Points of Care in Using Statistics in Method Comparison Studies

As clinical chemists and laboratory scientists, we are often concerned when personnel who have little laboratory training begin to perform laboratory tests, such as in point-of-care applications. It may be easy to perform such tests today with modern analytical systems, but there still are things that could go wrong. We hope that some kind of quality system is used to check that everything is working okay with point-of-care analyses.

Imagine how statisticians might feel about the powerful statistics programs that are now in our hands. It is so easy to key-in a set of data and calculate a wide variety of statistics—regardless what those statistics are or what they mean. There also is a need to check that things are done correctly in the statistical analyses we perform in our laboratories.

In this issue of the Journal, Stöckl et al. *(1)* provide an interesting discussion of linear regression techniques in method comparison studies, pointing out that the quality of the data may be more important than the quality of the regression technique (e.g., ordinary linear regression vs Deming regression vs Passing-Bablock regression). In this Journal, the standard method for analyzing the data from a method comparison experiment has been to prepare a "comparison plot" that shows the test method results on the *y*-axis and the comparative method results on the *x*-axis, and then to calculate regression statistics to determine the best line of fit for the data. Different regression techniques may be appropriate, depending on the characteristics of the data—particularly the analytical range that is covered relative to the test values that are critical for medical applications.

Elsewhere in the literature *(2)*, there is a movement to discourage the use of regression analysis altogether and replace it with a simple graphical presentation of method comparison data in the form of a "difference plot", which displays the difference between the test and comparative results on the *y*-axis vs the mean of the test and comparative results on the *x*-axis. This difference plot has become known as the Bland-Altman plot *(3)*. Hyltoft Petersen et al. *(4)*, writing in this Journal, have shown that a difference plot must be carefully constructed to make an objective decision about method performance. The difference plot is actually not so simple when an objective interpretation is to be made.

In spite of these recent reports and recommendations on the use of statistics, many analysts and investigators still have difficulties with method comparison data. We studied some of the problems 25 years ago *(5)* and for the most part, there are similar problems today—with the exception that the calculations are much easier to perform with today's computer programs. There has not been much improvement, if any, in the basic statistical knowledge and skills available in laboratories today, not only for method validation studies but also for statistical quality control. That does not mean there have not been improvements in the theory and recommendations appearing in the literature, but rather that the practices in laboratories have not really changed very much.

Therefore, we still need to exercise a great deal of care in collecting, analyzing, and interpreting method comparison data. Here are some points of care to consider:

Point 1: Use statistics to provide estimates of errors, not as indicators of acceptability. This is perhaps the most fundamental point for making practical sense of statistics in method validation studies. The statistics do not directly tell you whether the method is acceptable; rather they provide estimates of errors that allow you to judge the acceptability of a method. You do this by comparing the amount of error observed with the amount of error that would be allowable without compromising the medical use and interpretation of the test result. Method performance is judged acceptable when the observed error is smaller than the defined allowable error. Method performance is not acceptable when the observed error is larger than the allowable error. This decision-making process can be facilitated by mathematical criteria *(6)* or by graphic tools *(7)*.

Point 2: Recognize that the main purpose of the method comparison experiment is to obtain an estimate of systematic error or bias. The comparison of methods experiment is performed to study the accuracy of a new method. The essential information is the average systematic error, or bias. It is also useful to obtain information about the proportional and constant nature of the systematic error and to quantify the random error between the methods. The components of error are important because they relate to the things we can manage in the laboratory to control the total error of the testing process (e.g., reduce proportional systematic error by improved calibration). The total error is important in judging the acceptability of a method and can be calculated from the components.

Point 3: Obtain estimates of systematic error at important medical decision concentrations. The collection of specimens and choice of statistics can be optimized by focusing on the concentration (or concentrations) where the interpretation of a test result will be most critical for the medical application of the test. If there is only a single medical decision concentration, the method comparison data may be collected around that concentration (i.e., a wide range of data will not be necessary), and a difference plot should be useful (along with an estimate of bias from *t*-test analysis—see *point 8* below). If there are two or more decision concentrations, it is desirable to collect specimens that cover a wide analytical range, use a comparison plot to display the results, and calculate regression statistics to estimate the systematic error at each of the decision concentrations.

Point 4: When there is a single medical decision concentration, make the estimate of systematic error near the mean of the data. The main consideration when there is a single medical decision concentration is to collect the data around that medical decision concentration. The choice of statistics will not be critical when there is only one medical decision concentration of interest and it falls near the mean of the data. The bias statistic from paired *t*-test calculations and the systematic error calculated from regression statistics will provide the same estimate of the error.

[Note of explanation: The bias is the difference between the means of the two methods (bias = $Y_{av} - X_{av}$), which is also equivalent to the average of the paired differences from paired *t*-test calculations. With regression statistics, the systematic error (SE) is estimated at a critical concentration, X_C, as follows: SE = $Y_C - X_C$, where Y_C is calculated from the regression statistics by the equation $Y_C = a + bX_C$, where *a* is the *y*-intercept and *b* is the slope of the regression line. In ordinary linear regression, the slope is calculated first, and then the *y*-intercept is determined from $a = Y_{av} - bX_{av}$. When the decision concentration equals X_{av}, then SE = $(a + bX_{av}) - X_{av} = Y_{av} - bX_{av} + bX_{av} - X_{av} = Y_{av} - X_{av}$, i.e., the same estimate of SE will be obtained from regression statistics as from *t*-test statistics, even if the range of data is narrow and the values for the slope and intercept are not reliable.]

Point 5: When there are two or more medical decision concentrations, use the correlation coefficient, r, to assess whether the range of data is adequate for using ordinary regression analysis. As confirmed by Stöckl et al. *(1)*, when *r* is ≥0.99, the range of data should be wide enough for ordinary linear regression to provide reliable estimates of the slope and intercept. They recommend that when *r* is <0.975 ordinary linear regression may not be reliable and that data improvement or alternate statistics are now appropriate. Note that *r* is not used to judge the acceptability of method performance here, but to judge the acceptability of the concentration range of the data being used to calculate the regression statistics.

Point 6. When r is high, use the comparison plot along with ordinary linear regression statistics. The reliability of the slope and intercept are affected by outliers and nonlinearity, as well as the concentration range of the data. Outliers need to be identified, preferably at the time of analysis by immediately plotting the data on the comparison graph; discrepant results can then be investigated while the specimens are still available. Nonlinearity can usually be identified from visual inspection of the comparison plot, the range can be restricted to the linear portion, and the statistics recalculated. Stöckl et al. *(1)* recommend using the residual plot that is available as part of regression analysis and inspecting the sign-sequence of the residuals for making this assessment.

Point 7: When r is low, improve the data or change the statistical technique. Consider the alternatives of improving the range of data, reducing the variation from the comparison method by replicate analyses, estimating the SE at the mean of the data, dividing the data into subgroups whose means agree with the medical decision concentrations (which can then be analyzed by *t*-test statistics and the difference plot), or using a more complicated regression technique. Stöckl et al. *(1)* find that the Deming regression technique is more satisfactory than the Passing-Bablock technique. Note that these regression techniques are not standard in ordinary statistics programs; however, they are available in special programs designed for laboratory method evaluation studies.

Point 8: When r is low and a difference plot is used, calculate t-test statistics to provide a quantitative estimate of SE. Given the objective of estimating SE from the method comparison experiment, the usefulness of the difference plot by itself is questionable because visual interpretation will be mainly influenced by the scatter or random error observed between the methods. The bias, or average difference of paired sample results, should be calculated. Computer routines for calculating *t*-test statistics will provide this estimate, along with an estimate of the SD of the differences, which gives a quantitative measure of the scatter between the methods. Note that this scatter between the methods depends on the imprecision of the test method, the imprecision of the comparison method, and any interferences that affect individual samples differently by the two methods. The *t*-value itself is a ratio of systematic to random error and is mainly useful for determining if sufficient data have been collected to make a reliable estimate of the bias (again, avoid using a statistic as an indicator of the acceptability of the method). Although Bland and Altman *(3)* also recommend calculation of the mean difference and the SD of the differences and suggest that the mean difference ± 2 SD be drawn on the chart, it is wrong to judge the acceptability of the observed differences by comparison to themselves. Hyltoft Petersen et al. *(4)* provide an extensive discussion of judging method acceptability on the basis of the difference plot.

Point 9: When in doubt about the validity of the statistical technique, see whether the choice of statistics changes the outcome or decision on acceptability. Given the ease with which the calculations can be performed with computer programs, the effect of the statistical technique on the estimates of performance can be assessed by comparing the results from the different techniques. If the statistical technique affects your decision on the acceptability of the method, then be careful. Usually it will be best to collect more data and to be sure that these new data satisfy the assumptions of the data analysis technique.

Point 10: Plan the experiment carefully and collect the data appropriate for the statistical technique to be used. You can collect the data to fit the assumptions of the statistics, or you can change the statistics to compensate for limitations in the data. An understanding of the proper use and application of the statistics will help you plan the experiment and minimize the difficulties in interpreting the results. If you are establishing a standard method validation process in your laboratory, it may be best to put your efforts into collecting the appropriate data—my personal recommendation as the best approach for most healthcare laboratories. This emphasis on getting good data also involves collecting the right specimens under the right conditions, processing those specimens properly, storing the samples appropriately, operating the method or analytical system under representative conditions, and analyzing the patient samples with a process that is under statistical control. This point requires the most care and should have the highest priority. The statistics really do not matter if you do not take care with the data.

In quality management terms, the proper use of statistics is a chronic problem that will continue to flare up until the process is fixed. The process that needs fixing here is the education and training process in clinical chemistry, clinical pathology, and clinical laboratory science. There is a deficiency in a core competency—the ability to use basic statistics in method validation studies as well as for statistical quality control. Correcting this deficiency requires courses for students in undergraduate programs, continuing education workshops and seminars for professionals already in the field, and periodic articles in the scientific literature to remind investigators of the problems and difficulties. There also is a need for easy-to-use statistical software that is designed specifically to deal with the needs and applications in healthcare laboratories. It might even be appropriate for professional organizations such as the AACC or IFCC to support a continuing education curriculum to deal with this ongoing need. With today's Internet technology, basic training courses and improved software tools could be delivered to anyone, anywhere, anytime.

References

1. Stöckl D, Dewitte K, Thienpont M. Validity of linear regression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data? Clin Chem 1998;44:2340–6.
2. Hollis S. Analysis of method comparison studies [Editorial]. Ann Clin Biochem 1996;33:1–4.
3. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
4. Hyltoft Petersen P, Stöckl D, Blaabjerg O, Pedersen B, Birkemose E, Thienpont L, et al. Graphical interpretation of analytical data from comparison of a field method with a reference method by use of difference plots [Opinion]. Clin Chem 1997;43:2039–46.
5. Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method-comparison studies. Clin Chem 1973;19:49–57. [Available in PDF format at http://www.westgard.com/method1.htm, with permission from *Clinical Chemistry*.].
6. Westgard JO, Carey RN, Wold S. Criteria for judging precision and accuracy in method development and evaluation. Clin Chem 1974;20:825–33.
7. Westgard JO. A method evaluation decision chart for judging method performance. Clin Lab Sci 1995;8:277–83.

James O. Westgard
Department of Pathology
and Laboratory Medicine
University of Wisconsin
Medical School
Room D4/237
600 Highland Avenue
Madison, WI 53792
Fax 608-263-1568
E-mail jo.westgard@hosp.wisc.edu