Methods for evaluating the agreement between diagnostic tests

Charity J. Morgan, PhD,[a] and Inmaculada Aban, PhD[a]

[a] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL

Received May 3, 2015; accepted May 3, 2015
doi:10.1007/s12350-015-0175-7

See related article, pp. 499–510

During the course of the development of a new diagnostic test, it is often necessary to compare the performance of the new test to that of an existing method. When the tests in question produce qualitative results (e.g., a test that indicates the presence or absence of a disease), the use of measures such as sensitivity/specificity or percent agreement is well established. For tests that lead to quantitative results, different methods are needed. The paper by Dunet et al published in this issue provides an example of an application of some of these methods.[1]

At the heart of this issue is quantifying the agreement between the results of two (or more) tests. That is, the two tests should yield similar results when applied to the same subject. Here, we consider the setting where we have a sample of subjects tested using both tests. A natural starting point in assessing agreement between quantitative results would be to consider the differences between the test results for each subject. While the paired t test could then be used to test whether the mean difference significantly differs from zero, this test cannot provide evidence that there is agreement. That is, rejecting the null hypothesis of no difference between the two sets of test results would only allow us to say that the tests do not agree; failing to reject this hypothesis would not constitute proof that the tests agree.

While we cannot ''prove the null'' of no difference between the test results, we can, however, use equivalence testing to determine whether the average difference between test results is small enough to be considered (clinically) insignificant. Bland and Altman's limits of agreement (LOA) approach the problem in this way, by providing an estimate of a range in which 95% of differences between test results are expected to fall (assuming those differences are approximately normally distributed).[2,3] The LOA are calculated as $\bar{d} \pm 1.96 \cdot s_d$, where $\bar{d}$ is the sample mean of the differences and s_d is the sample standard deviation. If the range of the LOA includes what would be considered clinically significant differences, this result would suggest that agreement between the tests is not satisfactory. The LOA are also often represented graphically by plotting the average result for each subject against the difference between those results. Figure 1 is an illustration of this tool based on a hypothetical example. Bland and Altman caution that the LOA are only a meaningful result if the mean and variance of the differences between test results are constant across the range of the test results.[3] In other words, the LOA should not be used if the agreement between tests varies with the quantity being measured. Such a situation might arise if the tests yield similar results for subjects whose test results fall within a normal range, but have poor agreement for subjects outside of that range.

Another method of visually assessing whether two tests are in agreement is by constructing a scatterplot of the results from the first test against the results from the second test. If the two tests have good agreement, we should expect the points to fall on or near the 45° (i.e., $y = x$) line; departures from this line would indicate poor agreement. Although the Pearson correlation coefficient, ρ, may be used to assess the strength of a linear relationship between results of two tests, the Pearson correlation is not an appropriate means of assessing agreement: while it is true that the results from tests that agree well will have a high Pearson correlation, the converse is not always true as will be illustrated later.

An alternative to the Pearson correlation more suited for comparing diagnostic tests is the intraclass correlation coefficient (ICC). It was first proposed by Fisher[4] and is defined by assuming that results of the

Reprint requests: Charity J. Morgan, PhD, Department of Biostatistics, University of Alabama at Birmingham, 1720 Second Avenue South, Birmingham, AL, 35294-0022; *cjmorgan@uab.edu*
J Nucl Cardiol 2016;23:511–3.
1071-3581/$34.00
Copyright © 2015 American Society of Nuclear Cardiology.

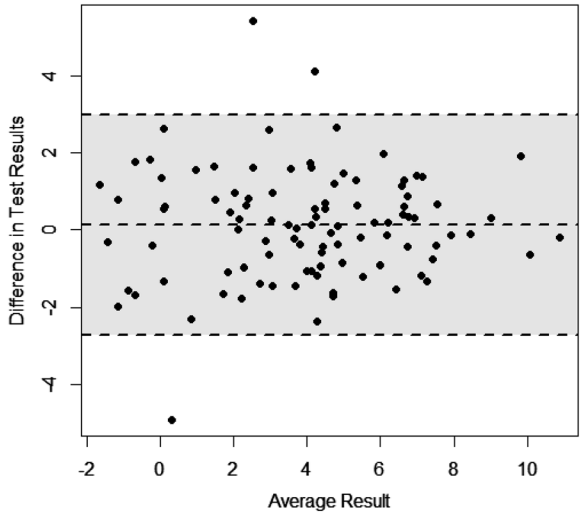

Figure 1. Example of a Bland–Altman plot. Limits of agreement are represented by the *shaded area*.

diagnostic tests follow a one-way ANOVA model with a random effect for subject. This random effect accounts for the repeated measures for each subject. The ICC is defined as the ratio of the between-subject variance (σ_α^2) to the total variance, which is composed of the between-subject variance and the within-subject variance (σ_e^2):

$$\mathrm{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}.$$

The ICC ranges from 0 (no agreement) to 1 (perfect agreement).

The ICC has since been extended for use in a variety of settings. Bartko[5] proposed using a two-way ANOVA model to account for rater effects, which can be either fixed (for a finite set of raters) or random (useful when the raters are selected from a larger pool). The model can also be extended to accommodate replicated measurements and/or subject by rater interactions. Different versions of the ICC have been defined for each of these varying ANOVA models. Shrout and Fleiss[6] provide a useful discussion of the different forms of the ICC (see also McGraw and Wong[7,8]).

Finally, the concordance correlation coefficient (CCC) proposed by Lin[9] is designed to assess the agreement between two measures without assuming an underlying ANOVA model. The CCC is defined as

$$\mathrm{CCC} = \frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \cdot \rho = C_\mathrm{b} \cdot \rho,$$

where μ_j and σ_j^2 are the mean and variance of the jth test. Note that C_b depends, in part, on the ''bias'' when the interest is to estimate the difference between the means of the two tests, i.e, $\mu_1 - \mu_2$. C_b is also referred to as the ''bias correction factor.''[9] The CCC can thus be conceptualized as the product of a measure of consistency

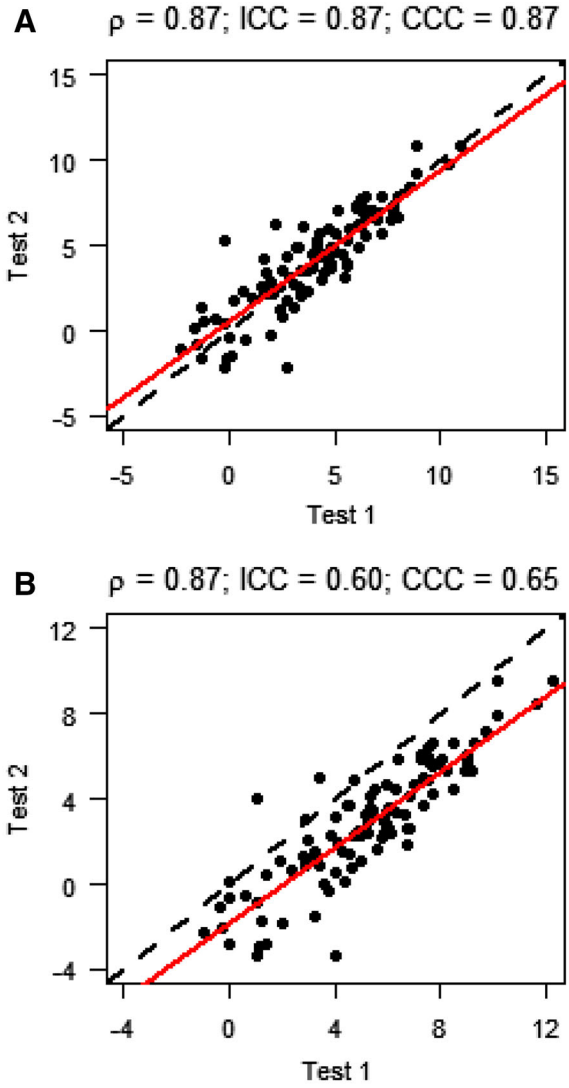

Figure 2. Scatterplots and correlation values for two tests.

(i.e., Pearson's correlation coefficient) and a measure of bias. That is, the CCC quantifies not only how closely observations fall on the regression line (through ρ), but also how close that regression line is to the 45° line of perfect agreement (via C_b).

Note that if one of the tests is a reference or ''gold standard,'' then the bias is based on the difference between the new test's result and the ''true value'' of the quantity being measured, and hence a measure of *accuracy*.[10] For these cases, the CCC can be said to measure accuracy as well as consistency. But when neither test is a gold standard, it is not appropriate to state that CCC also provides a measure of accuracy.

As with the ICC, the CCC can be modified to handle replications or repeated measurements.[11-13] Carrasco and Jover[14] have shown that, in the case of no replications, the CCC is identical to an ICC defined using a

two-way ANOVA model rather than a one-way ANOVA (see also Nickerson[15]). A detailed accounting of how the different versions of the CCC and ICC may be found in Chen and Barnhart.[12,13,16]

Scatterplots of simulated data based on two scenarios are displayed in Figure 2 to further illustrate the differences among these correlations. For the first scenario (see Figure 2A), the fitted regression line (solid red) falls close to the 45° line (dashed black), demonstrating that the tests tend to give similar results. The Pearson correlation coefficient (ρ), ICC, and CCC are all large, further indicating good agreement. The second scenario (see Figure 2B) shows that results from Test 2 are consistently greater than results for Test 1. The ICC and CCC are decreased, indicating poor agreement. The Pearson correlation coefficient is unchanged, as it is unable to detect this departure from the 45° line.

In their paper, Dunet et al[1] use both Bland and Altman's limits of agreement and Lin's concordance correlation coefficient to assess the agreement between software packages. These two methods provide complementary pieces of information. The limits of agreement are useful for determining, when test results differ, whether those differences are likely to be clinically significant; use of the CCC yields a concise summary of the consistency and bias.

Disclosure

The authors have no conflicts of interest to disclose.

References

1. Dunet V, Klein R, Allenbach G, Renaud J, deKamp R, Prior J. Myocardial blood flow quantification by Rb-82 cardiac PET/CT: A detailed reproducibility study between two semi-automatic analysis programs. J Nucl Cardiol. 2015. doi:10.1007/s12350-015-0151-2.
2. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
3. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-60.
4. Fisher RA. Statistical methods for research workers. London: Oliver and Boyd; 1925.
5. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19:3-11.
6. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
7. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30-46.
8. McGraw KO, Wong SP. Correction to McGraw and Wong (1996). Psychol Methods 1996;1:390.
9. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255-68.
10. US Food and Drug Administration. Guidance for industry: Bioanalytical method validation; 2001. http://www.fda.gov/downloads/Drugs/Guidances/ucm070107.pdf.
11. King TS, Chinchilli VM, Carrasco J. A repeated measures concordance correlation coefficient. Stat Med 2007;26:3095-113.
12. Chen CC, Barnhart HX. Assessing agreement with repeated measures for random observers. Stat Med 2011;30:3546-59.
13. Chen CC, Barnhart HX. Assessing agreement with intraclass correlation coefficient and concordance correlation coefficient for data with repeated measures. Comput Stat Data Anal 2013;60:132-45.
14. Carrasco J, Jover L. Estimating the generalized concordance correlation coefficient through variance components. Biometrics 2003;59:849-58.
15. Nickerson CAE. A note on ''A Concordance Correlation Coefficient to Evaluate Reproducibility''. Biometrics 1997;53:1503-7.
16. Chen CC, Barnhart HX. Comparison of ICC and CCC for assessing agreement for data without and with replications. Comput Stat Data Anal 2008;53:554-64.