A review of method comparison studies published in the American Journal of Clinical Pathology indicates that hematology evaluations are less rigorous than their chemistry counterparts and rely heavily on the correlation coefficient. While clinical chemistry evaluations depend more on linear regression, they tend to omit relevant, lesser-known statistics, such as the standard error of the estimate. The authors reiterate guidelines for the collection and statistical analysis of method comparison data and recommend that both hematology and chemistry evaluations be improved. (Key words: Instrument evaluation; Method evaluation; Accuracy; Precision; Allowable error) Am J Clin Pathol 1986; 86: 391–393

RECENTLY, Fraser and Singer pointed out that many method comparison studies published in the clinical chemistry literature either omit useful statistics or rely too heavily on weak or inappropriate statistical tests.8 Because a cursory inspection of method evaluation studies in the clinical pathology literature appeared to demonstrate these same limitations, we reviewed the articles published in a leading clinical pathology journal, the American Journal of Clinical Pathology. We studied all the evaluations of instruments and commercial quantitative assays published in the American Journal of Clinical Pathology from January 1984 to June 1985 inclusive (volumes 81–83) and tabulated the statistical tests employed by both clinical chemistry and hematology/coagulation evaluations.

Methods and Results

Because there are no well-accepted guidelines for the evaluation of qualitative and semiquantitative assays, we eliminated such studies from further consideration. Of the 13 articles that evaluated quantitative assays, 6 were in the general category of clinical chemistry and 7 in the category of hematology/coagulation. The statistics used by each of the articles were tabulated and are shown in Table 1.

Table 1 indicates that clinical chemistry evaluations are more likely to use linear regression and report the slope and y-intercept (four of six vs. two of seven). By contrast, hematology and coagulation evaluations tend to only present a scatter graph without regression analysis (five of seven) and rely heavily on the correlation coefficient (four of seven). While the clinical chemistry studies appeared to be well planned and better documented, several useful statistics were omitted. The standard error of the estimate (also known as the standard deviation of the regression line, $S_{y,x}$) was reported in only one study. The standard error of the slope and the standard error of the y-intercept were not reported in any of the articles. Only one study determined the total error of the test method and compared it with the allowable error at a medical decision level.9

Discussion

The recent proliferation of clinical chemistry evaluation protocols plus the requirements of journals such as Clinical Chemistry for specific method evaluation statistics2-4,11-20 have contributed to the increasing sensitivity of clinical chemists and chemical pathologists to the rigorous demands of method comparison studies. Our study confirms the findings of Fraser and Singer, that clinical chemistry evaluations still omit important statistics such as standard error of the estimate and standard deviation of the slope and intercept. In fact, estimates of within-run and between-run imprecision were provided for only one of the evaluations. There is little attempt to estimate total analytic error and compare it with the medically allowable error. Only one article, a calcium evaluation, evaluated the total error and compared it with the medically allowable error.9

In contrast to the clinical chemistry evaluations, the evaluations of quantitative hematology and coagulation methods are generally less rigorous, with approximately one-half employing the correlation coefficient as the primary criterion for test acceptability. While few of the
evaluations present imprecision data, we would like to believe that within-run and between-run imprecision studies were performed but not documented. The determination of imprecision by regression analysis is sporadic.

With the onset of the prospective payment system, the clinical laboratory has become a cost center. The laboratory cannot afford to produce data of uncertain analytic and clinical significance. Published method evaluations should be rigorous and provide estimates of imprecision, inaccuracy, and, also, total analytic error. Such evaluations will make it easier for laboratory personnel to avoid the evaluation and installation of instruments and methods with borderline performance.

**Guidelines for Method Comparison Studies**

Based on accepted recommendations for method comparison studies and tempered by our experience, we offer the following brief method comparison protocol, which is summarized in Table 2. A linearity study should always be done to confirm the analytic range (the range over which the method is applicable without modification). The calibration curve should be accurate and reproducible at the medical decision levels. Recovery and interference studies require some effort and should have been performed by the manufacturer before marketing. Nevertheless, selected experiments should be done to rule out proportional error and susceptibility to common interfering substances such as bilirubin.

Measurement of imprecision requires measurement of short-term (within-run or within-day) and long-term (between-day) imprecisions. The short-term imprecision studies should be carried out first. For these studies, we recommend that ten replicate analyses be performed on five to ten varied samples, chosen to represent different levels of medical interest, using both patient and control material. The between-day imprecision studies may be carried out at the same time as the comparison of methods experiment. The between-day imprecision studies should extend over at least one week in the absence of stable control material but preferably over 20 separate days, using two to three levels of control material. We recommend four measurements per day for 20 days. Using more observations per day or less than 20 days will underestimate the between-day imprecision. If it is necessary to conduct the between-day imprecision study over a short period of time, it may be necessary to use analysis of variance (ANOVA) to estimate the between-day imprecision. In any event, short- and long-term standard deviations should be calculated for each method.

Ideally, the new test method should be compared with a definitive or reference method. When such a method is unavailable or is too labor intensive, the new method should at least be compared with the one currently in use. Comparison of methods experiments that are to be published should include at least 100 patient samples selected so that they reflect the entire range of medical interest. These patient samples should be analyzed five per day over at least 20 days. The data should be plotted daily so that outliers can be identified, investigated, and, if indicated, eliminated from the final computation. It may be useful to analyze the samples in duplicate to aid in the identification of outliers. The duplicate values can be used later if an alternate method of linear regression analysis is necessary, such as the method of Deming.

Linear regression should be used to analyze the paired

---

**Table 1. The Use of Statistics in Recent Method Comparison Studies Published in the American Journal of Clinical Pathology**

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Clinical Chemistry</th>
<th>Hematology/Coagulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of studies</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Within-day imprecision</td>
<td>3/6</td>
<td>2/7</td>
</tr>
<tr>
<td>Between-day imprecision</td>
<td>3/6</td>
<td>1/7</td>
</tr>
<tr>
<td>Scatter plot</td>
<td>4/6</td>
<td>7/7</td>
</tr>
<tr>
<td>Slope of regression line</td>
<td>4/6</td>
<td>2/7</td>
</tr>
<tr>
<td>Standard deviation of the y-intercept</td>
<td>3/6</td>
<td>3/7</td>
</tr>
<tr>
<td>Standard deviation of the y-intercept</td>
<td>0/6</td>
<td>0/7</td>
</tr>
<tr>
<td>Standard error of the estimate</td>
<td>1/6</td>
<td>0/7</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>5/6</td>
<td>6/7</td>
</tr>
<tr>
<td>Bias</td>
<td>4/6</td>
<td>2/7</td>
</tr>
<tr>
<td>Bias (evaluated with t-test)</td>
<td>2/6</td>
<td>1/7</td>
</tr>
<tr>
<td>Total error at medical decision level</td>
<td>1/6</td>
<td>0/7</td>
</tr>
</tbody>
</table>

---

**Table 2. Recommended Studies and Pertinent Statistics to Be Used in Method Comparison Studies**

<table>
<thead>
<tr>
<th>Recommended Studies</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linearity</td>
<td>Determines the analytic range and should encompass the medical decision level(s)</td>
</tr>
<tr>
<td>Recovery</td>
<td>Estimate of proportional error</td>
</tr>
<tr>
<td>Interference</td>
<td>Estimate of error due to common interfering substances</td>
</tr>
<tr>
<td>Imprecision</td>
<td>Estimate of random error</td>
</tr>
<tr>
<td>Within-day imprecision</td>
<td>Ten replicates of 5–10 samples</td>
</tr>
<tr>
<td>Between-day imprecision</td>
<td>Conducted over a 20-day period</td>
</tr>
<tr>
<td>Paired sample comparison</td>
<td>At least 100 patient samples</td>
</tr>
<tr>
<td>Scatter plot</td>
<td>Proportional error</td>
</tr>
<tr>
<td>Linear regression analysis</td>
<td>Constant error</td>
</tr>
<tr>
<td>Slope</td>
<td>Intermethod random error</td>
</tr>
<tr>
<td>y-Intercept</td>
<td></td>
</tr>
<tr>
<td>Standard error of the estimate</td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td></td>
</tr>
<tr>
<td>t-Test</td>
<td></td>
</tr>
<tr>
<td>Estimate of total error</td>
<td></td>
</tr>
</tbody>
</table>
sample data because it identifies both proportional error (slope of the regression line) and constant error (y-intercept). Classical linear regression usually is employed and minimizes the distance between the paired data points and the regression line only in the y (vertical) direction. Alternate forms of regression, such as the technic of Deming, are required if the range of values is narrow (for example, serum sodium levels and mean corpuscular volumes) or if the comparative method is imprecise.5,17,18 Deming's technic minimizes the distances between the paired data points and the regression line in a direction perpendicular to the regression line and makes no assumptions about the relative imprecisions of the test and comparative methods. Since the correlation coefficient (r) is extremely sensitive to random error, outliers, and the range of concentration of analyte,21 it is not an appropriate estimate of the agreement between two methods. The correlation coefficient's most useful role in method comparison studies is in indicating the need for an alternate method of linear regression. The heavy reliance on the correlation coefficient as the criterion for accepting or rejecting a test method is disconcerting because it is not sensitive to constant or proportional error.

The standard error of the estimate, the standard deviation of the points around the regression line, provides a measure of the intermethod imprecision. The standard error of the slope of the regression line and the standard error of the y-intercept and their confidence intervals can be used to evaluate the magnitude of the differences of the slope and the y-intercept from their ideal values of 0 and 1, respectively.

Linear regression assumes that the variance of the observed test values (the y's) is constant throughout the range of the comparative values (the x's). This assumption can be verified by visual inspection of the scatter plot of the data. If this assumption of stable imprecision of the test method proves false, weighted regression or transformation of the data may be required for the analysis. In fact, the robust nature of linear regression usually guarantees adequate results for the evaluation of most clinical laboratory methods.

The importance of the paired t-test in analyzing the method comparison data is overemphasized. While the elements of the paired t-test (the bias and the standard deviation of the differences) offer measures of systematic and random error, these estimates tend to be valid only near the means of the data sets. In addition, the t-test may indicate a statistically significant bias in situations where the difference may not be clinically significant, especially when N is large.

The ultimate criterion for instrument or method selection should be whether the estimate of total analytic error exceeds the medically allowable error. The total analytic error is derived from the sum of the estimates of imprecision and inaccuracy.18,20 Estimates of medically allowable error for clinical chemistry tests have been suggested by Barnett, Elion-Gerritzen, and others6,16,18 and have been neatly summarized by Frazer.7 With the exception of hemoglobin, there are no accepted estimates of medically allowable error for hematology and coagulation tests. Westgard has provided a conceptually simple approach for comparing the total error with the medically allowable error.19,20

References
20. Westgard JO, de Vos DJ, Hunt MR, Quam EF, Carey RN, Garber CC: Method evaluation. Houston, TX, American Society of Medical Technology, 1978