Opinion

Proficiency Testing in a Medical-Needs Context
John A. Lott, Neal R. Manning, and Mary Kay Kyler

Proficiency testing by State and Federal agencies is an ongoing activity of clinical laboratories and an occasional source of anxiety; strict statistical evaluations of "snapshot" laboratory values are inappropriate and medical-needs criteria should be used. The quality of laboratory results largely depends on available technology. Fortunately, for most of the common clinical chemistry analytes there has been a steady reduction of imprecision during the past 20 years. Proficiency testing may have been the stimulus for this improved performance. Medical-needs criteria differ, depending on the testing goal. For proficiency testing, population screening criteria are appropriate, e.g., the College of American Pathologists fixed criteria for the common tests. Stricter criteria are needed for short-term (inpatient) and long-term (outpatient) monitoring of laboratory data. Explicit proficiency-testing limits are given here for nine of the common clinical chemistry tests for each of the three medical-needs criteria described above. The limits consider total error—i.e., bias from the believed correct value, and imprecision. Rather broad limits are acceptable for the commonly performed enzyme tests when used for screening purposes.

Proficiency testing (PT) is a major activity of clinical laboratories. Internal quality control programs and participation in one or more external surveys such as those of the American Association for Clinical Chemistry, the Association of Bioanalysts, or the College of American Pathologists (CAP) is part of the routine in most larger laboratories. Participation in external PT can also be a source of consternation and anxiety. State regulators (2, 3) use PT performance to grade laboratories and attempt to segregate the "good" from the "bad"; the latter can lose Medicare/Medicaid reimbursement and be threatened with closure. Among the requirements of the Clinical Laboratories Improvement Act of 1967 are the development of standards for clinical laboratories. Some view this as an urgent and largely unmet need (4), and public criticism of allegedly unreliable laboratories in the press is common (5).

The need of government agencies to grade, regulate, and "weed out" (2) laboratories has created an adversarial gestalt between laboratories and regulators, and considerable chicanery in the assay of PT specimens. Cembrowski and Vanderlinde (6) found special practices occurring with PT challenges, such as multiple assays, analysis with more than one instrument, and averaging of values. Patients' specimens are rarely given such treatment; thus the PT values do not reflect the typical performance of the laboratory. Many experienced laboratorians contend that external PT cannot be the sole indicator of performance; however, PT can indeed serve to trigger further investigation into a potential problem (7, 8).

The consensus of laboratorians at a CAP conference on PT (9) was that a single ("snapshot") laboratory-produced value is inadequate for assessing general laboratory performance. The same situation applies in patients: a single abnormal value is rarely used in making a diagnosis but generally, if untoward, is the impetus for further investigation.

Evaluations of PT results based on peer-group statistics have been the mainstay of CAP surveys since 1967 (10). Values outside of the mean ± 2 SD—i.e., outside of 1,2, for the peer group—are deemed to be unsatisfactory. An attractive feature is the utter simplicity for the evaluators of the survey; however, the arbitrary exclusion of about 5% of laboratories is wrong most of the time (11). Also, such peer-group evaluations are deficient, because they tend to ignore bias. We have come a long way since the findings of Belk and Sunderland in 1947 (12). Many currently available clinical analyzers such as the BMD Hitachi (Indianapolis, IN 46250), Du Pont aca (Wilmington, DE 19803), Kodak Ektachem (Rochester, NY 14650), and the Technicon RA 1000 and SMAC (Ardsley, NY 10591) consistently have CVs of <5% for most of the common analytes. Using peer-group statistics and snapshot, single values makes no scientific sense in labeling most of the laboratories that use these and similar instruments. Equally undesirable are the peer groups labeled as acceptable in the face of wide variability and (or) bias owing to the inappropiate use of the 1,2 rule for the peer group. Interesting alternatives to the 1,2 rule have been proposed by Ehrmeyer et al. (13). Their methods appear to be more powerful. However, they are still based on PT participants' statistics. Medical-needs criteria are not used in their scheme.

Intra-individual variability and use to which laboratory data are put should play a role in determining the requisite analytical quality of the laboratory data (10). The difficulties with medical-needs criteria are subjectivity, ex cathedra statements, and assumptions based on incomplete or unavailable data. Medical-needs criteria for laboratory performance are necessarily arbitrary and based on "soft" data, creating a difficult scenario for laboratory regulators and their apparent greater comfort in relying on "hard" data such as participant-derived statistics (2). Nevertheless, the general consensus of authors who have wrestled with this question is that medical-needs criteria should determine acceptable laboratory performance (10). The tool must fit the task, and purely statistical evaluations cannot be the sole arbiters of laboratory performance.

An Historical View of Testing Imprecision

We present here a somewhat different perspective on medically acceptable laboratory error. It is a complex and
multifaceted problem, and we have tried to consider the problems of the laboratory, the patient, the regulatory agencies, and society. What emerges at once is that there is no simple solution. However, some common-sense suggestions are possible.

A Look Back at One Laboratory

It is illustrative to review our internal PT data and show how technological changes have affected clinical laboratory testing. To a major extent, laboratory testing can only be as good as the currently available technology. Certainly, the performance of the laboratory personnel at all levels and certain other factors play a role in the quality of laboratory results (14); but inadequate methods and instruments can still have serious negative consequences on the quality of laboratory data.

Figure 1(a–e) depicts some of the internal quality-control experience in our laboratory over the last 21 years. Each figure shows a unique phenomenon, and all the common clinical chemistry tests performed by us have a pattern resembling one of these five. The concentrations of the analytes in the control sera were reasonably constant over the 21-year period and were at or near decision points. The reason we show these data is to illustrate time trends of imprecision. The data are not a product of special handling or repeat analyses. It was the best the laboratory could do with the available materials, equipment, and people.

**Bilirubin.** In Figure 1a, bilirubin at concentrations of 8 to 12 mg/L shows a pattern of improving, then worsening, precision. Between 1967 and 1979, assays were carried out with manual diazotized sulfanilic acid procedures; until 1972, it was that of Malloy–Evelyn (15) and then a version of the Jendrassik–Gröf method was used (16). The monthly mean CV was about 8% until 1979, when we began using the automated method on the DuPont acx. The latter gave consistently more precise results, and the CVs declined to a mean of about 4%, a marked improvement. In early 1984, we began assaying total bilirubin in the Astra instrument (Beckman, Brea, CA 92621), and the CV increased to about 15%. Part of this increase was ascribable to a lower concentration of bilirubin in the control, i.e., 8 mg/L vs 10 mg/L as used earlier. Although we were concerned about this deterioration of precision, we made a subjective decision to accept the values based on the pressure to reduce costs and increase automation. At the time, we judged that an uncertainty of 3 mg/L for bilirubin at a total concentration of about 10 mg/L was medically acceptable. We really had no basis for this conclusion.

**Calcium.** Calcium, at 1.8 to 2.2 mmol/L (Figure 1b), shows the second type of pattern of generally improving CVs that apparently plateau at about 3%. The manual Calcein titrimetric method was used until 1970 and was clearly unsatisfactory, but it was the least bad of all the then-available methods. An automated atomic absorption procedure (17) used between 1970 and 1982 was better, although the method showed large swings in precision from month to month. We used the acx between 1983 and 1985, which provided clearly lower CVs, and in 1985, we started using an Astra for calcium assays, with some minor loss of precision.

**Cholesterol.** Cholesterol at 1800 to 2200 mg/L (Figure 1c) shows the third and most desirable type of pattern: a steady decrease in imprecision. From 1967 to 1978, we used a manual FeCl₃ method (18), and the variability of the method was large indeed! There is also a suggestion of long-

---

**Fig. 1.** a–e. Twenty-year quality-control experience at The Ohio State University Hospitals with internal control sera for indicated analytes. Approximate concentrations are bilirubin, 10 mg/L; Ca, 2.0 mmol/L; cholesterol, 2000 mg/L; creatinine, 12 mg/L; and glucose, 700 mg/L. Each point represents the CV of about 30 assays of the same control for a given month. The dashed lines represent ± 2 SD for the data shown (n = 247). The first five months of 1967 had no data for any of these tests.

---

term cyclical fluctuations of the CVs with time. Some months had mean CVs exceeding 13%; the limits of Figure 1c do not show this. In 1977, we started assaying cholesterol with the ABA-100 instrument (Abbott Laboratories, North Chicago, IL 60064) and their version of the cholesterol oxidase method (19). The ABA-100 was a consistent and precise performer, and the decrease in imprecision was significant. We now have a mean monthly CV of about 3%, which meets a recently recommended criterion for cholesterol testing (20).

Creatinine. CV data for creatinine at concentrations of 10 to 12 mg/L (Figure 1d) probably reflect the best that most laboratories can do. Between 1967 and 1977, the test was performed in a Technicon AutoAnalyzer with their alkaline picrate method. The precision was fairly poor. Between 1978 and 1980, we used the aca; the imprecision was now only 2%. Since 1981, we have used the Beckman Astra and their rate picrate method, which has an imprecision significantly larger than that of the aca. A CV of about 5% to 7% at a creatinine of 10 mg/L is probably the best we can do at present.

Glucose. Finally, data for glucose at about 700 mg/L are given in Figure 1e. This pattern of slow and continuous improvement of the precision is quite common for other frequently done tests not described here. We used the AutoAnalyzer between 1967 and 1972 and their alkaline ferricyanide procedure, and the Technicon SMA-6 from 1972 to 1980 and our version of a glucose oxidase method (21). The imprecision decreased with the introduction of the latter instrument. Since 1981, we have used the Astra and their hexokinase procedure. With the exception of some recent decreases in imprecision, the mean CV has changed little since 1975.

Does our internal PT program have an effect on our imprecision for the above analytes? We think it has, but we have no direct evidence for this opinion. Over the years, our PT program has identified problem areas and stimulated remedial action. In the absence of within-laboratory vigilance, it is likely that our imprecision would be substantially larger.

A Look Back at Many Laboratories

Figure 2(a–e) shows CAP Comprehensive Chemistry Survey Data for 1967 to 1987 at five-year intervals. The five most popular methods for each year are given—and note that these are not the same for each year (see figure legends). Total bilirubin (Figure 2a) shows a steady and welcome decrease in imprecision between 1977 and 1987. Data for 1972 were not available. The CVs for calcium (Figure 2b) changed little between 1977 and 1982; the values in 1987 reflect an apparent improvement in the most popular methods. Cholesterol (Figure 2c) shows a steady...
and salutary decline in imprecision over the 15-year period. Perhaps laboratories can achieve CVs of <3% for cholesterol in the near future. The CVs for creatinine (Figure 2d) are little changed over the 15-year span. The CVs are rather large and will probably continue to be so for years to come. This test warrants further attention. Glucose (Figure 2e) shows a decline in imprecision that plateaued in 1982. A CV of 4% for glucose appears to be an attainable goal.

What Is Medically Acceptable Error?

Several authors have struggled with the problem of medically acceptable error, a topic summarized elsewhere (10, 22). The somewhat capricious but largely correct answer is: "it depends!" We divide this question into two subtopics: what is acceptable error for population screening—the least stringent limits—and what is allowable for detecting short-term trends in hospitalized patients—probably the most stringent need. Other types of medical needs are assumed to fall between these two extremes.

Medically Acceptable Error in Population Screening

What are the requirements of the laboratory to identify individuals with a biochemical disorder? Assume, as did Harris (23), that the well and sick have biochemical data with gaussian distributions and equal variances, and that the two groups overlap. (For example, see 10, Figure 6.) A gaussian distribution is the exceptional case (24). However, as Harris (23) has shown, it is an extremely useful assumption, and can be used to make generalizations of the effect of changes in C or (i.e., the %CV) on the clinical sensitivity and specificity. To be consistent with Harris' model, we use his terminology, i.e., C and C; the latter is the intrapersonal biological variation.

With some exceptions such as cholesterol (25), the population histograms of the well and sick are largely unknown. A decision value at the overlap point produces an equal number of false positives and false negatives when the numbers of well and of sick are the same. Also, the sensitivity, specificity, and diagnostic efficiency (DE) of the test at the overlap point are the same. DE is the percentage of all results that are true results. With this model, the effect on DE of changes in precision can be estimated. The effect of bias from the true value depends on the direction of the bias relative to the random error. If the bias and random error are in the same direction, the effect can be mild to disastrous, and can render the test result meaningless. Bias can usually be dealt with as is discussed below. Also, with knowledge of the prevalence of disease, the predictive values of an abnormal test result (PV+), and a normal test result (PV-) can be calculated, and the alterations of PV+ and PV- that are caused by changing C and Cs can be estimated. PV+ is the probability of disease with an abnormal test, and PV- is the probability of absence of disease with a normal test.

The general case. We followed Harris' method in making calculations of the effect of changing precision, and some general results are shown in Figure 3, a and b. Figure 3a illustrates the concept and is for equal populations of well and sick and the case where the overlap occurs at Z = 1.00, i.e., the overlap is exactly 1 SD from the means of each population. Here, there is substantial overlap, and a 16% false-positive and false-negative rate. Various percentages of intrapersonal biological variation (C) are shown, and the effect of changing precision is plotted.

Several conclusions can be drawn from Figure 3a: The greatest effect on DE of decreasing C occurs when C is small. When C is below about 5%, reducing C below 4% has little effect on DE. With increasing values of C, a plateau of no change in DE is reached despite decreasing values of C. At a C of 20%, a C of 10% is quite acceptable. A good example of extensive overlap of the "well" and "sick" (better, "patients at risk") is the cholesterol data obtained in the Framingham Study of those with and without coronary artery disease (25). Here the overlap is such that decreasing C would substantially improve the DE of serum cholesterol testing.

Figure 3b is like Figure 3a; however, the overlap of the well and sick is at Z = 1.64, and is much smaller, yielding only 5% false-positive and false-negative values at the intersection of the population histograms. A decrease in C affects DE only when C is less than about 5%. Above a C of about 5%, decreasing C does not improve DE, and efforts to reduce C would be wasteful.

The intuitive notion is correct that a large C is tolerable with well-separated populations of well and sick. It may be possible to make economic and diagnostic decisions based on data similar to those shown in Figure 3a and b. If C is 10% and C is 8% and there is a 5% overlap of the populations, little improvement in DE is achieved by trying to decrease C still further.

Effect of prevalence. In Figure 3, a and b, a 50% prevalence of the sick is assumed; the prevalence is always smaller in population biochemical screening programs. We show in Table 1 the effect of changing C and C at two population overlaps: Z = 1.00 and Z = 1.64. In Table 1, it is obvious that at Z = 1.00, C = 8 and C = 1; the PV+ is so small at a 1%
and 0.1% prevalence that the test has marginal or no value in discriminating between the well and sick and is no more valuable than a coin-flip decision. At a $Z = 1.64$—i.e., at a smaller overlap—the PV(+) still declines dramatically with decreasing prevalence at all values of $C_a$ and $C_b$. In the real world, the distributions of the well and sick would not be the same, but the DE would decrease with a increasing $C_a$, although less than shown here.

*Model applied to specific analytes.* The above technique can be applied to specific analytes to assess the effect of changing precision on PV(+).

**Calcium.** Calcium is a test that was identified at the 1976 CAP Aspen Conference (26) as requiring better precision in nearly all laboratories. Data from the six largest peer groups participating in the 1988 CAP “C-A” Comprehensive Chemistry Survey are shown in Table 2. The mean $C_a$ is about 3%. Does the $C_a$ meet medical needs for biochemical screening? The prevalence of true hypercalcemia detected by population screening is <1%. Boonstra and Jackson (27) found 12 (0.2%) confirmed cases of hypercalcemia in screening 12,000 outpatients. In a subsequent report (28), they found the prevalence of hypercalcemia owing to hyperparathyroidism to be about 0.1%. How well does the laboratory have to perform to identify patients with hypercalcemia?

Table 3 shows PV(+) data for serum calcium at various prevalences. With a $C_a$ of 2.4% (29), $Z = 1.00$ and prevalence of 0.2%, the $C_a$ must be 2% or smaller to identify correctly 80% of patients who have true hypercalcemia during a screening exercise. There is a large change in PV(+) in going from a $C_a$ of 4% to $Z = 1.00$ and at all prevalences. Assuming a $Z$ of 1.64 or larger, then a $C_a$ of 4% is probably satisfactory. Implicit here is the increasing PV(+) with a patient's increasing serum calcium. We chose the overlap of the well and sick as the decision point as a matter of convenience and to be in accord with the earlier work of Harris (23). Clearly, the higher the patient's serum calcium, the greater the allowed $C_a$ can be to detect true hypercalcemia; i.e., the medically acceptable error increases with increasing serum calcium.

**Table 1. Changes in PV(+) at Different Levels of Disease Prevalence, Laboratory Random Error, and Intrapersonal Biological Variation**

<table>
<thead>
<tr>
<th>$C_a$</th>
<th>$C_b$</th>
<th>PV(+) 50%</th>
<th>PV(+) 10%</th>
<th>PV(+) 1%</th>
<th>PV(+) 0.1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. $Z = 1.00$ (84% confidence interval)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.855</td>
<td>0.485</td>
<td>0.079</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.855</td>
<td>0.456</td>
<td>0.071</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.864</td>
<td>0.414</td>
<td>0.050</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.922</td>
<td>0.375</td>
<td>0.056</td>
<td>0.107</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.905</td>
<td>0.514</td>
<td>0.088</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.999</td>
<td>0.991</td>
<td>0.910</td>
<td>0.500</td>
</tr>
<tr>
<td>5</td>
<td>0.981</td>
<td>0.852</td>
<td>0.343</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.916</td>
<td>0.548</td>
<td>0.099</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>B. $Z = 1.64$ (95% confidence interval)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.855</td>
<td>0.485</td>
<td>0.079</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.855</td>
<td>0.456</td>
<td>0.071</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.864</td>
<td>0.414</td>
<td>0.050</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.922</td>
<td>0.375</td>
<td>0.056</td>
<td>0.107</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.905</td>
<td>0.514</td>
<td>0.088</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.999</td>
<td>0.991</td>
<td>0.910</td>
<td>0.500</td>
</tr>
<tr>
<td>5</td>
<td>0.981</td>
<td>0.852</td>
<td>0.343</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.916</td>
<td>0.548</td>
<td>0.099</td>
<td>0.011</td>
<td></td>
</tr>
</tbody>
</table>

* $C_a$ is random error and $C_b$ is intrapersonal biological variability. Note that at PV(+) 50%, the predictive value of an abnormal test at 50% disease prevalence, sensitivity = specificity = diagnostic efficiency.

**Table 2. $C_a$ for Calcium for the Six Largest Peer Groups in 1988 CAP Comprehensive Chemistry Survey**

<table>
<thead>
<tr>
<th>Instrument</th>
<th>No. labs in peer group</th>
<th>$C_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DuPont acet</td>
<td>621</td>
<td>2.7</td>
</tr>
<tr>
<td>Kodak Ektachem 400, 700</td>
<td>513</td>
<td>2.3</td>
</tr>
<tr>
<td>Hitachi (all)</td>
<td>364</td>
<td>2.8</td>
</tr>
<tr>
<td>Beckman Astra 4, 8</td>
<td>238</td>
<td>2.1</td>
</tr>
<tr>
<td>Technicon RA 1000</td>
<td>236</td>
<td>4.0</td>
</tr>
<tr>
<td>Baxter Paramax</td>
<td>233</td>
<td>2.3</td>
</tr>
</tbody>
</table>

*Peer group $C_a$'s for CAP Comprehensive Chemistry Survey Specimen C-A: 4961 laboratories participated in the survey.

**Table 3. Effect of Precision Changes on PV(+) for Calcium at Various Prevalences**

<table>
<thead>
<tr>
<th>$C_a$</th>
<th>PV(+) 50%</th>
<th>PV(+) 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.841</td>
<td>0.0507</td>
</tr>
<tr>
<td>8</td>
<td>0.891</td>
<td>0.0763</td>
</tr>
<tr>
<td>6</td>
<td>0.944</td>
<td>0.146</td>
</tr>
<tr>
<td>4</td>
<td>0.986</td>
<td>0.416</td>
</tr>
<tr>
<td>2</td>
<td>0.999</td>
<td>0.953</td>
</tr>
</tbody>
</table>

*Predictive value of positive test at prevalence shown.

**Table 4. Effect of Precision Changes on PV(+) for Cholesterol at Various Prevalences**

<table>
<thead>
<tr>
<th>$C_a$</th>
<th>PV(+) 50%</th>
<th>PV(+) 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.841</td>
<td>0.638</td>
</tr>
<tr>
<td>8</td>
<td>0.875</td>
<td>0.700</td>
</tr>
<tr>
<td>6</td>
<td>0.911</td>
<td>0.771</td>
</tr>
<tr>
<td>4</td>
<td>0.939</td>
<td>0.837</td>
</tr>
<tr>
<td>2</td>
<td>0.958</td>
<td>0.886</td>
</tr>
</tbody>
</table>

*Predictive value of positive test at prevalence shown.

**Cholesterol.** Data for serum cholesterol are shown in Table 4. Given the considerable recent publicity on serum cholesterol and the recommendations that clinical laboratories should strive for a $C_a$ of 3% or less, it is most appropriate to consider what is a medically acceptable error for cholesterol. In an unselected outpatient population of adult men and women ($n = 200$), we observed a 27% prevalence of serum cholesterol values $>240$ mg/dL; others observed a prevalence of 25% (30). The $C_a$ for cholesterol is about 6.6% (29), and the data for PV(+) at several prevalences are given in Table 4. The overlap of serum cholesterol values is substantial for those not at risk vs those at risk (29). An assumption of $Z = 1.00$ at the overlap is reasonable.
therefore the gain in PV(+) at a 25% prevalence in going
from a C₁ of 6% to 4% to 2% is not very large. A C₁
of 4% to 6% looks medically acceptable. Tighter
precision requirements appear to be unwarranted based on the above
considerations.

Creatinine. There is a less than 1% prevalence of undiagnosed pathology leading to a serum creatinine that exceeds
15 mg/L. In an outpatient population seen here, after
patients with chronic heart or kidney diseases are excluded,
we estimate that less than 2% of patients have a creatinine
>15 mg/L. Assuming a prevalence of 1% and a 5% overlap
(Z = 1.64), then a medically acceptable error is in the range of
4% to 6%. Currently, most laboratories in the CAP
Chemistry Survey report C₁'s of between 7% and 10% for
a creatinine value that is within normal limits. Serum creatinine
is a test with unmet medical needs, and the current
level of C₁ is too large in nearly all laboratories. More effort
is needed by the manufacturers and users of the creatinine
test to reduce imprecision.

Glucose. The prevalence of undiagnosed glucose intolerance
of any cause in population screening studies is about
2% to 4% (31). Assuming a prevalence of 5% and a 5% overlap
(Z = 1.64), then a medically acceptable error is about 5% to
10%. Most laboratories have C₁'s for glucose of <6% at 1000
mg/L. Serum glucose is a test where medical needs have
been largely met.

Trend Detection

The detection of within-patient trends in clinical analytes
places the greatest demands on the clinical laboratory. The
problem is estimating true within-patient trends in the
milieu of all the perturbations that can affect the final
results. We follow Harris' (23) approach in estimating
whether a trend is statistically significant, defined by us as
Z = 1.64 (P < 0.05) for a one-sided test.

After renal transplantation, the serum creatinine receives
continuous and critical scrutiny as an index to both acute
and chronic rejection (32). Assuming C₁ is 10% at a creatinine
of 15 mg/L, can a change from 15 to 17 mg/L be
detected? The statistical test we chose is whether Z
is greater than the critical value, 1.64. Z is calculated
from the change in creatinine and the analytical SD of the method as
follows: Z = |Δ| - 17 / (SD / √2). In the first example, Z₁ = 2(1.5) (1.41), or 0.946, which is much below the critical
value of 1.64. The C₁ in the laboratory is too large to detect
this change; the C₁ would have to be 5.7% or lower to
detect the change with a P < 0.05. If now a third value of 19 mg/L is
obtained, can the laboratory detect the trend? Comparing
the third value with the average of the first two, we obtain:
Z₂ = |19 - 16| / 1.5(√3/2), or 1.63, slightly above the critical
value. A fourth value of 20 mg/L gives Z₃ = |20 - 17| / 1.5 (√4/3),
where 17 is the mean of the three previous values.
Here, Z₃ = 1.73, which is definitely significant.
Clinical intuition that a creatinine series of 15, 17, 19, and
20 mg/L represents a real trend is also the common-sense
conclusion.

We developed some general tables, using the above
relationships to simplify the concept. For three values in a
trend, the percent difference between the mean of the first
two and the third value can be calculated readily as shown
above, and is defined here as "delta." Assuming certain
values of C₁, the Z statistic can be calculated, and the results
for three values in a series are shown in Table 5. If we use
1.64 as the critical Z value, the data reveal that to detect a
delta of 4%, a C₁ of <3% is required. At the other extreme, a
delta of 16% can be detected with a C₁ as large as about 8%.

Table 5. Critical Z Values for Percent Trend at Various
C₁ Values *

<table>
<thead>
<tr>
<th>C₁</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.82</td>
<td>1.63</td>
<td>2.45</td>
<td>3.27</td>
<td>4.08</td>
<td>4.90</td>
<td>5.72</td>
<td>6.53</td>
</tr>
<tr>
<td>4</td>
<td>0.41</td>
<td>0.82</td>
<td>1.23</td>
<td>1.63</td>
<td>2.04</td>
<td>2.45</td>
<td>2.86</td>
<td>3.27</td>
</tr>
<tr>
<td>6</td>
<td>0.27</td>
<td>0.54</td>
<td>0.82</td>
<td>1.09</td>
<td>1.36</td>
<td>1.63</td>
<td>1.91</td>
<td>2.18</td>
</tr>
<tr>
<td>8</td>
<td>0.20</td>
<td>0.41</td>
<td>0.61</td>
<td>0.82</td>
<td>1.02</td>
<td>1.23</td>
<td>1.43</td>
<td>1.63</td>
</tr>
<tr>
<td>10</td>
<td>0.16</td>
<td>0.33</td>
<td>0.49</td>
<td>0.65</td>
<td>0.82</td>
<td>0.98</td>
<td>1.14</td>
<td>1.31</td>
</tr>
</tbody>
</table>

*Table for three successive values with indicated percent unidirectional trend. For example, assume that the mean of a patient's first and second value differ from the third value by 6% (delta = 6), then the laboratory must have a C₁ of <4% to detect the trend with 95% confidence, i.e., Z = 1.64.

Table 6. Critical Z Values for Percent Trend at Various
C₁ Values *

<table>
<thead>
<tr>
<th>C₁</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.87</td>
<td>1.73</td>
<td>2.60</td>
<td>3.46</td>
<td>4.33</td>
<td>5.20</td>
<td>6.06</td>
<td>6.93</td>
</tr>
<tr>
<td>4</td>
<td>0.43</td>
<td>0.87</td>
<td>1.30</td>
<td>1.73</td>
<td>2.17</td>
<td>2.60</td>
<td>3.03</td>
<td>3.46</td>
</tr>
<tr>
<td>6</td>
<td>0.29</td>
<td>0.58</td>
<td>0.87</td>
<td>1.16</td>
<td>1.44</td>
<td>1.73</td>
<td>2.02</td>
<td>2.31</td>
</tr>
<tr>
<td>8</td>
<td>0.22</td>
<td>0.43</td>
<td>0.65</td>
<td>0.87</td>
<td>1.08</td>
<td>1.30</td>
<td>1.52</td>
<td>1.73</td>
</tr>
<tr>
<td>10</td>
<td>0.17</td>
<td>0.35</td>
<td>0.52</td>
<td>0.69</td>
<td>0.87</td>
<td>1.04</td>
<td>1.21</td>
<td>1.39</td>
</tr>
</tbody>
</table>

*Table for four successive values with indicated percent unidirectional trend. For example, assume that the mean of a patient's first, second, and third value differ from the fourth value by 5%, then the laboratory must have a C₁ of about 5% to less than detect the trend with 95% confidence, i.e., Z = 1.64.

Intuitively Acceptable Error

Acceptable error should be based on how tests are used
clinically. We describe a case here where expert opinion and
intuition can serve well in describing medically acceptable
error. In adults, results for the commonly performed serum
enzyme tests—e.g., alkaline phosphatase, aspartate aminotransferase,
and creatine kinase—when used in population
screening studies, are usually viewed by clinicians as "normal,"
slightly to moderately increased, or significantly
increased. For an enzyme test with result in the reference
range, a large C₁ is acceptable. If, for example, the upper
reference limit is 100 U/L and the patient's enzyme result is
40 U/L, an acceptable error could be a C₁ of 100%! This may
strike many as extreme error, but any values for the enzyme
test between 0 and 80 U/L would generally elicit the same
response from the clinician. At 100 U/L, we would allow a C₁
of 50%; any values between 50 and 150 U/L are expected to
receive the same response, i.e., normal or slightly increased.
At 200 U/L, we would allow a C₁ of 25%, and even this broad
limit may be too strict given how the enzyme tests are used
in screening studies.

The above intuitive logic is the basis of the fixed criteria
used to evaluate enzyme data in the CAP Comprehensive
Chemistry Survey (33). To require a C₁ of, say, 25% at 50
U/L for acceptable PT is overly rigid and unwarranted,
given how the results are used for screening populations.

Table 7. Recommended Fixed Limits for Proficiency Testing (mmol/L, or As Stated)

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Population screening*</th>
<th>Detecting long-term trends</th>
<th>Detecting short-term trends</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium</td>
<td>±0.25</td>
<td>±0.15</td>
<td>±0.10</td>
</tr>
<tr>
<td>Chloride</td>
<td>±5</td>
<td>±3</td>
<td>±2</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>±15%</td>
<td>±8%</td>
<td>±4%</td>
</tr>
<tr>
<td>Creatinine</td>
<td>±2 mg/L</td>
<td>±1.5 mg/L</td>
<td>±1 mg/L</td>
</tr>
<tr>
<td>Glucose</td>
<td>±10%</td>
<td>±5%</td>
<td>±5%</td>
</tr>
<tr>
<td>Urea N</td>
<td>±60 mg/L</td>
<td>±50 mg/L</td>
<td>±15 mg/L</td>
</tr>
<tr>
<td>Uric acid</td>
<td>±10%</td>
<td>±5%</td>
<td>±5%</td>
</tr>
<tr>
<td>Sodium</td>
<td>±4</td>
<td>±2</td>
<td>±1</td>
</tr>
<tr>
<td>Potassium</td>
<td>±9%</td>
<td>±5%</td>
<td>±3%</td>
</tr>
<tr>
<td>Sodium</td>
<td>±17%</td>
<td>±9%</td>
<td>±5%</td>
</tr>
</tbody>
</table>

*These values are the same as the Fixed Limits of the CAP Comprehensive Chemistry Survey. *Whichever is greater.

There is really no perceived difference between 0 and 100 U/L for nearly all the commonly performed enzyme tests assuming an upper reference limit of 100 U/L. For use other than population screening—e.g., assays for serum enzymes in suspected acute myocardial infarction cases—the above criteria are too liberal and CVs of 10% to 15% would be required.

Conclusions

A medically acceptable error for laboratory tests must be defined in the context of test use. Bias from the true value must be considered, and the combination of bias and random error constitutes the total medically acceptable error. Bias of certain common analytes can be dealt with by proper standardization. Definitive methods exist for serum Ca, Cl, Fe, K, Li, Mg, Na (34), cholesterol, glucose, urea, and uric acid (35), and excellent estimates of the true values for these analytes are produced by certain peer groups in the CAP Chemistry Survey. Such results can be used to decrease or eliminate bias in most instances.

For population screening, what factors demand the best accuracy that the laboratory can produce? Conversely, where does mandating a reduction in imprecision yield no benefit? If there is considerable overlap of the well and sick populations, then reducing imprecision yields improvements in the diagnostic efficiency. Also, if the biological variability is small, if the prevalence of disease is low, and if the cost to society of a missed diagnosis is unacceptably high, then the greatest accuracy possible is required of the laboratory.

More uncertainty in laboratory data is acceptable when the population distributions for healthy and sick are well separated, the intrapersonal biological variability is large, and the prevalence of disease is high. The laboratory is of course not omnipotent, and repeat determinations will improve the yield of diagnostic information.

For trend detection, inpatients (short-term trends) or outpatients (long-term trends), better accuracy is needed than for population screening. We have shown here that short-term trends can be detected as early as in the third test of a monotonous series if the laboratory imprecision is low. Consistent trends are generally obvious by the fourth or fifth test, and more error is acceptable.

Evaluation of PT results by peer-group statistics is invalid, as Rose (10) has amply shown. Evaluations based solely on statistical analyses of participants' results are easy to do but hard to justify; they ignore medical-needs criteria, the only reasonable yardstick today. We recommend that the CAP-fixed criteria for the common analytes in clinical chemistry (Table 7) be used for PT testing evaluation and as total error limits for population screening. We suggest that these limits be halved for long-term trend detection and approximately quartered for short-term trend recognition (see Table 7). These limits are arbitrary but reasonable, given the evidence shown here and elsewhere (10). We believe the three sets of limits define reasonable goals for within-laboratory proficiency and satisfy three types of medical needs.

References

7. Rippey JH, Williamson WE. The overall role of a proficiency testing program. Ibid., 1980;40-42.