Statistics for Laboratory Method Comparison Studies

Robert T. Magari

In laboratory testing, you may want to use a new analytical method for measuring a chemical substance. To do so, you must be able to prove that the new method gives the same results as the previous one. Not all statistical methods for making laboratory comparisons are equally valid.

Method comparison studies are used to assess the relative agreement between two laboratory analytical methods that measure the same chemical substance, primarily to assess the performance of a newly released method. Laboratory personnel must compare the new method with the one currently in use to see whether their measurements are indeed comparable. The current method of analysis is called the *reference* (the gold standard), and the new method is called the *test;* both methods are known to measure with a certain degree of imprecision. This is different from *calibration* approaches, in which known quantities are measured by the new method, and results are compared. The comparison of test and reference is based on paired data obtained from several independent subjects. Sometimes more than one replicate is obtained for each subject.

Using statistical analysis to determine agreement begins with plotting the data and obtaining a visual assessment of the relationship. Test method data plotted against reference method data should produce a scatter that lies along a straight line (usually called the *equality line*) passing through the origin in a 45° angle. This type of plot is very familiar, but it can be misleading: The greater the range of measurements, the better the agreement will appear to be. Measuring the differences between the two methods for each subject plotted against their means is a better way of assessing the relationship (1). Such a plot clearly shows the pattern of the individual differences and allows you to calculate the limits of agreement and to transform the data if necessary. A scatter of the differences along the zero-line is an indication of agreement between two methods. If the results fall into a bell-type scatter, then the differences are proportional to the means, and a logarithmic transformation is recommended.

Assessing Agreement

With nonsignificant bias between them, the test and reference methods agree. The bias is nonsignificant when the differences between the methods are due only to the amount of measurement imprecision. That imprecision needs to be assessed before any agreement testing.

The presence of bias may impair agreement between the two analytical methods. Sources of biases depend on the type of measurements and other issues, but they are generally classified in two groups: constant bias (*CB*) and proportional bias

Table 1. Simulated data with no bias and with constant bias only

		No Bias (*D* = *IP*)			Constant Bias (*D* = *CB*+ *IP*)		
Subjects	Reference Method	Test Method	*D*	*PB*	Test Method	*D*	*PB*
S1	5.0	5.6	0.62	0	8.6	3.62	0
S2	6.0	5.7	−0.31	0	8.7	2.69	0
S3	7.0	7.3	0.27	0	10.3	3.27	0
S4	8.0	6.8	−1.25	0	9.8	1.75	0
S5	9.0	9.3	0.34	0	12.3	3.34	0
S6	10.0	10.2	0.23	0	13.2	3.23	0
S7	11.0	10.6	−0.40	0	13.6	2.60	0
S8	12.0	12.3	0.35	0	15.3	3.35	0
S9	13.0	13.3	0.30	0	16.3	3.30	0
S10	15.0	15.0	0.00	0	18.0	3.00	0
CB				0			3

Note: *D* = difference between methods, *IP* = imprecision, *CB* = constant bias, *PB* = proportional bias

Robert T. Magari *is a staff scientist at Beckman Coulter Corporation, 11800 SW 147 Avenue, Miami, FL 33116-9015, 305.380.4239, fax 305.380.3699, robert_magari@coulter.com, www.beckmancoulter.com.*

Table 2. Simulated data with proportional bias only and with constant and proportional bias

		Proportional Bias (*D* = *PB* + *IP*)			Constant and Proportional Bias (*D* = *CB* + *PB* + *IP*)		
Subjects	Reference Method	Test Method	*D*	*PB*	Test Method	*D*	*PB*
S1	5.0	7.1	2.12	1.5	10.1	5.12	1.5
S2	6.0	7.0	0.99	1.3	10.0	3.99	1.3
S3	7.0	8.3	1.27	1.0	11.3	4.27	1.0
S4	8.0	7.6	−0.45	0.8	10.6	2.55	0.8
S5	9.0	9.8	0.84	0.5	12.8	3.84	0.5
S6	10.0	10.2	0.23	0.0	13.2	3.23	0.0
S7	11.0	10.1	−0.90	−0.5	14.5	2.55	−0.5
S8	12.0	11.5	−0.45	−0.8	14.5	2.55	−0.8
S9	13.0	12.3	−0.70	−1.0	15.3	2.30	−1.0
S10	15.0	13.5	−1.50	−1.5	16.5	1.50	−1.5
CB				0			3

Note: *D* = difference between methods, *PB* = proportional bias, *IP* = imprecision, *CB* = constant bias

(*PB*). *CB*s are systematic deviations estimated as the average differences between the test and the reference methods. The presence of *CB* indicates that the test method measures consistently higher or lower in comparison with the reference. Estimation and testing of *CB* are based on the average response of the two methods, and a systematic adjustment may correct for its presence. Two analytical methods can agree on average but exhibit biases in particular ranges of measurements. Those are *PB* biases. *PB* is usually related to the range of measurements, but it can be related also to the quality of measurement.

In our model, *PB* is regarded as a significant difference between two methods that exceeds the amount of *CB* for each subject. *PB* can be present for several reasons. The test method may measure lower than the reference in the lower range of measurements and higher than the reference in the upper range. That is an example of linear *PB* that, in practice, can be related to a lack of linearity of measurements. Another reason for the presence of *PB* is that the test method may measure higher (or lower) than the reference in both upper and lower ranges, or it may measure higher or lower in only one extreme range, and agree with the reference elsewhere. Test methods can also be more sensitive to the measuring environment or less discriminative to related substances. Sometimes it is important to interpret *PB* as a source of disagreement between the two methods.

Biases can be present in a data set, but how statistically significant they are is related to the level of precision. The same amount of bias between a test and a reference can be considered as statistically significant for highly precise methods of measurement (with a small statistical error), but insignificant for some less precise methods (with a large statistical error). The observed difference (*D*) between the two methods for each subject can be expressed in terms of biases and imprecision (*IP*) as

$$D = CB + PB + IP.$$

The two methods will agree with each other when $D = IP$.

Simulated data in Tables 1 and 2 show how the presence of biases affects agreement between two methods in ten subjects. Small *D*s (due to imprecision only) are observed when no bias is present. The values of *D* have no pattern and tend to average to zero for a large data set because imprecision errors are considered to be random and normally distributed with a mean of zero (Table 1). The presence of *CB* increases all the test method values by three (*CB*=3). The values of Tables 1 and 2 are also presented as difference plot (X-axis, means of test and reference method, and Y-axis, *D*) in Figure 1. The test method shows a linear *PB* by measuring higher in the lower range and lower in the higher range in comparison with the reference method.

Estimating Bias

A model for two methods randomly measuring a number of independent subjects in replicates can be written as,

$$Y = \mu + \alpha + \beta + \gamma + \epsilon,$$

where *Y* is the observed measurement (the response variable), μ is the overall mean, α is the subject parameter, β is the method parameter, and γ is the method-by-subject interaction. Because subjects are randomly selected to represent a population of interest, the subjects constitute a random component in our model, whereas the methods are considered as fixed. The errors (ϵ) are also random and assumed to be identically, normally distributed with the same expectation $E[\epsilon] = 0$ and variance $\mathrm{Var}[\epsilon] = \sigma^2_e$, over subjects and methods. The variance in the estimate of error represents the run-to-run variability and is an estimate of precision.

The expected difference between two methods for each subject is

$$E(D) = \mu_{iT} - \mu_{iR},$$

where μ_{iT} is the expected average across replicates of the test method for subject *i* ($1 \leqslant i \leqslant n$, *n* is the number of subjects), and μ_{iR} is the expected average across replicates of the reference method for the same subject.

Therefore, the expected difference for the two methods can be expressed as,

$$E(D) = \mu + \alpha_i + \beta_T + \gamma_{iT} - \mu - \alpha_i - \beta_R - \gamma_{iR},$$

where β_T and β_R are the parameters for test and reference methods respectively. Because βs are subject to the restriction

$$\beta_T + \beta_R = 0$$

(full rank reparameterization),

$$E(D) = 2\beta_1 + (\gamma_{iT} - \gamma_{iR}),$$

and

$$E(D) = CB + PB.$$

PB is a set of *n* parameters obtained as $\gamma_{iT} - \gamma_{iR}$ for each subject. Thus, the expected difference is composed of an average difference between methods (*CB*) and a specific difference for that measurement (*PB*), whereas the observed difference (*D*) has also an error term that is related to imprecision.

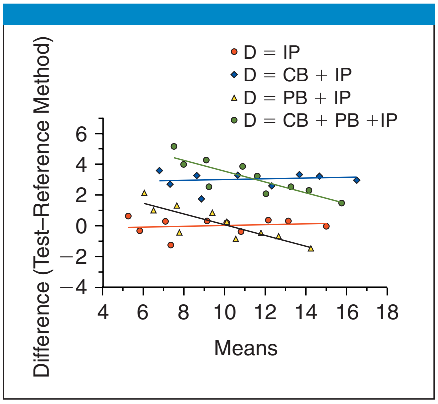

Figure 1. The effects of biases on agreement based on simulated data

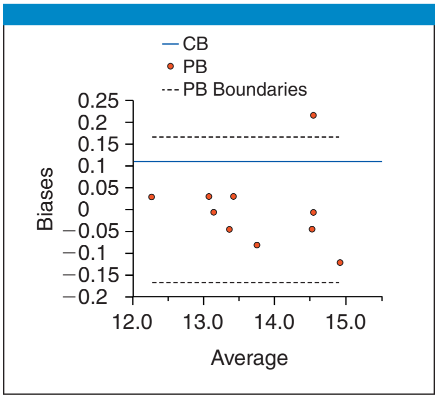

Figure 2. Constant and proportional biases plotted against the average of the two methods

The implementation of that model has been written about elsewhere (2,3).

The results from an experiment comparing hemoglobin measured by two methods in ten subjects are analyzed and presented in Table 3 using the process described. The average difference between two methods (*CB* = 0.11) is not statistically different from zero at 99% confidence (P-value = 0.0191). Overall PB is also not significant (P-value = 0.0594). *PB* is also partitioned for each individual subjects. Subject 10 shows a significant *PB* (P-value = 0.0013); none of the other individual *PB*s is statistically significant. *CB*, along with *PB*, and their 99% confidence boundaries are plotted against the averages of the two methods for each subject in Figure 2. Only subject 10 is outside the boundaries. The results indicate that the two methods for measuring hemoglobin agree with each other because no significant biases exist between them.

Regression Approach

Linear regression is probably the most popular approach in method comparison studies. Based on that approach, the regression of test method to reference method should yield a straight line nonsignificantly different from the equality line. The equality line is determined by two parameters: slope=1 and intercept=0. Deviation from the equality line indicates a lack of agreement between the two methods. Because error is always present in the statistical tests for regression, the parameters provide a test for agreement between the methods.

The least-squares technique is usually used to fit the regression data and to estimate the parameters. The least-squares linear regression line through a set of points is defined as the straight line that minimizes the squared deviations of the observed test method data from the regression line, when those deviations are drawn parallel to the Y-axis. The estimates of slope and intercept and their standard errors can be obtained from a computer program or calculated as shown in many basic statistical books. Student t-test can be used to test for slope=1 and intercept=0 (some software packages do not provide a test for slope=1). The calculated value of the t-test in this case is:

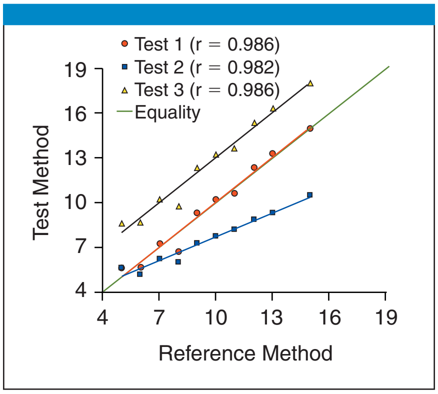

Figure 3. Correlation coefficient fails to measure agreement between the two methods

$$t = \frac{\beta - 1,}{S_{\beta}}$$

where β is the estimated slope and S_b is the standard error of the estimate. This value follows a t-distribution with $n-2$ degrees of freedom.

Other regression approaches. A concern when using the least-squares approach is that both the reference and the test methods contain error (they are both measured with some degree of imprecision), which violates one of the statistical assumptions in regression analysis that states that the independent variable (the reference method) should be measured without error. That assumption is necessary to obtain unbiased estimates of the slope and intercept. Several authors have concluded that it is, therefore, preferable to use alternatives to least squares approach (4,5).

One alternative is to use the Deming regression, which provides estimates of the slopes and intercept that are proven to be orthogonal least-squares estimates. The orthogonal least-squares technique minimizes the squared deviations of the observed test method data from the regression line (when those deviations are drawn perpendicular to the regression line). Thus, it provides unbiased regression estimates when both methods are measured with error.

Passing and Bablok have proposed a linear regression procedure with no special assumptions regarding the distribution of the data (5). This nonparametric method is based on ranking the observations so is computationally intensive. The result is independent of the assignment of the reference method as *X* (the independent variable) and the test method as *Y* (the dependent variable). Special software

Table 3. Estimates of the constant and proportional biases from hemoglobin experiments

Bias Parameter	Estimate	Standard Error	P-Value
Constant Bias	0.1100	0.03859	0.0191
Proportional Bias			
Overall	0.0111	0.00708	0.0594
Subject 1	0.0297	0.06056	0.6274
Subject 2	0.0297	0.06056	0.6274
Subject 3	−0.0074	0.06056	0.9032
Subject 4	−0.0445	0.06056	0.4676
Subject 5	−0.0074	0.06056	0.9032
Subject 6	0.0297	0.06056	0.6274
Subject 7	−0.0816	0.06056	0.1875
Subject 8	−0.0445	0.06056	0.4676
Subject 9	−0.1188	0.06056	0.0591
Subject 10	0.2153	0.06056	0.0013

Table 4. The estimates of agreement based on the regression approach for assessing agreement between two methods

	Least Squares			Deming			Passing and Bablok		
Parameters	**Estimate**	**SD Error**	**P-Value**	**Estimate**	**SD Error**	**P-Value**	**Estimate**	**Lower**	**Upper**
Slope	0.978	0.050	0.664	0.988	0.035	0.734	0.93	0.78	1.12
Intercept	0.417	0.682	0.558	0.281	0.487	0.580	1.03	−1.46	3.09
Residual SE	0.128			0.091					

packages are needed to compute either the Deming or the Passing and Bablok regressions (4,5).

Regression estimates using least squares, Deming, and Passing and Bablok approaches for hemoglobin data are given in Table 4. The values from the Deming and least-squares estimates are close to each other. Smaller residual standard errors are usually obtained by the Deming regression. All p-values are greater than the significant level ($\alpha = 0.01$) indicating that the two methods agree with 99% confidence. The Passing and Bablok approach provides lower and upper boundaries to the estimates of the slope and intercept. The two methods agree if 1 and 0 are within the boundaries for the slope and intercept respectively.

Regression parameters are not always related to inaccuracy biases, and the results cannot be interpreted for practical purposes. A slope that tests statistically different from 1 indicates the presence of linear *PB*. The value of intercept indicates a vertical displacement of the regression line, which can be related to *CB* if slope is not statistically different from one. The intercept has no practical interpretation when the slope is statistically different from one. The estimate of residual error depends on the data fit and is not a uniform minimum unbiased estimate. When nonlinear *PB* is present in the data, the portion of it uncaptured by the linear trend will inflate the residual error.

Correlation Coefficient

Sometimes when the results of the two methods have high correlation coefficient values it is interpreted as an indicator of agreement. However, high correlation does not necessarily mean agreement between methods. The correlation coefficient measures the strength of the relationship, and it is incorrect to interpret it as a measure of agreement. The methods agree when their scatter lies along the equality line, but high correlation can be obtained if the scatter lies along any straight line. The presence of constant bias does not change the value of the correlation coefficient either.

Simulated data for three test methods compared with the same reference method are presented in Figure 3. The three test methods have very high correlation with the reference, but only Test 1 agrees with the reference. Test 2 does not lie along the equality line, whereas Test 3 exhibited a *CB* in comparison with the reference. The use of the correlation coefficient to assess agreement in this case would be misleading. Statistical testing of the correlation coefficient is irrelevant to the question of agreement because two methods designed to measure the same chemical substance are expected to be related. The values of the correlation depend on the range of the data and also on the precision of measurements.

Concordance coefficient (6) and the gold-standard correlation (7) are two improved versions of the correlation coefficient.

Concordance coefficient indicates the strength of relationship between the two readings that fall on the 45° line through the origin and is calculated as

$$r_c = \frac{2S_{TR}}{S^2_T + S^2_R + \left(T - R\right)^2}$$

where S_{TR} is the covariance between the test and the reference method, $S^2{}_T$ is the variance of the test, $S^2{}_R$ is the variance of the reference, and *T* and *R* are the means of the test and reference methods respectively.

Gold-standard correlation is based on a constrained bivariate distribution and is calculated as

$$r_G = \frac{1}{1 + \dfrac{S_D}{S^2_R\left(n - 1\right)}}$$

where $S_D = \Sigma D^2$. Table 5 shows the values of the correlation coefficient r_C and r_G. Both r_C and r_G values are smaller than the correlation coefficient for Test 2 and Test 3 indicating a lack of agreement between the methods.

Paired t-Test

Sometimes a paired t-test is used to test the difference between the means of the two methods. Mean comparison may not the best way for comparing two methods because the means may agree on the average, but may not agree for specific ranges. Only CB can be detected by mean comparison, and that result can be misleading if a linear PB is present. Mean comparison by analysis of variance is performed for data with complex structures, and when several sources of variability are present.

Practical Considerations

Statistical methods for assessing agreement depend on the way data are collected. The quality of the analytical input data is crucial for interpretation of the method comparison study. Sometimes more than one comparison method is necessary to achieve a good understanding of the relationship between the methods. Subjects should be randomly chosen, and they should represent the intended reference range of the application. The number of subjects should be established before conducting the experiment to achieve the desired statistical power.

Table 5. The estimates of correlation coefficient, concordance, and gold-standard correlation for assessing agreement between a reference method and three different test methods

TM[a]	*CC*[b]	(r_C)[c]	(r_G)[d]
Test 1	0.986	0.986	0.986
Test 2	0.986	0.602	0.766
Test 3	0.982	0.666	0.705

[a] test method [b] correlation coefficient
[c] concordance [d] gold-standard correlation

A statistically significant bias may not always be clinically important, so tests for predetermined clinically important differences may be performed. However, a sound clinical significant difference should exist beforehand, in these cases, and the test for it should be suggested by the data and by experience.

References

(1) D.G. Altman and J.M. Bland, “Measurement in Medicine: The Analysis of Method Comparison Studies,” *Statistician* 32, 307–317 (1983).
(2) R.T. Magari, “A Statistical Approach for Hematology Comparison Studies,” *Lab. Hematol.* 4,199–203 (1998).
(3) R.T. Magari, “Evaluating Agreement Between Two Analytical Methods in Clinical Chemistry,” *Clin. Chem. Lab. Med.* 38, 1021–1025 (2000).
(4) K. Linnet, “Evaluation of Regression Procedures for Method Comparison Studies,” *Clin. Chem.* 39, 424–432 (1993).
(5) H. Passing and W. Bablok, “A New Biometrical Method Procedure for Testing the Equality of Measurements from Two Different Analytical Methods,” *J. Clin. Chem. Clin. Biochem.* 21, 709–720 (1983).
(6) L.K. Lin, “A Concordance Correlation Coefficient to Evaluate Reproducibility,” *Biometrics* 45, 255–268 (1989).
(7) R.T. St. Laurent, “Evaluating Agreement with a Gold Standard in Method Comparison Studies,” *Biometrics* 54, 537–545 (1998). **BP**