Therapeutic Innovation
& Regulatory Science
47(2) 242-247
© The Author(s) 2013
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2168479012474273
tirs.sagepub.com

Sample Size Calculation for Bioequivalence Studies Assessing Drug Effect and Food Effect at the Same Time With a 3-Treatment Williams Design

Jiacheng Yuan, PhD[1], Tiejun Tong, PhD[2], and Man-Lai Tang, PhD[2]

Abstract

The US Food and Drug Administration issued a guidance in 2002, "Food-Effect Bioavailability and Fed Bioequivalence Studies," in which it states "in addition to a BE [bioequivalence] study under fasting conditions, we recommend a BE study under fed conditions for all orally administered immediate-release drug products" for abbreviated new drug applications. This statement involves 3 studies: a BE study under fasting status, a food-effect (FE) study, and a BE study under fed status. In practice, when it is known that there is no FE with a reference (*R*) formulation, a sponsor may choose to run a BE study that assesses the drug effect and food effect with a test (*T*) formulation in a single study that includes 3 treatments: *R* formulation at fasting status, *T* formulation at fasting status, and *T* formulation at fed status. Such a study combines the fasting BE study and the FE study on the *T* formulation and may justify the waiver of the fed BE study if conclusions can be made that there is no FE with the *T* formulation after this combined study completes. This article discusses how to calculate the sample size for this kind of study with different primary analysis models. Also discussed are (1) sample size calculations with more general BE studies and (2) how they can be implemented using commercial software in a standard 2-treatment, 2-period, and 2-sequence crossover design, as well as (3) a related practical issue of how to retrieve residual intrasubject mean squared error from historical summary results in the literature.

Keywords
bioequivalence, crossover study, mixed-effects model, sample size, two 1-sided tests

1. Introduction

Bioavailability (BA) and bioequivalence (BE) studies are widely conducted in pharmaceutical companies. *Bioavailability* of a drug is defined as the rate and extent to which the active drug ingredient or therapeutic moiety is absorbed and becomes available at the site of drug action. *Bioequivalence* involves comparison of BA between a test (*T*) and a reference (*R*) drug product, where *T* and *R* can vary depending on the comparison to be performed (eg, generic drug vs reference-listed drug, extended vs immediate release, oral vs intravenous, different formulations or manufacturing processes, etc). *Food effect* (FE) refers to the impact on the BA when the drug is taken with food versus an empty stomach. The US Food and Drug Administration (FDA) issued a guidance in 2002, "Food-Effect Bioavailability and Fed Bioequivalence Studies," in which it states "in addition to a BE study under fasting conditions, we recommend a BE study under fed conditions for all orally administered immediate-release drug products" for abbreviated new drug applications.[1] This statement involves 3 studies: a BE study under fasting status, an FE study, and a BE study under fed status. Sometimes, it is known that there is no FE with *R* formulation, and it is reasonable to assume that there is no FE with the *T* formulation either. For example, a drug has already been approved in tablet form (*R* formulation), and it has already been established that there is no food effect with the tablet, and now in developing pediatric therapy, a suspension (*T* formulation) with the same effective drug substance is studied. However, although it is almost certain that there will not be any FE on *T* formulation, the regulatory agency would

[1] Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA
[2] Department of Mathematics, Hong Kong Baptist University, Hong Kong

Submitted 10-Jul-2012; accepted 4-Dec-2012.

Corresponding Author:
Jiacheng Yuan, PhD, Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA.
Email: jiacheng.yuan@novartis.com

Table 1. A 6 × 3 crossover design.

	Period		
Sequence	1	2	3
1	A	B	C
2	A	C	B
3	B	A	C
4	B	C	A
5	C	A	B
6	C	B	A

A = *R* under fasting status; B = *T* under fasting status; C = *T* under fed status.

still want to see data to substantiate it. In this situation, a sponsor may choose to run a BE study assessing the drug effect (in terms of exposure) and food effect with *T* formulation in a single study that includes 3 treatments: *R* formulation at fasting status, *T* formulation at fasting status, and *T* formulation at fed status. With such a design, assessment is made on the FE on the *T* formulation in addition to BE between *T* and *R* formulations (ie, the study would combine the fasting BE study and the FE study on the *T* formulation), and it may justify the waiver of the fed BE study if conclusions can be made that there is no FE with the *T* formulation after this combined study completes. If FE on the *T* formulation cannot be ruled out, a fed BE study is probably still necessary. In this situation, one can consider a Williams design with balanced 3 treatments, 3 periods, and 6 sequences (6 × 3) and can randomize subjects according to a schedule shown in Table 1, where A stands for *R* under fasting status, B stands for *T* under fasting status, and C stands for *T* under fed status.[2] Any permutation of the order of the 6 sequences will not affect the sample size calculations or BE/FE assessment. However, the design matrix depends on such permutation.

Method 1

Denote the treatment effect by F_A, F_B, and F_C. When it is expected that there is no FE on the test formulation, we suggest Method 1 to assess BE of *T* versus *R:* we calculate a 90% confidence interval (CI) for the difference $\theta_1 = (F_B + F_C) / 2 - F_A$, under a mixed-effects model, where the response variable is the log-transformed pharmacokinetic (PK) parameter (eg, area under the curve [AUC] or peak concentration [$C_{\max}$]) and the explanatory variables include treatment (sequence) period as fixed effects and subject within sequence as a random effect.[3,4] The sequence effect can be dropped from the list of explanatory variables because the variation due to sequence, if it exists, will then be incorporated into the intersubject random errors—not a problem for the purpose of the BE analysis, in which the intrasubject random error plays the important role. BE is declared if the back-transformed CI (anti-log of the CI derived from the model) falls completely within the interval (Δ_L, Δ_U), with default values $\Delta_L = 0.80$ and $\Delta_U = 1.25$ for all PK parameters under consideration.

Method 2

The second method to assess BE of *T* versus *R* is to construct a 90% CI for the difference $\theta_2 = F_B - F_A$ with the same mixed-effects model. BE (under fasting status) is declared if the back-transformed CI falls completely within the interval (Δ_L, Δ_U). This method is adopted when it is expected that there may be a food effect on the test formulation.

Both methods are to test the following hypotheses:

$$\begin{aligned} &H_0 : \{\mu T/\mu R \leq \Delta_L\} \cup \{\mu T/\mu R \geq \Delta U\} \text{ versus} \\ &H_1 : \Delta L < \mu T/\mu R < \Delta H, \end{aligned} \quad (1)$$

where μ_T and μ_R are the mean value of the PK parameter for the 2 formulations, respectively. In the log-scale, for Method 1, the drug effect is essentially reparameterized as follows:

$$F_A = F_R + F_{N/R}, F_A = F_T + F_{N/T}, F_C = F_T + F_{W/T},$$

with $F_R = \log(\mu_R)$ representing the direct *R* effect, $F_T = \log(\mu_T)$ the direct *T* effect, and $F_R + F_T = 0$; $F_{N/R} = 0$ the effect of fasting status under *R*; $F_{N/T}$ the fasting status under *T*; $F_{W/T}$ the fed status under *T*; and $F_{N/T} + F_{W/T} = 0$. For Method 2, in which the BE is assessed under fasting status only, $F_A = \log(\mu_R)$ and $F_B = \log(\mu_T)$.

The 2 formulations are declared bioequivalent if it can be demonstrated that the 90% CI of the ratio is wholly contained within the interval (Δ_L, Δ_U). The hypotheses in Equation 1 can be decomposed as two 1-sided tests at the $\alpha = 0.05$ level, with the null hypotheses as H_{01}: $\mu_T/\mu_R \leq \Delta_L$ and H_{02}: $\mu_T/\mu_R \geq \Delta_U$. If neither null hypothesis holds, then the alternative is accepted, $\Delta_L < \mu_T/\mu_R < \Delta_U$ (ie, a success requires winning both component tests).

The rest of this paper is organized as follows. Section 2 describes the statistical models. Section 3 elaborates the algorithm of the sample size calculation, accompanied with an example. Section 4 discusses more general situations. Section 5 discusses how to retrieve residual intrasubject mean squared error from historical summary results in the literature. Finally, we conclude the paper in Section 6 with a discussion.

2. Statistical Models

Suppose that there are *n* subjects in each sequence in the 6 × 3 crossover design in Table 1. Then the total sample size is $N = 6n$. We also assume that there are no carryover effects in the model because in practice, a washout period of sufficient length can be chosen to completely eliminate the residual effects from one dosing period to the next.[5] Note that the setting can include a different number of subjects in different sequences. Our focus is on the sample size calculation at the design stage, and it is more reasonable to have a balanced design with these 6 sequences. The model is as follows.

$$Y_{ijk} = \mu + S_{ik} + P_j + F_{(j,k)} + e_{ijk}, \text{ with}$$
$$i = 1, \ldots, {}_n;\ \mathrm{j} = 1, 2, 3;\ \textit{and } k = 1, \ldots, 6$$

where

Y_{ijk} is the response of the ith subject in the kth sequence at the jth period;

μ is the overall mean;

S_{ik} is the random effect of the ith subject in the kth sequence, $S_{ik} \overset{iid}{\sim} N(0, \sigma_S^2)$;

P_j is the fixed effect of the jth period, $P_1 + P_2 + P_3 = 0$;

$F_{(j,k)}$ is the direct fixed effect of the formulation in the kth sequence, which is administered at the jth period.

Under the design in Table 1,

$$F_A = F_{(1,1)} = F_{(1,2)} = F_{(2,3)} = F_{(3,4)} = F_{(2,5)} = F_{(3,6)},$$
$$F_B = F_{(2,1)} = F_{(3,2)} = F_{(1,3)} = F_{(1,4)} = F_{(3,5)} = F_{(2,6)},$$
$$F_c = F_{(3,1)} = F_{(2,2)} = F_{(3,3)} = F_{(2,4)} = F_{(1,5)} = F_{(1,6)},$$

and $F_A + F_B + F_C = 0$;

e_{ijk} is the intrasubject random error in observing Y_{ijk}, $e_{ijk} \overset{iid}{\sim} N(0, \sigma_e^2)$;

$\{S_{ik}\}$ and $\{e_{ijk}\}$ are mutually independent.

With Method 1, the target difference is $\theta_1 = F_B + F_C/2 - F_A$, and we estimate it with the following linear contrast of sequence-by-period means, $\bar{Y}_{\cdot jk} = \frac{1}{n}\sum_{i=1}^{n} Y_{ijk}$:

$$\Lambda_1 = \sum_{k=1}^{6}\sum_{j=1}^{3} C_{2jk}\bar{Y}_{\cdot jk}, \text{ with } \left(C_{1jk}\right)_{6\times3} = \frac{1}{6}\begin{pmatrix} -1 & 0.5 & 0.5 \\ -1 & 0.5 & 0.5 \\ 0.5 & -1 & 0.5 \\ 0.5 & 0.5 & -1 \\ 0.5 & -1 & 0.5 \\ 0.5 & 0.5 & -1 \end{pmatrix},$$

and the variance of $\wedge_1$ is as follows:

$$Var(\Lambda_1) = \frac{\sigma_e^2}{n}\sum_{k=1}^{6}\sum_{j=1}^{3} C_{1jk}^2 = \frac{1.5\sigma_e^2}{N}.$$

With Method 2, the target difference is $\theta_2 = F_\mathrm{B} - F_\mathrm{A}$, which we also estimate with a linear contrast:

$$\Lambda_2 = \sum_{k=1}^{6}\sum_{j=1}^{3} C_{2jk}\bar{Y}_{\cdot jk}, \text{ with } \left(C_{2jk}\right)_{6\times3} = \frac{1}{6}\begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{pmatrix},$$

and the variance of $\wedge_2$ is as follows:

$$Var(\Lambda_2) = \frac{\sigma_e^2}{n}\sum_{k=1}^{6}\sum_{j=1}^{3} C_{2jk}^2 = \frac{2\sigma_e^2}{N}.$$

Let V_h be the estimate of $Var(\mathrm{L}_h)$ for $h = 1, 2$. That is, $V_1 = \frac{1.5\hat{\sigma}_e^2}{N}$ and $V_2 = \frac{2\hat{\sigma}_e^2}{N}$, where $\hat{\sigma}_e^2$ is the intrasubject mean squared error obtained from the analysis of variance table. The test statistic for the hypotheses in Equation 1 is

$$T_h = \hat{\theta}_h / V_h^{1/2},$$

where $\hat{\theta}_h$ is the ordinary least squares estimate of θ_h, and $h = 1$ and 2, corresponding to the above two methods. T_h follows a noncentral t distribution with the degrees of freedom determined by $\hat{\sigma}_e^2$.

For a 6 × 3 design with a total sample size of N, the total sum of squares (SS) has $3N - 1$ overall degrees of freedom (DFs), and subject accounts for $N - 1$ DFs. Therefore, $2N$ DFs are left for within-subject SS. With treatment and period each accounting for 2 DFs, the residual intrasubject SS has $2N - 4$ DFs.

The sample size is decided by the following formula[6]:

$$\phi = PR_{NT}(-t_{\nu,\alpha}, \nu 6 \times 3, \tau_2) - PR_{NT}(-t_{\nu,\alpha}, \nu 6 \times 3, \tau_1), \quad (2)$$

where α is the significance level, $t_{\nu,\alpha}$ is the upper α-quantile of the Student t distribution with ν degrees of freedom, ϕ is the power, and $PR_{NT}(\cdot)$ represents probability from a noncentral t distribution, with

$$\nu_{6\times3} = 2N - 4, \tau_1 = [\log(\mu T/\mu R) - \log(\Delta L)]/V_h,$$
$$\tau_2 = [\log(\mu T/\mu R) - \log(\Delta_U)]/Vh, V_1 = 1.5\hat{\sigma}_e^2/N,$$
and $V_2 = 2\hat{\sigma}_e^2/N$.

3. Algorithm for Sample Size Calculation

Let $\hat{\sigma}_e^2$ denote the residual intrasubject MSE from a historical study; with Method 1, the sample size can be calculated with the following algorithm:

(i) Set values for α, ϕ, and μ_T/μ_R (default $\alpha = 0.05$, $\phi = 0.80$, and $\mu_T/\mu_R = 1$);

(ii) Select a range of sample size (N_1, N_2), for each $N \in [N_1, N_2]$, and do the following:
  a. $V_1 = 1.5\hat{\sigma}_e^2/N$
  b. $\tau_1 = [\log(\mu_T/\mu_R) - \log(\Delta_L)]\ /\ V_1$
  c. $\tau_2 = [\log(\mu_T/\mu_R) - \log(\Delta_U)]\ /\ V_1$
  d. Calculate $PR_{NT}(-t_{\nu,\alpha}, \nu_{6\times3}, \tau_2) - PR_{NT}(-t_{\nu,\alpha}, \nu_{6\times3}, \tau_1)$, where $PR_{NT}(\cdot)$ is the probability from a noncentral t distribution, and $\nu_{6\times3} = 2N - 4$

(iii) If Equation 2 is established with one N, then stop; otherwise, return to step ii.

Table 2. Numeric calculation results for the right-hand side (RHS) of Equation 2.

	RHS	
N	Method 1	Method 2
20	0.7188	0.5242
21	0.7489	0.5609
22	0.7760	0.5951
23	0.8004	0.6269
24	0.8223	0.6564
25	0.8420	0.6838
26	0.8596	0.7093
27	0.8754	0.7328
28	0.8894	0.7546
29	0.9020	0.7748
30	0.9132	0.7934
31	0.9232	0.8105
32	0.9321	0.8264
33	0.9400	0.8410
34	0.9470	0.8545
35	0.9532	0.8668
36	0.9587	0.8772

As FDA guidance requires a BE study have a sample size of at least 12, in practice, (N_1, N_2) can start with (12,100); for most BE studies, this range would be wide enough to cover the sample size with target power, as well as some with higher powers for consideration. In rare situations, if that range is not enough, then try a range with larger numbers—for example, (101,200), (201,300), and so on—until the desired power appears, which should happen within just a few of these intervals as normally BE studies are not large.

For Method 2, the algorithm for sample size calculation is similar: just replace $V_1 = 1.5\hat{\sigma}_e^2/N$ with $V_2 = 2\hat{\sigma}_e^2/N$ in step ii.

Example 1. Hypothetically, assume we have the information of intrasubject variability of historical study, and the residual intrasubject MSE $\hat{\sigma}_e^2 = 0.0862$. (This value is chosen because $CV_{\text{intra}} = \sqrt{\exp(\hat{\sigma}_e^2) - 1} = 0.30$.) (In practice, the residual intrasubject MSE is not often available. Later in this article, we will show how to retrieve it with other usual available information.) Assuming $\alpha = 0.05$, $\phi = 0.80$, $\mu_T/\mu_R = 1.0$, $\Delta_L = 0.80$, and $\Delta_U = 1.25$, then with different numbers of N, the right-hand side of Equation 2 is shown in Table 2.

Recall that the total sample size N needs to be a multiple of 6 for the adopted 6 × 3 crossover design. Thus, from Table 2, the total sample size required is 24 based upon Method 1 and 36 based upon Method 2, for providing a power of 80%. It is clear that Method 1 is more powerful than Method 2.

4. General Situation

In a general situation, a BE study could include multiple treatments formed by R and/or T administered under more conditions. (Examples with 4 treatments: a BE study with R and T each administered as both an intravenous and a subcutaneous dose, or a BE study with R and T each administered at both fasting and fed status.)

The primary interest is the assessment of BE between R and T. In a standard 2 × 2 design, the interesting contrast is $F_T - F_R = \vec{c}_0{}' \cdot \vec{F}_0$, where $\vec{c}_0 = (-1, 1)'$ denotes the vector of the contrast coefficients, $\vec{F}_0 = (F_R, F_T)'$ denotes the vector of the drug effects, and the notation $\vec{a}' \cdot \vec{b}$ means the inner product of vectors $\vec{a}$ and $\vec{b}$. In a BE study with $t(t > 2)$ treatments formed by R and/or T administered under more conditions, denote the treatment effect with $F_1, \ldots, F_t$, $\vec{F} = (F_1, \cdots, F_t)'$, and the interesting contrast (for the assessment of BE between R and T) is $\vec{c}' \cdot \vec{F}$, where $\vec{c} = (c_1, \cdots, c_t)'$ is the vector of the coefficients for the contrast of interest, in which all the R components sum to –1 and all the T components sum to 1. Then the total sample size for this multiple-treatment trial can be approximated with the total sample size for a standard 2 × 2 design with the same α, ϕ, Δ_L, Δ_U, and μ_T/μ_R, but replacing the residual intrasubject MSE, $\hat{\sigma}_e^2$, with an adjusted residual intrasubject MSE,

$$\hat{\sigma}_{\text{adj}}^2 = \hat{\sigma}_e^2 \cdot \left\|\vec{c}\right\|^2 / \left\|\vec{c}_0\right\|^2.$$

Example 1 (continued). $\hat{\sigma}_e^2 = 0.0862$, with Method 1, the vector of the coefficients for the contrast of interest is $\vec{c}_1 = (-1, 0.5, 0.5)'$, then $\hat{\sigma}_{\text{adj}}^2 = \hat{\sigma}_e^2 \cdot \left\|\vec{c}_1\right\|^2 / \left\|\vec{c}_0\right\|^2 = 0.75\hat{\sigma}_0^2 = 0.06465$. Therefore, the sample size can be approximated with the one for a standard 2 × 2 with the same input for all parameters but a smaller residual intrasubject MSE.

With Method 2, the vector of the coefficients for the contrast of interest is $\vec{c}_2 = (-1, 1, 0)'$, then $\hat{\sigma}_{\text{adj}}^2 = \hat{\sigma}_e^2 \cdot \left\|\vec{c}_2\right\|^2 / \left\|\vec{c}_0\right\|^2 = \hat{\sigma}_0^2 = 0.0862$. Therefore, the sample size can be approximated by that for a standard 2 × 2 design with the same input for all parameters.

5. Retrieving the Residual Intrasubject MSE

Usually, from a historical study in the literature, the following information is available:

(1) the design, eg, single-sequence drug-drug interaction (DDI), 2×2 BE, or 6×3 BE;
(2) the total sample size, N_0;
(3) the 100(1 – 2α)% CI, (lower limit, upper limit), for the geometric mean ratio (GMR), T/R, of the PK parameters.

Let $MOE = \frac{1}{2}[\log(UL) - \log(LL)]$ denote the margin of error, and let ν denote the DFs of the t statistic, and then the formula for the residual intrasubject MSE is

$$\hat{\sigma}_e^2 = \frac{N_0}{2} * MOE^2 / t_{\nu,\alpha}^2,$$

where ν is decided by the design and the sample size as follows:

$\nu = N_0 - 1$ for single-sequence DDI,
$\nu = N_0 - 2$ for a 2 × 2 BE,
$\nu = 2(N_0 - 2)$ for a 6 × 3 BE.

Example 2. Hypothetically, let's assume there are historical data from a single-sequence DDI study in which a total of 30 subjects took the studied drug at period 1 followed by another drug in addition to the studied drug at period 2. The 90% CI of the GMR of the PK parameters (of the studied drug) between the 2 periods is (0.83, 1.15), which means there is no DDI. Now, to design a new BE study for the studied drug, we want to know the intrasubject variability based on the historical data.

Solution:

With $N_0 = 30$, $MOE = [\log(1.15) - \log(0.83)] / 2 = 0.1630$, $\nu = N_0 - 1 = 29$, and $\alpha = (1 - 90\%) / 2 = 0.05$, we have

$$\hat{\sigma}_e^2 = \frac{N_0}{2} * MOE^2 / t^2_{\nu,\alpha} = 0.1381.$$

This derived intrasubject MSE (together with assumptions of significance level α, power ϕ, expected mean ratio between different formulations μ_T/μ_R, and the limits for equivalence Δ_L and Δ_U) can then be used to calculate the sample size for the new study.

Example 3. In this hypothetical example, the historical data are from a 2 × 2 BE study in which a total of 30 subjects (ie, 15 subjects per sequence) took one formulation at the first period and another at the second period, and the 2 sequences have a different order for the 2 formulations. The 90% CI of the GMR between the 2 formulations is (0.83, 1.15), which means the 2 formulations are bioequivalent. To know the intrasubject variability, we have the following.

Solution:

With $N_0 = 30$, $MOE = [\log(1.15) - \log(0.83)] / 2 = 0.1630$, $\nu = N_0 - 2 = 28$, and $\alpha = (1 - 90\%) / 2 = 0.05$, we have

$$\hat{\sigma}_e^2 = \frac{N_0}{2} * MOE^2 / t^2_{\nu,\alpha} = 0.1378.$$

Example 4. In this hypothetical example, the historical data are from a 6 × 3 BE (+ FE) study, the same design as has been introduced at the beginning of this article. The total sample size is 30 subjects (ie, 5 subjects per sequence). The 90% CI of the GMR between the 2 formulations is (0.83, 1.15), which means the 2 formulations are bioequivalent. To know the intrasubject variability, we have the following:

Solution:

With $N_0 = 30$, $MOE = [\log(1.15) - \log(0.83)] / 2 = 0.1630$, $\nu = 2N_0 - 4 = 56$, and $\alpha = (1 - 90\%) / 2 = 0.05$, we have

$$\hat{\sigma}_e^2 = \frac{N_0}{2} * MOE^2 / t^2_{\nu,\alpha} = 0.1425.$$

6. Discussion

6.1. Testing of FE in Method 1

With Method 1, a prudent statistician may feel that a formal test of the FE on the T formulation should be performed before testing the BE between R and T formulations with $\theta_1 = (F_B + F_C)/2 - F_A$, and the latter test should happen only after it has already been established that there is no FE with the T formulation.

As we stated in Section 1, Method 1 is not recommended for general situations but in the case when it is almost certain there is no FE on the T formulation with good scientific basis. Nevertheless, with a prudent statistical approach, we can perform the FE test first and then the BE upon successful establishment of no FE. From a hypothesis testing perspective, we can split the type I error rate between these 2 tests so that the family-wise type I error rate is maintained at the .05 level. Then the testing procedure is as follows.

First we test H_{01} with $\alpha = 0.025$:

$$\begin{aligned} &H_{01} : \{F_B - F_C \le \delta_L\} \cup \{F_B - F_C \ge \delta_U\} \text{ versus} \\ &H_{11} : \delta_L < F_B - F_C < \delta_U. \end{aligned} \tag{3}$$

If H_{01} is rejected, we then test H_{02} with $\alpha = 0.025$:

$$\begin{aligned} H_{02} : \{(F_B + F_C)/2 - F_A \le \delta_L\} \cup \{(F_B + F_C)/2 - F_A \\ \ge \delta_U\} \text{ versus } H_{12} : \delta_L < (F_B + F_C)/2 - F_A < \delta_{U.} \end{aligned} \tag{4}$$

If H_{01} cannot be rejected, we then test H_{03} with $\alpha = 0.025$:

$$\begin{aligned} &H_{03} : \{F_A - F_B \le \delta_L\} \cup \{F_A - F_B \ge \delta_U\} \text{ versus} \\ &H_{13} : \delta_L < F_A - F_B < \delta_U, \end{aligned} \tag{5}$$

where $\delta_L = \log(0.8)$ and $\delta_U = \log(1.25)$.

For this testing procedure, the sample size can be decided as the maximum of those required for tests 3 to 5, where the algorithm for Method 2 can be used for tests 3 and 5 with $\alpha = 0.025$, and the algorithm for Method 1 can be used for test 4 with $\alpha = 0.025$.

Finally, we note that in some situations, the washout period may not be sufficient to fully remove the carryover due to the limited resources. If this happens, one may need to include a carryover effect in the mixed-effect model or to conduct a test for the carryover effect before planning the sample size calculation for bioequivalence studies.

6.2. How to Use nQuery to Calculate the Sample Size in General Situations

Using nQuery to calculate the sample size for Method 1 in Example 1, we do the following:

(i) Click *File/New*, and tick the following: Goal *Means*, Number of Groups *Two*, Analysis Method *Equivalence*;

(ii) Select *TOST for ratio of means (log scale) for two group or cross-over*;

(iii) Select *TOST for equivalence for ratio of means (log scale) for cross-over*;

(iv) Input: $\alpha = 0.05$, $\Delta_L = 0.80$, $\Delta_U = 1.25$, $\mu_T/\mu_S = 1$, $\text{sqrt}(MSE) = \text{sqrt}(0.75\hat{\sigma}_0^2) = \text{sqrt}(0.06465) = 0.2543$.

Then we get the per-sequence size $n = 12$, so the total sample size is $N = 2n = 24$, which is the same as what we got with numerical calculation in Table 2 for Method 1.

For Method 2 in Example 1, steps i to iii remain the same as above, and in step iv, make the only change that $\text{sqrt}(MSE) = \text{sqrt}(\hat{\sigma}_0^2) = \text{sqrt}(0.0862) = 0.2936$. Then we get the sequence size $n = 16$; for a 2 × 2 study, the total sample size would be $N = 2n = 32$, which implies a total sample size of 36 for the 2 × 2 design, which is the same as what we got with numerical calculation in Table 2 for Method 2.

Using the above algorithm with Method 2, the calculated sample size is very close to the result from commercial software nQuery or PASS.[7,8] Both nQuery and PASS only give sample sizes per sequence, assuming a standard 2 × 2 crossover design. For a 6 × 3 design, one doubles the result from the software for the total sample size. The reason is because the sample size equation under a standard 2 × 2 crossover is as follows:

$$\phi = PR_{NT}(-t_{\nu,\alpha}, \nu_{2\times2}, \tau_2) - PR_{NT}(-t_{\nu,\alpha}, \nu_{2\times2}, \tau_1), \qquad (6)$$

where

$$\nu_{2\times2} = N - 2;$$
$$\tau_1 = [\log(\mu T/\mu R) - \log(\Delta_L)]/V_0;$$
$$\tau_2 = [\log(\mu T/\mu R) - \log(\Delta_U)]/V_0;$$
$$V_0 = 2\hat{\sigma}_e^2/N.$$

Practically, the total sample size N is not very small, and therefore the t distributions with $\nu_{2\times2} = N - 2$ and $\nu_{6\times3} = 2N - 4$ are very close. Since we also have $V_0 = V_2$, Equation 6 should give a result very similar to that from Equation 2 with Method 2. The rationale for $\nu_{2\times2} = N - 2$ is as follows. The total SS has $2N - 1$ overall DFs, and subjects account for $N - 1$ DFs. Therefore, N DFs are for within-subject SS. With treatment and period each accounting for 1 DF, the residual intrasubject SS has $N - 2$ DFs.

Acknowledgments

The authors are grateful to the editor, the associate editor, and 2 reviewers for their constructive comments and suggestions that have led to a substantial improvement in the article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Food and Drug Administration. Guidance for industry: food-effect bioavailability and fed bioequivalence studies. 2002. http://www.fda.gov/downloads/regulatoryinformation/guidances/ucm126833.pdf
2. Williams EJ. Experimental designs balanced for the residual effects of treatment. *Aust J Sci Res*. 1949;2:149-168.
3. Chinchilli VM, Esinhart JD. Design and analysis of intra-subject variability in cross-over experiments. *Stat Med*. 1996;15:1619-1634.
4. Food and Drug Administration. Guidance for industry: statistical approaches to establishing bioequivalence. 2000. http://www.fda.gov/downloads/Drugs/Guidances/ucm070244.pdf
5. Chow SC, Liu JP. *Design and Analysis of Bioavailability and Bioequivalence Studies*. 3rd ed. London, UK: Chapman & Hall/CRC; 2009.
6. Julious SA. Tutorial in biostatistics—sample sizes for clinical trials with normal data. *Stat Med*. 2004;23:1921-1986.
7. Elasho JD. *nQuery Advisor Version 4 User's Guide*. Los Angeles, CA: Statistical Solutions; 2000.
8. Hintz JL. *PASS 2000 User's Guide*. Kaysville, UT: NCSS; 2000.