Allowable Imprecision for Laboratory Tests Based on Clinical and Analytical Test Outcome Criteria

James O. Westgard¹, Julie J. Seehafer, and Patricia L. Barry

The allowable imprecision for laboratory tests has been estimated from criteria based on clinical and analytical test outcome. The analytical outcome criteria studied are the Clinical Laboratory Improvement Amendments (CLIA) criteria for proficiency testing. The clinical outcome criteria are estimates of medically significant changes in test results taken from a study in the literature. The estimates of allowable imprecision were obtained from quality-planning models that relate test outcome criteria to the allowable amount of imprecision and inaccuracy and to the quality control that is necessary to assure achievement of the desired outcome criteria in routine operation. These operating specifications for imprecision are consistently more demanding (require lower CVs) than the medically useful CVs originally recommended in the literature because the latter do not properly consider within-subject biological variation. In comparing estimates of allowable imprecision, the CLIA outcome criteria are more demanding than the clinical outcome criteria for aspartateaminotransferase (asymmetric patients), cholesterol, creatinine (asymmetric patients), glucose, thyroxine, total protein, urea nitrogen, hematocrit, and prothrombin time. The clinical outcome criteria are more demanding for bilirubin (acute illness), iron, potassium, urea nitrogen (acute illness), and leukocyte count. The estimates of allowable imprecision from analytical and clinical outcome criteria overlap for aspartateaminotransferase (acute illness), bilirubin (asymmetric patients), calcium, creatinine (acute illness), sodium, triglyceride, and hemoglobin.

Indexing Terms: laboratory management/quality control/proficiency testing

Test outcome criteria provide consumer-oriented requirements for quality, in contrast to performance goals for imprecision and inaccuracy, which provide laboratory specifications for operating a testing process. Outcome criteria define a limit for the total variation that is allowable in a laboratory test result, rather than setting a limit for an individual factor or component that contributes to the variation of a test result. Such outcome criteria can be formulated to reflect either analytical or clinical requirements for test performance.

Analytical outcome criteria describe the total analytical errors that would cause a test result to be judged as analytically unacceptable. For example, the 1988 Clinical Laboratory Improvement Amendments (CLIA) define the total error criteria for proficiency testing (PT) to be used in judging the acceptability of laboratory performance for ~80 tests that are regulated in the US (1).² Given that the origins of the CLIA PT criteria are not well-documented, we consider it of interest to understand how they compare with clinical outcome criteria.

Clinical outcome criteria can be formulated in terms of the total variation in a test result that would cause the medical interpretation to change. Almost 10 years ago, Skendzel et al. (2) surveyed physicians by use of clinical vignettes that focused on test interpretation in various situations, such as patients who are healthy and undergoing routine screening, patients having a variety of disorders, and patients being monitored to detect drug toxicity. The authors determined the average changes in laboratory test results that would be judged by physicians to be medically significant. These estimates reflected physicians' opinions of changes in test results that would cause them to "... take action such as ordering more tests, changing therapy, or considering another diagnosis."

The usefulness of clinical outcome criteria based on physicians' opinions vs the quality goals for imprecision derived from biological variation has been discussed (3), as has the use of analytical outcome criteria for PT vs quality goals for imprecision (4). Unfortunately, direct comparisons are not possible between clinical outcome criteria, analytical outcome criteria, and goals for imprecision and inaccuracy because each set of criteria describes limits for different factors that affect the variation in a test result. Clinical outcome criteria describe a limit for both preanalytical and analytical components of variation in the testing process. Analytical outcome criteria describe a limit for only components of the analytical process, e.g., imprecision, inaccuracy, and quality control (QC). Goals for imprecision and inaccuracy are operating specifications for individual components.

Comparisons can be made when clinical and analytical outcome criteria are used to derive operating specifications for the imprecision and inaccuracy that are allowable and for the QC that is necessary in routine operation of an analytical testing process (5, 6). Previously, we illustrated how proposed European specifications for imprecision and inaccuracy could be compared with US CLIA proficiency testing criteria (7). Here, we determine the maximum imprecision that is allowable based on the clinical outcome criteria of Skendzel et al. (2) and compare these estimates with the maximum

¹ Department of Pathology and Laboratory Medicine, Medical School, University of Wisconsin, 600 Highland Ave., Madison, WI 53792.
² Nonstandard abbreviations: PT, proficiency testing; CLIA, Clinical Laboratory Improvement Amendments (1988); QC, quality control; and NCEP, National Cholesterol Education Program.

Received April 4, 1994; accepted May 24, 1994.

imprecision that is allowable based on CLIA analytical outcome criteria.

Materials and Methods

Quality-Planning Models

The formulation of clinical and analytical quality planning models has been described earlier (8, 9). To apply the clinical model, we estimated the clinical decision interval (D_int) as the difference between the “from” and “to” values of physicians’ opinions of a significant change in test results shown in Table 1 of Skendzel et al. (2). D_int was then related to preanalytical and analytical factors by the following equation:

$$D_{int} = \text{bias}_{spec} + \text{bias}_{meas} + \Delta \text{SE}_{cont} \text{SE}_{meas}$$

$$+ z \sqrt{\frac{\text{VAR}_{\text{wu}}}{\text{n}_{\text{test}}} + \frac{\text{VAR}_{\text{spec}}}{\text{n}_{\text{spec}}} + \frac{(\Delta \text{RE}_{\text{cont}} \text{RE}_{\text{meas}})^2}{\text{n}_{\text{test}} \text{n}_{\text{spec}} \text{n}_{\text{samp}}}}$$

where bias_{spec} is the sampling bias, bias_{meas} is the stable measurement bias or analytical inaccuracy, ΔSE_{cont} is the sensitivity of the QC procedure for detecting systematic error, s_{meas} is the stable measurement standard deviation or analytical imprecision, z is related to the maximum defect rate allowable before stopping the process, s_{wu} is the within-subject biological variation, s_{spec} is the between-specimen sampling variation, ΔRE_{cont} is the sensitivity of the QC procedure for detecting random error or changes in imprecision, n_{test} is the number of tests performed, n_{spec} is the number of specimens drawn, and n_{samp} is the number of samples measured for each specimen.

When the preanalytical factors in the clinical quality-planning model are set to 0, only analytical factors remain, so that the outcome criterion becomes the total allowable analytical error, TE_a. For the condition that n_{test}, n_{spec} and n_{samp} are 1 (i.e., a single test with a single specimen drawn and a single aliquot tested), TE_a is related to the analytical factors as follows:

$$\text{TE}_a = \text{bias}_{meas} + \Delta \text{SE}_{cont} \text{SE}_{meas} + z \Delta \text{RE}_{cont} \text{RE}_{meas}$$

In applying the analytical model, the CLIA PT criteria for acceptable performance provide values for TE_a, which we indicate by the symbol TE_{PT}.

Estimation of Allowable Imprecision

Outcome criteria were applied at the medical decision levels or target values recommended by Skendzel et al. (2), which correspond to clinical situations for asymptomatic patients, acute illness, and drug monitoring (as identified in Table 1 of ref. 2).

Sampling bias (bias_{spec}) and between-specimen variation (s_{spec}) were assumed to be 0, analytical performance was evaluated for a single test (n_{test} = n_{spec} = n_{samp} = 1), z was 1.65 (to set a one-tailed or one-sided maximum defect rate of 0.05 or 5%), and control performance was optimized for detection of systematic error (ΔRE_{cont} = 1.0).

Estimates of within-subject biological variation for chemistry tests were based on Fraser's (10, 11) summaries of studies in the literature. The particular values we used represent the average of all values tabulated for studies that lasted >1 week. Results listed in the original summaries as “Neg” were excluded, as was one value for alkaline phosphatase that did not seem consistent with others. Estimates of within-subject variation for hematology tests were obtained from a single study by Fraser (12), and the estimate for prothrombin time was the mean value for the group from a study by Dot et al. (13).

Estimates of the maximum allowable imprecision were determined from the x-intercepts of the lines describing the allowable limits of imprecision and inaccuracy on charts of operating specifications (OPSpecs charts, 6), which were prepared with the QC Validator program (Westgard QC, Ogunquit, ME). These estimates represent the imprecision that would be allowable with commonly used QC procedures, such as 1.96 with N = 2, 2.58 with N = 2 and 4, 1.96 with N = 2 and 4, 1.96/2.58/4.8 for N = 2, and 1.96/2.58/4.8 for N = 4. OPSpecs charts for 90% analytical quality assurance were utilized to specify an error detection of 0.90 or 90% for critical systematic errors that would cause measurement performance to exceed the defined outcome criteria. The relative demands of clinical and analytical outcome criteria are evaluated by comparing these estimates of the clinical and analytical maximum allowable imprecision.

Results

Figure 1 shows the inaccuracy and imprecision that are allowable for a calcium method at a decision level of 85 mg/L when the medically significant change is 9 mg/L, or 10.58%. The top line on this chart describes the maximum limits that would be allowable for a perfectly stable measurement procedure that requires no QC. The lower lines describe the operating limits that are necessary to assure that the desired quality will be achieved in routine operation with use of the QC procedures that are commonly employed in clinical laboratories today. The key area at the right of the chart identifies the control rules and number of control measurements in the order of the lines from top to bottom. The operating point, which is seen to be off scale on this chart, represents the medically useful CV of 4.8% from Table 3 of Skendzel et al. (2). By comparison, when within-subject biological variation and the sensitivity of the QC procedures are accounted for, the clinical maximum allowable CV is 1.7–2.4%, as determined from the x-intercepts of the operating lines for the different QC procedures.

Figure 2 shows similar information for calcium for the CLIA PT criterion of 10 mg/L, which corresponds to 11.76% at a decision level of 85 mg/L. The analytical maximum allowable imprecision is 2.3–3.0%, as estimated from the x-intercepts of the operating lines for the different QC procedures. The operating point again shows that the medically useful CV of 4.8% exceeds the allowable limits.
Table 1 summarizes the allowable imprecision for those tests for which clinical outcome criteria are defined by Skendzel et al. (2) and analytical outcome criteria are defined by CLIA PT criteria (1). Column 4 shows the clinical outcome criteria in the form of the medically significant changes from Table 1 by Skendzel et al. (2). Column 5 shows the estimates of within-subject biological variation that were used to derive the clinical maximum allowable imprecision shown in column 6. These estimates of imprecision are to be compared with the values in column 7 for the medically useful CVs [taken from Table 3 of Skendzel et al. (2)] and with the values for the analytical maximum allowable imprecision (column 8), derived from the CLIA analytical outcome criteria shown in column 9.

These estimates of the maximum allowable imprecision show that the CLIA analytical outcome criteria are more demanding than the clinical outcome criteria for aspartate aminotransferase (asymptomatic patients), cholesterol, creatinine (asymptomatic patients), glucose, thyroxine, total protein, urea nitrogen, hematocrit, and prothrombin time. The clinical outcome criteria are more demanding than the analytical outcome criteria for bilirubin (acute illness), iron, potassium, urea nitrogen (acute illness), and leukocyte count. The estimates of allowable imprecision overlap for aspartate aminotransferase (acute illness), bilirubin (asymptomatic patients), calcium, creatinine (acute illness), sodium, triglyceride, and hemoglobin. All the operating specifications for allowable imprecision are more demanding than the medically useful CVs recommended by Skendzel et al. (2).

Discussion

The difficulty in comparing different requirements for analytical performance is illustrated in Table 1 by the iron test, where the medically significant change or clinical decision interval of 33% is almost entirely consumed by the within-subject biological variation of 19.8%, requiring analytical imprecision of <1% when biological variation is taken into account, rather than the medically allowable CV of 17.2% recommended by Skendzel et al. (2). Similar situations are seen for bilirubin (acute illness) and urea nitrogen (acute illness), where the within-subject biological variation by itself will cause changes that exceed the intervals for medi-
Table 1. Estimates of allowable imprecision from clinical and analytical outcome criteria.

<table>
<thead>
<tr>
<th>Test</th>
<th>Decision level</th>
<th>Units</th>
<th>Medically significant change, (D_{\text{med}}) (%)</th>
<th>Within-subject variation, (s_{\text{within}}) (%)</th>
<th>Clin. maximum allowable imprecision, (s_{\text{max}}) (%)</th>
<th>Skenechel's medically useful CV, (s_{\text{med}}) (%)</th>
<th>Analytical maximum allowable imprecision, (s_{\text{an}}) (%)</th>
<th>US CLIA PT Criteria, TE&lt;sub&gt;PT&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspartate aminotrans.</td>
<td>30</td>
<td>U/L</td>
<td>13 (43.3)</td>
<td>10.8</td>
<td>6.5-9.0</td>
<td>26.3</td>
<td>3.9-5.1</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td></td>
<td>20 (28.6)</td>
<td></td>
<td>2.9-4.2</td>
<td>14.3</td>
<td>3.9-5.1</td>
<td></td>
</tr>
<tr>
<td>Bilirubin, total</td>
<td>8</td>
<td>mg/L</td>
<td>6 (75.0)</td>
<td>18.1</td>
<td>11.2-15.7</td>
<td>23.4</td>
<td>9.7-13.0</td>
<td>20% or 4 mg/L&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td></td>
<td>20 (13.3)</td>
<td></td>
<td>0.4</td>
<td>5.4</td>
<td>3.9-5.1</td>
<td></td>
</tr>
<tr>
<td>Calcium</td>
<td>85</td>
<td>mg/L</td>
<td>9 (10.6)</td>
<td>2.0</td>
<td>1.7-2.4</td>
<td>4.8</td>
<td>2.3-3.0</td>
<td>10.0 mg/L</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>2100</td>
<td>mg/L</td>
<td>100 (33.3)</td>
<td>6.1</td>
<td>5.6-7.6</td>
<td>12.3</td>
<td>1.9-2.6</td>
<td>10%</td>
</tr>
<tr>
<td>Creatinine</td>
<td>10</td>
<td>mg/L</td>
<td>5 (50.0)</td>
<td>4.9</td>
<td>9.1-12.4</td>
<td>17.2</td>
<td>5.7-7.7</td>
<td>15% or 3 mg/L&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td></td>
<td>5 (16.7)</td>
<td></td>
<td>2.2-3.1</td>
<td>10.1</td>
<td>2.9-3.8</td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td>1000</td>
<td>mg/L</td>
<td>300 (30.0)</td>
<td>12.2</td>
<td>2.7-3.9</td>
<td>11.2</td>
<td>1.9-2.6</td>
<td>10% or 60 mg/L&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Iron</td>
<td>1500</td>
<td>(\mu)g/L</td>
<td>500 (33.3)</td>
<td>19.8</td>
<td>0.2-0.3</td>
<td>17.2</td>
<td>3.9-5.1</td>
<td>20%</td>
</tr>
<tr>
<td>Potassium</td>
<td>3.8</td>
<td>mmol/L</td>
<td>0.4 (10.5)</td>
<td>4.9</td>
<td>0.7-1.0</td>
<td>4.8</td>
<td>2.6-3.4</td>
<td>0.5 mmol/L</td>
</tr>
<tr>
<td>Sodium</td>
<td>125</td>
<td>mmol/L</td>
<td>5 (4.0)</td>
<td>0.8</td>
<td>0.7-0.9</td>
<td>1.7</td>
<td>0.6-0.8</td>
<td>4 mmol/L</td>
</tr>
<tr>
<td></td>
<td>135</td>
<td></td>
<td>6 (4.4)</td>
<td></td>
<td>0.7-1.0</td>
<td>2.7</td>
<td>0.6-0.8</td>
<td></td>
</tr>
<tr>
<td>Thyroxine, total</td>
<td>60</td>
<td>(\mu)g/L</td>
<td>20 (33.3)</td>
<td>5.8</td>
<td>5.7-7.6</td>
<td>17.2</td>
<td>3.9-5.1</td>
<td>20% or 10 (\mu)g/L&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total protein</td>
<td>70</td>
<td>g/L</td>
<td>15 (21.4)</td>
<td>3.3</td>
<td>3.8-5.1</td>
<td>8.3</td>
<td>1.9-2.6</td>
<td>10%</td>
</tr>
<tr>
<td>Triglyceride</td>
<td>1300</td>
<td>mg/L</td>
<td>600 (46.2)</td>
<td>19.8</td>
<td>3.4-5.5</td>
<td>16.1</td>
<td>4.8-6.4</td>
<td>25%</td>
</tr>
<tr>
<td>Urea N</td>
<td>180</td>
<td>mg/L</td>
<td>100 (56.6)</td>
<td>12.2</td>
<td>8.7-12.0</td>
<td>18.7</td>
<td>2.2-2.8</td>
<td>9% or 20 mg/L&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td></td>
<td>80 (20.0)</td>
<td></td>
<td>0.0</td>
<td>12.2</td>
<td>1.8-2.3</td>
<td></td>
</tr>
<tr>
<td>Hematocrit</td>
<td>420</td>
<td>m/L</td>
<td>50 (11.9)</td>
<td>2.5</td>
<td>1.9-2.6</td>
<td>5.4</td>
<td>1.2-1.5</td>
<td>6%</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>150</td>
<td>g/L</td>
<td>12 (8.0)</td>
<td>2.4</td>
<td>1.1-1.5</td>
<td>3.6</td>
<td>1.4-1.8</td>
<td>7%</td>
</tr>
<tr>
<td>Leukocyte count</td>
<td>(5 \times 10^8)</td>
<td>cell/L</td>
<td>1.6 (\times 10^6) (32)</td>
<td>15.6</td>
<td>1.7-2.6</td>
<td>16.4</td>
<td>2.9-3.8</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>(25 \times 10^9)</td>
<td>cell/L</td>
<td>7 (\times 10^9) (28)</td>
<td></td>
<td>0.6-1.0</td>
<td>14.0</td>
<td>2.9-3.8</td>
<td></td>
</tr>
<tr>
<td>Prothrombin time</td>
<td>12.0</td>
<td>s</td>
<td>3 (25.0)</td>
<td>2.3</td>
<td>5.1-7.1</td>
<td>15.2</td>
<td>2.9-3.8</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>20.0</td>
<td></td>
<td>5 (25.0)</td>
<td></td>
<td>5.1-7.1</td>
<td>15.2</td>
<td>2.9-3.8</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Preanalytical variation exceeds the medically significant change.

<sup>b</sup> Whichever is greater.

The rigorous demands for analytical imprecision result from the design objective to achieve 90% detection of critical-sized systematic errors by using only the low numbers of control measurements \(N = 2-4\) that are deemed acceptable by professional practice and regulatory guidelines. The 90% objective is based on the desire to detect medically important analytical errors in the first run in which they occur (expected average run length ARL = \(1/P_{\text{ARL}}\) or 1/0.90 = 1.1). Use of low \(N\) values limits the power of statistical QC procedures and makes it difficult to detect critical systematic errors that are \(< -3\ CV\) of the method. Improving precision (reducing the CV) has the effect of changing those critical systematic errors to higher multiples of the CV, enabling greater error detection with commonly used QC procedures.

These rigorous demands for analytical performance are achievable in routine operation, as demonstrated in an earlier study (16), where 90% error detection was obtained for 15 out of 18 tests on a multistart chemistry analyzer with use of \(N = 2\) and single-rule QC procedures. Lower probabilities for error detection may be...
acceptable for very stable analyzers that have a low frequency of problems (17), but it is generally advisable to first attempt to achieve 90% detection and then, if necessary, optimize performance for the observed or expected frequency of problems. When appropriate, such optimization can be aided by OPSpecs charts with 50% or 25% analytical quality assurance specifications (available from the QC Validator program) or by QC selection grids (18).

In summarizing recommendations for allowable imprecision, our use of the x-intercepts from the OPSpecs curves assumes that inaccuracy is zero (bias_nonzero = 0.0%). The effects of non-zero biases can be determined from the OPSpecs curves themselves and would lead to the need for smaller CVs. Even smaller CVs would be required if other preanalytical components were significant. Except for within-subject biological variation, we assumed other preanalytical variables were zero because of the difficulty of obtaining this information for the many analytes considered here; however, the quality-planning models provide the capability of dealing with additional variables, such as sample variation and specimen bias.

When within-subject biological variation is accounted for in the clinical model and QC sensitivity is included in both the clinical and analytical models, the estimates of allowable imprecision from the clinical and analytical outcome criteria are actually in good agreement for about one-third of the tests studied, the analytical outcome criteria are more demanding than the clinical for another half, and the clinical are more demanding than the analytical for the remaining sixth. These findings suggest that it should be possible to establish a coherent set of clinical and analytical outcome criteria that would provide consistent operating specifications for the imprecision, inaccuracy, and QC needed by laboratory testing processes.

Boone discussed the advantages of using analytical outcome criteria of the total error format for PT purposes but also stressed the need for PT criteria to reflect medical usefulness (4). This would seem to be achievable by relating clinical and analytical outcome criteria through the operating specifications required to assure the desired outcome. Estimates of medically important changes and within-subject biological variation are the primary information from which the total analytical error would be determined. Other preanalytical components of error could be considered when necessary. PT criteria in the format of an allowable total analytical error could then be established that would provide consistent operating specifications for imprecision, inaccuracy, and QC. This approach would permit PT criteria to reflect medical needs and provide a rationale for deriving, explaining, and verifying such criteria.

Experience with cholesterol testing in the US is cited by Boone (4) as evidence that medically relevant goals for analytical performance are helpful in setting realistic criteria for PT and QC. National Cholesterol Education Program (NCEP) guidelines (19) for use and interpretation of a cholesterol test have established an interval from 2000 to 2400 mg/L as the medically significant change, whereas it is the change from 2100 to 2800 mg/L that represents medical practice according to the study by Skendzel et al. (2). The NCEP reduction of the earlier decision interval of 700 mg/L to 400 mg/L places considerably greater demands on test performance.

Wiebe and Westgard (20) also identify cholesterol as a model system for relating medical needs to analytical performance and demonstrate that current NCEP clinical requirements for data interpretation and CLIA PT criteria are not consistent. The NCEP guidelines for interpretation of a single cholesterol test correspond to a decision interval of 20%, which leads to an estimate of 2.4% to 3.5% for the maximum allowable imprecision, compared with the 1.9% to 2.6% that is allowable for a 10% CLIA PT criterion. A CLIA PT criterion of 13% would provide operating specifications that are consistent with the NCEP clinical requirement for data interpretation. Thus, clinical guidelines for the interpretation of a cholesterol test could be used to establish operating specifications for imprecision, inaccuracy, and QC, which in turn could be related to an analytical outcome criterion that requires consistent operating specifications.

This cholesterol model illustrates the importance of understanding the hierarchy between clinical and analytical outcome criteria, the relationships between outcome criteria and operating specifications, and the need to include the performance of the QC procedures in establishing operating specifications. All of these characteristics are important in managing the quality of laboratory testing processes, but each has a particular role to play in a system for total quality management.

In summary, this study demonstrates how the imprecision that is allowable for a laboratory test can be determined from either clinical or analytical outcome criteria and suggests that medically relevant criteria for proficiency testing can be established by using coherent clinical and analytical quality-planning models. In conjunction with our earlier study that demonstrated how analytical goals for imprecision and inaccuracy can be compared with analytical outcome criteria (7), it should also be apparent that those analytical goals can similarly be compared with operating specifications derived from clinical outcome criteria. Together, these studies illustrate an approach for relating, comparing, and making some sense of the many different quality requirements, recommendations, goals, and specifications for laboratory tests.

References
3. Fraser CG, Hyltoft Petersen P. Desirable standards for labora-
tory tests if they are to fulfill medical needs. Clin Chem 1993;39:1447–53.