January 27, 2015

Abstract

Center for Quality and Applied Statistics Kate Gleason College of Engineering Rochester Institute of Technology Technical Report 20053 May 12, 2005 A Study of the Bland-Altman Plot and its Associated Methodology Joseph G. Voelkel Bruce E. Siskowski Center for Quality and Applied Statistics Reichert, Inc. Rochester Institute of Technology bsiskowski@reichert.com joseph.voelkel@rit.edu

1

A Study of the Bland-Altman Plot and its Associated Methodology

by Joseph G. Voelkel and Bruce E. Siskowski

- Consider the situation in which two measurement devices are compared but no true value exists. For this problem, when a set of subjects are measured with both de- vices, Bland and Altman have strongly criticized several methods of analysis. They advocated a plot of the differences in each subjects readings versus the average of such readings and statistics associated with it.

- They recommended this plot both for checking assumptions, such as homogeneity of variances, and for assessing the agree- ment between the devices.

- They also advocated several other methods associated with such comparisons. We argue that a sound comparison of devices should do more than measure agreementfor example, every comparison of devices should also com- pare the repeatability of the devices, a measure that cannot be obtained by simply measuring agreement. We show how the Bland-Altman plot and associated meth- ods can itself be misleading through several examples.

- We propose another method, structural equation modeling, as the mathematical framework for such studies. We critique another method suggested by Bland and Altman for measuring agreement when the devices are measuring the phenomenon on different scales that are thought to be linearly

related, and suggest alternative methods. We continue to advocate the use of the Bland-Altman plot, when used cautiously, as one of several ways for checking model adequacy.

1 Introduction

An important consideration in many areas of medicine is the comparison of measurement devices, especially when no true value can be measured. As Bland and Altman (1986) noted, Clinicians often wish to have data on, for example, cardiac stroke volume or blood pressure where direct measurement without adverse effects is difficult or impossible. The true values remain unknown. Instead indirect methods are used, and a new method has to be evaluated by comparison with an established technique rather than with the true quantity. If the new method agrees sujciently well with the old, the old may be replaced, This is very different from calibration, where known quantities are measured by a new method and the result compared 1

with the true value or with measurements made by a highly accurate method.

- When two methods are compared neither provides an unequivocally correct measurement, so we try to assess the degree of agreement. But how?" (Italics ours.) The idea of agreement plays a key role in Bland and Altman analyses of device comparisons, a point to which we will return later.

- Earlier, Altman and Bland (1983) had addressed this question because of the statistically incorrect methods of analysis they had witnessed in the medical literature, including correlation analysis, regression analysis, and hypothesis tests to compare means.

- To correct these deciencies, they proposed a simple graphical technique that has become widely used in the medical literature. A set of subjects is selected, preferably at random from the population of interest, but at least to cover the range of values over which the devices should be compared.

- Each subjects feature is measured on each of the two devices in such a way that it is reasonable to compare the measure- ments. A graph is then made of (a) the differences Y X between the two readings vs (b) the average (X + Y) /2 of the two readings. Note that this basic plot is based on exactly one measurement from each device for each subject we refer to this as the one-measurement case.

- This plot, commonly called the Bland-Altman plot, has gained wide acceptance in the medical literature. Similarly, their recommendations for investigating the data using this plot, as well as related recommendations, may be called the Bland-Altman method.

- The plot and the method underlying it does essentially examine the agreement between the two techniqueslarge differences (where large should be clinically determined) indicate the two device do not agree well.

- Altman and Bland (1983) continued their argument by stating [t]he main emphasis in method comparison studies clearly rests on a direct comparison of the results obtained by the alternative methods.

- The question to be answered is whether the methods are comparable to the extent that one might replace the other with suicient accuracy for the intended purpose of measurement. (Italics ours.)

- By using this plot instead of a plot of Y vs. X, Altman and Bland (1983) noted that it is much easier to assess the magnitude of the disagreement (both error and bias) as well as other features in the data such as outliers and to see if there is any trend, for example an increase in Y X for high values. (Bland and Altman used A and B instead of Y and X, but we substitute our terminology in their quotes for consistency in this paper.)

- If such an increase (or decrease) takes place, they suggested attempting a transformation of the raw data. They continue, [i]n the absence of a suitable transformation it may be reasonable to describe the differences between the method by regressing Y X on (X + Y) /2.

- The same point is stated in Bland and Altman (1995): *There may also be a trend in the bias, a tendency for the mean difference to rise or fall with increasing magnitude. . . In [the] gure . . . for example, therc is an increase in bias with magnitude, shown by the positive slope of the regression line.*

- They again suggest that a transformation may eliminate this effect. The model they appear to have considered in their article will be shown in Section 3.

The 1983 article, which included a number of excellent insights on measurement studies from the view of applied statistics, had a strong impact on the ways in which such studies were analyzed. The purposes of our article are to review this plot and the associated methodology proposed by Bland and Altman, to note some difficulties with this methodology, and to recommend alternative analyses.

2 On Comparing Measurement Devices

In our experience, a measurement-device study is usually conducted to address the following questions:

1. Are the devices identical in their ability? This is equivalent to the word inter- changeability used in elds of testing and instrumentation.

2. If they are not identical as is, can they be made so after calibration? Or (b) are they measuring the same features but with different precision?

3. If they are not identical even after calibration or allowing for different precision, are they measuring the same features on the subjects, after calibration and allowing for different precision?

4. If they are not measuring the same features, to what extent are they measuring dierent features on the subjects?

From the italicized section of the BlandAltman quotes that we have cited, and from the extensive writings of Bland and Altman, their key emphasis is to try to assess the extent to which the devices agree. The question to be answered is whether the methods are comparable to the extent that one might replace the other with sufficient accuracy for the intended purpose of measurement.

Whether two measurements are close enough must be determined from clinical considerations, so suppose that close has been dened. If a study is conducted that can address the four questions we pose, then it can address the questions posed by Bland and Altman. However, if the study only addresses the question posed by Bland and Altman, it can not necessarily address the questions we have posed.

Consider the one-measurement case. The assumption that the current device provides consistent readings cannot be ascertained from the one-measurement case, so it is possible that the current device does not agree very well with (is not close to) itself. If the current measurement device is not particularly precise, it is possible using the BA methodthat the new device is more precise than the old one, but cannot be used to replace it.

3 Mathematical Model

The model we use in this paper is the common and useful structural model (e.g. Fuller(l987)). This is a natural model for the comparison of two devices. See, e.g., Mandel (1984), who used this technique on data from the U.S. National Institute of Standards and Technology (NIST). This model is a simple example of structural equations with latent variables, e.g. Bollen (1989).

Consider the case of two measuring devices. Let 90, represents the longterm average (true) value of the measurement for the it, subject when measured on the ac-device (y-device) at a some xed point in time. (The xed point in time" is needed in the medical eld, because the underlying values often change over time for subjects.) The structural model assumes that these yi) values lie on a straight line this means that the two devices are measuring the same feature on a given subject except that a possible linear calibration needs to be made.