Assessing Agreement: A Graphical Approach

Paul Hshieh*, Ph. D.
paul.hshieh @fda.hhs.gov

Tie Hua Ng*, Ph. D.
tiehua.ng@fda.hhs.gov

Center for Biologics Evaluation and Research
US Food and Drug Administration
10903 New Hampshire Avenue, Silver Spring, MD 20993

Abstract

Altman and Bland (1983) criticized the use of correlation, regression and the mean difference in the comparison of two methods of measurement. They proposed to plot the difference between the two methods against the average (Bland-Altman plot). A scatter plot and Q-Q plot are useful graphical tools to assess the agreement of two methods of measurement.

In this paper, we proposed a new "procedure" in assessing the agreement of a new method to an old method using three graphical approaches: scatter plot, Bland-Altman plot and Q-Q plot. In this procedure, subjects are measured twice using the Old method (Old1, Old2) and once using the New method (New). In each graphical approach, the plot of Old1 vs. Old2 is used as a norm in which the plots of New vs. Old1 and New vs. Old2 can be compared with. If the new method is comparable to the old, then the plots of New vs. Old1 and New vs. Old2 should each appear similar to the plots of Old1 vs. Old2.

A simple measurement error model is used in a simulation study to learn how the systematic bias and unequal variance are revealed in these plots. Such learning can then be used to assess the agreement of two methods in real life. A real example is used to illustrate the new "procedure".

Key Words: Bland-Altman plot, Q-Q plot, graphical approach, systematic bias.

Introduction

Traditionally, the assessment of agreement between two methods of measurement had been done by several statistical methods such as correlation, regression and paired t-test. Altman and Bland (1983) criticized the use of these statistical methods and proposed to plot the difference between the two methods against the average (Bland-Altman plot). A scatter plot and Q-Q plot are useful graphical tools to assess the agreement of two methods of measurement. In this paper, we proposed a new "procedure" in assessing the agreement of a new method to an old method using three graphical approaches: scatter plot, Bland-Altman plot and Q-Q plot.

* The views expressed in this paper are those of the authors and do not necessarily reflect the perspective of the U.S. Food and Drug Administration.

We first introduce the real data and then describe the concept of Quantile-Quantile (Q-Q) plot. Scatter plot, Bland-Altman plot and Q-Q plot of simulated data will be presented. We then learn how the systematic bias and unequal variance are revealed in these plots. Such learning is then used to assess the agreement of two devices in the real example.

The statistical software R was used to generate simulated data and draw all plots.

A Real Example

A new non-invasive device has been developed to measure the hemoglobin concentration in blood donation settings. This new device is compared to a predicate device in a study in which 253 subjects were enrolled. For each subject, the quantitative determinations of hemoglobin were measured twice using each of new device (denoted by New1 and New2) and an old device (denoted by Old1 and Old2). The results of this study are summarized in Table 1.

Table 1. Summary Results

Variable	Mean	Standard Deviation
Old1	12.91	1.46
Old2	12.93	1.46
New1	13.17	1.37
New2	13.07	1.45

How can one assess the agreement between the two devices?

What is Q-Q plot?

A Q-Q plot (Kotz S and Johnson NL, 1981) is a visual technique to inspect how closely the data fits an underlying distribution. The quantiles from a real data set are plotted against the corresponding quantiles from an underlying distribution. If the plotted points fall close to the identity line (y=x), then the data fits the underlying distribution well.

Following this concept, the quantiles of the data from the first measurement method is plotted against the quantiles of the data from the second measurement method. If the plotted points fall close to identity line (y=x), then the two measurement methods has a certain degree "closeness" based on a visual assessment.

Simulation: A Simple Measurement Error Model

In the simulation study, the data were generated using a simple measurement error model with k raters or methods of measurement:

$$Y_{ij} = X_i + \varepsilon_{ij},\ i = 1, \ldots, n;\ j = 1, \ldots, k,$$

where

n : the sample size
k : the number of raters or methods of measurement
X_i: unknown characteristic of i^{th} subject randomly selected from a population.
Y_{ij}: Measurement of the i^{th} subject by j^{th} rater,
μ_j: systematic bias for the j^{th} rater.

For each i, X_i is assumed to be normally distributed with mean μ and variance σ^2. For each i and j, the measurement error ε_{ij} is assumed to be normally distributed with mean μ_j and variance σ_j^2. Finally, X_i and ε_{ij} are assumed to be independent.

Simulated Data

A sample size of 250 is used (n = 250) with k = 2. The parameters μ and σ are chosen so that it mimics the real example data. Therefore, we set $\mu = 13$, and $\sigma = 1.5$ based on Table 1. Three sets for measurements were simulated: two replicates for old method (Old1, Old2) and one replicate for new method (New). Although there are two replicates for the new device for the real data, only one replicate for the new device is simulated because the new "procedure" does not need replication of the new device.

The old method measurement is considered to be reproducible, and the replicate measurement from old method should demonstrate strong agreement between them. The plots of Old1 vs. Old2 are used as a norm in which the plots of New vs. Old1 and New vs. Old2 can be compared with.

We assume that the measurement error for the old device is unbiased ($\mu_1 = 0$) with a small variance ($\sigma_1 = 0.15$) relative to population variance ($\sigma = 1.5$) and the measurement error for the new device could be unbiased ($\mu_2 = 0$) or biased ($\mu_2 = 2$) with equal ($\sigma_2 = 0.15$) or unequal variance ($\sigma_2 = 1$) resulting in the following four cases:

Case 1: No systematic bias and equal variance ($\mu_1 = \mu_2 = 0$; $\sigma_1 = \sigma_2 = 0.15$)
Case 2: Systematic bias and equal variance ($\mu_1 = 0$, $\mu_2 = 2$; $\sigma_1 = \sigma_2 = 0.15$)
Case 3: No systematic bias and unequal variance ($\mu_1 = \mu_2 = 0$; $\sigma_1 = 0.15$, $\sigma_2 = 1$)
Case 4: Systematic bias and unequal variance ($\mu_1 = 0$, $\mu_2 = 2$; $\sigma_1 = 0.15$, $\sigma_2 = 1$)

For each combination of four simulated data sets and three graphical approaches, three figures, namely, (a) Old1 vs. Old2 (b) New vs. Old1 (c) New vs. Old2 are shown in Figure 1.1 through Figure 4.3.

Figure 1.1 Case 1: No systematic bias and equal variance ($\mu_1 = \mu_2 = 0$; $\sigma_1 = \sigma_2 = 0.15$)
Scatter Plots

(a) Old1 vs. Old2 (b) New vs. Old1 (c) New vs. Old2

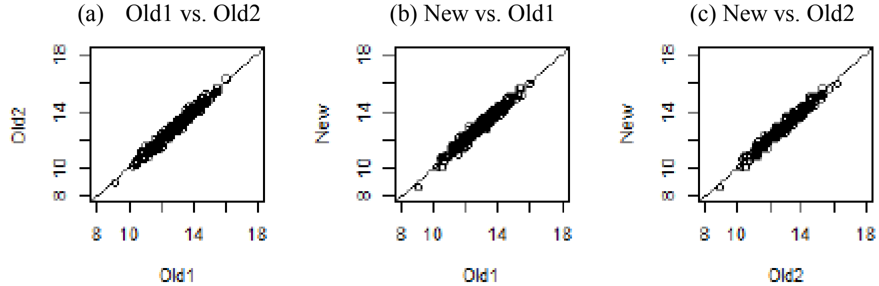


Figure 1.2 Case 1: No systematic bias and equal variance ($\mu_1 = \mu_2 = 0$; $\sigma_1 = \sigma_2 = 0.15$) Bland-Altman Plots

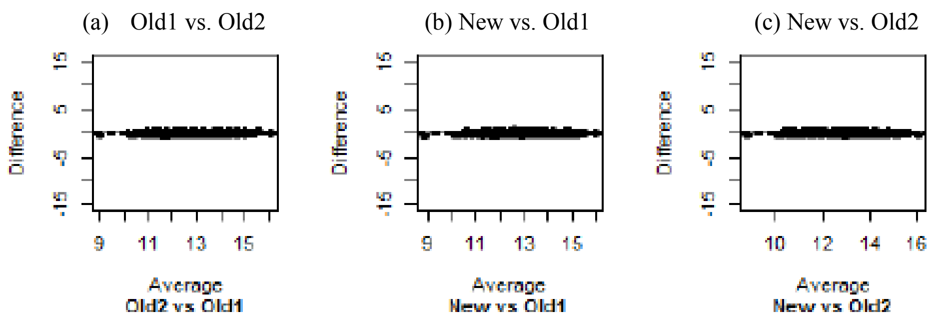


Figure 1.3 Case 1: No systematic bias and equal variance ($\mu_1 = \mu_2 = 0$; $\sigma_1 = \sigma_2 = 0.15$) Q-Q Plots

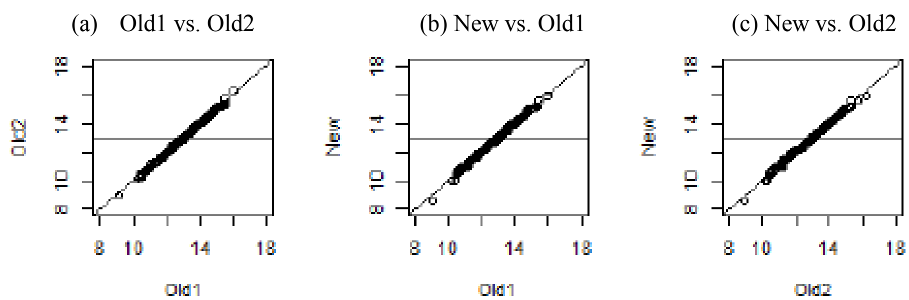


Figure 2.1 Case 2: Systematic bias and equal variance ($\mu_1 = 0$, $\mu_2 = 2$; $\sigma_1 = \sigma_2 = 0.15$) Scatter Plots

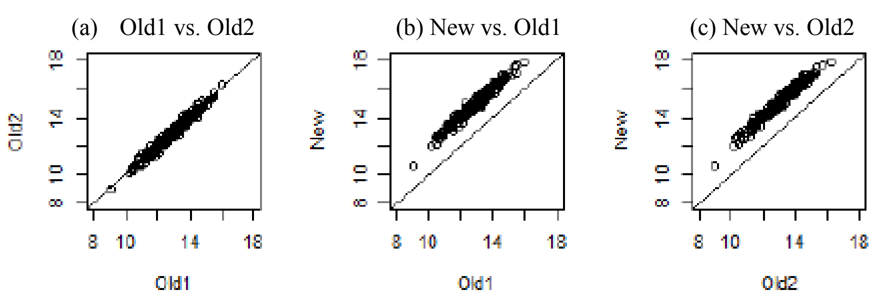


Figure 2.2 Case 2: Systematic bias and equal variance ($\mu_1 = 0$, $\mu_2 = 2$; $\sigma_1 = \sigma_2 = 0.15$)
Bland-Altman Plots

(a) Old1 vs. Old2 (b) New vs. Old1 (c) New vs. Old2

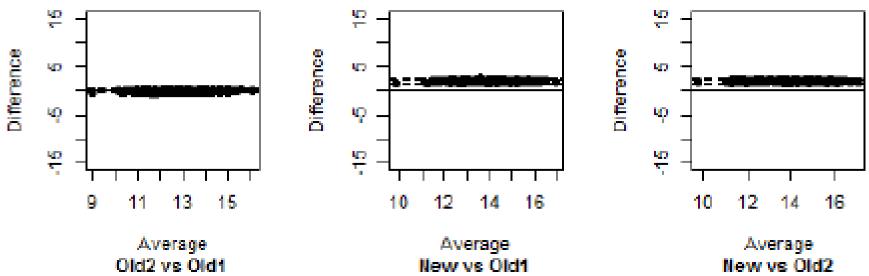


Figure 2.3 Case 2: Systematic bias and equal variance ($\mu_1 = 0$, $\mu_2 = 2$; $\sigma_1 = \sigma_2 = 0.15$)
Q-Q Plots

(a) Old1 vs. Old2 (b) New vs. Old1 (c) New vs. Old2

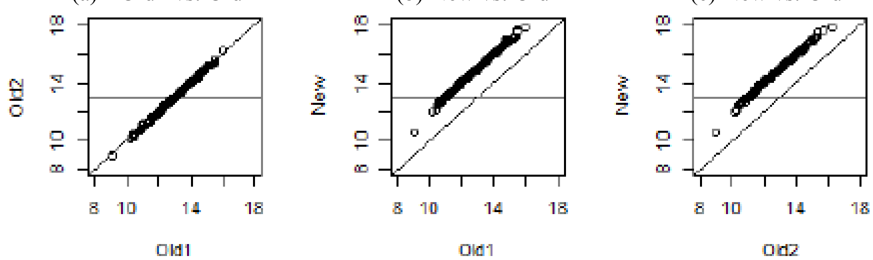


Figure 3.1 Case 3: No systematic bias and unequal variance ($\mu_1 = \mu_2 = 0$; $\sigma_1 = 0.15$, $\sigma_2 = 1$)
Scatter Plots

(a) Old1 vs. Old2 (b) New vs. Old1 (c) New vs. Old2

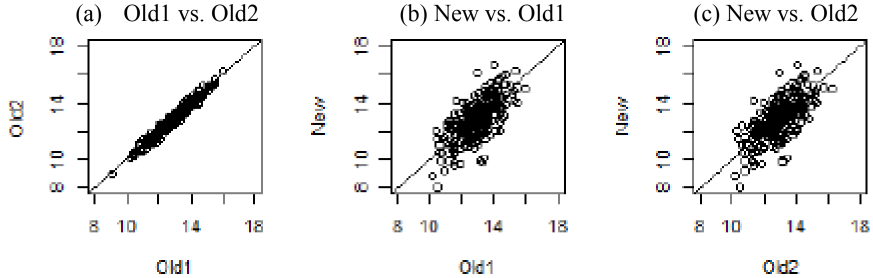


Figure 3.2 Case 3: No systematic bias and unequal variance ($\mu_1 = \mu_2 = 0$; $\sigma_1 = 0.15$, $\sigma_2 = 1$)
Bland-Altman Plots

(a) Old1 vs. Old2 (b) New vs. Old1 (c) New vs. Old2

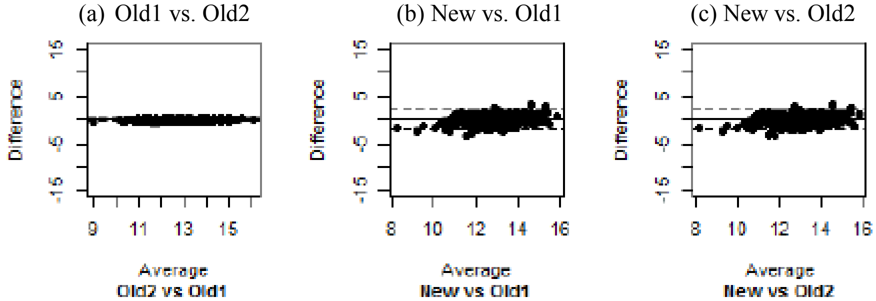


Figure 3.3 Case 3: No systematic bias and unequal variance ($\mu_1 = \mu_2 = 0$; $\sigma_1 = 0.15$, $\sigma_2 = 1$)
Q-Q Plots

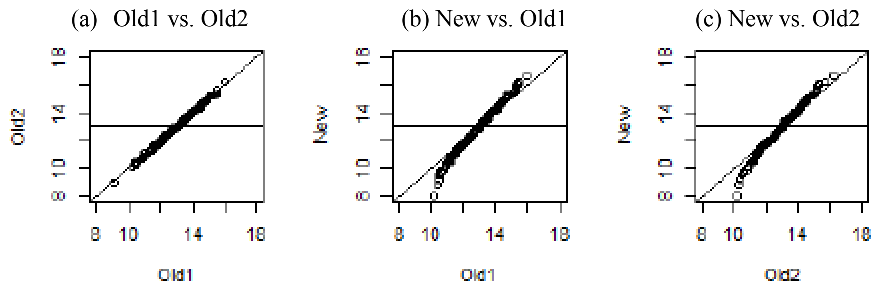


Figure 4.1 Case 4: Systematic bias and unequal variance ($\mu_1 = 0$, $\mu_2 = 2$; $\sigma_1 = 0.15$, $\sigma_2 = 1$)
Scatter Plots

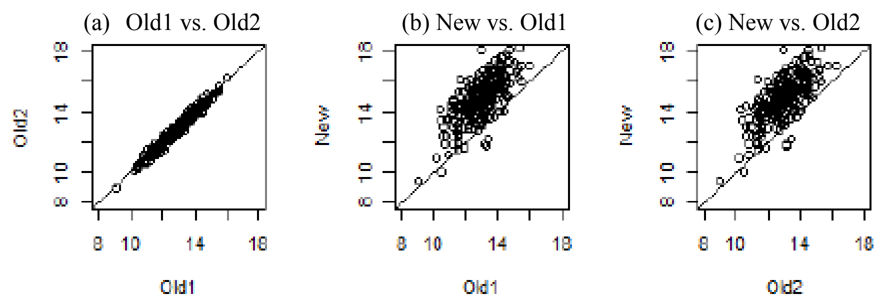


Figure 4.2 Case 4: Systematic bias and unequal variance ($\mu_1 = 0$, $\mu_2 = 2$; $\sigma_1 = 0.15$, $\sigma_2 = 1$)
Bland-Altman Plots

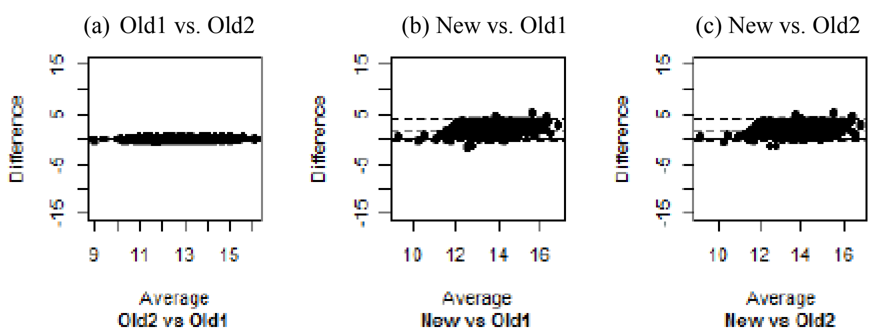
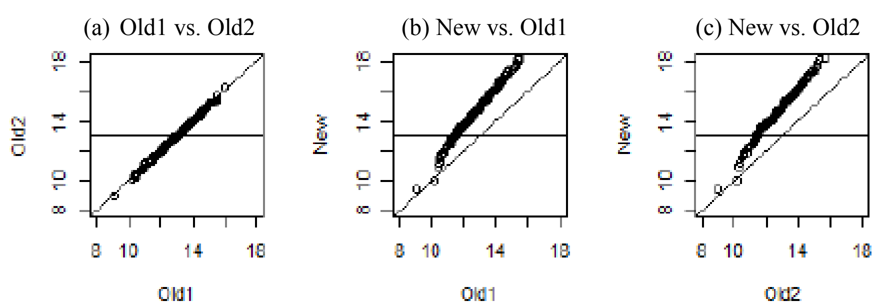


Figure 4.3 Case 4: Systematic bias and unequal variance ($\mu_1 = 0$, $\mu_2 = 2$; $\sigma_1 = 0.15$, $\sigma_2 = 1$)
Q-Q Plots



The agreement of a new method to an old method may be assessed by comparing Figures (b) New vs. Old1 and (c) New vs. Old2 to Figure (a) Old2 vs. Old1. Since Figure (a) is a norm, “Significant” difference of figures (b) and (c) from the (a) indicates the disagreement between two methods due to bias and/or unequal variance.

Comparing Figure (b) and (c) to Figure (a) in Figure 1.1 through Figure 4.3, we have the following observations:

1. Figures (b) and (c) looks very similar to Figure (a) indicating good agreement between the two methods (see Figures 1.1-1.3)
2. The systematic bias can be “detected” by all three graphical approaches based on Figures (b) and (c) without referring to Figure (a). Figures (b) and (c) show that all data points are shifted away from the 45^0 degree line (y=x) for the Scatter Plot and Q-Q Plot, and shifted away from the y=0 horizontal line for the Bland-Altman Plot when there is a systematic bias as in Cases 2 and 4.
3. The new “procedure” can “detect” unequal variance in all three graphical approaches by comparing Figures (b) and (c) with Figure (a). With unequal variance (Cases 2 and 4), Figures (b) and (c) show that the data points are spread out more than those in Figure (a) (see Figures 2.1, 2.2, 4.1 and 4.2). The Q-Q plots (see Figures 2.3 and 2.4) also indicate unequal variance for the “skewness”.

These observations will be used to assess the agreement in the real example given earlier in this paper.

An application to the Real Data

In each of three graphical approaches, we have the following five figures: (a) Old1 vs. Old2, (b) New1 vs. Old1, (c) New1 vs. Old2 (d) New2 vs. Old1 and (e) New2 vs. Old2, as shown in Figures 5.1-5.3.

Figure 5.1: Scatter Plots

(a) Old2 vs. Old1 (b) New1 vs. Old1 (c) New1 vs. Old2 (d) New2 vs. Old1 (e) New2 vs. Old2

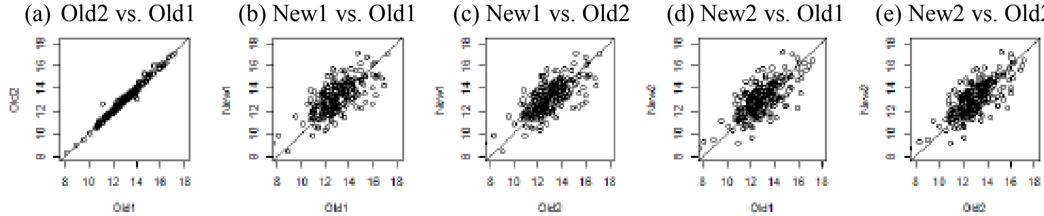


Figure 5.2: Bland-Altman Plots

(a) Old2 vs. Old1 (b) New1 vs. Old1 (c) New1 vs. Old2 (d) New2 vs. Old1 (e) New2 vs. Old2

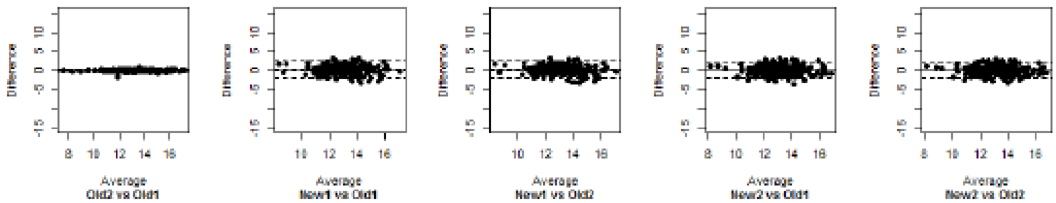
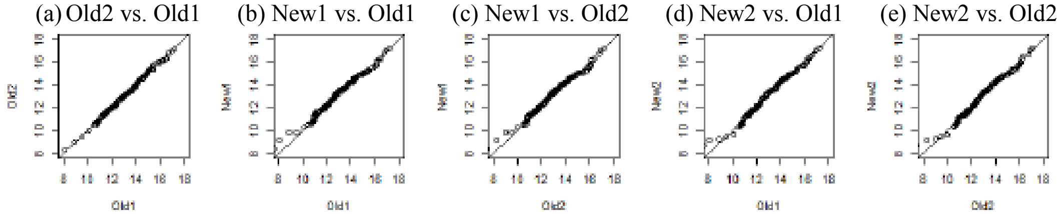


Figure 5.3: Q-Q Plots

(a) Old2 vs. Old1 (b) New1 vs. Old1 (c) New1 vs. Old2 (d) New2 vs. Old1 (e) New2 vs. Old2



Based on what we learn from the simulation study, it appears that there is no systematic bias but the two devices may have unequal variances.

Discussion

Since Figures (d) and (e) are very similar to Figures (b) to (c), respectively, the second replicate for the new device (New2) is not necessary to assess the agreement.

When visually assessing the agreement between two devices, we normally plot the data from New vs. the data from Old. However, in such plots there is no norm which can be used in the assessment. We have shown that our graphical methods helpful in assessing unequal variance.

References

Altman DG and Bland JM (1983). "Measurement in medicine: the analysis of method comparison studies." *Statistician*, 32: 307-17.

Kotz S and Johnson NL (1981). Encyclopedia of Statistical Sciences, Volume 7, Page 233. John Wiley & Sons, Inc.,

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.