REVIEW ARTICLE

Statistics in Nuclear Cardiology: Optimizing Assessment of Agreement between Like Continuous Measurements

David N. Williams, PhD[1], David M. Harrild, MD, PhD[2], Kathryn A. Williams, MS[3] and Michael Monuteaux ScD[4]

Received: June 21, 2017/Revised manuscript received: July 25, 2017/Accepted: July 25, 2017
J-STAGE Advance published: August 23, 2017

Abstract

Clinical researchers often need to quantify the degree of agreement, the "closeness," between two quantitative methods of measurement taken on the same subject and same variable, either when a new or revised method is considered for use in place of an established one or when measurements are separated by time. An ideal model of agreement would capture the degree of exactly equal outcomes. Analysis should be based on differences between two methods of measurement. Inappropriate statistical methods, including correlation, e.g. Pearson's *r*, are often used. But correlation, *r*, measures the degree of linear relationship between two variables, not the extent of agreement. A more informative and appropriate tool, the Bland and Altman (B&A) plot analysis, uses a graphical approach to present and assess differences between two measures. The differences between paired measurements are plotted on the Y-axis against the mean of the paired measures on the X-axis. The B&A chart allows for assessment of bias, both in size and consistency, over the range of measurements; it does not specify whether the agreement limits are acceptable; acceptable limits must be defined before. Additional reference lines can facilitate interpretation of the scatter plot.

Keywords: Agreement, Altman, Bland, Clinical research, Product-moment correlation coefficient
Ann Nucl Cardiol 2017；3（1）：48−52

Clinical researchers often need to assess relative agreement between two quantitative methods of measurements. For example, the introduction of a new or revised method for measurement requires assessing whether the new method can be used interchangeably with, or in place of, an established one (1, 2). When two methods are compared and neither is a gold standard, the degree of agreement between them is the key measure for showing the accuracy of the new technique. Agreement is defined as the "closeness" between quantitative measurements on the same subject assuming that nothing changed other than the time or the method of the measurements (3). But two readings will not generally be identical because measurement always includes some degree of random, unavoidable error (3); any quantification of agreement must recognize random error. In this review, we will focus on statistical methods for measuring agreement between like, continuous measurements.

Bland and Altman plot analysis (B&A) and the product-moment correlation coefficient (***r***) are the two most commonly used methods (4) to measure absolute agreement between continuous measurements. In this article, we compare ***r*** with B&A plot analysis to illustrate the two analysis methods, their outcomes and how they may be interpreted.

doi: 10.17996/anc.17-00030

1) David N. Williams
Senior Biostatistician, Biostatistics and Research Design Core, Institutional Centers for Clinical and Translational Research, Department of Adolescent Medicine, Boston Children's Hospital, and Instructor of Pediatrics, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115, USA
E-mail: david.williams3@childrens.harvard.edu
2) David M. Harrild
Attending Physician, Department of Cardiology, Assistant Professor of Pediatrics, Harvard Medical School
3) Kathryn A. Williams
Senior Biostatistician, Biostatistics and Research Design Core, Institutional Centers for Clinical and Translational Research, Boston Children's Hospital
4) Michael Monuteaux
Assistant Professor of Pediatrics, Harvard Medical School, Assistant Director, Biostatistics and Research Design Core, Institutional Centers for Clinical and Translational Research, Boston Children's Hospital

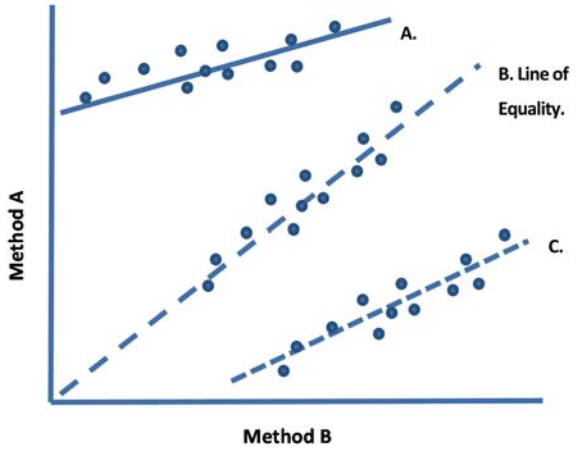

Fig. 1 Pearson's product moment correlation scatter plots (***r***).

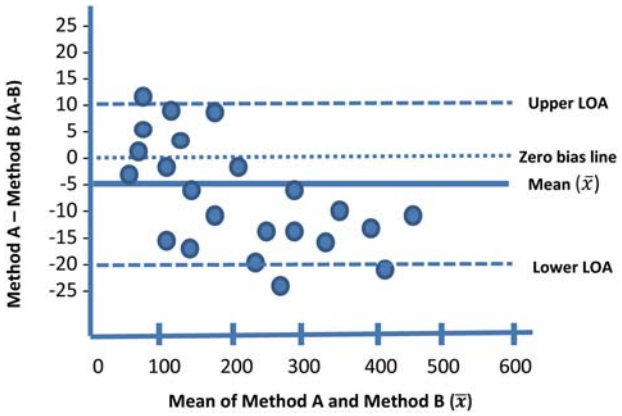

Fig. 2 Bland-Altman (B&A) plot.

Product-moment correlation

The product-moment correlation coefficient (***r***), even though commonly used, is not appropriate for assessing agreement (5). While an ideal analysis of agreement should reflect the extent of agreement of measurements on the same subject using two methods (or a single method at two different times) (1), correlation, ***r***, does not do this; it represents the degree of linear relationship between two measures (5). Two measures may be strongly correlated, such as height and weight, but have a poor degree of absolute agreement. In the example of height and weight (which have an approximate ***r***[1)] of 0.72) the two measures will be very different. Outcomes from two measures of the same patient and same variable may have perfect correlation if the points lie along any straight line (Fig. 1 presents 3 highly correlated examples) but perfect agreement is only gained when the points lie along the "line of equality (B)" which represents absolute agreement between the two methods.

Bland Altman plot

Bland and Altman (6) developed a graphical method to describe agreement between two quantitative, continuous measurements by assessing the average of the differences between the paired data. The basic concept of Bland and Altman's method (B&A) is the calculation of the difference of the paired measurements (A-B) made by the two methods, and plotting these differences (A-B) on the Y-axis against the mean of the two readings, (A+B)/2, on the X-axis (2) (Fig. 2).

Additional reference lines may be added to the scatter plot to facilitate interpretation. Bland and Altman indicate that approximately 95% of the data points should lie within ±2 standard deviations of A-B (Sdiff). To reflect this:

- A mean difference line ($\bar{x}$) line reflects the mean of differences (method A-method B).
- Upper and lower Limits of Agreement (LOA) are added; these are shown as the mean difference +/- 1.96 × Sdiff.
- A zero-bias line (at '0') allows for clear understanding of the degree and direction of bias in the plotted points. The gap between the zero bias and mean lines reflects the amount and direction of bias. In Fig. 2, the average of the differences (reflected in the Mean line) is about -5; this suggests that, on average, the second method (B) measures 5 units higher than the first one (A). Fig. 2 also illustrates an evident systematic bias: A-B differences take on negative values (B becomes consistently larger than A) as mean values (A+B/2) increase.

The B&A plot does not present whether the degree of agreement between the two methods is sufficient or suitable to use the methods interchangeably. Clinical or analytical guidelines should be used to define, before-hand, the maximum acceptable differences. Whether that difference is clinically important should be based on clinical considerations such as the overall distribution of data; for example: a difference of several percentage points may be unimportant in cardiac ejection fraction but unacceptable for a measurement of patient blood chemistries. A priori limits should be established that reflect the maximum acceptable clinically important difference. Bland and Altman suggest that 95% [i.e.(1-α)%)] of the differences should be within these a priori defined acceptable limits (3, 5).

The statistical significance of the bias (i.e. the gap between the Mean and Zero Bias lines) can be determined by calculating a confidence interval (CI) to reflect the precision of the estimated mean difference lines (Fig. 3). If the zero bias line falls outside the CI around the mean line (as it does in Fig. 3) then it may be concluded that there is a statistically significant systematic difference between the two methods (5).

Differences can also be expressed on the Y axis as

[1)] General approximation based on range of studies.

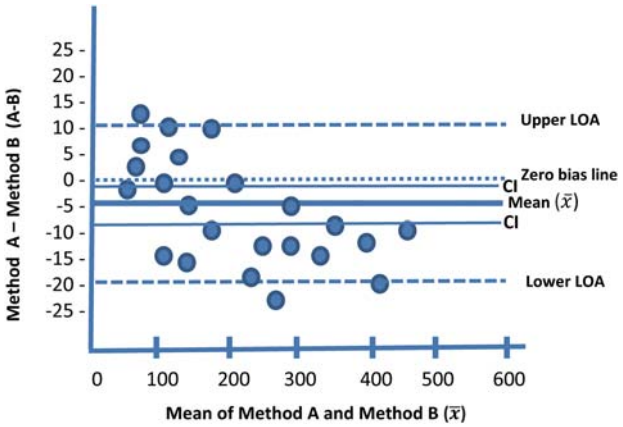

Fig. 3 Bland-Altman plot with confidence intervals around Mean (x).

percentages of the values, that is proportionally to the magnitude of measurements [(Method A-Method B)/mean %)] (5). This alternative is useful when variability of the differences increases as the size of the measurement increases.

An important assumption in the B&A assessment is that the differences (A-B) are normally distributed. If this assumption is not valid the upper and lower LOA estimates may not be accurate and need to be estimated using a nonparametric method. While a visual assessment (such as with a histogram or Q-Q plot) can offer a reasonable evaluation, a test for normal distribution (such as Kolmogorov-Smirnov) is recommended. Data can be transformed if differences seem not to be normally distributed (5).

A clinical application of agreement analysis

In the previous issue of the *Annals of Nuclear Cardiology*, the study by Dimitriu-Leen and colleagues considered a direct comparison method for cardiac ^{123}I-MIBG imaging in which 4 hour cardiac counts, mediastinal counts, heart-to-mediastinum ratio and washout rates were estimated from measurements taken at 1, 2 and 3 hours (7); these estimates were compared with actual 4 hour values which was considered as standard measurement. Optimal estimation would have been represented by the estimated and actual values being in full agreement. Both correlation (Pearson's ***r***) and Bland and Altman analysis results were given. These findings offer an excellent comparison of the results of both tests[2)].

The first three scatter plots in Fig. 4 present the linear relationship between the two measures. Beginning with estimates based on H1 (left) to H3 (right), ***r*** values remain consistently high and almost identical. The scatterplot and regression lines move closer to the equivalence line (running at a 45 degrees angle through the graphs). Yet while the correlations between the three estimates/actual counts are almost equal, the degrees of agreement are substantially different across the 3 plots. The B&A charts for these same comparisons more effectively illustrate the changing levels of agreement (Fig. 5).

From the B&A charts we can see that agreement markedly improves between estimates of cardiac counts based on H1 counts to those based on the H3 counts. The mean difference (i.e. bias) decreases from approximately -11.4 at H1, to -5.8 at H2 and -2.5 at H3; suggesting that, on average, the second method, actual count, measures approximately 11.4 units higher at H1 and approximately 2.5 units higher at H3, a substantial improvement that is not reflected in the ***r*** values. The B&A chart also illustrates an evident systematic bias; with H1 estimates mean differences increase substantially as cardiac counts increase; this trend decreases with H2 and is almost nonexistent with H3 estimates. Dimitriu-Leen also present a ***r*** value given in lower left box which represents the linear relationship between y-axis differences and x-axis mean (A-B/2). While not a standard feature of B&A charts, it is further evidence of unequal variance between the two measures resulting in the mentioned trends (an ***r*** of 0 would suggest that the differences in the A and B methods do not vary over the observed range of A and B) (9).

A review of Washout Rate (WR) scatter plots from the same study by Dimitriu-Leen et al. presents more marked findings. ***r*** is consistently strong, e.g. >0.73 for all comparisons, yet degree of agreement as reflected in the B&A charts is clearly different between the three estimates. Agreement improves substantially from H1 to H2 but decreases slightly with H3, even though data points in H3 clearly fall more closely to the Line of Equality (Fig. 6).

The B&A charts present more insightful information (Fig. 7). For H1 estimates, bias (i.e. mean differences) is 22.7, for H2 9.2, H3 1.8. We can say that estimates based on H1 data were approximately 21.7 points higher than H4 findings; for H2: 9.2 and for H3: 1.8. Trends in bias appear relatively mild compared with earlier examples. Dimitriu-Leen concluded that estimates derived from H3 acquisitions most accurately estimated H4 measures. We concur with the authors' conclusions but do so based on the information provided in the B&A charts rather than on the scatter plots and ***r*** values

[2)] If one of the two measurements represents a gold standard, i.e. the actual hour 4 measurement may be considered the standard or reference method, it has been suggested that the standard values be used instead of the mean of the two measurements (8), yet this suggestion is controversial (1); further evaluation should be carried out prior to use. The Dimitriu-Leen et al. clinical example given here is given for purpose of comparing scatter plots and ***r*** values to B&A charts; questions regarding alternative statistical methods are beyond the scope of this article.

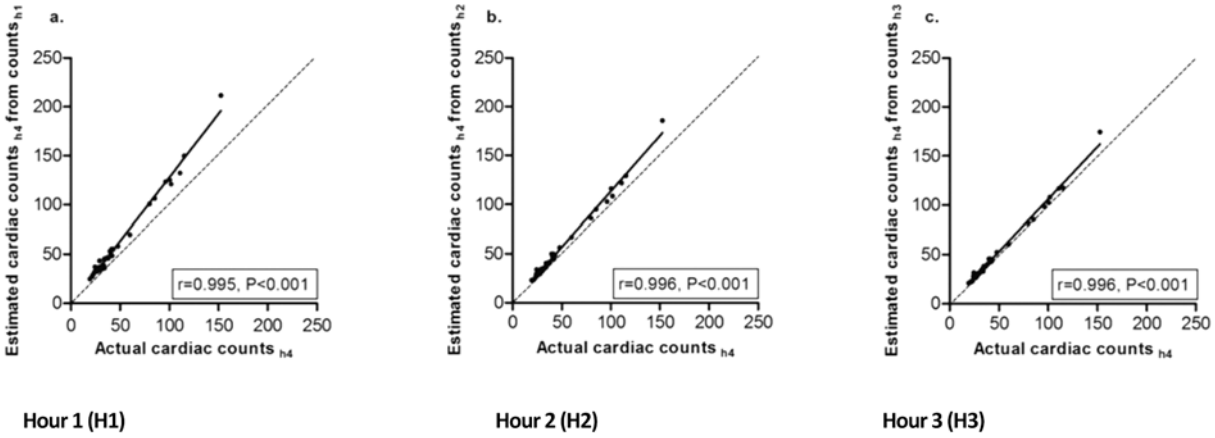

Fig. 4 Linear relationship between estimated and actual cardiac counts at 3 different hours.

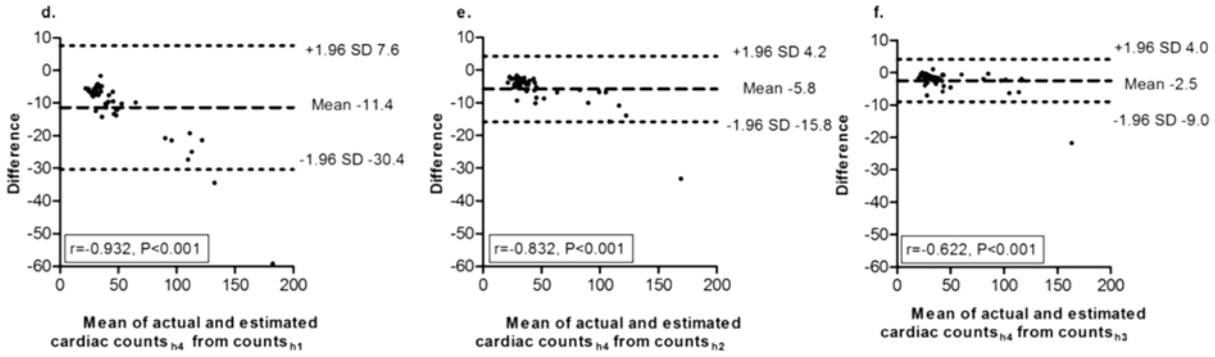

Fig. 5 B&A charts presenting levels of agreement between estimated and actual cardiac counts at 3 different hours.

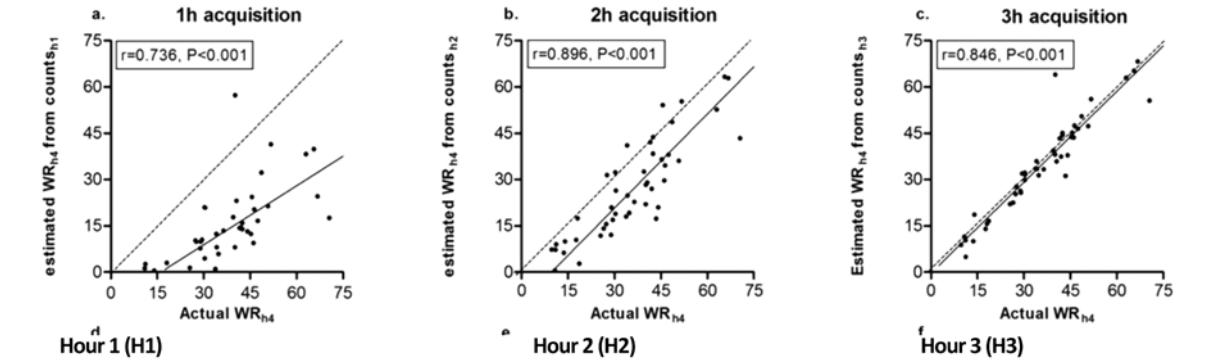

Fig. 6 Linear relationship between estimated and actual Washout Rate at 3 different hours.

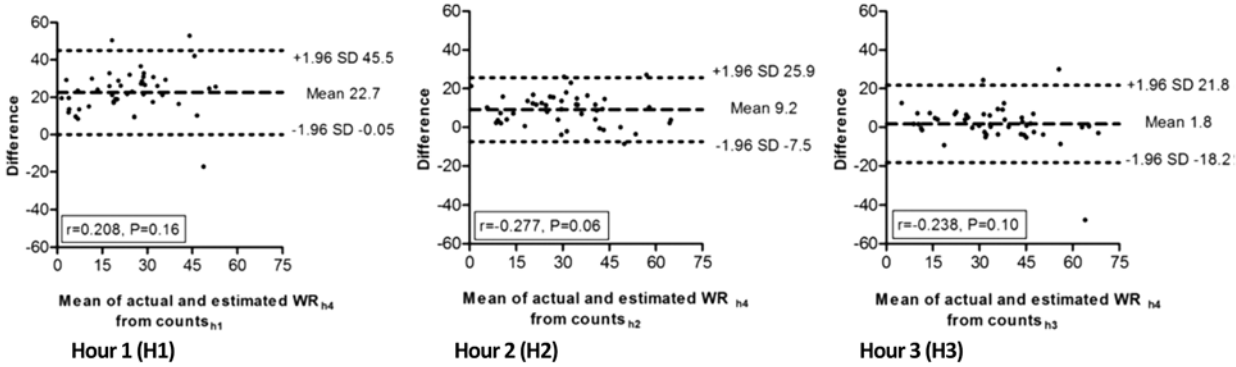

Fig. 7 B&A charts presenting levels of agreement between estimated and actual Washout Rate at 3 different hours. Permission obtained from publisher (Ref 7: Dimitriu-Leen AC, Gimelli A, van Rosendael AR, et al. Cardiac ^{123}I-MIBG Parameters at 4 Hours Derived from Earlier Acquisition Times. Ann Nucl Cardiol. 2016; 2(1): 21-9).

presented. Correlation between methods is often misleading as exemplified in the Dimitriu-Leen examples; a near perfect linear relationship may not truly represent agreement. B&A charts offer insights into both the strength and nature of agreement between two measurement methods not available in measures of correlation.

Conclusions

Comparison of two methods of measurement on the same subject and variable, in order to determine if they can be used interchangeably, is common in clinical research. "Agreement" is a critical concept, different from "relationship," that must be directly assessed. Pearson's ***r*** is often misapplied as a measure of agreement and will commonly produce irregular or misleading results. Bland and Altman proposed a graphical presentation method, the B&A plot, that allows for formal assessment of agreement between two measures. While it requires modification for use with various types of data, e.g. repetitive measures and agreement against a gold standard, the method represents an efficient and insightful approach for evaluation of agreement between two measures.

Acknowledgments

I would like to acknowledge the contributions of my fellow biostatisticians, Peter Forbes and Paul Mitchell at Boston Children's Hospital, who freely gave their knowledge and critical feedback for this paper.

Sources of funding

None.

Conflicts of interests

None.

Reprint requests and correspondence:
David N. Williams, PhD
Senior Biostatistician, Biostatistics and Research Design Core, Institutional Centers for Clinical and Translational Research, Department of Adolescent Medicine, Boston Children's Hospital, and Instructor of Pediatrics, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115, USA
E-mail: david.williams3@childrens.harvard.edu

References

1. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995; 346: 1085-7.
2. Fernandez R, Fernandez G. Validating the Bland and Altman Method of Agreement Western Users of SAS Software Conference 2017. http://www.wuss.org/, Western Users of SAS; 2017 [cited 2017 4/21/2017]. An evaluation of Bland-Altman method of agreement.
3. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. J Biopharm Stat 2007; 17: 529-69.
4. Zaki R, Bulgiba A, Ismail R, et al. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PLoS One 2012; 7: e37908.
5. Giavarina D. Understanding Bland Altman analysis. Biochem Med 2015; 25: 141-51.
6. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307-10.
7. Dimitriu-Leen AC, Gimelli A, van Rosendael AR, et al. Cardiac ^{123}I-MIBG parameters at 4 hours derived from earlier acquisition times. Ann Nucl Cardiol 2016; 2: 21-9.
8. Krouwer JS. Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. Stat Med 2008; 27: 778-80.
9. Seed P. Comparing several methods of measuring the same quantity. Stata Technical Bulletin 2001; 10: Available from: http://EconPapers.repec.org/RePEc:tsj:stbull:y:2001:v:10:i:55:sbe33.