PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.294

Sequential design approaches for bioequivalence studies with crossover designs‡

MAIN PAPER

Diane Potvin[1], Charles E. DiLiberti[2], Walter W. Hauck[3,*,†], Alan F. Parr[4], Donald J. Schuirmann[5] and Robert A. Smith[6]

[1] *Theratechnologies Inc., Montréal, Qué., Canada*
[2] *Barr Laboratories, Inc., Woodcliff Lake, NJ, USA*
[3] *US Pharmacopeia, Rockville, MD, USA*
[4] *GlaxoSmithKline, Inc., Research Triangle Park, NC, USA*
[5] *Center for Drug Evaluation and Research, US Food and Drug Administration, USA*
[6] *Bristol-Myers Squibb Pharmaceutical Research Institute, Princeton, NJ, USA*

The planning of bioequivalence (BE) studies, as for any clinical trial, requires a priori specification of an effect size for the determination of power and an assumption about the variance. The specified effect size may be overly optimistic, leading to an underpowered study. The assumed variance can be either too small or too large, leading, respectively, to studies that are underpowered or overly large. There has been much work in the clinical trials field on various types of sequential designs that include sample size reestimation after the trial is started, but these have seen only little use in BE studies. The purpose of this work was to validate at least one such method for crossover design BE studies. Specifically, we considered sample size reestimation for a two-stage trial based on the variance estimated from the first stage. We identified two methods based on Pocock's method for group sequential trials that met our requirement for at most negligible increase in type I error rate. Copyright © 2007 John Wiley & Sons, Ltd.

Keywords: ***sequential design; sample size reestimation; adaptive design; bioequivalence***

1. INTRODUCTION

If a new pharmaceutical formulation for the US market (e.g. a new formulation developed by the innovator or a generic version of the innovator product) is intended to be therapeutically

*Correspondence to: Walter W. Hauck, US Pharmacopeia, 12601 Twinbrook Parkway, Rockville, MD 20852-1790, USA.
†E-mail: wh@usp.org
‡The views expressed in this publication have no endorsement by the US Food and Drug Administration.

interchangeable with a predicate product termed, for generic substitution, the Reference Listed Drug, the new pharmaceutical formulation may be required to be shown to be bioequivalent to the Reference Listed Drug. Bioequivalence (BE) studies are performed based on the requirements set forth in part 320 of section 21 of the Code of Federal Regulations (CFR) and guidance given by the US Food and Drug Administration's (FDA's) Center for Drug Evaluation and Research (CDER) [1].

Planning BE studies, as is the case in planning any clinical trial, requires some prior estimate of the appropriate variance and a decision regarding the effect size at which to determine power (probability of meeting the preestablished BE criteria). BE studies are most commonly conducted using crossover designs, so the variance needed is *within* subject. The effect size is specified by choosing a ratio of the geometric means of the two formulations. Minimum sample size is obtained if the 'effect size' is chosen to be 100%, i.e. perfect equivalence. It is common practice to allow for some departure from perfect equivalence, with a resulting statement like, 'We will have power of at least 80% to conclude the two formulations to be bioequivalent as long as the true ratio of geometric means lies between 95.0% and 105.3%.' BE studies are also no different from other clinical trials in that prior information regarding the variance may be poor or nonexistent. The choice of the effect size may also be overly optimistic. If the variance used in the power calculation is too low or the chosen effect size overly optimistic, the study is underpowered; conversely, if the variance estimate is too high, the sample size may be unnecessarily large.

Some regulatory agencies (Canada [2], Japan [3], and the World Health Organization (WHO) [4]) permit 'add-on' designs. With these designs, if the failure to declare the two formulations bioequivalent appears to be due to insufficient power, it is permissible to add additional subjects and pool results of the additional subjects with the original trial. For example, the WHO and Japanese guidelines allow an add-on study provided the additional sample size is at least 50% of that of the original study. Statistically, add-on designs cannot preserve the nominal type I error rate if a test is conducted at the nominal level following the completion of the initial trial and then again after the additional subjects are included. (WHO does indicate that the level of consumer risk must be considered; Canada does not.) The argument in favor of add-on designs is that they do make use of the data already collected, and the inflation of the type I error rate is 'acceptable.' Add-on designs are not accepted by CDER or the Australian Therapeutic Goods Administration (TGA). The TGA discourages sequential designs and indicates that, if used, must be in the protocol and a Bonferroni correction applied [5]. In the US, the study may be repeated, but the data from the second study are not pooled with those from the first study. The second study has to stand on its own merits.

In the general clinical trials field, these problems are addressed in at least two ways. First, there are group sequential designs, which permit stopping the trial early if the results are sufficiently strong (for declaring in favor of the alternative) or weak (futility for rejecting the null). More recently, there are methods for sample size reestimation after the trial is partially completed. The reestimation may depend on the variance or the effect size or both as observed part way through the trial.

In a group sequential trial, interim analyses are conducted on the data available at one or more intermediate stages, where the sample size (n_i) and allowed type I error rate (α_i), at each stage are preestablished according to some rules. The size of each stage is fixed *a priori* and is not permitted to change based on interim results. The key feature of the group sequential approach is that the α_i are chosen so that the upper bound on the overall probability of incorrectly concluding success at any stage is still fixed at $\alpha = 0.05$. By permitting early stopping, group sequential approaches provide some protection against unnecessary use of resources if the planned total sample size was based on an overestimated variance. However, because these approaches make no provision for changing the maximum number of subjects, they provide less opportunity for protecting against

DOI: 10.1002/pst

either over- or underpowered studies due to a poorly anticipated value for the variance.

Application of group sequential approaches to BE studies differs from their application to most other types of clinical studies because the former generally involves crossover designs, testing of equivalence hypotheses, and testing based on *t*-distributions, whereas the latter generally involves parallel designs with testing of difference hypotheses. At the *i*th stage of a group sequential BE trial, data are analyzed from the first n_i of the planned maximum number of subjects n, and the trial is stopped and BE is concluded if the $(1-2\alpha_i)\times 100\%$ confidence intervals for the test-to-reference ratios are entirely contained within the interval [80%, 125%] for both $C_{\max}$ (maximum drug concentration) and AUC (area under the drug concentration versus time curve); otherwise the trial continues to the next stage. Hauck *et al.* [6] described how to apply the α-spending functions of Lan and DeMets [7] for this purpose. The α_1^*-spending function they considered approximates the O'Brien–Fleming [8] group sequential method in which the α_i at the interim analyses are conservatively small so that the final analysis of the full n subjects is conducted at a significance level near the full 0.05. Although that provides the greatest chance of ultimately concluding BE, it also reduces the chance of concluding BE at any of the earlier analyses. According to Table II from Hauck *et al.*, the type I error rate can go as high as 6% for moderate variances using this spending function, suggesting that some modifications of existing stopping rules may be necessary for BE applications. An alternative is the Lan and DeMets α_2^*-spending function, which approximates the less conservative Pocock [9] method. With this spending function, the α_i at the interim analyses are larger, providing greater chances of stopping the trial early to conclude BE, but smaller than with O'Brien–Fleming at the final stage, thus with more risk of not concluding BE at the final sample size, n. Gould [10] argues for enhancing the probability of reaching an early decision. His approach, based on that of Jennison and Turnbull [11], is similar to Pocock's, but uses equal critical values rather than equal values for the α_i's.

Sample size reestimation methods have not typically been considered up to now for BE studies. There are several discussions [12–20] in the literature for reestimation based on the observed variance in an initial sample when using *t*-tests, going back to the original proposal of Stein [17]. The general approach is to complete a portion of the trial (the 'internal pilot') and then analyze the data to determine the variance estimate and proceed to determine the sample size anew. If the new sample size exceeds the size of the stage already completed, the trial is continued to the new sample size, otherwise the trial is stopped. The literature is consistent regarding the fact that there is *some* inflation of the type I error rate when data from all stages are used to obtain the final variance estimate, but is not consistent as to whether the degree of inflation is important. For example, Wittes and Brittain [18] and Birkett and Day [12] find that the actual type I error could be 5.2–5.3% but consider that to be an unimportant inflation. Later, Wittes *et al.* [19] chose to be more conservative and concluded that some adjustment to the critical values is needed to retain the overall type I rate. Regardless, none of the methods are validated for crossover studies and two one-sided *t*-tests, so there is a need to start from the beginning in considering these approaches.

Another possible application of sample size reestimation is to adjust the sample size based on the observed *effect size* seen in an initial sample. For example, one might choose a sample size based on the assumption that the ratio of geometric means is between 0.95 and 1.053, but the estimated ratio seen in an initial sample may be, say, 0.92. One might be tempted, in this case, to recalculate the sample size assuming that the true ratio of population means was 0.92. However, simulations carried out by Cui *et al.* [21] indicate that if this procedure is applied naïvely, the overall type I error rate may be inflated by 30% or more. More complicated statistical procedures have been proposed to allow sample size reestimation based on the observed effect size. For a review see, for example, Jennison and Turnbull [22] and Chen *et al.* [23]. While these more complicated procedures could certainly be considered for a BE

DOI: 10.1002/pst

study, they lack the feature of using the usual test statistic formulae (possibly with a simple adjustment of the nominal type I error rate), and they have not been validated for use with two one-sided tests and unknown variance. We will not consider them further here.

Both group sequential and reestimation approaches have potential use for BE trials. In order to combine the best features of each, we studied an *adaptive sample size sequential method*, which is defined here as a hybrid of the two methods described above. In an adaptive design (1) the final sample size required can be reevaluated after each stage; and (2) the endpoint can be evaluated *more than once* with the possibility of early stopping when criteria are met. There are many possible variants to this general adaptive sample size sequential approach.

The Product Quality Research Institute (PQRI) formed a subcommittee to its Biopharmaceutics Technical Committee to investigate these methods and make recommendations for BE studies. This paper reports the results of this work. The general goal of the subcommittee was to determine whether these methods could be validly applied to BE trials and, if so, to establish the principle that they are appropriate design choices for BE. More specifically, the goal was to identify and validate at least one particular method that could be accepted by regulatory agencies and used by sponsors without further validation. Toward that end, the committee established a number of desirable properties that we would seek for such a method. First and most prominent was that the method had to either not inflate the type I error rate or inflate it with a determined maximum inflation judged negligible. We considered an overall type I error rate of less than 5.2% for a nominal 5% test to be a negligible inflation. Other desirable properties for the work here were:

- Number of stages permitted = 2.
- Allow stopping (with passing results) after completing either stage if criteria are met.
- No requirement for blinding to treatment effect at each stage, since data would routinely be unblinded in pharmacokinetic (PK) BE studies.
- Should have better properties than current practice (which, in effect, allows multiple trials, each time tested at 5%).
- Should provide a unique, unambiguous result: i.e. there should only be one outcome from any given data set. This would rule out, for example, bootstrap procedures.
- Does not require poolability criteria (or at least should know whether results from both stages are poolable *before* sample analysis, i.e. base poolability on study conduct such as subject demographics, temporal considerations, use of same protocol, use of same site, etc., rather than a statistical test of poolability).

2. METHODS

We considered four methods for two-stage BE trials. The first, Method A, is a simple naïve sample size reestimation method using the nominal α ($=0.05$) at whichever stage the test is carried out and using data from both stages (if two stages are carried out) to compute the final variance estimate. Wittes *et al.* [19] proved that, when data from both stages are used to compute the final variance estimate, we will tend to underestimate the variance. Coffey and Muller [14] proved that the actual significance level of such a naïve reestimation method will generally be greater than the nominal level. The purpose of studying Method A was to determine whether there was *nonnegligible* inflation of the type I error rate in the BE testing context.

Methods B, C, and D are three variations of adaptive methods. Although the methods proposed in this paper did not use equal sample sizes at each stage, the α of 0.0294, as originally proposed by Pocock [9] for equal sample sizes at each stage, was considered a good starting point in order to control the overall type 1 error to a maximum of 5%. Method D uses a slightly smaller (and hence more conservative) α at each stage based on preliminary work (not shown). In all this work, we assume that the results from the two stages are poolable.

Our results consider only a single endpoint, i.e. AUC or C_{max}. This choice is based on CDER

DOI: 10.1002/pst

policy. At this time, CDER requires that statistical tests for assessing BE preserve the type I error rate at 0.05 for each individual endpoint (e.g. AUC, C_{max}) considered [24]. An argument that a particular method, while it does not preserve the type I error rate for each individual endpoint, does preserve the overall type I error rate for all endpoints taken together, or similar argument, would not be accepted by CDER at this time. It is not the objective of this manuscript to explore this issue.

For all four methods below:

- All calculations are performed following natural log transformation (ln) of the original data.
- The variance used in the calculations is the residual variance from the appropriate analysis of variance (ANOVA) (see Section 2.3 for details).
- All power calculations are performed at the specified α level using a ratio of population geometric means (on the original scale) of 95%. The value of 95% was chosen as typical of values used for designing BE studies; i.e. allowing for some difference from reference.
- A BE criterion at a specified α level means that the two-sided confidence interval for the ratio of geometric means (Test/Reference, on the original scale) at the $(1-2\alpha)$ level falls entirely within 80–125%. Thus, the BE criterion at the $\alpha = 0.05$ level corresponds to the conventional BE criteria of the 90% two-sided confidence interval for the ratio of geometric means falling within 80–125%.
- Wherever a second stage is required, the number of subjects to be used in that second stage is the minimum *even* number of subjects required for the combined data from stages 1 and 2 to have at least 80% power. That is, each stage was assumed to be balanced with respect to sequence.
- All four methods stop (i.e. lead to a pass or fail conclusion) after either one or two stages – there is no provision for any more stages.
- No poolability criteria were applied (i.e. pooling data from stages 1 and 2 were always allowed, even if there was a statistically significant difference between the results from the two stages).
- The minimum size for stage 2 was 2 (if the decision rule determined that the study should continue to stage 2) and there was no upper limit to the size of stage 2.
- No 'futility' criterion was applied, i.e. if the algorithm specified that stage 2 should be performed, then it was simulated, regardless of how unlikely it would be to meet the BE criteria or how large the stage 2 sample size was determined to be.

2.1. Sample size reestimation method (method A)

For the sample size reestimation method, the sample size was estimated based on the variance estimate after stage 1. The following method was evaluated:

A. Evaluate the power at stage 1 using the variance estimate from stage 1 and an α level of 0.05. If the power is $\geqslant 80\%$ at stage 1, evaluate BE at stage 1, using an α level of 0.05 and stop whether BE is concluded or not. If the power is less than 80%, calculate the sample size based on the variance estimated at stage 1 using an α level of 0.05 and continue to stage 2.

Evaluate BE at stage 2 using data from both stages and an α level of 0.05. Stop here whether BE is met or not and regardless of what power was achieved. Note that BE is evaluated only once in any event. This is sometimes referred to as an internal pilot study design.

2.2. Adaptive sample size sequential methods (methods B, C, and D)

For the adaptive sample size sequential method, the following three methods were considered:

B. Evaluate BE at stage 1 using an α level of 0.0294, regardless of the power achieved. If the BE criterion is met, stop. If the BE criterion is not met, calculate the sample size based on the variance estimated at stage 1 and an α level of 0.0294. If stage 1 already has at least 80% power, then stop. If not, continue to stage 2. Evaluate BE at stage 2 using data from both stages and an α level of

DOI: 10.1002/pst

0.0294. Stop here whether BE is met or not and regardless of the power achieved.

C. Evaluate the power at stage 1 using the variance estimate from stage 1 and an α level of 0.05. If the power is greater than or equal to 80%, evaluate BE at stage 1 using an α level of 0.05 and stop whether BE is met or not. If the power is less than 80%, evaluate BE using an α of 0.0294. If the BE criterion is met, stop. If the BE criterion is not met, calculate the sample size based on the variance estimated at stage 1 and an α level of 0.0294 and continue to stage 2. Evaluate BE at stage 2 using data from both stages and an α level of 0.0294. Stop here whether BE is met or not and regardless of the power achieved.

D. Evaluate the power at stage 1 using the variance estimate from stage 1 and an α level of 0.05. If the power is greater than or equal to 80%, evaluate BE at stage 1 using an α level of 0.05 and stop whether BE is met or not. If the power is less than 80%, evaluate BE using an α of 0.028. If the BE criterion is met, stop. If the BE criterion is not met, calculate the sample size based on the variance estimated at stage 1 and an α level of 0.028 and continue to stage 2. Evaluate BE at stage 2 using data from both stages and an α level of 0.028. Stop here whether BE is met or not and regardless of the power achieved.

Figures 1–4 illustrate the different approaches. In summary, methods B, C, and D all begin with a test of BE at stage 1. They differ in that method B always tests stage 1 at an α level of 0.0294, whereas methods C and D determine the α level to be applied to the stage 1 results based on the power achieved at stage 1 (i.e. stage 1 α is either 0.05 or 0.0294 for method C and either 0.05 or 0.0280 for method D). Each method permits three possible decisions following stage 1, stopping and concluding BE, stopping and not concluding BE (if stage 1 has sufficient power) or continuing with stage 2 (if stage 1 has insufficient power and BE was not concluded). Methods B, C, and D power stage 2 and then test for BE on combined stages 1 and 2 data based on α levels of 0.0294, 0.0294, and 0.028, respectively.

In light of the tendency to underestimate the true variance when data from the first stage are used to determine the size of the second stage and data from all stages are used to compute the final variance estimate, we could have considered a variation on methods B, C, and D whereby we would only use stage 1 data to compute the final variance estimate (similar to the proposal by Stein [17]). However, we wanted to examine procedures that used the same computational methods and formulae that are used for nonsequential studies. Methods that only use stage 1 data to compute the final variance estimate are a topic for future research.

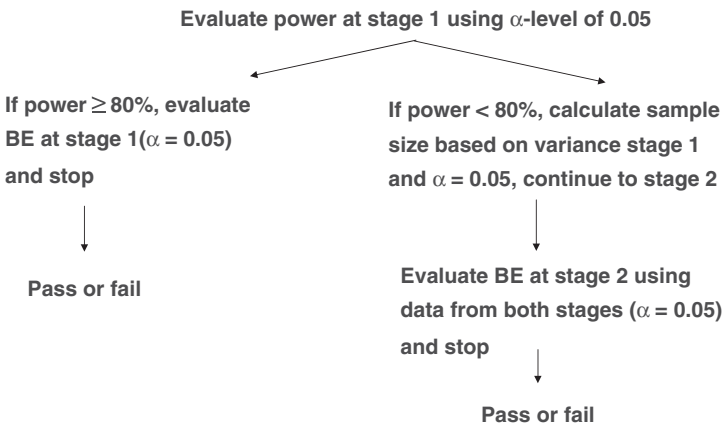

Figure 1. Sample size reestimation method A.

DOI: 10.1002/pst

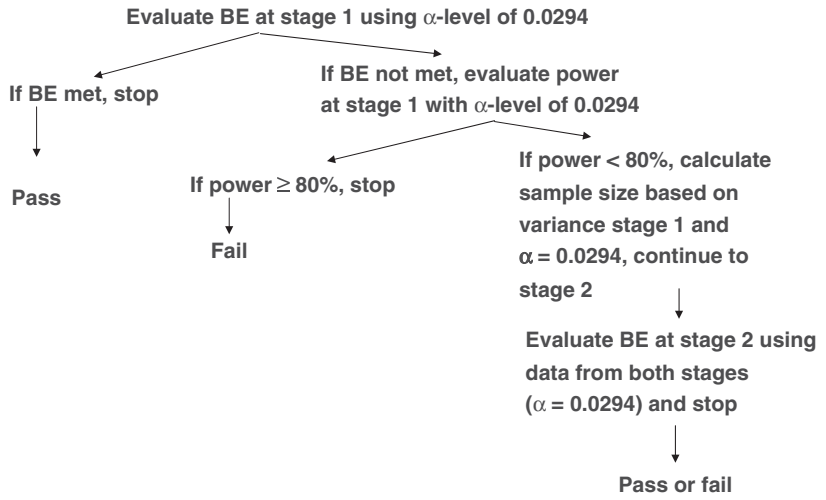

Figure 2. Adaptive sequential sample size method B.

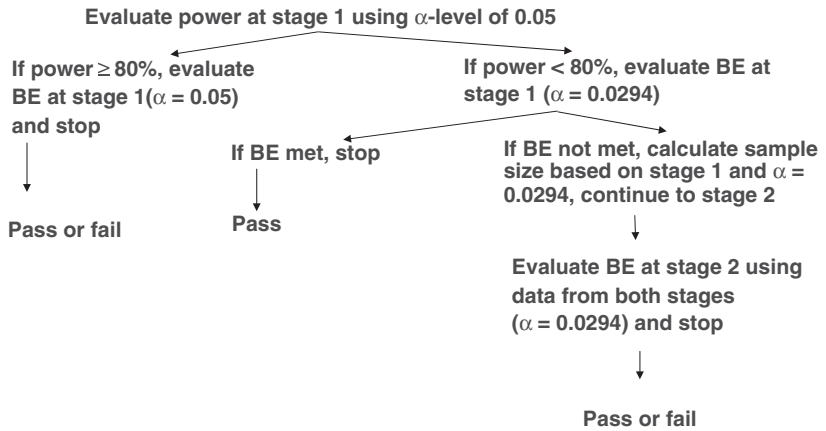

Figure 3. Adaptive sequential sample size method C.

2.3. Simulation methodology and formulas

Two-way crossover BE studies with two stages were simulated using methods A, B, C, and D described above. The individual ln(Test)−ln(Reference) differences from a two-way crossover BE design were simulated as normally distributed values with mean $\ln(\theta)$ and variance $2\sigma^2$, where θ was the true ratio of the Test to the Reference geometric means and σ^2 was the true intrasubject variance of the drug. Reference is also made to the true intrasubject *coefficient of variation* (CV), based on an assumed log normal distribution [25,26], given by

$$\text{intrasubject CV}(\%) = 100\sqrt{e^{\sigma^2} - 1}$$

For each simulated study, an even number of subjects at each stage, n_1 and n_2 (when applicable) were simulated. BE was evaluated using the two one-sided *t*-tests [27]. The power was calculated using the modification of Hauschke *et al.* [26]

DOI: 10.1002/pst

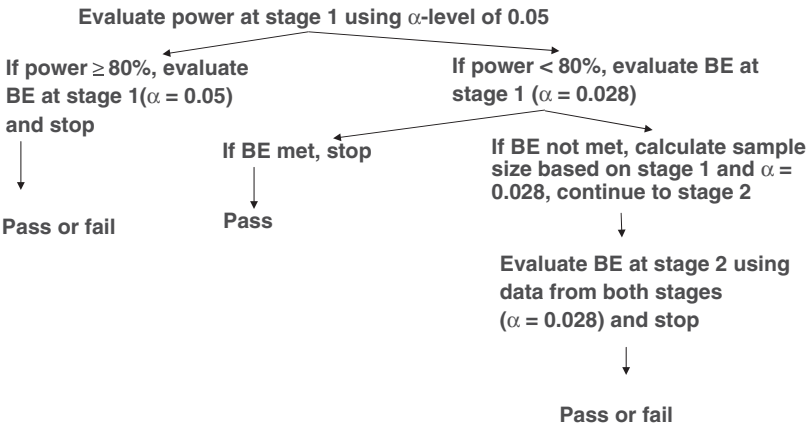

Figure 4. Adaptive sequential sample size method D.

expressions as

$$1-\beta = F_t\left(\frac{\ln(1.25/\theta)}{s\sqrt{2/n}} - t_{1-\alpha,DF}, DF\right) - F_t\left(\frac{-\ln(1.25 * \theta)}{s\sqrt{2/n}} + t_{1-\alpha,DF}, DF\right)$$

where $1-\beta$ represents the power, s is the sample standard deviation (i.e. the estimate of σ), DF is the degrees of freedom associated with the error, $F_t(x, DF)$ is the cumulative distribution function of Student's t-distribution with DF degrees of freedom, and $t_{1-\alpha,DF}$ is the $(1-\alpha)$th percentile of a Student's t density function.

For the sample size calculation, iterations were performed to find the smallest even n that was needed to attain the desirable power, $1-\beta$.

At the first stage, the estimate, s_1^2 of the true σ^2 was calculated as follows.

Let X_{ijk} (assumed to be normally distributed) represent the ln(Test)−ln(Reference) difference for the kth subject in Sequence j ($j = 1$ or 2) of Stage i ($i = 1$ or 2), then the formula for the error sum of squares (SS) for Stage 1 ($SS1$) is

$$SS1 = \frac{1}{2}\sum_{j=1}^{2}\sum_{k=1}^{n_1/2}(X_{1jk} - \bar{X}_{1j\cdot})^2 = \frac{1}{2}\sum_{j=1}^{2}\left[\left(\sum_{k=1}^{n_1/2}X_{1jk}^2\right) - \frac{(\sum_{k=1}^{n_1/2}X_{1jk})^2}{n_1/2}\right]$$

where

$$\bar{X}_{1j\cdot} = \frac{2}{n_1}\sum_{k=1}^{n_1/2}X_{1jk}$$

and then $s_1^2 = SS1/(n_1 - 2)$.

The division by 2 in the formula for $SS1$ is to put things on a 'per-observation' basis; that is, σ^2, the quantity generally referred to as 'within-subject variance,' is *one-half* the variance of a within-subject Test−Reference difference. This quantity is in fact a linear combination of the 'pure' intrasubject variance of the Test product, the 'pure' intrasubject variance of the Reference product, and any contribution to the variance from *Subject-by-Treatment interaction*.

At the second stage, the estimate, s_2^2 of the true σ^2 was calculated as follows:

$$\begin{aligned} s_2^2 &= SS1/(n_1 - 2) \quad \text{with } n_1 - 2 \text{ degrees of freedom} \quad \text{if } n_2 = 0 \\ &= (SS1 + SSmean)/(n - 3) \quad \text{with } n - 3 \text{ degrees of freedom} \quad \text{if } n_2 = 2 \\ &= (SS1 + SSmean + SS2)/(n - 3) \quad \text{with } n - 3 \text{ degrees of freedom} \quad \text{if } n_2 > 2 \end{aligned}$$

DOI: 10.1002/pst

where

$$SS2 = \frac{1}{2}\sum_{j=1}^{2}\sum_{k=1}^{n_2/2}(X_{2jk} - \bar{X}_{2j\cdot})^2$$

$$= \frac{1}{2}\sum_{j=1}^{2}\left[\left(\sum_{k=1}^{n_2/2} X_{2jk}^2\right) - \frac{(\sum_{k=1}^{n_2/2} X_{2jk})^2}{n_2/2}\right]$$

$$\bar{X}_{2j\cdot} = \frac{2}{n_2}\sum_{k=1}^{n_2/2} X_{2jk}$$

and

$$SSmean = \frac{(\bar{X}_{1\cdot\cdot} - \bar{X}_{2\cdot\cdot})^2}{(2/n_1 + 2/n_2)}$$

s_1^2 and s_2^2 were used to calculate the two one-sided t-tests, the power and the sample size calculations, at stages 1 and 2, respectively. (Note that these formulae are applicable to the *balanced* studies we simulated, but would have to be modified for *unbalanced* studies.)

At the end of stage 1, the $1-2\alpha$ confidence interval is

$$\bar{X}_{1\cdot\cdot} \pm t_{1-\alpha(n_1-2)}\sqrt{s_1^2\frac{2}{n_1}}$$

If we go on to stage 2, the $1-2\alpha$ confidence interval at the end of stage 2 is

$$\bar{X}_{\cdot\cdot\cdot} \pm t_{1-\alpha(n-3)}\sqrt{s_2^2\frac{2}{n}},$$

where $n = n_1 + n_2$.

It is useful to note that $SS1$ also represents the error SS from a general linear model (GLM) ANOVA on the original ln-transformed observations (i.e. not ln(Test)−ln(Reference) differences) from stage 1 only, with sequence, subject(sequence), period, and treatment effects in the model. Similarly, $SS2$ represents the error SS from a GLM ANOVA on the original ln-transformed observations from stage 2 only. *SSmean* represents the stage × treatment effect SS from a GLM ANOVA model including sequence, treatment, stage, period, stage × treatment, subject(sequence × stage) on the original ln-transformed observations from stages 1 and 2 combined.

A very simple way for calculating the appropriate s^2 term for each of the three cases above ($n_2 = 0$, $n_2 = 2$, $n_2 > 2$) is to perform GLM ANOVA on the ln(Test)−ln(Reference) differences (from all of the data) using stage, sequence, and stage × sequence effects in the model. (If only one stage is conducted this model reduces to just a sequence effect.) Then the appropriate s^2 term is given directly by (stage SS + error SS)/($2DF$), where DF is the appropriate degrees of freedom, as given in the equations above. If the individual ln-transformed data are to be used in the analysis instead of the differences, then the error term derived from the GLM ANOVA model including sequence, stage, period(stage), treatment, subject(sequence × stage) will give the appropriate s^2 term.

The simulations were performed using Compaq Visual Fortran, Standard Edition 6.1.0. A different randomly selected seed was used for each scenario. A scenario was defined as a specific combination of ratio of geometric means, intrasubject CV, n_1 and Method (A, B, C, or D). One million BE studies each consisting of $n = (n_1 + n_2)$ Test−Reference differences were simulated for each different scenario. Simulations were performed with true ratio of geometric means of 0.95 and 1.25, intrasubject coefficients of variation from 0.10 to 1.00, power of 80%, type I error at each stage of 0.028, 0.0294, or 0.05 (as described in methods A, B, C, and D above), n_1 varying between 12 and 60. With one million simulated studies at a 0.05 significance level, the standard error of estimated type I error rates is 0.0002. Standard errors of power estimates and percentages in phase 2 are each not more than 0.0005.

3. RESULTS

3.1. Type I error rates and powers

Table I shows the estimated type I error rates for methods A, B, C, and D and estimated powers for methods B and C.

The type I error for Method A is seen to be inflated in certain scenarios. For example, with a

DOI: 10.1002/pst

Table I. Estimated type I error rates and powers.

		Estimated type I error rate				Estimated power	
		Method				Method	
Sample size stage 1 (n_1)	Intrasubject CV (%)	A	B	C	D	B	C
12	10		0.0297	0.0496	0.0498	0.9772	0.9890
24	10		0.0294	0.0500	0.0500	0.9999	1.0000
36	10		0.0294	0.0500	**0.0504**	1.0000	1.0000
48	10		0.0292	0.0501	0.0502	1.0000	1.0000
60	10		0.0294	**0.0504**	0.0501	1.0000	1.0000
12	20	***0.0584***	0.0463	**0.0510**	0.0499	0.8429	0.8473
24	20	**0.0505**	0.0320	0.0490	0.0493	0.8810	0.9097
36	20	0.0497	0.0294	0.0499	0.0499	0.9550	0.9750
48	20	0.0500	0.0292	0.0495	0.0497	0.9885	0.9944
60	20	0.0500	0.0297	0.0500	0.0500	0.9973	0.9989
12	30	***0.0575***	0.0437	0.0441	0.0415	0.7857	0.7860
24	30	***0.0550***	0.0475	0.0492	0.0475	0.8305	0.8314
36	30	***0.0523***	0.0397	0.0477	0.0471	0.8379	0.8470
48	30	0.0502	0.0324	0.0494	0.0495	0.8548	0.8873
60	30	0.0498	0.0296	0.0502	0.0499	0.8997	0.9362
12	40		0.0344	0.0346	0.0330	0.7507	0.7505
24	40		0.0433	0.0435	0.0408	0.8033	0.8036
36	40		0.0485	0.0489	0.0464	0.8236	0.8239
48	40		0.0458	0.0469	0.0455	0.8303	0.8309
60	40		0.0409	0.0470	0.0456	0.8312	0.8361
12	50		0.0309	0.0311	0.0296	0.7359	0.7357
24	50		0.0338	0.0339	0.0319	0.7829	0.7828
36	50		0.0420	0.0418	0.0395	0.8052	0.8050
48	50		0.0484	0.0480	0.0456	0.8192	0.8192
60	50		0.0483	0.0483	0.0461	0.8245	0.8249
12	60		0.0297	0.0299	0.0284	0.7286	0.7286
24	60		0.0307	0.0307	0.0290	0.7747	0.7740
36	60		0.0333	0.0331	0.0315	0.7901	0.7908
48	60		0.0399	0.0399	0.0373	0.8045	0.8044
60	60		0.0466	0.0472	0.0442	0.8145	0.8150
12	70		0.0294	0.0294	0.0280	0.7246	0.7250
24	70		0.0299	0.0298	0.0287	0.7708	0.7703
36	70		0.0306	0.0308	0.0290	0.7850	0.7851
48	70		0.0328	0.0325	0.0306	0.7939	0.7933
60	70		0.0381	0.0380	0.0359	0.8016	0.8017
12	80		0.0292	0.0292	0.0276	0.7211	0.7223
24	80		0.0298	0.0301	0.0285	0.7676	0.7682
36	80		0.0303	0.0299	0.0286	0.7827	0.7828
48	80		0.0303	0.0302	0.0288	0.7895	0.7893
60	80		0.0318	0.0319	0.0301	0.7942	0.7945

DOI: 10.1002/pst

Table I (*continued*)

		Estimated type I error rate				Estimated power	
		Method				Method	
Sample size stage 1 (n_1)	Intrasubject CV (%)	A	B	C	D	B	C
12	90		0.0289	0.0285	0.0278	0.7206	0.7204
24	90		0.0298	0.0298	0.0284	0.7668	0.7669
36	90		0.0296	0.0296	0.0284	0.7803	0.7805
48	90		0.0297	0.0298	0.0287	0.7874	0.7873
60	90		0.0300	0.0301	0.0286	0.7926	0.7922
12	100		0.0291	0.0290	0.0274	0.7195	0.7194
24	100		0.0298	0.0295	0.0283	0.7659	0.7657
36	100		0.0298	0.0297	0.0282	0.7798	0.7801
48	100		0.0297	0.0297	0.0285	0.7864	0.7864
60	100		0.0301	0.0297	0.0281	0.7906	0.7903

Note: Values in italics are considered to be nonnegligible inflation of the type I error rate. The standard errors for type I error rates of 0.05 are 0.0002. Estimated type I error rates of 0.05036 or greater are statistically significantly greater than 0.05 by a one-sided 5% test and are shown in bold. The standard errors of power estimates are not more than 0.0005. Power is estimated at a true ratio of geometric means of 0.95.

first stage of 12 subjects, the inflation in type I error rate is up to about 16% (to 0.058). A first-stage sample size of 8 had estimated type I error rates of up to 0.066 (results not shown). Because Method A clearly did not satisfy our requirement for at most a negligible increase in type I error rate, we conducted only a portion of the type I error simulations and did not consider Method A further.

Method C is a variation of Method B, in which a test at $\alpha = 0.05$ is allowed at the first stage if the power at that stage is at least 80%. That is, if n_1 provides sufficient power, the study will be, in effect, like a conventional single-stage study with the customary α of 0.05. This variation necessarily must have higher type I error rate than Method B. Method D is a variation of Method C designed to be slightly more conservative than Method C; the maximum estimated type I error rate for Method D in Table I was 0.0504 compared to 0.0510 for Method C. Further, all estimated type I error rates for Method C, including some not shown for CVs of 20–30% in 1% increments, never exceeded 0.052, our criterion for nonnegligible inflation of type I error rate. Because Method D is more conservative and would require a larger average sample size than Method C (data not shown), we do not consider it further in this manuscript.

Type I error rates for Method B are consistently conservative; those for Method C nearly so. Method B is comparatively more conservative than Method C in cases of lower intrasubject CV (from 10% to 30% intrasubject CV). Power results favor Method C over Method B, though the difference is seldom large. The largest difference in this set of comparisons is 0.8997 for B vs 0.9362 for C (intrasubject CV of 30% and $n_1 = 60$).

Table II shows the impact of a two-stage design in terms of sample size for Methods B and C. The average total sample size, the median, and 5th and 95th percentiles for the total sample size, and the percentage of studies that will continue to the second stage are listed. For a given CV, the average total sample sizes are approximately constant when the initial sample size (n_1) is too small for the actual CV (i.e. first stage is under-powered). For larger initial sample sizes, the average total sample size increases, reflecting that the initial sample sizes were larger than needed. The 5th and 95th percentiles demonstrate that there can be substantial variability in the total sample size. Comparing the two methods, as with the power comparison, Method C goes to the second stage less often and has smaller average total sample sizes than Method B for cases where

DOI: 10.1002/pst

Table II. Total sample sizes and proportion requiring second stage.

		Ratio = 1.25				Ratio = 0.95			
				% Studies				% Studies	
		Mean n total (5th, 50th, 95th)		In stage 2		Mean n total (5th, 50th, 95th)		In stage 2	
Sample size stage 1 (n_1)	Intra-subject CV (%)	Method B	Method C	Method B	Method C	Method B	Method C	Method B	Method C
12	10	12.1 (12, 12, 12)	12.0 (12, 12, 12)	3.8	0.9	12.0 (12, 12, 12)	12.0 (12, 12, 12)	0.6	0.2
24	10	24.0 (24, 24, 24)	24.0 (24, 24, 24)	0.0	0.0	24.0 (24, 24, 24)	24.0 (24, 24, 24)	0.0	0.0
36	10	36.0 (36, 36, 36)	36.0 (36, 36, 36)	0.0	0.0	36.0 (36, 36, 36)	36.0 (36, 36, 36)	0.0	0.0
48	10	48.0 (48, 48, 48)	48.0 (48, 48, 48)	0.0	0.0	48.0 (48, 48, 48)	48.0 (48, 48, 48)	0.0	0.0
60	10	60.0 (60, 60, 60)	60.0 (60, 60, 60)	0.0	0.0	60.0 (60, 60, 60)	60.0 (60, 60, 60)	0.0	0.0
12	20	23.2 (12, 22, 40)	23.1 (12, 22, 40)	88.1	80.0	20.6 (12, 18, 40)	20.6 (12, 18, 40)	56.4	53.8
24	20	26.0 (24, 24, 34)	25.4 (24, 24, 34)	34.5	14.2	24.6 (24, 24, 28)	24.4 (24, 24, 24)	8.6	4.4
36	20	36.0 (36, 36, 36)	36.0 (36, 36, 36)	0.7	0.0	36.0 (36, 36, 36)	36.0 (36, 36, 36)	0.1	0.0
48	20	48.0 (48, 48, 48)	48.0 (48, 48, 48)	0.0	0.0	48.0 (48, 48, 48)	48.0 (48, 48, 48)	0.0	0.0
60	20	60.0 (60, 60, 60)	60.0 (60, 60, 60)	0.0	0.0	60.0 (60, 60, 60)	60.0 (60, 60, 60)	0.0	0.0
12	30	47.1 (20, 44, 84)	47.1 (20, 44, 84)	98.0	97.4	46.5 (12, 44, 84)	46.5 (12, 44, 84)	93.1	92.7
24	30	46.9 (24, 46, 72)	46.7 (24, 46, 72)	95.0	90.4	39.9 (24, 38, 70)	39.9 (24, 38, 70)	58.3	56.6
36	30	47.6 (36, 46, 66)	46.5 (36, 46, 66)	81.1	58.0	40.7 (36, 36, 62)	40.5 (36, 36, 62)	29.0	22.7
48	30	51.3 (48, 48, 64)	49.9 (48, 48, 64)	39.2	11.7	48.9 (48, 48, 56)	48.6 (48, 48, 48)	9.4	3.4
60	30	60.3 (60, 60, 62)	60.1 (60, 60, 60)	5.6	0.3	60.1 (60, 60, 60)	60.0 (60, 60, 60)	1.0	0.1
12	40	79.1 (32, 74, 142)	79.2 (32, 74, 142)	99.6	99.5	79.1 (32, 74, 142)	79.0 (32, 74, 142)	99.0	99.0
24	40	78.7 (44, 76, 120)	78.7 (44, 76, 120)	98.4	98.2	76.1 (24, 76, 120)	76.2 (24, 76, 120)	90.2	90.2
36	40	78.3 (48, 78, 112)	78.3 (48, 78, 112)	97.0	96.1	67.3 (36, 70, 112)	67.3 (36, 70, 112)	66.3	65.9
48	40	78.5 (50, 78, 108)	78.1 (48, 78, 106)	95.7	89.3	64.1 (48, 48, 104)	63.9 (48, 48, 104)	45.9	43.8

DOI: 10.1002/pst

60	40	79.1 (60, 78, 104)	77.3 (60, 78, 104)	88.6	64.4	67.4 (60, 60, 98)	66.9 (60, 60, 98)	31.0	24.3
12	50	117.5 (48, 110, 212)	117.6 (48, 110, 212)	99.9	99.9	117.5 (48, 110, 212)	117.4 (48, 110, 212)	99.8	99.8
24	50	117.4 (66, 114, 180)	117.4 (66, 114, 180)	99.6	99.6	117.0 (66, 114, 180)	117.1 (66, 114, 180)	98.7	98.7
36	50	116.8 (74, 116, 166)	116.8 (74, 116, 166)	98.5	98.5	112.7 (36, 116, 166)	112.6 (36, 116, 166)	90.8	90.8
48	50	116.1 (76, 116, 160)	116.1 (76, 116, 160)	97.4	97.4	102.3 (48, 110, 160)	102.3 (48, 110, 160)	73.9	73.8
60	50	116.1 (78, 116, 154)	116.1 (80, 116, 154)	97.0	96.4	95.5 (60, 98, 152)	95.4 (60, 98, 152)	57.8	57.6
12	60	161.0 (66, 150, 292)	160.8 (66, 150, 292)	100.0	100.0	160.8 (66, 150, 292)	160.7 (66, 150, 292)	100.0	100.0
24	60	160.9 (92, 156, 246)	160.8 (92, 156, 246)	99.9	99.9	160.7 (92, 156, 246)	160.7 (92, 156, 246)	99.8	99.8
36	60	160.7 (104, 158, 228)	160.7 (104, 158, 228)	99.6	99.6	159.9 (102, 158, 228)	159.9 (102, 158, 228)	98.6	98.6
48	60	159.9 (108, 158, 218)	159.9 (108, 158, 218)	98.8	98.8	155.0 (48, 158, 218)	155.0 (48, 158, 218)	92.5	92.6
60	60	159.0 (110, 158, 212)	158.9 (110, 158, 212)	97.8	97.7	144.0 (60, 154, 212)	144.1 (60, 154, 212)	80.1	80.2
12	70	207.7 (84, 194, 378)	207.8 (84, 194, 378)	100.0	100.0	207.6 (84, 194, 378)	207.7 (84, 194, 378)	100.0	100.0
24	70	207.7 (118, 202, 318)	207.5 (118, 202, 318)	100.0	100.0	207.6 (118, 202, 318)	207.6 (118, 202, 318)	100.0	100.0
36	70	207.6 (134, 204, 296)	207.7 (134, 204, 296)	99.9	99.9	207.6 (134, 204, 296)	207.5 (134, 204, 296)	99.8	99.8
48	70	207.4 (142, 204, 282)	207.4 (142, 204, 282)	99.7	99.7	206.7 (142, 204, 282)	206.5 (142, 204, 282)	98.8	98.8
60	70	206.6 (148, 206, 274)	206.7 (146, 206, 274)	99.0	99.0	201.7 (60, 206, 274)	201.6 (60, 206, 274)	94.3	94.2
12	80	256.8 (102, 240, 468)	257.1 (104, 240, 468)	100.0	100.0	256.9 (102, 240, 468)	257.1 (104, 240, 468)	100.0	100.0
24	80	257.0 (146, 250, 394)	256.8 (146, 250, 394)	100.0	100.0	256.7 (146, 250, 394)	256.9 (146, 250, 394)	100.0	100.0
36	80	257.0 (164, 252, 366)	256.9 (164, 252, 366)	100.0	100.0	257.0 (164, 252, 366)	256.9 (164, 252, 366)	100.0	100.0
48	80	256.9 (176, 254, 350)	256.8 (176, 254, 350)	100.0	100.0	256.9 (176, 254, 350)	256.8 (176, 254, 350)	99.8	99.8
60	80	256.6 (184, 254, 340)	256.7 (184, 254, 340)	99.8	99.8	255.7 (184, 254, 340)	255.8 (184, 254, 340)	99.0	99.0
12	90	307.5 (124, 288, 560)	307.6 (122, 288, 560)	100.0	100.0	307.4 (122, 288, 560)	307.7 (124, 288, 560)	100.0	100.0
24	90	307.6 (174, 298, 474)	307.6 (174, 298, 472)	100.0	100.0	307.6 (174, 298, 472)	307.4 (174, 298, 472)	100.0	100.0
36	90	307.5 (198, 302, 438)	307.6 (196, 302, 438)	100.0	100.0	307.6 (196, 302, 438)	307.6 (196, 302, 438)	100.0	100.0
48	90	307.6 (212, 304, 420)	307.4 (212, 304, 418)	100.0	100.0	307.6 (212, 304, 418)	307.5 (212, 304, 420)	100.0	100.0
60	90	307.6 (220, 304, 406)	307.5 (222, 304, 406)	100.0	100.0	307.3 (220, 304, 406)	307.4 (220, 304, 406)	99.9	99.9
12	100	358.6 (144, 336, 654)	358.9 (144, 336, 654)	100.0	100.0	358.9 (142, 336, 654)	358.9 (144, 336, 654)	100.0	100.0
24	100	358.9 (202, 348, 552)	358.8 (202, 348, 552)	100.0	100.0	358.7 (202, 348, 550)	358.7 (202, 348, 552)	100.0	100.0
36	100	358.8 (230, 352, 512)	358.9 (230, 352, 512)	100.0	100.0	358.9 (230, 352, 512)	358.7 (230, 352, 512)	100.0	100.0
48	100	358.8 (246, 354, 490)	358.7 (246, 354, 490)	100.0	100.0	358.7 (246, 354, 490)	358.9 (246, 354, 490)	100.0	100.0
60	100	358.7 (258, 354, 474)	358.7 (258, 354, 474)	100.0	100.0	358.8 (258, 354, 474)	358.7 (258, 354, 474)	100.0	100.0

DOI: 10.1002/pst

Table III. Sample sizes for single-stage designs.

Intrasubject CV (%)	n
10	8
20	20
30	40
40	66
50	98
60	134
70	174
80	216
90	258
100	300

Note: Smallest even sample sizes for at least 80% power at a true ratio of 0.95. These are the sample sizes if designing a single-stage trial for various presumed values of the intrasubject CV.

the initial sample size is reasonable for the CV. When the initial sample size, n_1, is small for the actual CV, both methods will go to the second stage almost all the time, and the average total sample sizes of the two methods are similar.

In comparison to Table II, Table III shows the sample sizes for a preplanned single-stage design with *a priori* knowledge of the population intrasubject CV. The average sample sizes in Table II, with a few exceptions, are somewhat larger than for the corresponding single-stage designs. There is some cost to using a two-stage design when the *a priori* variance assumption is correct. If the initial sample size, n_1, is too small for the actual CV, the average total sample size for the two-stage design will be about 20% higher than that for the single-stage design with a correct choice of CV for determining the sample size.

4. EXAMPLES

The data for these examples are included in Appendix A.

Example 1

Method B:

1. Based on the stage 1 data, the mean difference ($\ln(T)$–$\ln(R)$) is 0.16785, $SS1$ is 0.20977, and s_1^2 is 0.020977. Using these to test BE at the $\alpha =$ 0.0294 level, with $DF = 10$, we obtain a two-sided CI for the ratio (T/R) of geometric means of 104.27−134.17%, which does not meet the 80–125% criterion.
2. The power with 12 subjects at an α of 0.0294, a presumed ratio of geometric means (T/R) of 0.95, and 10 DF, is 75.6%. Because this is less than 80%, we will need to conduct a second stage. Using the stage 1 data, we calculate the smallest even number of subjects required to achieve at least 80% power assuming a population ratio (T/R) of geometric means of 0.95, an α of 0.0294, and $(n_1 + n_2 - 3)$ degrees of freedom to be 14 subjects (power = 83.1%). Since we have already used 12 subjects in stage 1, stage 2 will be conducted with only two subjects.
3. Using the data from both stages, we find that the mean ($\ln(T)-\ln(R)$) is 0.14401 and $SSmean$ is 0.023868. $SS1$ is based on stage 1 data only, and so is still 0.20977. Further, $s_2^2 = (SS1 + SSmean)/11 = 0.0212240$. Using these to test BE at the $\alpha = 0.0294$ level with $DF = 11$, we obtain a two-sided CI for the ratio (T/R) of geometric means of 102.83−129.71%, which still does not meet the 80–125% acceptance criterion. We stop here and are unable to show BE.

Method C:

1. Using the same data from stage 1 only, we find that the power, assuming a population mean (T/R) ratio of 0.95, an α of 0.05, and $DF = 10$, is 84.1%, which precludes proceeding to stage 2.
2. Using the data from stage 1 only to test BE at the $\alpha = 0.05$ level, with $DF = 10$, we obtain a two-sided CI for the ratio (T/R) of geometric means of 106.26−131.66%, which does not meet the 80–125% criterion. We stop here and are unable to show BE.

Example 2

Method B:

1. Based on the stage 1 data, the mean ($\ln(T)$–$\ln(R)$) is 0.08396, $SS1$ is 0.32634, and s_1^2 is 0.032634. Using these to test BE at the $\alpha =$

DOI: 10.1002/pst

0.0294 level, with $DF = 10$, we obtain a two-sided CI for the ratio (T/R) of geometric means of 92.93−127.28%, which does not meet the 80–125% criterion.
2. The power with 12 subjects with a presumed population geometric mean (T/R) ratio of 0.95, 10 DF, and an $\alpha = 0.0294$ is 50.5%. Because this is less than 80%, we will need to conduct a second stage. Using the stage 1 data, we calculate the smallest even number of subjects required to achieve at least 80% power assuming a population ratio (T/R) of geometric means of 0.95, an α of 0.0294, and $(n_1 + n_2 - 3)$ degrees of freedom to be 20 subjects (power = 82.4%). Since we have already used 12 subjects in stage 1, stage 2 will be conducted with eight subjects.
3. Using the data from both stages, we find that the mean $(\ln(T) - \ln(R))$ is 0.014439, $SSmean$ is 0.072493, and $SS2$ is 0.38140. $SS1$ is based on stage 1 data only, and so is still 0.32634. Continuing, $s_2^2 = (SS1 + SSmean + SS2)/17 = 0.045896$. Using these to test BE at the $\alpha = 0.0294$ level with $DF = 17$, we obtain a two-sided CI for the ratio (T/R) of geometric means of 88.45−116.38%, which meets the 80–125% acceptance criterion. We stop here and conclude BE, irrespective of the fact that we have not yet achieved the desired power of 80% (power = 66.3%).

Method C:

1. Using the same data from stage 1 only, we find that the power, assuming a population mean (T/R) ratio of 0.95, $DF = 10$, and an α of 0.05, is 64.9%. Because this is less than 80%, we evaluate the BE at an α of 0.0294 and $DF = 10$ and obtain two-sided confidence intervals for the ratio (T/R) of geometric means of 92.93−127.28%. Because this does not meet the BE criteria of 80–125%, we must conduct a second stage.
2. Using the data from stage 1, we calculate the smallest even number of subjects required to achieve at least 80% power with a presumed population ratio (T/R) of geometric means of 0.95, an α of 0.0294, and $(n_1 + n_2 - 3)$ degrees of freedom. We find that 20 subjects would give us 82.4% power. Because we have already used 12 subjects in stage 1, we conduct stage 2 with eight subjects.
3. Using the data from both stages, we find that the two-sided CI for the ratio (T/R) of geometric means at an α level of 0.0294 and $DF = 17$ is 88.45−116.38%, which meets the 80–125% acceptance criterion. We stop here and conclude BE, irrespective of the fact that we have not yet achieved the desired power of 80% (power = 66.3%).

5. DISCUSSION AND RECOMMENDATIONS

The goal of our group was to validate at least one method for two-stage designs that could be used for BE studies. Methods B and C meet our criteria of not more than minimal inflation of type I error rate. We recommend that regulatory agencies accept either. It is our understanding that the FDA has accepted studies with designs like those considered here.

For sponsors, there is a small power advantage to Method C over Method B, so we consider Method C as the method of choice. Another advantage of Method C is that it was designed so that if the study were found to have adequate power at the first stage, the α for that study would be the same as if it were designed to be single-stage study.

We must note that the intention to employ a group sequential method for an interim analysis with possible early termination and conclusion of BE *must* be described in the study protocol. Also the method should be prespecified.

The more difficult question is when to use two-stage designs instead of single-stage (standard) designs. There are costs involved with two-stage designs, the first of which is the need by the laboratory to analyze stage 1 samples before deciding on the need for a second stage. A second is the need to stop and then perhaps restart the study while maintaining sufficient continuity so

DOI: 10.1002/pst

that results from the second stage may be reasonably combined with those from the first. Last, the maximum possible sample size with a two-stage design will be larger than with a single-stage design *for which the right variance is used in planning*. (To see this, compare the 95th percentiles from Table - II with the corresponding entries in Table III.) The benefit of the two-stage designs considered here is that they provide insurance for not using the right variance in the planning with a penalty of about 20% in average sample size if the variance used for planning was correct. An alternate type of insurance is to use an initial variance estimate that is larger than required, with the attendant cost of a possibly larger than needed study. That cost can be balanced against the costs of two-stage designs.

This study did not seek to find the *best* possible two-stage design, but rather to find *good* ones that could be used by sponsors without further validation. The authors look forward to contributions from others in this area. More work needs to be done to examine the extension to more than two stages, issues associated with pooling results from two or more stages, inclusion of a futility rule, minimum sample sizes for second stage if performed, and upper limits on the total sample size.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Mr Jean Lavigne, from MDS Pharma Services, for his programming support and thank the PQRI for its support and encouragement in this work.

REFERENCES

1. Food and Drug Administration, Center for Drug Evaluation and Research (CDER). *Guidance for Industry*: *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products – General Considerations*. March 2003.
2. Health Canada, Ministry of Health – Health Products and Food Branch. *Guidance for Industry*: *Conduct and Analysis of Bioavailability and Bioequivalence Studies – Part A*: *Oral Dosage Formulations Used for Systemic Effects*, 1992.
3. Japan National Institute of Health Sciences, Division of Drugs. *Guideline for Bioequivalence Studies of Generic Products*, 1997.
4. World Health Organization Expert Committee on Specifications for Pharmaceutical Preparations. Fortieth report, annex 7, multisource (generic) pharmaceutical products: guidelines on registration requirements to establish interchangeability. *WHO Technical Report 937*, 2006; 347–390.
5. Australia Department of Health and Aging. *Australian Regulatory Guidelines for Prescription Medicines*, *Appendix* 15: *Biopharmaceutic Studie*s, 2004.
6. Hauck WW, Preston PE, Bois FY. A group sequential approach to crossover trials for average bioequivalence. *Journal of Biopharmaceutical Statistics* 1997; **7**:87–96.
7. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
8. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrika* 1979; **35**:549–556.
9. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199.
10. Gould AL. Group sequential extensions of a standard bioequivalence testing procedure. *Journal of Pharmacokinetics and Biopharmaceutics* 1995; **23**:57–86.
11. Jennison C, Turnbull BW. Sequential equivalence testing and repeated confidence intervals, with application to normal and binary responses. *Biometrics* 1993; **40**:225–230.
12. Birkett MA, Day SJ. Internal pilot studies for estimating sample size. *Statistics in Medicine* 1994; **13**:2455–2463.
13. Coffey CS, Muller KE. Exact test size and power of a Gaussian error linear model for an internal pilot study. *Statistics in Medicine* 1999; **18**:1199–1214.
14. Coffey CS, Muller KE. Controlling test size while gaining benefits of an international pilot design. *Biometrics* 2001; **57**:625–631.
15. Denne JS, Jennison C. Estimating the sample size for a *t*-test using an internal pilot. *Statistics in Medicine* 1999; **18**:1575–1585.
16. Kieser M, Friede T. Re-calculating the sample size in internal pilot designs with control of type I error rate. *Statistics in Medicine* 2000; **19**:901–911.
17. Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* 1945; **16**:243–258.
18. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficacy of clinical trials. *Statistics in Medicine* 1990; **9**:65–72.
19. Wittes J, Schabenberger O, Zucker D, Brittain E, Proschan M. Internal pilot studies I: type I error

DOI: 10.1002/pst

rate of the naïve *t*-test. *Statistics in Medicine* 1999; **18**:3481–3491.
20. Zucker DM, Wittes JT, Schabenberger O, Brittan E. Internal pilot studies II: comparison of various procedures. *Statistics in Medicine* 1999; **18**:3493–3509.
21. Cui L, Hung MJ, Wang S-J. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**:853–857.
22. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22**:971–993.
23. Chen YHJ, DeMets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine* 2004; **23**:1023–1038.
24. Food and Drug Administration, Center for Drug Evaluation and Research (CDER). *Guidance for Industry*: *Statistical Approaches to Establishing Bioequivalence*, January 2001.
25. Johnson NL, Kotz S. *Continuous univariate distributions* – 1. Wiley: New York, 1970.
26. Hauschke D, Steinijans VW, Diletti E, Burke M. Sample size determination for bioequivalence assessment using a multiplicative model. *Journal of Pharmacokinetics and Biopharmaceutics* 1992; **20**:557–561.
27. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 1987; **15**:657–680.

APPENDIX A: DATA FOR EXAMPLES

Example 1

Subject	Stage	Sequence	TRT *T*	TRT *R*	ln (*T*)	ln (*R*)	ln (*T*)−ln (*R*)
Stage 1 data							
1	1	TR	3.3	2.9	1.19392	1.06471	0.12921
2	1	TR	5.3	6.1	1.66771	1.80829	−0.14058
3	1	TR	6.2	7.1	1.82455	1.96009	−0.13555
4	1	TR	2.0	1.2	0.69315	0.18232	0.51083
5	1	TR	3.1	3.2	1.13140	1.16315	−0.03175
6	1	TR	0.8	0.8	−0.22314	−0.22314	0.00000
7	1	RT	1.6	1.5	0.47000	0.40547	0.06454
8	1	RT	11.8	6.9	2.46810	1.93152	0.53658
9	1	RT	4.1	3.2	1.41099	1.16315	0.24784
10	1	RT	1.7	1.2	0.53063	0.18232	0.34831
11	1	RT	4.4	3.5	1.48160	1.25276	0.22884
12	1	RT	3.1	2.4	1.13140	0.87547	0.25593
Stage 2 data							
13	2	TR	7.3	8.5	1.98787	2.14007	−0.15219
14	2	RT	1.4	1.2	0.33647	0.18232	0.15415

Example 2

Subject	Stage	Sequence	TRT *T*	TRT *R*	ln (*T*)	ln (*R*)	Difference
Stage 1 data							
1	1	TR	4.8	4.6	1.56862	1.52606	0.04256
2	1	TR	1.4	1.3	0.33647	0.26236	0.07411

DOI: 10.1002/pst

Subject	Stage	Sequence	TRT *T*	TRT *R*	ln (*T*)	ln (*R*)	ln (*T*)−ln (*R*)
3	1	TR	3.9	4.4	1.36098	1.48160	−0.12063
4	1	TR	1.5	2.4	0.40547	0.87547	−0.47000
5	1	TR	3.8	2.9	1.33500	1.06471	0.27029
6	1	TR	3.3	2.5	1.19392	0.91629	0.27763
7	1	RT	1.5	1.7	0.40547	0.53063	−0.12516
8	1	RT	1.9	1.3	0.64185	0.26236	0.37949
9	1	RT	2.4	2.1	0.87547	0.74194	0.13353
10	1	RT	13.9	11.4	2.63189	2.43361	0.19828
11	1	RT	2.3	1.5	0.83291	0.40547	0.42744
12	1	RT	2.4	2.6	0.87547	0.95551	−0.08004
Stage 2 data							
13	2	TR	4.0	4.2	1.38629	1.43508	−0.04879
14	2	TR	0.7	0.5	−0.35667	−0.69315	0.33647
15	2	TR	10.3	11.9	2.33214	2.47654	−0.14439
16	2	TR	4.7	4.8	1.54756	1.56862	−0.02105
17	2	RT	5.5	5.9	1.70475	1.77495	−0.07020
18	2	RT	5.4	3.8	1.68640	1.33500	0.35140
19	2	RT	2.1	4.3	0.74194	1.45862	−0.71668
20	2	RT	2.4	3.6	0.87547	1.28093	−0.40547

DOI: 10.1002/pst