for moderate or high risk and ≥6.21 mmol/L (≥240 mg/dL) for high risk of developing cardiovascular disease (22). Of the 646 patients with TC results, only 2 patients (0.3%) had discrepant results: one result was <5.17 mmol/L, whereas the other was ≥6.21 mmol/L (Table 5). For one of these two patients, the medical record noted that the TC concentration had markedly improved since the patient’s previous visit only 3 months earlier. The reported and observed S1 results were both 4.55 mmol/L, but the result for the previous visit had been

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Overall rate</th>
<th>Rate specific to TTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>1.2%</td>
<td>0.6%</td>
</tr>
<tr>
<td></td>
<td>(0.0–3.4%)</td>
<td>(0.0–2.2%)</td>
</tr>
<tr>
<td>K</td>
<td>3.0%</td>
<td>2.7%</td>
</tr>
<tr>
<td></td>
<td>(0.0–24.0%)</td>
<td>(0.0–24.0%)</td>
</tr>
</tbody>
</table>

* The ranges in parentheses correspond to those obtained for the six and five participating laboratories performing TC and K tests, respectively.

---

**Table 3. Standard-based discrepancy rates classified by cause.*a**

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Criterion</th>
<th>Isolated</th>
<th>Attributed to variability</th>
<th>Attributed to bias</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>CLIA PT</td>
<td>1.4%</td>
<td>0.0%</td>
<td>3.3%</td>
<td>4.6%</td>
</tr>
<tr>
<td></td>
<td>Biological</td>
<td>1.1%</td>
<td>0.0%</td>
<td>0.3%</td>
<td>1.4%</td>
</tr>
<tr>
<td>K</td>
<td>CLIA PT</td>
<td>1.9%</td>
<td>4.0%</td>
<td>2.7%</td>
<td>8.6%</td>
</tr>
<tr>
<td></td>
<td>Biological</td>
<td>1.9%</td>
<td>4.1%</td>
<td>2.7%</td>
<td>8.7%</td>
</tr>
</tbody>
</table>

*a The ranges for the discrepancy rate (in parentheses) refer to the ranges for the six and five laboratories performing TC and K tests, respectively.

---

**Table 4. Rate of documented problems, overall and specific to the TTP.**

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Overall rate</th>
<th>Rate specific to TTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>1.2%</td>
<td>0.6%</td>
</tr>
<tr>
<td></td>
<td>(0.0–3.4%)</td>
<td>(0.0–2.2%)</td>
</tr>
<tr>
<td>K</td>
<td>3.0%</td>
<td>2.7%</td>
</tr>
<tr>
<td></td>
<td>(0.0–24.0%)</td>
<td>(0.0–24.0%)</td>
</tr>
</tbody>
</table>

*a The ranges in parentheses correspond to those obtained for the six and five participating laboratories performing TC and K tests, respectively.
5.28 mmol/L. SS testing yielded S2, A1, A2, and A3 results of 6.16–6.49 mmol/L (Table 5). Subsequent review of the S1 results for other study patients in the same clinic identified an S1 value of 6.10 mmol/L from another patient, whereas the corresponding S2 and audit sample results were all between 4.22 and 4.63 mmol/L. In these and a third case in which such a clinical impact criterion was met, only the S1 result was discrepant when compared with the S2, A1, A2, and A3 results.

### Discussion

Laboratory problems (also termed laboratory errors, blunders, mistakes, or defects) have generally been defined as incidences that significantly affect the accuracy of one or more laboratory results after they have been validated/verified and released (3–5, 23). Here we defined a problem as an incidence anywhere within the TTP that had either a “significant” effect on the accuracy and availability of laboratory results or caused a deviation from standard operating procedures. Most publications dealing with various TTP problems and their rates of occurrence are based on intra- or interinstitutional quality assurance (3–11, 23). Few of these studies use other means of assessing laboratory problems, such as an SS experimental design. Problem rates observed by the usual methods are obviously a function of the effectiveness of the existing laboratory problem monitoring systems as well as how a problem is defined (4, 5, 9, 23). Many of the steps in the TTP occur outside the laboratory; therefore, optimal quality improvements require broader institutional as opposed to laboratory-specific monitoring efforts.

The reasons for using an SS methodology as a laboratory quality-assurance system include its detection of problems not observed by other TTP quality-monitoring systems, which makes it a complementary means of quality assessment, and its objectivity once the criterion for result discrepancy has been defined. The SS methodology, however, should be considered in view of its limitations, which include (a) possible bias in both SS results from the laboratory whose performance is being assessed, or inaccurate results from the RL in cases in which, as in this study, the results from that laboratory are used for comparison; (b) insensitivity to TTP problems that may occur before specimens are collected and after results are reported; and (c) insensitivity to problems not impacting laboratory results, such as a switch between two specimens exhibiting similar analyte concentrations or composition. This study revealed that result discrepancy rates higher than ≏1.0%, as seen in published reports of laboratory error rates, were observed for both analytes. This comparison, however, should be made in light of the fact that our study is based on result discrepancy rates involving the referral and participating laboratories and that these rates cannot be equated to errors in laboratory test results. Our study design is different from virtually all other published reports in that (a) objective result discrepancy criteria are used on all split specimens obtained, whereas the other studies assessing laboratory test errors are highly dependent on the effectiveness of existing quality-assurance systems and detect testing problems in only a portion of the study population; (b) the definitions of a laboratory testing problem are probably different; (c) different parts of the TTP are probably monitored; and (d) our use of a different laboratory (RL), in contrast with most SS studies, avoids the possible existence of the same measurement bias in each SS analysis done by the same facility, which contributes to increased but valid discrepancy rates. We did address this issue successfully with the TC test because the bias in measurements from the RL could be evaluated by use of a referee laboratory participating in the CDC Lipid Standardization Program.

The first cited limitation of our SS design was evaluated by assessing the efficiency, predictive values, sensitivity, and specificity of the methodology, using a decision tree algorithm that was based on S1, S2, A1, A2, and A3 values. Although the absence of an SS result discrepancy was highly predictive of the S1 result not being in error (negative predictive value of 93–100%), the presence of an SS result discrepancy was much less predictive of the S1 result being in error (positive predictive value of 43–67%), which led to an efficiency of 93–98% for TC and 79–84% for K (Table 2). Adjusting the discrepancy rates for the predictive values of the SS design had little or no effect on two standard-based discrepancy rates, it decreased the CLIA PT-based discrepancy rate for TC (by 24%) from 4.6% to 3.5%, and increased the biologically-based discrepancy rate for K (by 22%) from 8.7% to 10.6% (Table 1).

It is critical that the operational logistics associated with the SS methodology be such that the actual testing process (from collection of specimens to reporting of results) is monitored and is not a variation of this process induced by the methodology itself. In our study, 14 of the

---

**Table 5. Discrepant TC concentrations (mmol/L) based on NCEP\(^a\) risk categorization.**

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>S1</th>
<th>S2</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6.10</td>
<td>4.60</td>
<td>4.22</td>
<td>4.63</td>
<td>4.55</td>
</tr>
<tr>
<td>A</td>
<td>4.55</td>
<td>6.34</td>
<td>6.16</td>
<td>6.49</td>
<td>6.36</td>
</tr>
<tr>
<td>B</td>
<td>4.78</td>
<td>8.87</td>
<td>8.85</td>
<td>8.79</td>
<td>8.97</td>
</tr>
</tbody>
</table>

\(^a\) NCEP, National Cholesterol Education Program.

\(^b\) The NCEP TC screening threshold is ≏5.17 mmol/L (≥200 mg/dL) for moderate or high risk and ≏6.21 mmol/L (≥240 mg/dL) for high risk of cardiovascular disease (22). The first case (S1 = 6.10 mmol/L) nearly, but not strictly, satisfied the definition of result discrepancy based on NCEP risk categorization.
40 problems (35%) that could be documented by retrospective review of medical and laboratory records were study-induced. We surmise that most of these problems emanated from the lack of familiarity of the participating facilities with the SS process. Although the SS design, as it was implemented here, had different laboratories analyzing the various specimens, facilities using the SS methodology for monitoring of the testing process may use this procedure differently in that all specimens are probably analyzed in the same laboratory. The major drawback of such a system, as stated earlier, is that SS result discrepancies are likely to be underestimated if the same laboratory is involved in testing because processes that lead to a biased (quantitative) or incorrect (qualitative) result may be operating during testing processes involving both specimens. It is unlikely, however, especially during the current period of increasing fiscal restraint, that medical facilities will use other laboratories for analyzing the specimens or audit samples for quality assurance of their testing processes.

The operational logistics associated with an SS design should be such that the procedure would not induce variation in the routine testing process. This was not quite the case in this study in that of the 40 problems documented in medical and laboratory records, 14 (35%) were not caused by routine processes within the TTP, but were study-induced problems. SS designs should, therefore, be constructed to minimize study-induced problems so that detected result discrepancies reflect actual problems within the testing process itself. Of the 26 TTP problems, only two (8%) could not be related to a stage of the TTP. Of the other 24 problems, none was of an analytical nature, two (8%) were related to the preanalytical stage of the TTP, and the remainder (92%) were postanalytical problems. This is in agreement with the observation by others that most problems are related to the nonanalytical stage of the TTP (5, 9–11). Such classification of problems is valuable in that knowledge of both the type of mistakes and the stage at which they occur may assist in maximizing problem detection so that measures may be taken to minimize their occurrence. However, what is missing from all published reports, including this one, is an evaluation of the impact of each problem type on test result accuracy and precision, turnaround time, medical decision making, disease and health management, and eventually, health outcome. Furthermore, studies that investigate problems in each step of the TTP should additionally assess the medical impact of these problems so that quality improvement efforts may focus on the more critical stages of the testing process.

In this study, problems in the TTP were identified directly only through retrospective medical and laboratory record review. However, the identification of result discrepancies is obviously motivated by the hypothesis that these discrepancies are indicative of TTP problems as well. Therefore, a strong association between result discrepancies and document review problems would lend support to this hypothesis; and, in fact, this study did reveal a significant association as evidenced by the odds ratio of 3.30, \( P = 0.0125 \). However, little redundancy was found between result discrepancies, as identified by at least one of the three discrepancy criteria, and problems identified by document review. Of the 105 result discrepancies, only 5 (5%) were also associated with document review problems. Of the 24 documented TTP problems for which complete results were available, only 5 (21%) were also associated with a result discrepancy. Therefore, used alone, either of these two quality-assessment measures (retrospective document review and result discrepancy analysis) is not likely to constitute as effective a means of identifying TTP problems as when they are used in combination.

The major emphasis of this study was to implement and evaluate an SS design, in conjunction with retrospective medical and laboratory record reviews, and to identify the nature and extent of problems within the TTP. As such, this report was not meant to describe how an SS design should be implemented but rather what can be learned by the implementation and evaluation of such an experimental design. Our goal was not to actually determine the frequency and type of problems within the TTP with any generalizable certainty. These results should be considered in light of laboratory selection bias attributable to the small sample size (11 facilities) and the oversampling of certain types of facilities in a limited geographic region. We found widely varying result discrepancy and documented problem rates among facilities, which also contributed to the uncertainty associated with any reported overall problem rate. Logistically, we encountered numerous difficulties in recruiting medical clinics and hospital and office laboratories to participate in this study. We faced a general resistance by clinicians as well as some laboratorians, which stemmed from complaints that they were too busy, overburdened with too much paperwork, understaffed, or that there was insufficient monetary incentive to participate. Once facilities agreed to participate, we received excellent cooperation, which allowed us to attain 92% of our collection goal of 1500 sets of split specimens.

In summary, our findings indicate that the SS methodology and its logistics could be implemented and evaluated, that the SS design used could provide a measure of result discrepancies with an overall efficiency (as defined in Materials and Methods) of 79–98% for the two analytes studied, and that in combination with retrospective review of medical and laboratory records (and perhaps other effective TTP quality-assurance systems), this methodology can serve as a monitor for a portion of the TTP (from collection of specimens to reporting of laboratory results).
We thank D. Joe Boone, Carlyn L. Collins, and Thomas L. Hearn of the Division of Laboratory Systems, Public Health Practice Program Office, CDC, for critical review of this manuscript and many insightful comments throughout the performance of this work. We also thank Philip J. Thompson, editor of the Public Health Practice Program Office, for his expert comments and suggestions, which have substantially improved the clarity and readability of this article. Finally, we are indebted to many of our colleagues from CDC, Analytical Sciences, Inc., and other institutions for contributions to various aspects of this project during its inception, implementation, and completion. These include D. Joe Boone, Carlyn L. Collins, James H. Handsfield, Thomas L. Hearn, and Harold W. Muir, Jr. from CDC, and Donald A. Holzworth, Maria N. Horner, Sylvia S. Hughes, and Margaret M. Sexsmith from Analytical Sciences, Inc. We also greatly appreciate the members of an expert advisory panel to Analytical Sciences, Inc., the contractor for this study, for their comments, input, and suggestions during the entire study period. The members of this panel were Dennis D. Boos (North Carolina State University, Raleigh, NC), George S. Cembrowski (University of Alberta, Edmonton, Alberta, Canada), John A. Koepe (Duke University Medical Center, Durham, NC), Laura C. Leviton (University of Alabama School of Public Health, Birmingham, AL), Charles W. Rhodes (Mount Pleasant Family Physicians, Mount Pleasant, NC), Bernard E. Statland (Statland Laboratory Consulting, Nashville, TN), and Peter Wilding (University of Pennsylvania Medical Center, Philadelphia, PA).

References