A CRITICAL ASSESSMENT OF THE TECHNIQUES USED TO DETERMINE AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT

Madeline J. Boyle-Whitesel, H. James Norton, and William Anderson, Carolinas Medical Center
Madeline J. Boyle-Whitesel, Department of Biostatistics, ROB Rm. 312, Carolinas Medical Center, PO Box 32861, Charlotte, NC 28323

Key Words: Agreement; Concordance; Bland and Altman

INTRODUCTION

The problem of assessing agreement for quantitative variables in medical studies is frequently addressed, although correct statistical analysis for determining if agreement truly exists is rare. Often, linear regression and Pearson's correlation coefficient (r) are used, but these methods do not fully address the issue of agreement. "Pearson's r only measures the association between two variables, and does not provide information about agreement. For example, we may observe a correlation coefficient of nearly 1 when one measure is approximately twice a second measure" (Muller and Buttner, 1994). Paired t-tests are also a common statistical method used to determine if agreement exists, but they can be misleading when attempting to confirm agreement between two quantitative variables. They will, however, reveal if any bias is evident within the variables. A more appropriate technique can be used in conjunction with these statistical tests to establish if agreement exists between two variables. The Bland and Altman technique, as it is known, is a statistical method for assessing agreement, based on graphical techniques and simple calculations (Bland and Altman, 1986).

First, we will give examples to display the various problems that can arise while attempting to assess agreement using certain statistical tests. Next, we will discuss the use of the Bland and Altman technique, in combination with other statistical tests, for assessing concordance. Finally, we will summarize the statistical tests used by authors of some papers addressing the agreement issue.

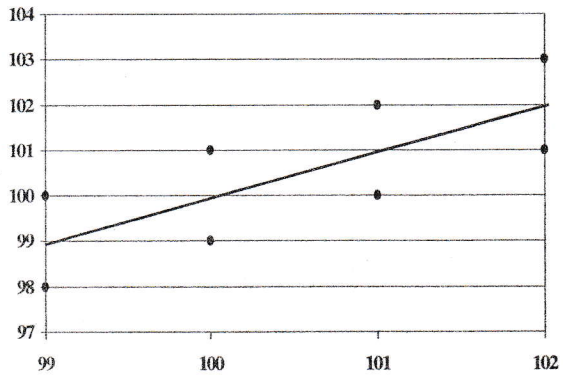
SAMPLE DATA

This section provides sample data to illustrate problems that may arise when using certain statistical tests to assess the agreement between two methods of clinical measurement. In the examples that follow, we shall suppose that a clinician is interested in determining if an oral temperature taken on a child agrees with a tympanic temperature taken on the same child. Both measurements will be assumed to have been taken within a reasonable amount of time of each other.

The A data in Table 1 has a Pearson's correlation coefficient of 1.00 and the p-value from the paired t-test is 1.00, indicating that the difference between the two thermometers is not statistically different from zero. However, as you can see from the data, one thermometer is higher one-half of the time and the other is higher the remainder. Obviously these thermometers do not agree; but if the clinician relied only upon the results from the correlation coefficient and the paired t-test, they may have mistakenly concluded strong agreement.

In Table 1 we see that data B shows consistently higher temperatures from thermometer 2 than those recorded from thermometer 1, a clear case of non-agreement. The correlation coefficient indicates a strong association between the measurements of the two thermometers (r =0.949), but the results from the paired t-test show that there is indeed systematic bias (p=0.0001). Here the clinician would only be mislead if the correlation coefficient was the only statistical technique used to evaluate the agreement between the measurements of the two thermometers.

Figure 1: Plot of Data C



The regression equation from data C has a slope of 1 and an intercept of 0 with a correlation coefficient of 0.745. The regression line is the unity line, which is expected if two measurements agree, but the lack of perfect correlation indicates that the two thermometers do not agree. If one were to look at a plot of the data (see Figure 1), it would be obvious that the measurements do not agree, but do fall on either side of the unity line, thus the regression results.

The final example is from data D in Table 1, which is the same as data C, with an additional observation recorded. Although a temperature of 106 degrees is unreasonable, the observation is recorded simply to illustrate a point. With the outlier included in this data set, the regression line remained unchanged from before, with a slope of 1 and an intercept of 0. However, the correlation coefficient is now 0.906, which shows a much stronger association between the measurements of the two thermometers than before. This may lead the clinician to incorrectly conclude that the measurements from the two thermometers agree, but on further observation of the data this is not the case. The points are again lying on either side of the unity line, with one outlier lying on the unity line.

These examples are indicative of the problems that clinicians may encounter as they attempt to interpret the results of such statistical techniques. All of the above techniques fail to indicate if the two clinical measurements are equivalent; further analysis of the data is needed to determine if agreement exists between these measurements.

TECHNIQUES TO ADDRESS AGREEMENT

Bland and Altman (1986) suggest using plots as a method of determining if agreement exists between two quantitative variables. This approach does not give one single measure of agreement, but instead allows for interpretation of the possible agreement between the measurements. The obvious first step, which according to Bland and Altman should be mandatory, is to plot the data; this plot can include a regression and unity line. However, "it is preferable to plot the difference between the methods (A-B) against (A+B)/2, the average", because "...it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers and see whether there is any trend, for example an increase in (A-B) for high values" (Bland and Altman, 1986). Another informative plot is a histogram of the difference between the methods (A-B), which allows the clinician to see the magnitude of disagreement.

A modified Bland and Altman graph can be used when one of the methods under study is the gold standard, or reference. Plotting the difference of the measurements (A-B) versus the gold standard can facilitate the clinician's comprehension of the results. However, if a gold standard is not being studied, then using one of the methods as such in the plot may be misleading and misrepresent the true results. The relationship between initial blood pressure prior to treatment, and the fall in pressure after treatment, has been examined and shown to be positively correlated, even when random numbers were substituted for the initial blood pressures (Gill, 1985). "There is consequently a logical and mathematical relation between the change in blood pressure and pre-treatment blood pressure which is not attributable to any treatment effect" (Gill, 1985). Oldham suggested that a better correlation would be between the change of blood pressure and the mean of the pre-treatment and post-treatment measurements; this correlation would avoid any "spurious associations" (Gill, 1985). For this reason, Bland and Altman suggest plotting the difference in the measurements of the two methods versus the mean measurements, because the mean will give a better estimate of the true unknown value. However, if one method is already established as the gold standard, then it follows that this measurement is the best estimate of the true value, and it would be appropriate to use this method in the graph.

The intraclass correlation coefficient can be used to measure agreement between two continuous variables when it is impossible to designate one variable as X and the other as Y (Zar, 1996). In these situations the comparison of two methods is to determine if the methods are consistent and therefore interchangeable; higher positive correlations indicate that the two methods have increasing agreement. Intraclass correlations account for correlations as well as systematic bias, which makes them preferable over the standard correlation coefficients (Lowenstein, 1993). This technique can be applied to determining intraobserver agreement, as well as to 3 or more methods or observers with a modest increase in complexity, but those procedures are beyond the scope of this paper (Zar, 1996).

For determining concordance among procedures that have a well-defined gold standard or reference, sensitivity and specificity can be used. "Sensitivity is the percent agreement of the test with the standard under the condition that the standard is positive (the disease is present), whereas specificity represents the percent agreement when the standard is negative (the disease is absent)" (Kramer and Feinstein, 1981). Although these

techniques do not correct for the agreement expected by chance, they could be used in conjunction with other techniques to determine concordance.

MEDICAL EXAMPLE

The agreement relationship between the arterial and venous measurements of three variables was analyzed using correlation coefficients, linear regression, paired t-tests, and graphs of the differences by the averages. The measurements were taken from rabbits at baseline and after a treatment was given.

Figures 2 and 3 display Bland and Altman type graphs where the average measurement is on the x-axis and the difference between the measurements is on the y-axis. Figure 2 shows the treatment data for carbon dioxide (CO_2) and figure 3 shows treatment data for base deficit (BD). In figure 2 we can clearly see an example of non-agreement, since the venous measurements of CO_2 are consistently higher than the arterial measurements. The graph in figure 3, however, cannot be interpreted as clearly. In this case the clinician should decide if the amount of the disagreement seen in the graph is acceptable for the measurement of BD.

The p-value from the paired t-test for carbon dioxide was statistically significant; while for BD it was not significant. The correlation coefficient from the regression for CO_2 was 0.39 (p=0.093), and for BD it was 0.62 (p=0.0039). These results, coupled with the graphs, show the agreement between the arterial and venous measurements for BD, and the lack of agreement for CO_2.

AGREEMENT IN LITERATURE

Twenty-six articles selected from a MEDLINE search (references for these available upon request), addressing agreement between two clinical measurements, were read. The most common techniques used to determine agreement were recorded as absent or present for each article; the results of these counts are given in table 2.

Of the twenty-six papers read, only nine (35%) discussed or even mentioned agreement. Even fewer papers (12%) used a Bland and Altman graph to represent the amount of agreement, or lack thereof, in their data. The majority of papers used correlation coefficients and/or paired t-tests to assess the agreement in their data.

Very few authors seemed to understand the issue of concordance, and most of those that mentioned or discussed agreement did so improperly or incompletely. Interpretations from these papers must be closely scrutinized, since the conclusions the authors derived from their data may be incorrect or inconclusive. Most of the papers simply did not perform the proper analyses to determine if the methods are interchangeable or not.

CONCLUSIONS

Comparisons of two clinical measurements are often performed in medical studies but analyzed inappropriately. This has been illustrated by showing some problems that may arise with the most frequently used techniques for determining if two measurements agree. As seen in this literature search of twenty-six published articles, most often agreement is incorrectly assessed. Therefore, interpretations and conclusions from these types of articles must be carefully examined.

REFERENCES

Bland, J.M., and Altman, D.G. (1986), "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement," *The Lancet*, February 8.

Gill, J.S., Beevers, D.G., Zezulka, A.V., and Davies, P. (1985), "Relation Between Initial Blood Pressure and its Fall with Treatment," *The Lancet*, March 9.

Kramer, M.S., and Feinstein, A.R. (1981), "Clinical Biostatistics: The Biostatistics of Concordance," *Clinical Pharmacology and Therapeutics*, 29 (1), 111-123.

Lowenstein, S.R., Koziol-McLain, J., and Badgett, R.G. (1993), "Concordance Versus Correlation," *Annals of Emergency Medicine*", 22, 269.

Muller, R., and Buttner, P. (1984), "A Critical Discussion of Intraclass Correlation Coefficients," *Statistics in Medicine*, 13, 2465-2476.

Zar, J.H. (1996), Biostatistical Analysis 3rd Edition, New Jersey: Prentice-Hall, Inc.

Table 1: Results from Measurements of Thermometer One and Two

A Data		B Data		C Data		D Data	
1	2	1	2	1	2	1	2
99	103	100	103	99	100	99	100
99	103	99	101	99	98	99	98
99	103	98	101	100	101	100	101
99	103	97	99	100	99	100	99
103	99	101	104	101	102	101	102
103	99	100	102	101	100	101	100
103	99	98	100	102	103	102	103
103	99	97	100	102	101	102	101
						106	106

Table 2: Counts of Statistical Techniques Used to Assess Agreement

Technique	Number of Times Used
Paired t-test	10
Correlation	17
Graphs of Data with Regression Line	5
Graphs of Data with Unity Line	2
Graphs of Data with both Lines	4
Slope and Intercept Reported	8
Histograms of Differences	4
Bland and Altman Graph	3
Sensitivity and Specificity Reported	8
Agreement Mentioned or Discussed	9

Figures 2 and 3