Journal of Clinical Monitoring and Computing (2008) 22:257–259
DOI: 10.1007/s10877-008-9127-y

A MEASURE OF CONFIDENCE IN BLAND–ALTMAN ANALYSIS FOR THE INTERCHANGEABILITY OF TWO METHODS OF MEASUREMENT

David Preiss, PhD MASc and Joseph Fisher, MD FRCP(C)

Preiss D, Fisher J. A measure of confidence in Bland-Altman analysis for the interchangeability of two methods of measurement.
J Clin Monit Comput 2008; 22:257–259

ABSTRACT. Bland-Altman (B–A) analysis has largely replaced the correlation coefficient as the predominant tool for evaluating the interchangeability of two methods of clinical measurement. However, we contend that B–A analysis might lead to erroneous conclusions when the data range is small. We provide an example to illustrate this and explore a possible analysis technique to address this limitation.

KEY WORDS. Bland-Altman, agreement, correlation.

INTRODUCTION

In order to determine if a new method of measurement is equivalent to one that is currently accepted, a ''comparison'' study must be conducted. This involves making simultaneous measurements of the same biological, chemical or physical parameter and comparing the results obtained from both the new and experimental methods. The question of how to compare them has been a subject of controversy for some time [1, 2]. For many years, investigators had used the correlation coefficient to determine whether the methods of measurement can be used interchangeably. However, some authors have pointed out that correlation is a measure of association rather than agreement and assigning a threshold value for widespread acceptance is difficult [3, 4]. 'Bland-Altman' (B–A) analysis has become widely applied in the determination of the extent of agreement between two methods [5] leaving the threshold for interchangeability to the investigator. B–A analysis consists of the calculation of both the average difference between each pair of measurements (bias), and the standard deviation (SD) of the differences (precision). In 'clinically acceptable' conditions for interchangeability of two methods, both values will be low, i.e., a small bias and a small SD of the differences.

Despite its popularity, there remains reluctance to rely completely and exclusively on B–A analysis as indicated by the persistence of correlation reports in the literature for these types of studies. This might be, as Bland and Altman suggest, unwillingness to kick 'old habits' [4], but it might also reflect that B–A analysis alone does not provide sufficient confidence in assessing interchangeability of methods. We wish to explore a potential limitation of B–A analysis and suggest an approach to address it.

From the Department of Anesthesia, Toronto General Hospital, University Health Network, 200 Elizabeth Street 7EN-242, Toronto, ON, Canada M5G 2C4.

Received 17 March 2008. Accepted for publication 2 May 2008.

Address correspondence to D. Preiss, Department of Anesthesia, Toronto General Hospital, University Health Network, 200 Elizabeth Street 7EN-242, Toronto, ON, Canada M5G 2C4. E-mail: david.preiss@utoronto.ca

ILLUSTRATIVE PROBLEM

Consider the following illustrations regarding the comparison of a new experimental method (E) of measuring cardiac output to the 'gold-standard' thermodilution technique (T) in 50 subjects. Figure 1a and b shows the B–A analysis from two nearly identical studies. The only difference was that Study #1 reflects data from 'healthy subjects sitting comfortably', whereas Study #2 reflects data from 'patients with heart failure performing some physical activities'. The average difference between each pair of measurements – the bias – is very similar in both cases (0.03 and 0.11 l/min for Studies #1 and #2, respectively). The precision calculated from both experiments (0.93 and 0.96 l/min, respectively) are equally reassuring about the interchangeability of T and E, according to predetermined, field specific standards [6]. Of note, the *R* values in these studies are quite different (0.02 in Study #1 and 0.90 in Study #2) but as stated earlier, correlation cannot be used to evaluate agreement.

In these examples, each data point collected using method T was paired with its corresponding, simultaneously-made measurement using the new method E. Consider, however, what would happen if we intentionally mispaired the T and E measurements in each study. Remarkably, the newly calculated precision for Study #1 may still show support of the interchangeability of the methods! This effect arises from the fact that the range of the data is similar to that of the accepted precision for cardiac output measurements. Thus, no matter how many ways the pairing is rearranged – in other words, if there was no association between the two methods at all – the precision may still remain in the clinically acceptable range. If the data from Study #1 are repeatedly mispaired 1,000 times, the probability distribution of precision shown in Figure 2a is produced.

However, the same is not true for Study #2. If the data are randomly mispaired in this case, the precision increases (a bad thing, vide supra) indicating that the original precision calculated from the properly paired data is statistically unlikely, i.e., only occurs in a small fraction of deliberately 'mispaired' combinations.

DISCUSSION

We used this technique of mismatching the pairings to illustrate that the reliability of the calculated precision is based partially on the range of the data collected. If the data are collected over a small range, then *any pairing* of data points collected from T or E will produce a clinically acceptable precision and applying B–A is unenlightening.

One of three conclusions can be drawn from Study #1. The first is that the data range was too narrow for B–A analysis to be helpful in assessing interchangeability of methods. The second is that the clinically acceptable precision must be reduced to a smaller, stricter value in order to strengthen the conclusion from the analysis. The final conclusion is that the two methods do, in fact, statistically agree with one another but this agreement is clinically uncompelling.

Therefore, we suggest that a further requirement be fulfilled before B–A analysis may be applied in a comparison study: the bias and precision must not only meet clinically acceptable requirements, but they must also be unique to the original data set pairing. In other words, it

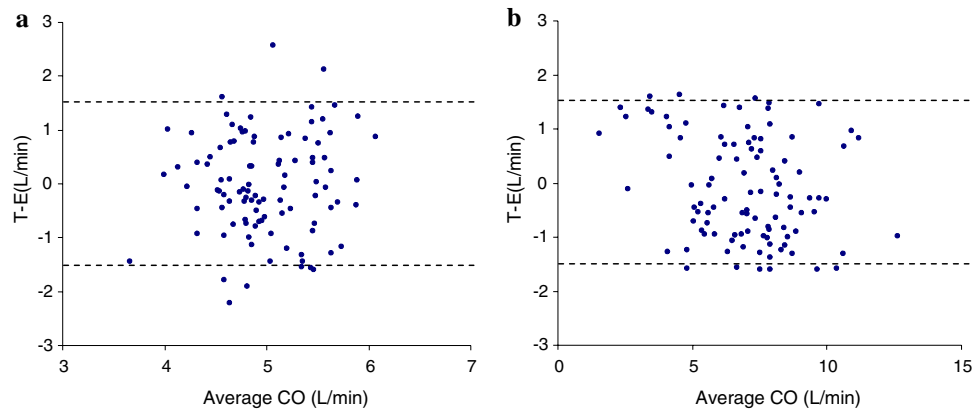

Fig. 1. a (left) and b (right). Bland Altman graph showing results of two experiments (Study #1 on the left and Study #2 on the right) comparing thermodilution (T) to an experimental (E) method for determining cardiac output. In Study #1 the range of cardiac output over which the observations are made is small compared to Study #2. The x-axes are deliberately scaled to make the experiments seem similar at first glance. In both graphs, dashed lines represent 95% limits of agreement.

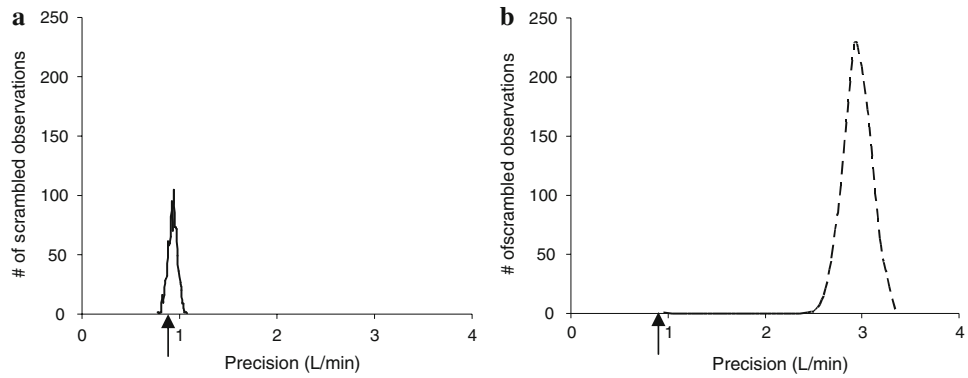

Fig. 2. a (left) and b (right): Distribution of precision versus frequency from 1,000 random mispairings of data presented from Study #1 (left) and Study #2 (right). The arrow represents the precision of the original pairing (approximately the same in both studies).

needs to be shown that less than 5% of all combinations of pairing of data meets the clinically acceptable limits of agreement. (If more than 5% of randomly paired combinations meet this requirement then the precision may have been met by chance alone.) One can then proceed with B–A analysis and evaluation of agreement.

If applied to the studies described earlier, one would find that the fraction of repeated pairings producing acceptable precision values in Study #1 was 0.63, far too many to be confident that B–A analysis could provide valid results. In Study #2, however, less than 5% of combinations produced clinically acceptable precision. Therefore, the results of B–A analysis are valid for Study #2 but not Study #1.

This problem could have been avoided by anticipating the low power of detecting agreement when collecting data over a narrow range. Since the precision of randomly paired data is:

$$\text{Precision} = \sqrt{2} * \frac{(SD_T + SD_E)}{2}$$

where SD_T and SD_E are the standard deviations of the data collected from methods E and T, respectively, it is clear that the precision is dependent on the SD (or range) of the original measurements. Therefore, the investigators designing Study #1 could have foreseen that so long as the collected cardiac output data lay within a low range, a clinically acceptable precision would inevitably be produced.

The procedure of randomly pairing data seems to be useful in evaluating the validity of B–A analysis in two ways. An investigator can use this technique to validate the results of a prior study by producing the probability distribution of precision from a known data set and seeing if the clinically acceptable precision lies within the 5% tail. A further application of this technique is to determine the necessary data range of a future study such that the clinically acceptable precision might reliably be produced. We provide the reader a utility to perform this procedure on his or her own data, at (www.isocapnia.com/statistics).

CONCLUSION

The outcome of B–A analysis is dependent on the range of the measurements such that narrow ranges will necessarily produce good agreements. This pitfall can be avoided by determining the probability that such observed agreement occurred in pairing of unassociated measurements. Standard B–A analysis can then proceed without concern about finding good agreement from unassociated methods.

REFERENCES

1. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol Med 1989; 19(1): 61–70.
2. Lee J. Evaluating agreement between two methods for measuring the same quantity: a response. Comput Biol Med 1992; 22(5): 369–371.
3. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990; 20(5): 337–340.
4. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003; 22(1): 85–93.
5. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1(8476): 307–10.
6. Critchley LA, Critchley JA. A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques. J Clin Monit Comput 1999; 15(2): 85–91.