Agreement and Individual Bioequivalence: A New Look

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

Ng and Hshieh (2016) proposed graphical approaches using scatter plot, Bland-Altman plot and Q-Q plot to assess (i) individual bioequivalence of a generic drug and (ii) interchangeability of a proposed biosimilar product to a reference product. In this paper, concordance correlation coefficient (Lin, 1989) is used instead of the plots. Simulated and real data are used to illustrate the new method.

Key Words: Generics, Biosimilars, individual bioequivalence, interchangeability, concordance correlation coefficient, permutation test.

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). Two other useful tools are scatter plot and Q-Q plot.

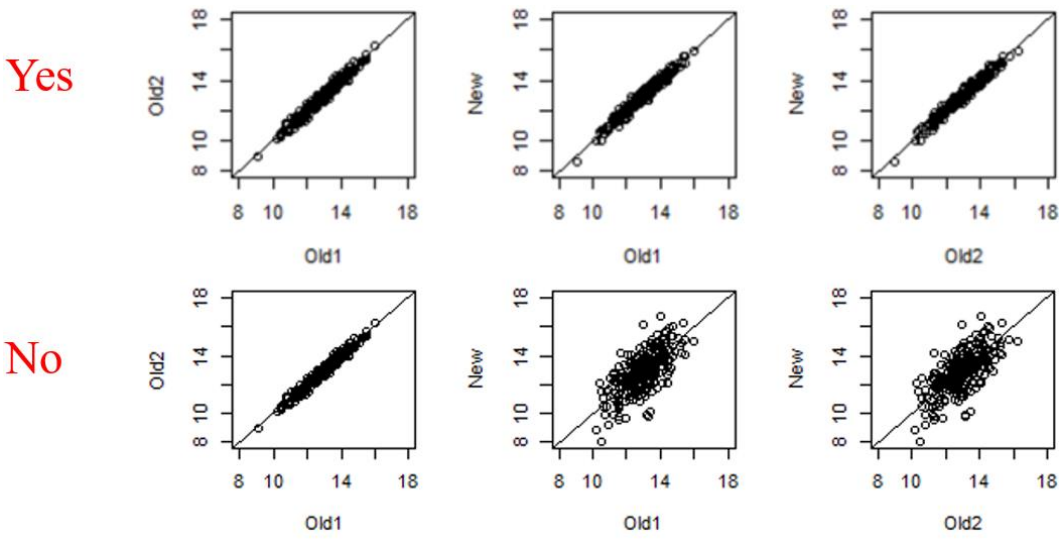
Hshieh and Ng (2015) proposed a new "procedure" in assessing the agreement of a new method to an old method using three graphical approaches: scatter plot, Bland-Altman plot and Q-Q plot. In this procedure, subjects are measured twice using the Old method (Old1, Old2) and once using the New method (New). 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 just like the old method, then the plots of New vs. Old1 and New vs. Old2 should each appear similar to the plots of Old1 vs. Old2.

Is the New method/product as good as the Old?

Figure 1 shows the use of this graphical approach using two simulated data sets to answer the following question: Is the New method/product as good as the Old? For the first data set (see the upper panel), the answer is yes. For the second data set (see the lower panel), the answer is no because there is a higher variability of the Old than the New, although there is no systematic bias.

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

Figure 1. Example: Scatter plot (simulated data)



Can we switch from Reference to Test?

Ng and Hshieh (2016) used similar approaches to assess (i) individual bioequivalence of a generic drug and (ii) interchangeability of a proposed biosimilar product to a Reference. Here, the Reference corresponds to the Old method and the generic drug (or the proposed biosimilar product) corresponds to the New method. They proposed the 2-sequence 3-period crossover design to assess switching from Reference (R) to Test (T) (see Table 1). The plots of R_2 vs. R_1 (or Old2 vs. Old1) is used as a norm. The plots of T vs. R_1 (or New vs. Old1) is to compare with the norm.

For illustration purposes, Ng and Hshieh (2016) extracted the data (Log-transformed AUC and C_{max}) from a bioequivalence study with 2-sequence 4-period crossover design (see Table 2) and the plots are shown in Figure 2. Visually, the two plots for each of the three types of plots look similar for Log(AUC) (see Figure 2a) but not for C_{max} (see Figure 2b).

Table 1. Study Design

	Period		
	1	2	3
Sequence 1	R_1	R_2	T
Sequence 2	R_1	T	R_2

R_1, R_2: Reference; T: Test

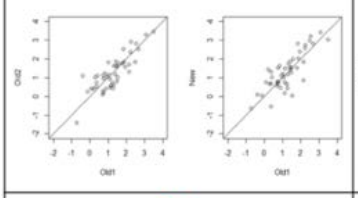
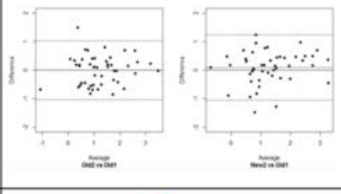
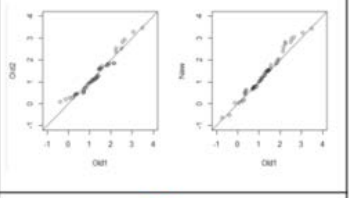
Table 2. Data Extraction

Study Design		# of Subjects		Data Extraction
		Enrolled	Completed	
Sequence 1	TRRT	27	23	TR_1R_2T
Sequence 2	RTTR	27	24	R_1TTR_2

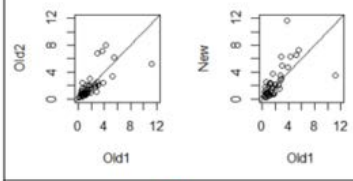
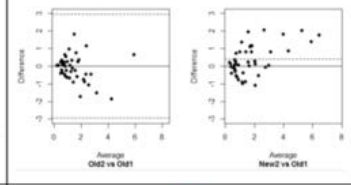
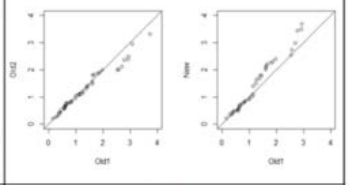
R_1, R_2: Reference; T: Test

Figure 2. Switching from Reference to Test

(a) Log-transformed AUC

Scatter Plot	Bland-Altman Plot	Q-Q Plot
		
Yes	Yes	Yes

(b) C_{max}

Scatter Plot	Bland-Altman Plot	Q-Q Plot
		
No	No	No

A New Look

Although the graphic is a useful tool, it is very subjective. To overcome this deficiency, use Concordance Correlation Coefficient (CCC, Lin, 1989) instead of the plots. To do so, let

$CCC(R_2R_1)$ denotes the CCC between R_2 and R_1, and
$CCC(TR_1)$ denotes the CCC between T and R_1.

If T is just "like" R (or R_2), then $CCC(TR_1)$ should be close to $CCC(R_2R_1)$, where $CCC(R_2R_1)$ is used as a norm. A test statistic

$$S = \log_{10}[CCC(TR_1)/CCC(R_2R_1)]$$

may be used to test the null hypothesis that

H_0: T is just "like" R,

against the alternative hypothesis that

H_1: T is different from R.

Derivation of the null distribution may be very complicated. Therefore, a permutation test is proposed instead. To do so, we determine the distribution of S under 2^n permutations of each pair of (R_2, T), where n is the sample size. The p-value is then calculated as

> p-value = (number of permutations such that S is as extreme as s, the observed value of S)/2^n.

The null hypothesis will be rejected if p-value is less than 0.05. A simulated data and real data are used to illustrate this method.

Note that one cannot conclude agreement (or equivalence) when we fail to reject H_0. It is being discussed here as a work in progress toward developing the equivalence testing that would be appropriate (see the last section).

Simulated Data

The data is simulated 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,
ε_{ij}: Measurement error of the i^{th} subject by 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, where μ_j where is the systematic bias for the j^{th} rater. Finally, X_i and ε_{ij} are assumed to be mutually independent.

With k = 3 and a different notation for Y, a data set is simulated as follows:

X: $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$; $n = 20$, $\mu = 13$, $\sigma = 1.5$,
R_1: $r_{1i} = x_i + \varepsilon_{1i}$, $\varepsilon_{1i} \sim N(0,1)$, $i = 1,\ldots, n$,
R_2: $r_{2i} = x_i + \varepsilon_{2i}$, $\varepsilon_{2i} \sim N(0,1)$, $i = 1,\ldots, n$,
T: $t_i = x_i + \varepsilon_{ti}$, $\varepsilon_{ti} \sim N(\mu_t, \sigma_t^2)$, $i = 1,\ldots, n$,

where x's and ε's are mutually independent. There are two scenarios for T:

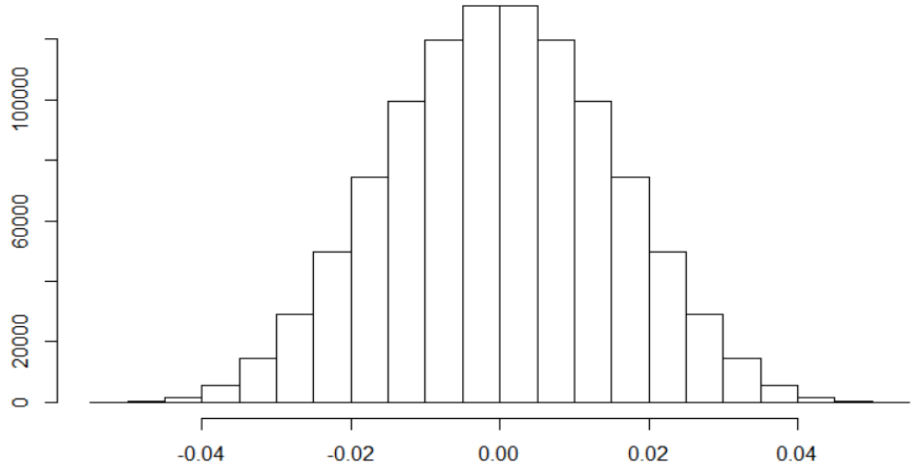
(a) Under H_0: $\mu_t = 0$, $\sigma_t^2 = 1$
(b) Under H_1: $\mu_t = 1$, $\sigma_t^2 = 4$

Scenario (a) is under the null hypothesis that T is just "like" R and Scenario (b) is under the alternative hypothesis that T is different from R with a systematic bias and higher variability. The permutation distributions (2^{20} = 1,048,576 permutations) of S, together with the observed s and p-value, under scenarios (a) and (b) are given in Figures 3a and

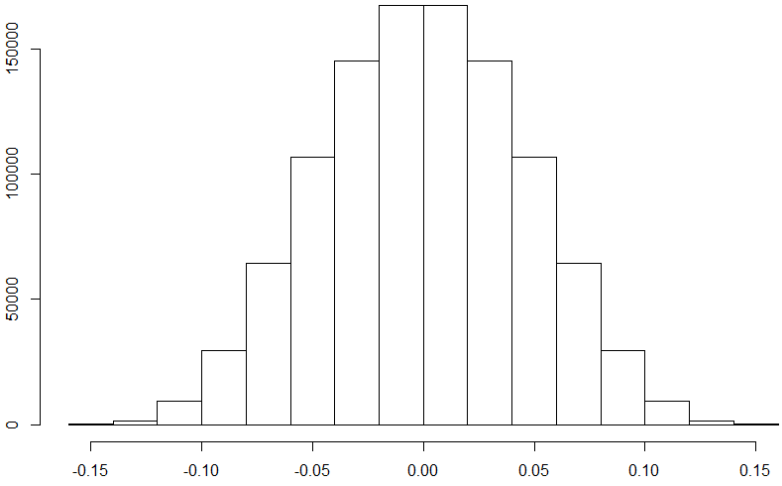
3b, respectively. So, we fail to reject the null hypothesis under scenario (a) (p = 0.89), while the null hypothesis is rejected under scenario (b) (p = 0.0094).

Figure 3. Permutation Distribution of S

(a) Under the Null Hypothesis: s = -0.002237, p = 0.89



(b) Under the Alternative Hypothesis: s = 0.1097, p = 0.0094



An application to a Bioequivalence Study

The new method is applied to a data set extracted from a bioequivalence study (see Table 2). The sample size is 47. The number of all possible permutations is huge (over 140 trillion). So, in this application, 100,000 permutations are randomly selected. The simulated permutation distributions (100,000 permutations) of S, together with the observed s and p-value, based on $\log_{10}$(AUC) and C_{max} are given in Figures 4a and 5b, respectively. We fail to reject the null hypothesis for both variables (p > 0.05).

Figure 4a. Permutation Distribution of S: $Log_{10}(AUC)$

s = 0.017; p = 0.33

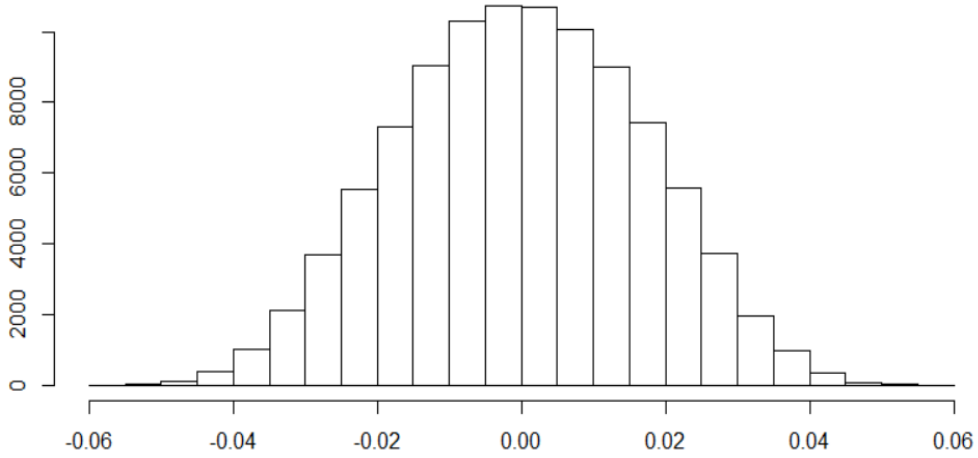
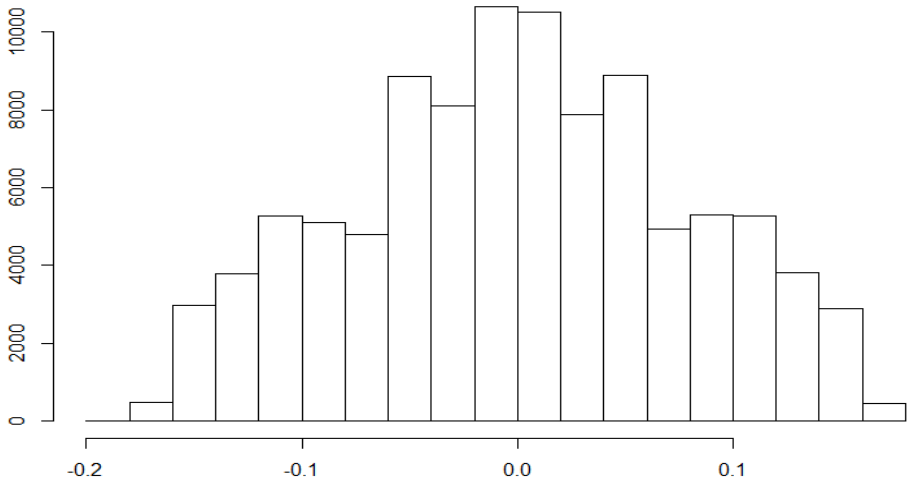


Figure 4b. Permutation Distribution of S: C_{max}

s = 0.071; p = 0.39



Discussion and Further Research

Since failing to reject H_0 does not mean 'T is just "like" R', the hypotheses are formulated as follows:

H_0*: T is different from R by a "lot"
H_1*: T is not too much different from R

One needs to define "not too much different" in terms of mean difference and the ratio of the two standard deviations. To test H_0*, construct the confidence region for these two parameters as

$\{(m, v) \mid H_0$ is not rejected when T is transformed to $v(T - m)\}$.

We then reject H_0* if the confidence region is within "not too much different".

References

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

Hshieh, P., and Ng, T-H. (2015). "Assessing Agreement: A Graphical Approach" In *JSM Proceedings*, Biopharmaceutical Section. Alexandria, VA: American Statistical Association.

Lin, L. (1989) A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45:255-268.

Ng, T-H., and Hshieh, P. (2016). "Individual Bioequivalence: A Graphical Approach" Poster presentation at ASA (American Statistical Association) Biopharmaceutical Section Regulatory-Industry Statistics Workshop.

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.