OPEN ACCESS Freely available online

PLOS ONE

Sample Size Determination for Individual Bioequivalence Inference

Chieh Chiang[1,3☯], **Chin-Fu Hsiao**[1,3☯], **Jen-Pei Liu**[1,2,3*]

1 Division of Biometry, Department of Agronomy, National Taiwan University, Taipei, Taiwan, **2** Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan, **3** Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Abstract

Statistical criterion for evaluation of individual bioequivalence (IBE) between generic and innovative products often involves a function of the second moments of normal distributions. Under replicated crossover designs, the aggregate criterion for IBE proposed by the guidance of the U.S. Food and Drug Administration (FDA) contains the squared mean difference, variance of subject-by-formulation interaction, and the difference in within-subject variances between the generic and innovative products. The upper confidence bound for the linearized form of the criterion derived by the modified large sample (MLS) method is proposed in the 2001 U.S. FDA guidance as a testing procedure for evaluation of IBE. Due to the complexity of the power function for the criterion based on the second moments, literature on sample size determination for the inference of IBE is scarce. Under the two-sequence and four-period crossover design, we derive the asymptotic distribution of the upper confidence bound of the linearized criterion. Hence the asymptotic power can be derived for sample size determination for evaluation of IBE. Results of numerical studies are reported. Discussion of sample size determination for evaluation of IBE based on the aggregate criterion of the second moments in practical applications is provided.

Citation: Chiang C, Hsiao C-F, Liu J-P (2014) Sample Size Determination for Individual Bioequivalence Inference. PLoS ONE 9(10): e109746. doi:10.1371/journal.pone.0109746

Editor: Shyamal D. Peddada, National Institute of Environmental and Health Sciences, United States of America

Received May 30, 2014; **Accepted** September 11, 2014; **Published** October 13, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: This research was supported by a grant from Taiwan's National Science Council (NSC 101-2118-M-002-002-MY2) to JPL. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: jpliu@ntu.edu.tw

☯ These authors contributed equally to this work.

Introduction

The traditional criterion for evaluation and approval of small-molecular chemical generic drug products is based on average bioequivalence (ABE). ([1] – [4]) On the other hand, biosimilar drugs and most of targeted drugs are biological products which are fundamentally different from traditional small-molecular chemical generic drugs in size, functional structure, physiochemical properties, impurities, immunogenicity and manufacturing processes. However, ABE considers only equivalence between population means and completely ignores the variability of the drug products and that of formulation effects between patients. Therefore, ABE is not an adequate criterion for evaluation of the generic copies of targeted drugs and biosimilar drug products. On the other hand, individual bioequivalence (IBE) simultaneously takes differences in population means, subject-by-formulation interaction, and within-subject variability into account. ([1], [4]) As a result, IBE may be more appropriate for evaluation of generic targeted drugs and biosimilar products. ([5], [6]).

The U.S. Food and Drug Administration (FDA) Guidance for Industry "*Statistical Approaches to Establishing Bioequivalence*" recommends replicated crossover designs for IBE studies [1]. The linearized criterion for IBE evaluation suggested in the U.S. FDA guidance is the linear combination of the squared mean difference, variance of subject-by-formulation interaction, and the difference in within-subject variances between the generic and innovative products. The U.S. FDA guideline proposes the upper confidence bound for the linearized form of the IBE criterion derived by the modified large sample (MLS) method as a testing procedure for evaluation of IBE. In other words, generic and innovative products are claimed to be IBE if the MLS $100(1-\alpha)\%$ upper confidence bound of the linearized criterion is less than zero. Despite a vast literature on various methodologies for evaluation of IBE, literature on sample size determination for evaluation of IBE is scarce. Under the two-sequence and four period (2×4) crossover design, we derive the asymptotic distribution of the MLS $100(1-\alpha)\%$ upper confidence bound and the asymptotic power for sample size determination for the IBE evaluation. Our approach is to determine the sample size to provide the asymptotic power for which the MLS $100(1-\alpha)\%$ upper confidence bound for the IBE criterion smaller than zero is greater than $1-\beta$.

In the next section, the method for construction of the MLS upper confidence bound for the IBE criterion for the 2×4 crossover design is reviewed. Our proposed methods of sample size determination for IBE evaluation based on the asymptotic distribution of the MLS upper confidence bound are then presented. The results of numerical studies, including numerical examples and simulation studies, are provided in the next section. Numerical examples illustrate applications of our proposed

method in practical scenarios. Simulation studies were conducted to investigate the impact of magnitudes of means differences, variance of subject-by-formulation interaction, and within-subject variances on sample sizes. In addition, empirical powers obtained from simulation studies are compared with the asymptotic powers to examine whether the sample sizes determined by our proposed methods can provide sufficient power. Discussion and final remarks are given in the last section.

Methods

Criterion for Individual Bioequivalence

In what follows, unless otherwise specified, all parameters and statistics are on the log-scale. Let μ_T and μ_R be the mean for test (generic product) and reference (innovative product) formulations, respectively. In addition, σ^2_{WT} and σ^2_{WR} denote the within-subject variance for the test and reference formulation, respectively, and let σ^2_D be the variance of the subject-by-formulation interaction. The IBE criterion [1,4,7] is defined as

$$\theta=\frac{(\mu_T-\mu_R)^2+\sigma^2_D+\sigma^2_{WT}-\sigma^2_{WR}}{\max\{\sigma^2_{W0},\sigma^2_{WR}\}}, \tag{1}$$

where σ^2_{W0} is the specified constant within-subject variance, which the U.S. FDA guidance suggests that it be set at 0.04 [1]. Based on the IBE criterion given in Equation (1), the null hypothesis of non-IBE and the alternative hypothesis of IBE are respectively given as

$$H_0:\theta\geq\theta_0 \text{ vs. } H_a:\theta<\theta_0, \tag{2}$$

where θ_0 is the upper limit of the IBE criterion, which is set as 2.4948 in the U.S. FDA guidance [1].

Hyslop et al [5] suggested the following linearized IBE criterion for assessment of IBE:

$$\begin{aligned}\eta=&(\mu_T-\mu_R)^2+\sigma^2_D+\sigma^2_{WT}\\&-\sigma^2_{WR}-\theta_0\max\{\sigma^2_{W0},\sigma^2_{WR}\}.\end{aligned} \tag{3}$$

To avoid direct estimation of σ^2_D, the linearized IBE criterion in Equation (3) can be re-expressed as [8].

$$\eta=\delta^2+\sigma^2_{a,b}+(1-a)\sigma^2_{WT}-(1+b)\sigma^2_{WR}-\theta_0\max\{\sigma^2_{W0},\sigma^2_{WR}\},$$

where $\delta=\mu_T-\mu_R$, $\sigma^2_{a,b}=\sigma^2_D+(a\sigma^2_{WT}+b\sigma^2_{WR})$ for $a>0$ and $b>0$. For the 2×4 crossover design, a = b = 0.5. Hence the linearized IBE criterion becomes

$$\eta=\delta^2+\sigma^2_{0.5,0.5}+0.5\sigma^2_{WT}-1.5\sigma^2_{WR}-\theta_0\max\{\sigma^2_{W0},\sigma^2_{WR}\}.$$

When $\sigma^2_{WR}\geq(<)\sigma^2_{Wo}$, η is referred to as the linearized reference (constant)-scaled criterion. Consequently, the IBE hypotheses given in Equation (2) can be reformulated using the linearized criterion as follows:

$$H_0:\eta\geq0 \text{ vs. } H_a:\eta<0. \tag{4}$$

Upper Confidence Bound by the Modified Large Sample (MLS) Method

Under the 2×4 crossover design (TRTR, RTRT) given in File S1, a MLS $100(1-\alpha)\%$ upper confidence bound [7] is given as

$$\hat{\tau}=\hat{\eta}+\sqrt{U}, \tag{5}$$

where $\hat{\eta}=\hat{\delta}^2+\hat{\sigma}^2_{0.5,0.5}+\frac{1}{2}\hat{\sigma}^2_{WT}-\left(\frac{3}{2}+\theta_0\cdot\phi\right)\hat{\sigma}^2_{WR}-\theta_0(1-\phi)\sigma^2_{W0}$, and

$$\begin{aligned}U=&\left[\left(|\hat{\delta}|+t_{1-\alpha,2(n-1)}\sqrt{\frac{1}{2n}\hat{\sigma}^2_{0.5,0.5}}\right)^2-\hat{\delta}^2\right]^2\\&+\hat{\sigma}^4_{0.5,0.5}\left(\frac{2(n-1)}{\chi^2_{\alpha,2(n-1)}}-1\right)^2+\frac{1}{4}\hat{\sigma}^4_{WT}\left(\frac{2(n-1)}{\chi^2_{\alpha,2(n-1)}}-1\right)^2\\&+\left(\frac{3}{2}+\phi\cdot\theta_0\right)^2\hat{\sigma}^4_{WR}\left(\frac{2(n-1)}{\chi^2_{1-\alpha,2(n-1)}}-1\right)^2.\end{aligned} \tag{6}$$

with $\phi=1$ if $\sigma^2_{WR}\geq\sigma^2_{W0}$ and $\phi=0$ if $\sigma^2_{WR}<\sigma^2_{W0}$. Here n is the sample size (the number of subjects) per sequence, and $t_{p,r}$ and $\chi^2_{p,r}$ are the $100p$th percentiles of the central t and central chi-square distributions, respectively, with r degrees of freedom. Estimators $\hat{\delta}$, $\hat{\sigma}^2_{0.5,0.5}$, $\hat{\sigma}^2_{WT}$, $\hat{\sigma}^2_{WR}$ and derivation of the MLS $100(1-\alpha)\%$ upper confidence bound for the linearized IBE criterion are given in File S1. The null hypothesis is rejected and the IBE is concluded at the α significance level if the MLS $100(1-\alpha)\%$ upper confidence bound given in Equation (5) is less than zero.

Sample Size Determination

By the delta method, $\hat{\tau}$ is asymptotically normal with mean $\mu_{\hat{\tau}}$ and variance $\sigma^2_{\hat{\tau}}$. Proof of the asymptotic normality of $\hat{\tau}$ and derivations of $\mu_{\hat{\tau}}$ and $\sigma^2_{\hat{\tau}}$ are given in File S2 [8,9]. Let $\tilde{\mu}_{\hat{\tau}}$ and $\tilde{\sigma}^2_{\hat{\tau}}$ be some specified values of $\mu_{\hat{\tau}}$ and $\sigma^2_{\hat{\tau}}$ respectively in the alternative hypothesis. An asymptotic power based on the MLS upper confidence bound using the normal distribution can be computed as

$$P(\hat{\tau}<0)=P\left(\hat{\tau}<\tilde{\mu}_{\hat{\tau}}+z_{1-\beta}\tilde{\sigma}_{\hat{\tau}}\right), \tag{7}$$

where z_p is the $100p$th percentile of standard normal distribution. Based on the mean value theorem, the derivatives of $\tilde{\mu}_{\hat{\tau}}$ and $\tilde{\sigma}_{\hat{\tau}}$ with respect to n for a small constant Δ are given as

$$\frac{d}{dn}\tilde{\mu}_{\hat{\tau}}(n)=\frac{\tilde{\mu}_{\hat{\tau}}(n+\Delta)-\tilde{\mu}_{\hat{\tau}}(n)}{\Delta} \text{ and } \frac{d}{dn}\tilde{\sigma}_{\hat{\tau}}(n)=\frac{\tilde{\sigma}_{\hat{\tau}}(n+\Delta)-\tilde{\sigma}_{\hat{\tau}}(n)}{\Delta}.$$

It follows that the smallest n can be derived as n converges at the (l+1)th iteration, where

$$n^{l+1}=n^l-\frac{\tilde{\mu}_{\hat{\tau}}(n^l)+z_{1-\beta}\tilde{\sigma}_{\hat{\tau}}(n^l)}{\frac{d}{dn}\tilde{\mu}_{\hat{\tau}}(n^l)+z_{1-\beta}\frac{d}{dn}\tilde{\sigma}_{\hat{\tau}}(n^l)}.$$

Since Equation (7) is derived directly from the asymptotic power, there exists only one solution for sample size determination with respect to the required power. Equation (7) can be evaluated by the numerical method. File S3 provides a SAS macro in PROC NLP (nonlinear programming) by the quasi-Newton method. This SAS macro is flexible to allow users to specify the significance level, the required power, the upper IBE equivalence limit, σ^2_{W0}, and the mean difference, the variance of subject-by-formulation interaction and the within-subject variances for the test and reference formulations.

Simulation Setup

The first objective of simulation studies is to determine the sample sizes for the nominal 80% power at the 5% significance level under different specifications for various combinations of parameters under the 2×4 crossover design (TRTR, RTRT). The second objective is to investigate the impact of magnitudes of means differences, variance of subject-by-formulation interaction, and within-subject variances on sample size The third objective is to compare the empirical power obtained from simulation studies with the asymptotic power obtained by Equation (7) and the nominal power of 80%. Because there are four parameters, a four-factor factorial simulation study with three levels for each factor was employed. Simulation studies were performed separately for the constant-scaled criterion and reference-scaled criterion. Four levels of the within-subject reference variance were used for the reference-scaled criterion. It follows that 3×3×3×3 and 3×3×4×3 factorial simulation studies were employed in simulation studies for the constant-scaled and reference-scaled criteria, respectively. The values of mean difference are set to be 0, 0.05, and 0.1. For the constant-scaled criterion, the magnitudes of the reference within-subject variance are specified to be 0.01, 0.02, and 0.03. They are 0.04, 0.09, 0.16, and 0.25 for the reference-scaled criterion. In order to investigate the impact of an increasing or reduction of the test within-subject variance on the sample size, the differences in the magnitude of the within-subject variance between the test and reference formulations are set to be −0.005, 0, and 0.005 for the constant-scaled criterion and −0.02, 0, and 0.02 for the reference-scaled criterion. The values of the variance of the subject-by-formulation interaction were selected in proportion to the magnitude of the within-subject variances. They are set to be 0.0001, 0.001, and 0.0225.

Table 1 provides the specifications of various combinations of the four parameters. The sample size for each of a total of 189 combinations given in Table 1 was determined by the proposed method. Under the model of the 2×4 crossover design in Equation (S1.1) in File S1, 10,000 random samples are generated according to the sample size obtained by the proposed method and the specification of the magnitudes for a particular combination of parameters. The MLS $100(1-\alpha)$% upper confidence bound for the IBE linearized criterion is then computed for each generated random sample, according to Equation (5). The empirical power is calculated as the proportion of the random samples with the MLS $100(1-\alpha)$% upper confidence bounds smaller than zero. For 10,000 random samples, it implies that the 95% of the empirical powers would be greater than 0.7934 if the sample size obtained by the proposed method can provide sufficient power with respect to the nominal power of 80%.

Results

Numerical Examples

For the purpose of illustration, examples of sample size determination under the 2×4 crossover design (TRTR, RTRT) for both the linearized constant-scaled criterion and reference-scaled criterion are provided. Under the linearized constant-scaled criterion, the specifications of the parameters for the sample size are $\delta=0.1$, $\sigma^2_D=0.0225$ and $\sigma^2_{WT}=\sigma^2_{WR}=0.03$. It follows that

$$\begin{aligned}\sigma^2_{0.5,0.5}&=\sigma^2_D+(0.5\sigma^2_{WT}+0.5\sigma^2_{WR})\\&=0.0225+(0.03+0.03)/2=0.0525\end{aligned}$$

and

$$\begin{aligned}\eta&=\delta^2+\sigma^2_{0.5,0.5}+0.5\sigma^2_{WT}-1.5\sigma^2_{WR}\\&\quad-\theta_0\max\{\sigma^2_{W0},\sigma^2_{WR}\}\\&=0.1^2+0.0525+0.03/2-3/2\times0.03\\&\quad-2.4948\times0.04\approx-0.0673.\end{aligned}$$

Using the SAS macro given in File S3, the sample size for the nominal power of 80% at the 5% significance level is 16 subjects per sequence. Since the asymptotic mean in Equation (S2.5) in File S2 is $\tilde{\mu}_{\tilde{\tau}}(16)=-0.0240$ and variance in Equation (S2.6) of File S2 is $\tilde{\sigma}^2_{\tilde{\tau}}(16)=0.0007$, the corresponding asymptotic power in Equation (7) is given as

$$P\left(Z<\frac{0+0.0240}{\sqrt{0.0007}}\right)=P(Z<0.9071)=0.8178.$$

Suppose that both δ and σ^2_D are kept the same and both σ^2_{WT} and σ^2_{WR} increase to 0.05. Since $\sigma^2_{WR}=0.05>0.04=\sigma^2_{W0}$, the linearized reference-scaled criterion is used. It follows that

$$\sigma^2_{0.5,0.5}=0.0225+(0.05+0.05)/2=0.0725$$

and

$$\begin{aligned}\eta&=0.1^2+0.0725+0.05/2\\&\quad-(3/2+2.4948)\times0.05\approx-0.0922.\end{aligned}$$

Under this scenario, the sample size is 29 subjects per sequence for the nominal power of 80% at the 5% significance level with $\tilde{\mu}_{\tilde{\tau}}(29)=-0.0303$, $\tilde{\sigma}^2_{\tilde{\tau}}(29)=0.0013$, and an asymptotic power

$$P\left(Z<\frac{0+0.0303}{\sqrt{0.0013}}\right)=P(Z<0.8404)=0.7997.$$

If σ^2_{WT} increases to 0.07 and σ^2_{WR} remains as 0.05, the sample size per sequence is increased to 53 with $\eta\approx-0.0722$, $\tilde{\mu}_{\tilde{\tau}}(53)=-0.0242$, $\tilde{\sigma}^2_{\tilde{\tau}}(53)=0.0008$ and an asymptotic power of 0.8039. However if σ^2_{WT} decreases to 0.03 and σ^2_{WR} remains the same, the sample size is reduced to 18 subjects per sequence with $\eta\approx-0.1122$, $\tilde{\mu}_{\tilde{\tau}}(18)=-0.0374$, $\tilde{\sigma}^2_{\tilde{\tau}}(18)=0.0019$ and the corresponding asymptotic power 0.8046. Therefore, if the test

Table 1. Specifications of parameters for simulation studies.

Parameters	Specifications
δ	0, 0.05, 0.1
σ_D^2	0.0001, 0.01, 0.0225
Constant-scaled criterion	
σ_{WR}^2	0.01, 0.02, 0.03
$\sigma_{WT}^2=\sigma_{WR}^2-0.005$	0.005, 0.015, 0.025
$\sigma_{WT}^2=\sigma_{WR}^2$	0.01, 0.02, 0.03
$\sigma_{WT}^2=\sigma_{WR}^2+0.005$	0.015, 0.025, 0.035
Reference-scaled criterion	
σ_{WR}^2	0.04, 0.09, 0.16, 0.25
$\sigma_{WT}^2=\sigma_{WR}^2-0.02$	0.02, 0.07, 0.14, 0.23
$\sigma_{WT}^2=\sigma_{WR}^2$	0.04, 0.09, 0.16, 0.25
$\sigma_{WT}^2=\sigma_{WR}^2+0.02$	0.06, 0.11, 0.18, 0.27

doi:10.1371/journal.pone.0109746.t001

formulation has a better quality by reducing the within-subject variability, fewer subjects are required for evaluation of IBE under the 2×4 crossover design.

Results of Simulation Studies

Figure 1 provides a graphic 3×3 presentation of the sample sizes of all 81 combinations considered for the linearized constant-scaled criterion. The three vertical panels are arranged by the mean difference in an ascending order from left to right. The three horizontal panels are presented by the within-subject variance of the test formulation in a descending order from top to bottom. For each of the nine cells, the vertical axis is the sample size, while the horizontal axis is the within-subject variance of the reference formulation. Three lines within each cell represent the sample sizes obtained from different values of the variance of subject-by-formulation interaction.

Figure 1 and Table S1 reveal that sample size ranges from 3 to 19 subjects per sequence for all 81 combinations considered under the linearized constant-scaled criterion and the 2×4 crossover design. However, the linearized constant-scaled criterion in Equation (3) is an increasing function of mean difference, variance of the subject-by-formulation interaction, and the difference in within-subject variances between the test and reference formulations. Figure 1 reveals that the sample size is also an increasing function of mean difference and variance of the subject-by-formulation interaction. For our simulation studies, the difference in within-subject variances between the test and reference formulations is set to be -0.005, 0, and 0.005. It follows that the linearized constant-scaled criterion is a function only of mean difference and the variance of the subject-by-formulation as long as the difference in within-subject variances between the test and reference formulations is a constant. In other words, since $\theta_0\sigma_{W0}^2$ is a constant, η, as shown in Table S1, is also a constant for any fixed specification of δ and σ_D^2. However, Figure 1 also reveals that sample size increases as the reference within-subject variance σ_{WR}^2 increases. This phenomenon may be due to the fact that the upper confidence bound in Equation (5) is an increasing function of the estimated within-subject variance of the reference formulation $\hat{\sigma}_{WR}^2$. On the other hand, a reduction of sample size can be achieved if the within-subject variance of the test formulation is smaller than that of the reference formulation. Otherwise, more subjects are required.

Sample sizes of all 108 combinations for the linearized reference-scaled criterion are also presented in a 3×3 graphical display in Figure 2. Sample sizes given in Figure 2 and Table S2 for the linearized reference-scaled criterion range from 8 to 84 per sequence. As a result, the range of the sample sizes for the linearized reference-scaled criterion is much wider that those of the linearized constant-scaled criterion because σ_{WR}^2 is confined to a narrow range between 0 and 0.04 for the constant-scaled criterion. Similar to the results of the linearized constant-scaled criterion, the sample size for the linearized reference-scaled criterion is an increasing function of mean difference and variance of the subject-by-formulation interaction, and fewer subjects are needed when the within-subject variance of the test formulation is smaller than that of the reference formulation.

However, a striking difference in the trend of sample sizes between Figure 1 and Figure 2 is that except for the specification when $\sigma_{WR}^2=\sigma_{W0}^2=0.04$, the sample size for the linearized reference-scaled criterion is a deceasing function of the within-subject variance of the reference formulation as depicted in Figure 2. This is due to the fact that η is a decreasing function of σ_{WR}^2. Except for the specification of $\sigma_{WR}^2=\sigma_{W0}^2=0.04$, Table S2 shows that η deceases from -0.2044 to -0.6436. On the other hand, the range of η for the linearized constant-scaled criterion is only from -0.0623 to -0.1047, as given in Table S1. For any fixed specification of δ and σ_D^2, the maximum of η occurs when $\sigma_{WR}^2=\sigma_{W0}^2=0.04$. As a result, as shown in Figures 1 and 2 and Tables S1 and S2, when $\sigma_{WR}^2=\sigma_{W0}^2=0.04$, the required sample size per sequence is the largest for any fixed specification of δ and σ_D^2.

Tables S1 and S2 also provide the asymptotic and empirical powers for a total of 189 combinations. Only 2 of the 189 empirical powers (1.05%) are below 0.7934. This demonstrates that with respect to the nominal power of 80%, the sample size obtained by our proposed method can provide sufficient power for evaluation of IBE under the 2×4 crossover design. Because of a narrow range of η and σ_{WR}^2, 60 of 81 sample sizes (84.5%) for the linearized constant-scaled criterion are smaller than 10. Due to the discrete nature of the sample size, both asymptotic and empirical powers are from 0.8107 to 0.9598, which are larger than the

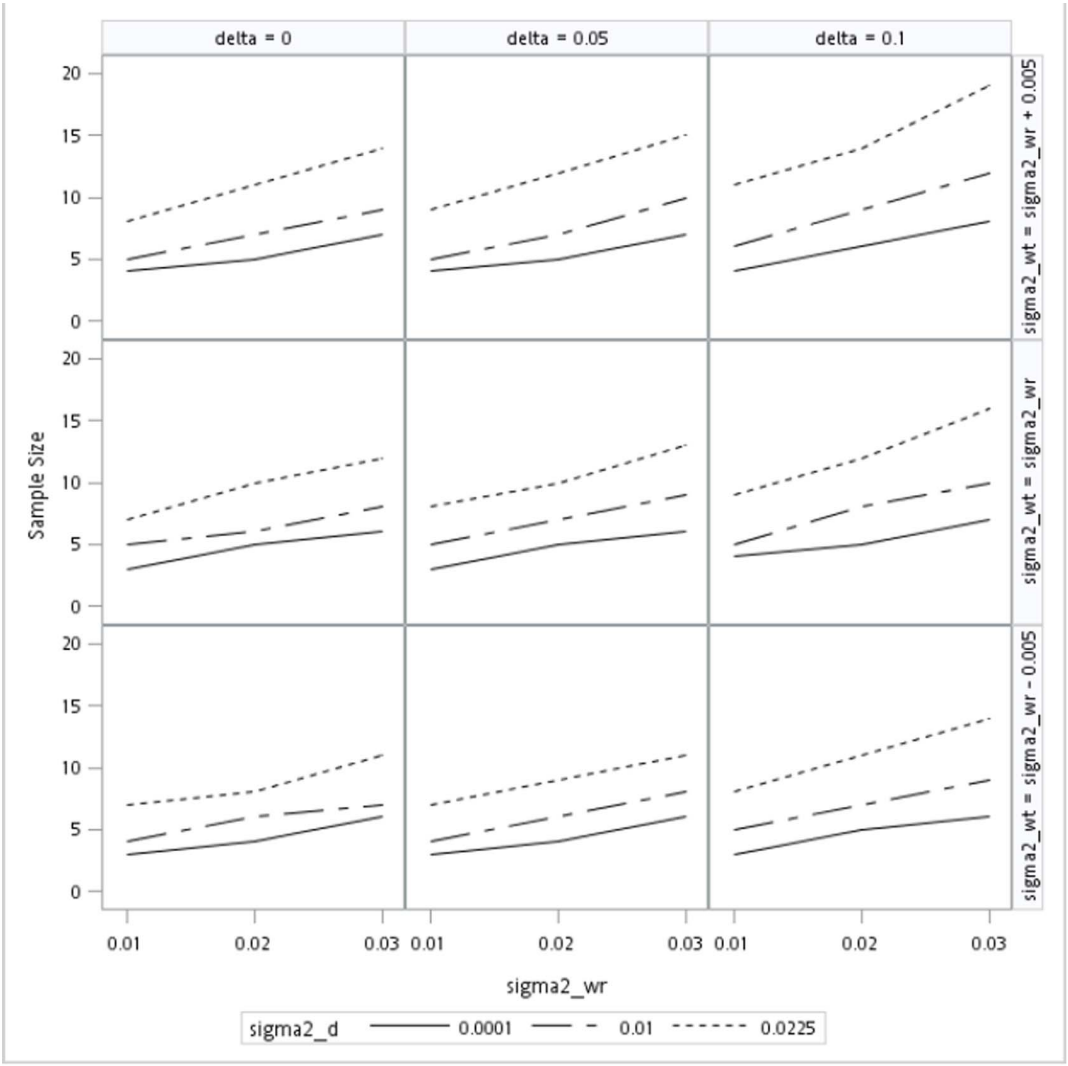

Figure 1. Sample size per sequence for responses with $\alpha = 0.05$ and power = 0.8 under constant-scaled criterion.
doi:10.1371/journal.pone.0109746.g001

nominal power of 80%. On the other hand, for the linearized reference-scaled criterion, only 2 out of 108 sample sizes (1.85%) are below 10. Consequently, the range of 108 empirical powers is from 0.7905 to 0.8289. It follows that with respect to a nominal power of 80%, the sample size obtained by the proposed method for the linearized reference-scaled criterion provides neither insufficient nor excessive power. Moreover, the maximum of absolute differences between empirical power and asymptotical power is 0.0258. In addition, only 29 of 189 absolute differences (15.3%) are greater than 0.01. This shows that the asymptotic power used for the sample size determination by the proposed method is quite accurate, as verified by the empirical power obtained by simulation studies.

Discussion

Although the upper confidence bound constructed by the MLS method for the linearized criterion has been used for evaluation of IBE, literature on analytical determination of sample size is scarce. We showed that the MLS upper confidence bound converges asymptotically to a normal distribution. Hence, we propose an analytical procedure for sample size determination for evaluation of IBE based on the approximate power derived from the asymptotic normal distribution of the MLS upper confidence bound of the linearized criterion under the 2×4 crossover design. Extensive simulation studies show that the sample sizes obtained by our proposed method can provide sufficient and yet not excessive power. In addition, the results of simulation studies also reveal that the approximation of the asymptotic power is quite accurate, as verified by the empirical power. Simulation studies also investigated the impact of magnitudes of the four parameters on sample sizes. Our numerical studies found that smaller sample sizes can be obtained if the within-subject variance of the reference formulation is less than 0.04 or the within-subject variance of the test formulation is smaller than that of the reference formulation.

For any fixed specification of δ, σ_D^2 and σ_{WT}^2, η is a decreasing function of σ_{WR}^2. However, the decreasing rate for the linearized constant-scaled criterion is -1 in a narrow range from 0 to 0.04 with a constant constraint of $\theta_0\sigma_{W0}^2$. On the other hand, η for the linearized reference-scaled criterion has a much faster decreasing rate of $-(1+\theta_0\sigma_{WR}^2)$. Therefore, the maximum sample size for evaluation of IBE occurs when the within-subject

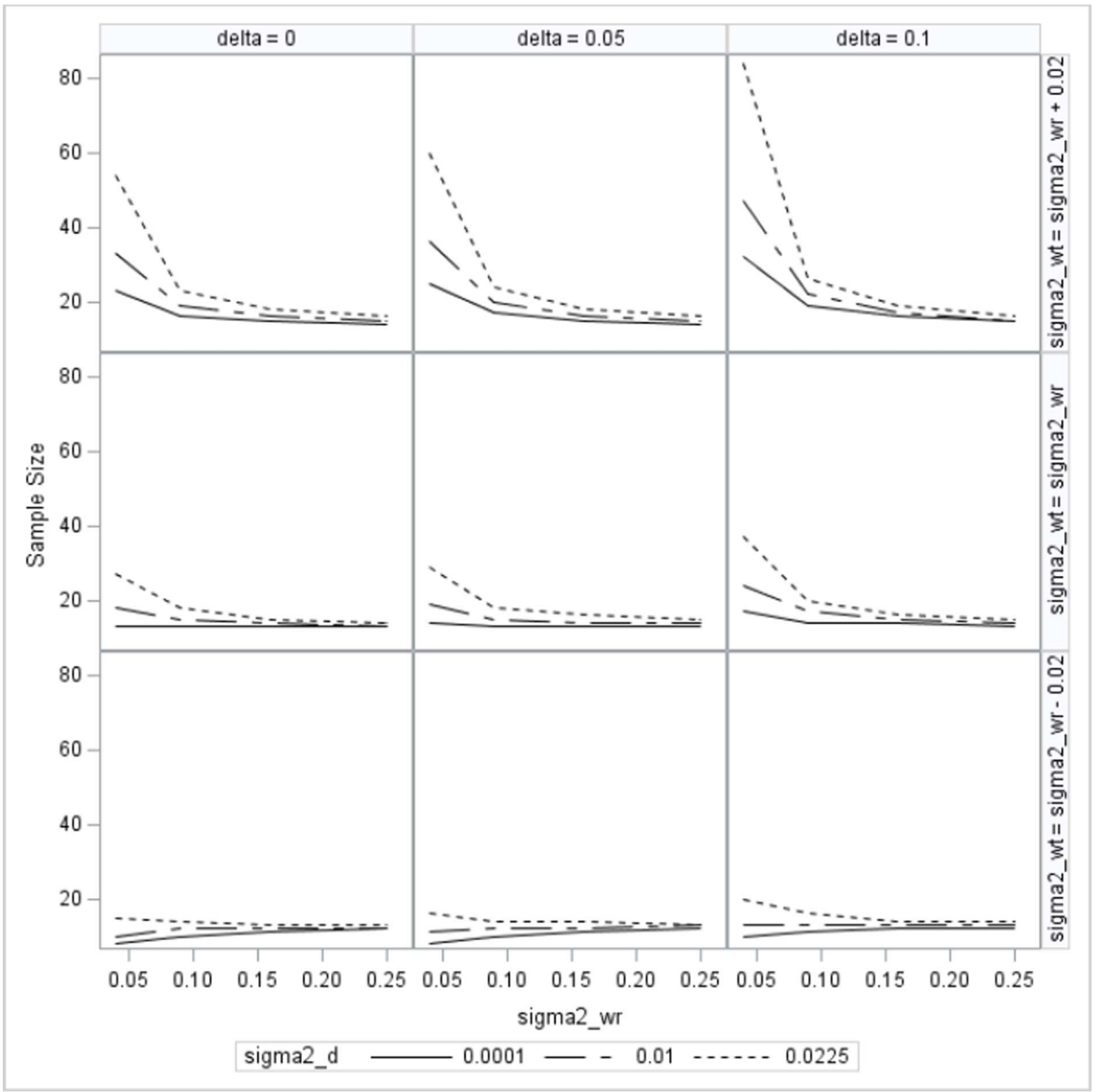

Figure 2. Sample size per sequence for responses with $\alpha = 0.05$ and power = 0.8 under reference-scaled criterion.
doi:10.1371/journal.pone.0109746.g002

variance of the reference formulation is at the boundary point of σ^2_{W0}.

The objective of the specified constant within-subject variance σ^2_{W0} in the constant-scaled criterion is to avoid a larger upper confidence bound of the IBE criterion when the reference product exhibits extremely small within-subject variability to prevent approval of any generic products. Sample sizes of all 81 combinations for the linearized constant-scaled criterion are smaller than 20 per sequence. This demonstrates that the IBE evaluation by the constant-scaled criterion can be accomplished with a reasonable sample size with respect to a nominal power of 80% at the 5% significance level if the within-subject variance of the reference formulation is smaller than 0.04. On the other hand, when $\sigma^2_{WR} > 0.04$, the sample size is a decreasing function of σ^2_{WR}. This contradicts the usual intuition that a larger variability requires a larger sample size.

The proposed method can also be easily adapted to other crossover designs such as the 2×3 crossover design (TRT, RTR). Table S3 compares the sample sizes required between the 2×3 crossover design (TRT, RTR) and the 2×4 crossover design (TRTR, RTRT) for a nominal power of 80% at the 5% significance level. Table S3 reveals that the number of subjects required for the 2×3 crossover design increases from 71% to 107% over that required by the 2×4 crossover design. Each subject in the 2×3 crossover design yields 3 observations per subject as compared to 4 observations per subject by the 2×4 crossover designs. However, the total number of observations for the 2×3 crossover design is still greater than that of the 2×4 crossover design. Therefore, although the duration of the 2×3 crossover design is shorter, the 2×4 crossover design is still more efficient for evaluation of IBE than the 2×3 crossover design.

In practice, one of the key issues is selection of the reference-scaled criterion or constant-scaled criterion for evaluation of IBE.

Three methods have been proposed. The first method is referred as to the estimation method (EST) suggested by Hyslop et al. [7]. The estimation method recommends using the reference-scaled criterion or constant-referenced criterion according to $\hat{\sigma}^2_{WR} \geq \sigma^2_{W0}$ or $\hat{\sigma}^2_{WR} < \sigma^2_{W0}$. The second method is the test method (TEST) which tests the hypothesis of $\sigma^2_{WR} \geq \sigma^2_{W0}$ vs. $\sigma^2_{WR} < \sigma^2_{W0}$ to decide which criterion should be used. [8] If $\hat{\sigma}^2_{WR}(n_1+n_2-2)/\chi^2_{0.05,n_1+n_2-2} \geq \sigma^2_{W0}$, then the reference-scaled criterion should be used; otherwise the constant-scaled criterion should be used. The third method (OPT) assumes that we know whether $\sigma^2_{WR} \geq \sigma^2_{W0}$. [10] Chow et al. [10] conducted simulation studies to compare the three methods. When $\sigma^2_{WR} \geq \sigma^2_{W0}$, all the three methods perform equally well in controlling the type I error rate. However, when $\sigma^2_{WR} < \sigma^2_{W0}$, the tests using the estimation method for choosing the reference-scaled criterion or constant-scaled criterion slightly inflate the type I error rate but only up to 0.06. On the other hand, the test using the test method is conservative when $\sigma^2_{WR} < \sigma^2_{W0}$. When $\sigma^2_{WR} = \sigma^2_{W0}$, the test method performs slightly better than the estimation method.

We also conducted additional simulation studies to compare empirical powers of the three methods when $\sigma^2_{WR} = \sigma^2_{W0} = 0.04$. The results are provided in Table S4. Most of differences in empirical powers between the three methods and the asymptotic powers are in the second or third decimal point. Except for only two cases, the difference between the empirical power by the estimation method and the asymptotic power does not exceed 10%. From Table S4, we reconfirm that the test method should be used when $\sigma^2_{WR} = \sigma^2_{W0}$ because, except for one case, all differences are in the third decimal point. In summary, when $\sigma^2_{WR} \neq \sigma^2_{W0}$, the estimation method should be used to select the criterion. On other hand, when $\sigma^2_{WR} = \sigma^2_{W0}$ or $\hat{\sigma}^2_{WR} \simeq \sigma^2_{W0}$, the test method should be used to choose the reference-scaled criterion or constant-scaled criterion.

Supporting Information

Table S1 Sample size per sequence, asymptotical power, and empirical power for the linearized constant-scaled criterion with respect to a nominal power of 80% at the 5% significance level.
(DOC)

Table S2 Sample size per sequence, asymptotical power, and empirical power for the linearized reference-scaled criterion with respect to a nominal power of 80% at the 5% significance level.
(DOC)

Table S3 Comparison of sample sizes required between the 2×3 and 2×4 crossover designs with respect to a nominal power of 80% at the 5% significance level.
(DOC)

Table S4 Comparison between different methods for determining the scaled criterion with respect to a nominal power of 80% at the 5% significance level, where $\delta = 0.1$, $\sigma^2_D = 0.0225$, and $\sigma^2_{WT} = \sigma^2_{WR}$.
(DOC)

File S1 Derivation of the $100(1-\alpha)\%$ upper confidence bound under 2×4 crossover design by the modified large sample method.
(DOC)

File S2 Derivation of the asymptotic normality of the upper confidence bound.
(DOC)

File S3 SAS macro code.
(DOC)

Acknowledgments

We would like to sincerely thank the anonymous reviewer for the careful, constructive, thorough and thoughtful review which greatly improves the contents and presentation of our work.

Author Contributions

Conceived and designed the experiments: CC JPL. Performed the experiments: CC JPL. Analyzed the data: CC JPL. Contributed reagents/materials/analysis tools: CC CFH JPL. Contributed to the writing of the manuscript: CC JPL.

References

1. U.S. Food and Drug Administration (FDA) (2001) Guidance for Industry: Statistical Approaches to Establishing Bioequivalence. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD, U.S.A. 7–45.
2. European Agency for the Evaluation of Medicinal Products (EMEA) (2001) Note for Guidance on the Investigation on Bioavailability and Bioequivalence. London, U.K. 2–18.
3. World Health Organization (WHO) (2005) Multisource (Generic) Pharmaceutical Products: Guidelines on Registration Requirements to Establish Interchangeability – Draft Revision. Geneva, Switzerland. 4–39.
4. Chow SC, Liu JP (2010) Design and Analysis of Bioavailability and Bioequivalence Studies. 3rd ed., New York, U.S.A.: CRC/Chapman & Hall. 335–419.
5. US Food and Drug Administration (FDA) (2012) Draft Guidance on Scientific Considerations in Demonstrating Biosimilarity to a Reference Product. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD, U.S.A. 12–20.
6. Chow SC, Wang J, Endrenyi L, Lachenbruch PA (2013) Scientific considerations for assessing biosimilar products. Statistics in Medicine 32: 370–381.
7. Hyslop T, Hsuan F, Holder DJ (2000) A small sample confidence interval approach to assess individual bioequivalence. Statistics in Medicine 19: 2885–2897.
8. Serfling RJ (1980) Approximation Theorems of Mathematical Statistics, Wiley. 120, 190.
9. Barndorff-Nielsen OE, Cox DR (1989) Asymptotic Techniques for Use in Statistics. London, U.K.: Chapman & Hall. 118–119.
10. Chow SC, Shao H, Wang HS (2002) Individual bioequivalence testing under 2×3 designs. Statistics in Medicine 21: 629–648.