How Good Are Clinical Laboratories? An Assessment of Current Performance

Robert Re1,2 and Richard W. Jenny3

The Clinical Laboratory Improvement Act of 1967 and Amendments of 1988 (CLIA '67 and CLIA '88) were enacted to ensure that clinical laboratories within the U.S. provide a quality of service that meets clinical needs for good patient care. Approved proficiency-testing programs are to judge the quality of laboratory testing by promulgated performance criteria. We examine the quality of analytical results reported in 1991 to the New York State Department of Health Proficiency Testing program in light of these criteria and analytical goals, based on medical usefulness. Analytical performance is examined for cholesterol, potassium, sodium, calcium, glucose, aspartate aminotransferase, digoxin, and theophylline. In general, proposed CLIA '88 performance standards are compatible with the current state of practice for the population of laboratories examined. Exceptions appear to be digoxin and sodium (failure rate exceeding average) and most therapeutically substances (low failure rate). Sources of analytical bias relative to an accuracy-based target value must be characterized as method-, laboratory-, or matrix-dependent if regulatory programs are to achieve the objective of improving analytical accuracy across all testing sites.

Additional Keyphrases: legislation · evaluation studies · accreditation · license · standards · proficiency testing

The question "How good are clinical laboratories?" requires examination of the term good in such a query. The multiple definitions (1) of this commonly used adjective include "adapted to the end designed or proposed; sufficient or satisfactory for its purpose." An even clearer connotation for our application is found in the definition "of comparative excellence in its kind; approaching the standard; commendable." Similarly, quality has also been defined broadly as "conformance to requirements" (2).

In the context of laboratory services, these requirements, standards, and goals must be the needs defined by patient care (3). Concepts of industrial total quality control, continuous improvement, and quality management are now stressed in all aspects of manufacturing, including adoption by the health-care industry (4, 5) and further application by the clinical laboratory (3, 6–9). The need to continually monitor and improve all aspects of the process, from test selection, patient preparation, and specimen collection and handling to the accurate analysis of specimens and timely reporting of test findings, is well recognized. Although quality control practices have long been conventional within clinical laboratories, there is a renewed interest in furthering the process (3).

Quality exists at two levels: quality in fact and quality in perception (10). Quality in fact is achieved when the laboratory performs up to its own specifications. Quality in perception is achieved when the laboratory performance specifications are consistent with clinical requirements for good patient care; that is, the service is believed to be as good, or better than, the customer expects. Media crusades exposing testing errors (11, 12) changed public perception of laboratory quality and at least a portion of regulatory programs have had their stimuli from such campaigns (7). The Clinical Laboratory Improvement Amendments of 1988 (CLIA '88) were enacted to update and consolidate the patchwork of existing regulations and to establish uniform performance and proficiency-testing requirements for laboratory assessment (13). These performance measures have been promulgated and are due to be implemented in the 1990s; similar standards (14), directed at laboratories currently regulated under the original Clinical Laboratory Improvement Act of 1967 (CLIA '67), have also been published in final form.

Proficiency testing, an important component of both CLIA '67 and CLIA '88, has been demonstrated as an effective mechanism for characterizing analytical performance (15). However, proficiency testing is usually insensitive to non-analytical processes and has limitations in completely characterizing the suitability of analytical performance (16). Appropriate design and implementation of a proficiency-testing program is essential if the perception of quality achieved by the regulatory program is to be, in fact, a reliable estimation of the actual quality of testing.

External quality-control programs that use accuracy-based target values and fixed evaluation limits established in terms of medical requirements provide a technological basis for process control of interlaboratory accuracy (17). They also judge quality of testing in such terms as method staff who must use the test findings for patient care. Performance criteria based solely on peer group statistics are insensitive to method error and merely describe relative performance.

The large number of critical comments concerning proposed regulations (ref. 13, p 7003) attests to the fact that many view CLIA '88 as a recrudescence measure that will fail to improve genuine quality. Several studies have evaluated the impact of proposed criteria on the probability of laboratory success in passing the challenge (18–20). Desirable standards of analytical perfor-

1 Wadsworth Center for Laboratories and Research, New York State Department of Health. 2 School of Public Health, State University of New York at Albany, Albany, NY 12201-0509. 3 Nonstandard abbreviations: CLIA, Clinical Laboratory Improvement Act/Amendments; NHLBI, National Heart, Lung, and Blood Institute; and NCEP, National Cholesterol Education Program.
mance based on medical need (analytical goals) have also been proposed (21–26), though few studies have evaluated the appropriateness of proposed performance interlaboratory requirements in terms of medical requirements (24, 27). In this paper, we characterize the state-of-practice of interlaboratory assay performance as judged by the New York State Department of Health Proficiency Testing program, and examine the quality of testing in light of proposed CLIA '88 and CLIA '67 criteria and proposed analytical goals based upon medical requirements for good patient care.

Materials and Methods

The laboratory population examined comprised those holding clinical laboratory permits issued by the New York State Department of Health and included ~700 laboratories, about 80% of which were located within New York State. Hospital-based laboratories represented about 40% of this population; others were commercial, government, or physicians’ office laboratories. Each laboratory did not perform all analyses for each of the analytes studied; individual numbers of participants are specified in Results and Discussion. Unless otherwise stated, all results were obtained in the routine proficiency testing administered by our laboratory over the period from August to September 1991.

Specimens were prepared from human plasma that was defibrinated and dialyzed (Aalto Scientific, Ltd., Vista, CA). Each specimen was prepared in lots of about 9 L, and individual analytes were supplemented by our laboratory to desired concentrations and allowed to mix at 4 °C for ≥12 h. Serum was filtered through a series of microporous filters (ranging to a diameter of 0.22 μm). Specimens were aseptically dispensed into 10- or 15-mL aliquots into sterile serum vials and stored at <−70 °C until shipped to participants. Results were returned via approved postal carriers within 11 days of shipment. Robust estimates of location (mean) and dispersion (standard deviation) were calculated by the method of Paulson and Nicklin (28).

Analyte target values were established by either recognized reference methods or by the robust estimate of the mean value of participant results as specified in the text. Interlaboratory performance was characterized as the cumulative frequency of the absolute bias, or percentage bias, of the participant results from the target value. Analytical goals stated are those proposed by earlier studies (as cited in Results and Discussion); we consider the goals most appropriate for our application to be those derived by using statistical postulates (29, 30) adopted by the 1976 Conference on Analytical Goals in Clinical Chemistry (21). State-of-practice was determined to be the error level (bias) demonstrated at the 95th percentile of participant results. That is, 5% of the laboratories in the examined population exceeded the error level found. Laboratory testing quality was judged by comparing the state-of-practice to proposed CLIA '88 performance requirements and to proposed analytical goals.

Results and Discussion

Accuracy and precision of interlaboratory measurements of cholesterol have been a focus in both scientific journals (31–34) and the lay press (11, 12). Figure 1 presents a cumulative frequency distribution of cholesterol results from 370 laboratories as a function of error. Error was defined as the absolute value of the percentage difference between the result reported and the cholesterol concentration [4.73 mmol/L (185 mg/dL)] in the specimen as determined by the National Reference Method for cholesterol (35). The robust mean of participant data was 4.78 mmol/L (185 mg/dL), in substantial agreement with the reference method value. Several “performance standards” are also provided in Figure 1 (points a–h). The current CLIA '88 and CLIA '67 criteria (13, 14) for acceptable performance (±10%) are indicated by (e). The initial National Heart, Lung, and Blood Institute (NHBLI)/National Cholesterol Education Program (NCEP) criterion (1987) for total allowable error of ±15% (33, 34) is shown by (g); the NHBLI/NCEP goal for 1992 of ±9% is shown by (c); and the NHBLI/NCEP recommended limits (33) for external surveys (±9.5%) by (d). A popular press article (12) reproached clinical laboratories that exceeded 5% (b). Based upon medical needs, a consensus group in 1976 (21) promulgated an analytical goal of ±20% (a CV of 10%) for population-based comparisons or group screening (Figure 1, point h). The same panel (21) recommended that an allowable imprecision of ±4.8% (a CV of 2.4%) be required for the sequential monitoring of individuals (Figure 1, point a).

External estimates of national interlaboratory performance for cholesterol measurement have demonstrated significant improvement during the past two decades (34). At the 95% confidence limit (2 × CV), interlaboratory imprecision was 37% in 1969, 22% in 1980, and 12% in 1986. Comparable estimates were obtained (36) for
laboratories within New York State in the 1970s. In 1984, interlaboratory imprecision was 14% for this population, also in accord with national estimates.

The 95% confidence limit (2 × CV) for laboratory performance measured in our survey in 1991 was 6.3%, demonstrating further improvement in performance and surpassing all proposed professional performance standards. The introduction of enzymatic procedures in 1975 has likely contributed to the significant improvement in interlaboratory performance over the period from 1975 to 1985. The further reduction in interlaboratory imprecision from 1984 to 1991 may well be the consequence of national campaigns to improve the accuracy base for cholesterol measurements in the U.S. (33, 34).

Figure 2 relates the analytical goal for theophylline to the state-of-practice and to the proposed CLIA '88 performance standard. The proposed performance standard of ±25% around the target value (point b) greatly exceeds the analytical goal indicated by (a); that is, a CV of 4% or a total allowable error of 8% (26). The CLIA standard is nearly twofold greater than the state-of-practice (±15%) defined as the error limit not exceeded by more than 5% of the laboratories.

Assay precision depends upon analyte concentration as noted by the cumulative frequency of the percentage difference of test results from the target values of 5 and 20 mg of theophylline/L (28 and 110 µmol/L). Prescribed dosage regimens and patient non-compliance are manifest by steady-state drug concentrations in serum that may range from subtherapeutic to toxic. Figure 2 demonstrates improved performance at higher analyte concentration, an observation that we have also found for many other analytes. Performance standards of proficiency-testing programs that do not take into account the concentration-dependent assay performance may unfairly penalize laboratories when they are challenged to measure drug concentrations in the low subtherapeutic range. Although our data (Figure 2) show that performance evaluations at an allowable error of ±25% are not sensitive to the concentration-dependent assay precision, the CLIA standard (±25%) appears quite stringent when contrasted to current laboratory capabilities at either concentration (±15%) and to medical usefulness criteria (±8%).

Cholesterol and theophylline performance criteria defined by CLIA are based on an allowable proportional error (target ± a percentage value). Table 1 provides a compilation of these criteria for these two and eight other analytes applying that scheme. The proportion of unsatisfactory results ranged from 0.2% (ethanol) to 2.1% (magnesium) of reported results when these criteria for the five specimens distributed were used. An analyte event failure is defined by CLIA regulations as two or more unsatisfactory results for that analyte within a single testing event. The numbers of laboratories encountering such an analyte event failure ranged from none, for theophylline and phenytoin, to 10 (4.1% of participants) for iron.

Another CLIA evaluation scheme requires that analytical performance meet a defined criterion independent of analyte concentration. Four analytes utilizing these fixed criteria are shown in Table 2. The incidence of unsatisfactory results and analyte event failures is approximately that found for analytes listed in Table 1. Of particular interest is the discrepancy in classification of sodium and potassium. These two analytes were measured by 386 laboratories, each utilizing identical instrumentation for both analytes within each laboratory (96% ion-specific electrodes, 4% flame emission photometers). Presumably, measurements were performed at the same time, with similar calibration and quality-control procedures, yet the failure rate for sodium exceeded that for potassium by threefold and analyte failures were four times greater for sodium. Figures 3 and 4 present cumulative frequency distributions for these two analytes. From these data, a revised performance standard for sodium of ±5 mmol/L appears more consistent with the current standard for potassium.

Larger numbers of laboratories were unable to meet the fixed criteria at higher concentrations for both analytes (Figures 3 and 4), owing to the use of fixed rather than proportional criteria. Laboratory performance, estimated by the maximum percentage error demonstrated by 95% of the laboratories, was either reduced or unchanged at higher analyte concentration. For potassium, the errors were 9.3% at 3.1 mmol/L and 6.3% at 7.9 mmol/L; for sodium measurements, percentage error was 3.4% at both 120 and 145 mmol/L.

Figure 5 presents cumulative frequency distributions for calcium. Suggested analytical goals are far more demanding than either current performance or the CLIA '88 performance standard. Only 35% of the participant results were within stated analytical goals; interlaboratory agreement has not improved significantly for this analyte.
Table 1. Laboratory Performance for Representative Analytes According to Proportional Criteria Proposed by CLIA '88

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Performance level</th>
<th>Results</th>
<th>Failure %</th>
<th>“Analyte event failure”: no. of labs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total protein</td>
<td>Target ≥ 10%</td>
<td>1805</td>
<td>10</td>
<td>(0.5)</td>
</tr>
<tr>
<td>Urate</td>
<td>Target ≥ 17%</td>
<td>1780</td>
<td>25</td>
<td>(1.4)</td>
</tr>
<tr>
<td>Albumin</td>
<td>Target ≥ 10%</td>
<td>1745</td>
<td>25</td>
<td>(1.4)</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>Target ≥ 10%</td>
<td>1825</td>
<td>25</td>
<td>(1.4)</td>
</tr>
<tr>
<td>Iron</td>
<td>Target ≥ 20%</td>
<td>1167</td>
<td>33</td>
<td>(1.8)</td>
</tr>
<tr>
<td>Magnesium</td>
<td>Target ≥ 25%</td>
<td>1336</td>
<td>29</td>
<td>(2.1)</td>
</tr>
<tr>
<td>Phenytoin</td>
<td>Target ≥ 25%</td>
<td>1577</td>
<td>3</td>
<td>(0.2)</td>
</tr>
<tr>
<td>Theophylline</td>
<td>Target ≥ 25%</td>
<td>1619</td>
<td>6</td>
<td>(0.4)</td>
</tr>
<tr>
<td>Ethanol</td>
<td>Target ≥ 25%</td>
<td>943</td>
<td>2</td>
<td>(0.2)</td>
</tr>
<tr>
<td>Phenobarbital</td>
<td>Target ≥ 20%</td>
<td>1464</td>
<td>26</td>
<td>(1.7)</td>
</tr>
</tbody>
</table>

Table 2. Laboratory Performance for Representative Analytes According to Fixed Criteria Proposed by CLIA '88

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Performance level</th>
<th>Results</th>
<th>Failure, %</th>
<th>“Analyte event failure”: no. of labs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potassium</td>
<td>Target ± 0.5 mmol/L</td>
<td>1915</td>
<td>15</td>
<td>(0.8)</td>
</tr>
<tr>
<td>Calcium</td>
<td>Target ± 0.250 mmol/L (1.0 mg/dL)</td>
<td>1897</td>
<td>23</td>
<td>(1.3)</td>
</tr>
<tr>
<td>Sodium</td>
<td>Target ± 4.0 mmol/L</td>
<td>1879</td>
<td>46</td>
<td>(2.4)</td>
</tr>
<tr>
<td>pH</td>
<td>Target ± 0.04</td>
<td>987</td>
<td>13</td>
<td>(2.4)</td>
</tr>
</tbody>
</table>

A third CLIA evaluation scheme requires that analytical performance meet a defined criterion independent of analyte concentration up to a threshold concentration. Above this concentration a proportional criterion would be applied. Five analytes utilizing these performance criteria are shown in Table 3. Failure rates and analyte failures are approximately those found for those analytes, with the exception of digoxin, listed in either Table 1 or 2.

The imprecision of digoxin assays is markedly concentration dependent, and the CLIA standard applies, appropriately, a fixed absolute performance evaluation criterion to proficiency challenges at concentrations of digoxin below 1.0 μg/L (target value ± 0.25 mmol/L (0.2 μg/L)) and a proportional criterion (+20%) to challenges at and exceeding 1.3 nmol/L (>1.0 μg/L). When laboratory performance for digoxin is judged by an accuracy-based target value that is <1.3 nmol/L (<1.0 μg/L), we find that about 10% of the laboratories fail to meet the criterion for acceptable performance (Figure 6, point c).

Fig. 3. Cumulative frequency of the absolute difference of potassium proficiency testing results from the target value

The cumulative frequency of the absolute bias of proficiency testing data from the target concentration was determined from test results reported for specimens containing potassium at 3.1 mmol/L (——) and 7.9 mmol/L (-----); 386 laboratories reported test results. A proposed analytical goal (21) for the total allowable error of potassium assays is represented by (a); (b) is the 95% performance level in 1975; and (c) is the proposed CLIA '88 performance standard (19) for the determination of potassium concentration in serum (± 0.5 mmol/L)

Fig. 4. Cumulative frequency of the absolute difference of sodium proficiency testing results from the target value

The cumulative frequency of the absolute bias of proficiency testing data from the target concentration was determined from test results reported for specimens containing sodium at 121 mmol/L (-----) and 145 mmol/L (——); 386 laboratories reported test results. A proposed analytical goal (21) for the total allowable error of sodium assays is represented by (a); (b) is the 95% performance level in 1975; and (c) is the proposed CLIA '88 performance standard (19) for the determination of sodium concentration in serum (± 4 mmol/L)
The standard deviation of test results among the analytical methods that are most frequently used ranged from 0.12 to 0.21 mmol/L (0.09 to 0.16 μg/L). The imprecision of interlaboratory test results suggests that proficiency testing programs that evaluate performance by peer-group statistics will also find a relatively high failure rate. Furthermore, laboratories will be at relatively high risk for unsuccessful performance if a test event includes two or more samples with digoxin concentrations ≥1.3 mmol/L (≥1.0 μg/L).

Stewart and Fraser (25) have proposed a CV of 5.2% (total allowable error 10.4%) as an analytical goal for the determination of digoxin concentration in serum around the upper limit of the therapeutic range (Figure 6, point a). The CLIA '88 performance standard of ±20% is about twice the proposed analytical goal. However, the performance standard is consistent with the state-of-practice.

This scheme of evaluation is also used for glucose (Table 3); the cumulative frequency distribution of results at the threshold concentration [3.44 mmol/L (62 mg/dL)] and an increased concentration of glucose [7.33 mmol/L (132 mg/dL)] are shown in Figure 7. In both cases, current performance exceeds statutory requirements but remains short of analytical goals.

Enzyme activity is a property of a specimen, not an amount; thus external evaluation of measurements is considerably more complicated than for other analytes. However, reasonable agreement of current analytical systems with the International Federation of Clinical Chemistry Reference Methods, where available, appears to have markedly improved performance. In 1975, we estimated within-method imprecision for aspartate aminotransferase at ≥20%; intermethod imprecision far exceeded 50% (37). In the current survey, we found that ~78% of all results were within ±20% of a single target value (Figure 8). The cumulative frequency diagram demonstrates a significant bimodal distribution with

---

**Table 3. Laboratory Performance for Representative Analytes According to Proportional and Fixed (Mixed) Criteria Proposed by CLIA '88**

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Performance level</th>
<th>Results</th>
<th>&quot;Analyte event failure&quot;: no. of labs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO₂</td>
<td>Target ± 8% or</td>
<td>Pass</td>
<td>991</td>
</tr>
<tr>
<td>Glucose</td>
<td>Target ± 5 mmHg</td>
<td>Fall</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Target ± 10% or</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Target ± 0.33 mmol/L (6 mg/dL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urea nitrogen</td>
<td>Target ± 9% or</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Target ± 0.33 mmol/L (2 mg/dL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creatinine</td>
<td>Target ± 15% or</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Target ± 26.5 μmol/L (0.3 mg/dL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digoxin</td>
<td>Target ± 20% or</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Target ± 0.26 mmol/L (0.2 μg/L)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1214 CLINICAL CHEMISTRY, Vol. 38, No. 7, 1992
approximately 20% of the laboratories reporting results ~25% greater than the other group.

We attribute this difference to two major populations of laboratories: those utilizing methods that include exogenous pyridoxal phosphate (higher results) and those that fail to add coenzyme (lower results).

The fourth performance standard is based on population statistics (target ± 3 SD). As expected, it evaluates the majority of results as acceptable (Table 4).

**Envol**

In our evaluations of laboratory performance, we judged the quality of testing by comparing participants' results with an accuracy-based target value determined either from the consensus robust mean of participant results or from Reference Method values. Ross (17) presented an elegant analysis of the merits of using accuracy-based target values and clinically relevant evaluation limits as a mechanism to improve and control interlaboratory accuracy. The current consensus strategy for the determination of desirable performance standards based on clinical requirements for patient care estimates tolerable analytical error from biological variation. For single-point diagnostic testing, the analytical imprecision should be less than or equal to one-half the within-individual variation. This clinical setting poses the greatest demand for reliability of testing and is used appropriately to establish performance standards that will satisfy, or exceed, requirements for all clinical situations. The fraction (one-half) has been derived by empirical means, but analytical methods that meet this goal contribute a maximum of about 12% to the overall variability of patients' results.

The intralaboratory performance of many state-of-the-art analytical systems usually fulfills quality goals for imprecision. However, as noted by our findings and by other studies, interlaboratory performance usually fails to meet desirable performance standards. Method bias and laboratory-specific bias are major components of overall variability of proficiency testing and interlaboratory results. Many analytical methods currently in use are based on sound analytical principles and are reasonably well characterized regarding the origin and the extent of nonspecific assay response. A predominant source of analytical bias is likely the lack of traceability of analyte values assigned to calibrator and reference quality-control materials to definitive or reference

---

**Table 4. Laboratory Performance for Representative Analytes According to Population-Based Criteria Proposed by CLIA '88**

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Performance level</th>
<th>Results</th>
<th>“Analyte event failure”: no. of labs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>α₂ Thromboplastin</td>
<td>Target ± 3 SD</td>
<td>984</td>
<td>16 (1.6)</td>
</tr>
<tr>
<td>Thyrotropin</td>
<td>Target ± 3 SD</td>
<td>1490</td>
<td>25 (1.6)</td>
</tr>
<tr>
<td>Chorionic gonadotropin</td>
<td>Target ± 3 SD</td>
<td>1213</td>
<td>22 (1.8)</td>
</tr>
<tr>
<td>Thyroxine</td>
<td>Target ± 3 SD</td>
<td>1507</td>
<td>33 (2.1)</td>
</tr>
<tr>
<td>Triiodothyronine</td>
<td>Target ± 3 SD</td>
<td>908</td>
<td>32 (3.4)</td>
</tr>
<tr>
<td>Free thyroxine</td>
<td>Target ± 3 SD</td>
<td>772</td>
<td>43 (5.2)</td>
</tr>
</tbody>
</table>

---

**CLINICAL CHEMISTRY, Vol. 38, No. 7, 1992 1215**
methods. Great success has been realized by the NCEP and the National Reference Method Laboratory Network by providing the mechanism to ensure traceability of analytical methods for cholesterol to the Reference Method (33--35). Cholesterol was the only analyte in our study for which interlaboratory performance exceeded most recommended analytical goals, and dramatic improvement in interlaboratory performance was observed over the past decades.

In the introduction to this article we referred to the multiple definitions of the word "good" in the context of laboratory performance. One of the definitions that we believe best fits this adjective in this framework is "true" (I) and accuracy-based target values in interlaboratory surveys will undoubtedly aid this process. The recent editorial of Bowers (38) succinctly outlines the significance of a system of clinical laboratory measurements and their evaluation based upon true values.

Figures 1-8 and Tables 1-4 present interlaboratory bias from a single target value in proficiency-testing surveys. Our data suggest that unique, accuracy-based target values can be successfully used in external surveys. It is, however, well known that some analytical systems may provide aberrant results with materials that are prepared in a matrix that is less complete than, or with properties that differ from, a patient's serum specimen. Procedures such as lyophilization result in physical changes in serum specimens and addition of stabilizers or other additives further alters the properties of the specimen (39-45). Even with the use of processed serum distributed in liquid form in our program, some peer-group evaluation is required.

Such anomalies are often considered to be "matrix effects" (46). Unfortunately this categorization is often accompanied by an implicit acceptance of these aberrant results, whereas true differences among analytical techniques are obscured. The exact nature of such matrix effects is often unknown and is difficult to determine. Furthermore, laboratorians are seldom certain that an analytical technique that is sensitive to matrix effects in quality-control specimens is not similarly affected by such interferences with authentic clinical specimens (45).

All matrix effects must be broadly considered as the consequence of the lack of specificity of the analytical technique in question. The IFCC definition (47) of specificity is the "ability of an analytical method to determine solely the component(s) it purports to measure." Clearly, other components of the matrix of the specimen that affect the result imply an inherent lack of specificity of the method. The concern is the point at which a proficiency-testing sample differs so fundamentally from the normally encountered specimen that the specificity of even the most robust clinical method is unfairly challenged. Added preservatives, effects of processes such as lyophilization, and peptide hormones or enzymes of animal derivation characterize specimens that are not encountered in routine practice. We would not necessarily require analytical procedures to provide analytically correct results with such specimens, however desirable that might be (45).

Two decades ago, a report from our laboratory addressed this problem (40), and we coined the term "commutability" to identify the properties of internal and external quality-control preparations that could potentially distinguish these surrogate samples from authentic patients' specimens. Although we originally applied this concept to enzyme activity measurements, the term has been accepted by others (46, 48) and can be broadly applied to all clinical chemical measurements. Following the original definition (40), commutability refers to the ability of quality-control materials to show interassay properties comparable with those demonstrated by authentic clinical specimens. We reviewed some of these concepts previously (39) and have suggested multivariate statistical techniques that could elucidate the intermethod behavior of clinical specimens and reference preparations (39, 41).

Only in cases where the external quality-control fluid can be demonstrated to possess characteristics aberrant from patients' specimens do we believe that method-specific target values can be set for external surveys. All data presented in Figures 1-8 were determined with a single target value based upon either a reference method or the robust mean of the participant data.

The use of target values established by reference or definitive methods would be an ideal quality goal for all proficiency programs. It has been established that the consensus mean of proficiency testing results agrees sufficiently well with definitive methodology to provide reliable reference points for the evaluation of laboratory performance (49). Our data are generally consistent with this assessment; the robust mean for reported cholesterol results by participants differed from the reference method by ~ + 1% (Figure 1). Such use of accuracy-based target values and the use of evaluation limits based on clinical requirements, serve to assess both the clinical and relative performance of methods and could ideally be quality goals for the U.S. in the 1990s. The recent opinion by Tietz et al. (50) presents arguments that further support our view.

Our study further suggests that the proposed CLIA '88 criteria for acceptable performance with accuracy-based target values are consistent with the state-of-practice in laboratories participating in the New York State Proficiency Testing Program. Exceptions include theophylline, digoxin at subtherapeutic concentrations, and sodium. Revision of performance criteria should be recurrent and, in our opinion, should match analytical goals for patient needs. Unfortunately, the punitive sanctions of CLIA '88 may prevent the setting of desirable quality goals rather than the current limits that are more likely to maintain the status quo. Nonetheless, the current CLIA '88 rules could serve as a timely first step in achieving genuine interlaboratory comparability of analytical results. Continued efforts are essential to improve analytical accuracy if national quality goals in laboratory testing are to be achieved.

References
11. Bogdanich W. Risk factor: inaccuracy in testing cholesterol hamper war on heart disease—some diagnoses are skewed by glitches such as use of ill-calibrated lab gear, missing the mark by 100%. USA Today: St. J, February 3, 1987: 1.
26. Jenny RW. Analytical goals for determinations of theophyl-
29. Harris EK. Statistical principles underlying analytic goal-setting in clinical chemistry. In ref. 21, 115-35.