Les Cahiers du GERAD

G-2015-70
July 2015

CITATION ORIGINALE / ORIGINAL CITATION

J. Xu, C. Audet, C.E. DiLiberti,
W.W. Hauck, T.H. Montague,
A.F. Parr, D. Potvin, D.J. Schuirmann, Optimal
adaptive sequential designs
for crossover bioequivalence studies,
Pharmaceutical Statistics, 15(1), 15–27, 2016.

GERAD HEC Montréal
3000, ch. de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Les Cahiers du GERAD ISSN: 0711–2440

Optimal adaptive sequential designs for crossover bioequivalence studies

J. Xu, C. Audet, C.E. DiLiberti,
W.W. Hauck, T.H. Montague,
A.F. Parr, D. Potvin, D.J. Schuirmann

G–2015–70

July 2015

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Optimal adaptive sequential designs for crossover bioequivalence studies

Jialin Xu [a]
Charles Audet [b]
Charles E. DiLiberti [c]
Walter W. Hauck [d]
Timothy H Montague [e]
Alan F. Parr [f]
Diane Potvin [g]
Donald J. Schuirmann [h]

[a] *Merck, Inc., Upper Gwynedd, PA USA*

[b] *GERAD & Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7*

[c] *Montclair Bioequivalence Services, LLC, Montclair, NJ, USA*

[d] *Sycamore Consulting LLC, New Hope, PA, USA*

[e] *GlaxoSmithKline, Inc., King of Prussia, PA, USA*

[f] *GlaxoSmithKline, Inc., Research Triangle Park, NC, USA*

[g] *Excelsus Statistics, Montréal (Québec) Canada*

[h] *Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA*

charles.audet@gerad.ca

July 2015

Les Cahiers du GERAD

G–2015–70

Abstract: In prior works, this group demonstrated the feasibility of valid adaptive sequential designs for crossover bioequivalence studies. In this paper, we extend the prior work to optimize adaptive sequential designs over a range of geometric mean test/reference ratios (GMRs) of 70–143% within each of two ranges of intra-subject coefficient of variation (10–30%, and 30–55%). These designs also introduce a futility decision for stopping the study after the first stage if there is sufficiently low likelihood of meeting bioequivalence criteria if the second stage were completed, as well as an upper limit on total study size. The optimized designs exhibited substantially improved performance characteristics over our previous adaptive sequential designs. Even though the optimized designs avoided undue inflation of type I error and maintained power at $\geq 80\%$, their average sample sizes were similar to or less than those of conventional single stage designs.

Key Words: Sequential design, sample size re-estimation, adaptive design, bioequivalence.

Acknowledgments: The authors would like to thank the Product Quality Research Institute (PQRI) for its financial support and encouragement in this work.

This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

Introduction

One of the key considerations in planning a bioequivalence (BE) study is study (sample) size. The sponsor seeks neither to incur unnecessary expense with too large a sample size nor to unduly risk study failure with too small a sample size. In the case of a crossover design study, the sample size decision is highly dependent on the within-subject variability (intra-subject coefficient of variation, or ISCV), a quantity for which reliable prior estimates may not be available for the drug to be studied. While an innovator planning a BE study may have good ISCV estimates from prior studies it has conducted on the same drug, a generic manufacturer might not have access to reliable ISCV estimates for the drug to be studied.

This group has published two papers [1,2] with several solutions to this design problem for crossover studies, namely adaptive two-stage designs, allowing for re-estimation of the second-stage sample size based on first-stage results. Elements of the designs were drawn from Pocock's approach [3] for group sequential parallel designs in having similar significance levels at the two stages. The designs presented and validated (in terms of preserving the type I error rate) in those two papers were in some ways demonstrations of what is possible with two-stage designs. For example, methods B, C, and D attempt to test BE using stage 1 data with an alpha level more stringent than 0.05 (method B) or at 0.05 level with sufficient power (methods C and D). If it is necessary to continue to stage 2 due to insufficient power at stage 1, a sample size re-estimation based on observed ISCV and assumed GMR were used. There was no constraint on maximum sample size. These three methods differ in that method B always tests stage 1 at an alpha level of 0.0294, whereas methods C and D determine the alpha level to be applied to the stage 1 results based on the power achieved at stage 1 [1]. However, there was no attempt to optimize their performance. The purpose of this paper is to develop and present adaptive two-stage designs for two-period crossover BE studies that are optimal within certain design spaces that are described below. Two new features that we introduce into our optimized designs are an upper limit for overall study size as well as a futility criterion (also considered in other forms by other authors [4,5,6]), which allows for the abandonment of a study after the first stage if there would be little hope of meeting BE criteria if the second stage were to be conducted.

While the two-stage adaptive designs we explored in our earlier publications did offer good protection against a fair degree of uncertainty in ISCVs, there were limitations in the breadth of the ISCV ranges over which they performed well in terms of power and sample size. Therefore, with these types of designs, a one-size-fits-all approach that would be tolerant of an extremely wide range of possible ISCVs is not possible. Accordingly, in this paper we optimized these methods across two ISCV ranges that we thought would be most useful practically. This is a somewhat different approach from those reported in other publications that have sought to improve on our original work [5,6].

One of the factors that we considered in selecting the most useful ISCV ranges over which to optimize our methods was the availability of the scaled average BE (SABE) approach [7,8],[1] which has attractive performance characteristics if the expected ISCV is about 30% or higher. For those cases in which a sponsor would have the option of conducting a scaled average BE study, there would generally be little motivation to conduct an adaptive sequential design study if the ISCV were expected to exceed 30%. Consequently, an adaptive sequential design would be attractive only if the risk of observing an ISCV $\geq$ 30% were low. To address this type of situation, we optimized our methods for a low-variability ISCV range of 10–30%.

However, the scaled average BE approach is not always permitted or feasible, so adaptive two-stage designs may still be attractive for drugs which are expected to be highly variable (i.e., having ISCVs $\geq$ 30%). For those cases, we chose to optimize our methods over as broad an ISCV range as possible, starting at an ISCV of 30% and extending the range upward as far as acceptable power and sample size performance permitted. Because we observed acceptable performance characteristics up to an ISCV of 55%, we chose to optimize our adaptive sequential methods over a high-ISCV range of 30–55% in order to address such situations.

[1] FDA's Scaled Average Bioequivalence (SABE) approach [7] is actually a mixed-scaling approach. If the ISCV observed in the pivotal biostudy is $<$ 30%, conventional unscaled average bioequivalence criteria apply; if it is $\geq$ 30%, scaling is applied. Therefore, the SABE (mixed scaling) design could be applied to cases in which the ISCV is $<$ 30%, although there would be no benefit to the sponsor for doing so, and even some disincentive, as the SABE design requires replication and has minimum sample size requirements ($n \geq 24$) that could make it less attractive than a conventional 2-way crossover design for ISCVs under 30%. For ISCVs well above 30%, however, the sample size reduction afforded by the SABE approach can be quite substantial.

In Methods we describe our simulation and optimization approaches. The comparator procedure is a standard, single-stage, fixed sample size design. We introduce two new procedures, E and F, based on methods B, C, and D of the prior papers. For a grid of values of ISCV (within the two ranges described above) and geometric mean ratios (GMRs) spanning 0.70–1.00, and thus covering 0.70 to 1.43, we estimate, by simulation, the sample size, power and type I error rate. The comparison of sample sizes to those of single-stage designs is based on a cost function that heavily penalizes designs with sample sizes greater than those of the corresponding single-stage designs, but modestly rewards those with sample sizes smaller than those of the corresponding single stage designs. Four design parameters, described below, were then varied by a state-of-the-art optimization method to minimize the cost function while still maintaining an acceptable type I error rate and power.

Methods

Designs

Single-stage design

A conventional single-stage BE study design was used as a comparator against which the performance characteristics of the two-stage designs were evaluated. The single-stage design assumed conventional BE criteria requiring that the 90% confidence interval (corresponding to a significance level of 0.05 for the equivalent two one-sided tests [9]) for the ratio of geometric means of the test and reference products fall within 0.80–1.25. For each value of the population ISCV tested with the two-stage designs, the minimum number of subjects required to achieve at least 80% power, assuming a population geometric mean ratio (GMR) of 0.95, was determined for the single-stage design by an adaptation of the method of Hauschke et al. [10] which is described below in Equation (1).

Adaptive sequential two-stage design method E

Method E (Figure 1) is a two-stage adaptive group sequential BE design, similar to method B proposed by Potvin et al. [1]. Specifically, method E begins by testing BE first, and allows early stopping for BE with a significance level of α_1 (smaller than 0.05), regardless of the actual power at stage 1. If the BE criterion is met, the algorithm terminates and concludes BE. If the BE criterion is not met at this point, the algorithm specifies the calculation of the stage 1 power at a second significance level, α_2. If the stage 1 power is $\geq 80\%$ with significance level α_2, the stage 1 data are retested for BE at the α_2 significance level, and the algorithm terminates with a conclusion of BE or non-BE.

If the stage 1 power at the α_2 significance level is $< 80\%$, then the algorithm specifies checking whether running stage 2 would be futile, based on the calculation of a 90% confidence interval for ln(GMR), [LCL; UCL], and application of a futility rule (described below) using data from stage 1. If the futility rule is met, the algorithm terminates and concludes non-BE. Otherwise, the algorithm specifies continuation to stage 2, using a stage 2 sample size (n_2) that is calculated as described below.

The algorithm then specifies evaluating BE using combined data from both stages at a significance level of α_2. At this point the algorithm terminates regardless of the power achieved or whether or not BE is concluded.

Adaptive sequential two-stage design method F

Method F (Figure 2) is an adaptive two-stage group sequential BE design, similar to methods C and D proposed by Potvin et al. [1]. Method F first specifies evaluating the power at stage 1 at a significance level of 0.05. If the stage 1 power is at least 80%, it then specifies evaluating BE at stage 1 at a significance level of 0.05 and the algorithm terminates regardless of whether BE or non-BE is concluded. If, however, the stage 1 power is less than 80%, the algorithm then specifies testing for BE at the significance level α_1 (< 0.05), and terminating if the BE criteria are met. If the BE criteria are not met, then the algorithm specifies checking whether running stage 2 would be futile, based on the calculation of a 90% confidence interval for ln(GMR),

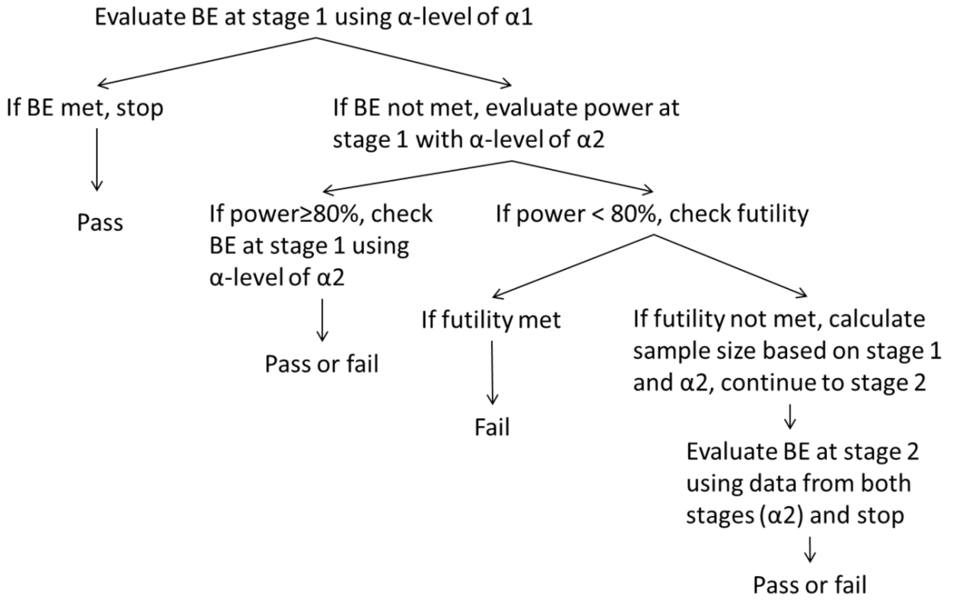

Evaluate BE at stage 1 using α-level of α1
If BE met, stop
Pass
If BE not met, evaluate power at stage 1 with α-level of α2
If power≥80%, check BE at stage 1 using α-level of α2
Pass or fail
If power < 80%, check futility
If futility met
Fail
If futility not met, calculate sample size based on stage 1 and α2, continue to stage 2
Evaluate BE at stage 2 using data from both stages (α2) and stop
Pass or fail

Figure 1: Adaptive sequential sample size method E.

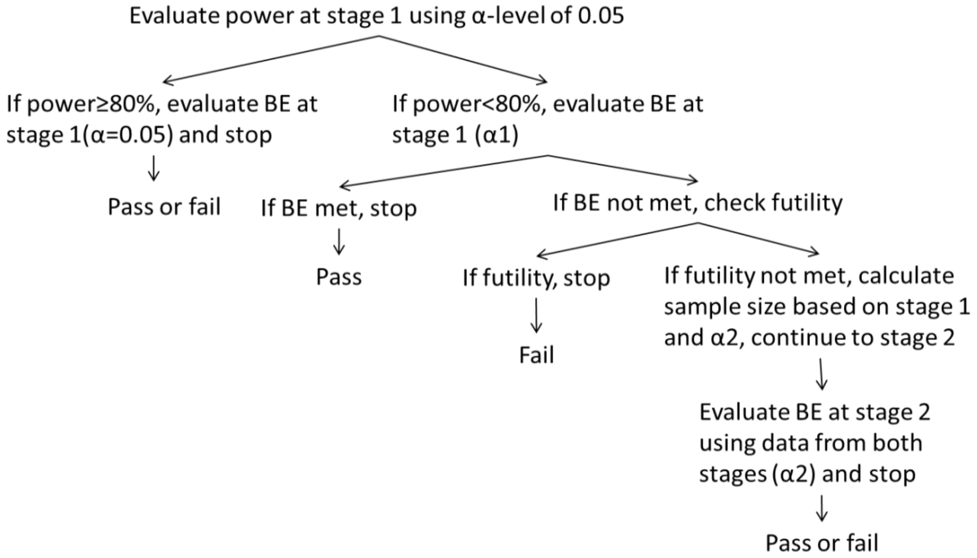

Evaluate power at stage 1 using α-level of 0.05
If power≥80%, evaluate BE at stage 1(α=0.05) and stop
Pass or fail
If power<80%, evaluate BE at stage 1 (α1)
If BE met, stop
Pass
If BE not met, check futility
If futility, stop
Fail
If futility not met, calculate sample size based on stage 1 and α2, continue to stage 2
Evaluate BE at stage 2 using data from both stages (α2) and stop
Pass or fail

Figure 2: Adaptive sequential sample size method F.

[LCL; UCL], and application of a futility rule (described below) using data from stage 1. If the futility rule is met, the algorithm terminates and concludes non-BE. Otherwise, the algorithm specifies proceeding to stage 2 using n_2 subjects calculated as described below using α_2. After completion of stage 2, the algorithm specifies evaluating BE at significance level α_2 using the combined data from both stages and concluding BE or non-BE, regardless of power.

Common features of the methods described above:

- **BE criterion:** A BE criterion at a specific significance level α means that the two-sided confidence interval for the estimated GMR (Test/Reference, on the original scale) at the $(1 - 2\alpha)$ level falls entirely within 80–125%.

- **General notes:** All power calculations were performed using the specified significance level, and assuming a population GMR of 0.95 and a population variance equal to the sample variance (ANOVA MSE) observed in stage 1. All overall type I error rate simulations were performed at a population GMR of 0.80. Because GMRs of 0.80 to 1.00 are symmetric to GMRs of 1.00 to 1.25, our simulations only explored population GMRs ≤ 1.00 without loss of generality.
- **Futility rule:** The futility rule was defined as follows. After stage 1, the lower and upper limits (LCL and UCL) of the 90% two-sided confidence interval around the ln(GMR) estimate were calculated. A futility criterion, $(f \geq 0)$, was one of the design parameters that was optimized. If the 90% confidence interval [LCL; UCL] on the natural-log scale calculated from the stage 1 data for a given study was completely outside the region of $[-f, f]$, running stage 2 was deemed to be futile, and the study was abandoned after stage 1 with a conclusion of non-BE. This futility rule is equivalent to rejecting the study for futility if the 90% confidence interval of the GMR after stage 1 (transformed back to the original scale) is completely outside $[\exp(-f)\ ,\ \exp(f)]$. Conceptually, this type of criterion concludes that further testing would be futile if, based on the stage 1 data, there is a sufficiently high degree of confidence that the population GMR is not close to 100%. This is superior to a simple test of the point estimate, in that it does not abandon poorly powered studies prematurely. A small value for f will tend to stop studies more frequently for futility, whereas a high value will tend to stop studies less frequently for futility. This futility criterion (f) is determined simultaneously with the alpha allocation (α_1, α_2) and first stage sample size (n_1). Design variants that do not inflate type I error and yield adequate power were used for optimal design search. Because of the existence of futility, the resulting nominal alpha allocation (α_1, α_2) may be higher than it would have been in the absence of futility [1].
- **Stage 2 sample size (n_2):** For methods E and F, if the decision rule determined that the study should continue to stage 2, the minimum size for stage 2 was 2, and the maximum size was $(n_{\max} - n_1)$, where $n_{\max}$ was the pre-determined maximum total allowable study size for stages 1 and 2 combined. The initial estimate of n_2 was the minimum even number of subjects required for the combined data from stages 1 and 2 to have at least 80% power, assuming a population GMR of 0.95, a population variance equal to the observed sample variance (ANOVA MSE) from stage 1, and a significance level of α_2. If this initial estimate for n_2 exceeded $(n_{\max} - n_1)$, then n_2 was set equal to $(n_{\max} - n_1)$ so as to constrain the maximum possible study size to no more than $n_{\max}$. The detailed power calculation formula is shown below in Equation (1).

Simulations

Simulation design spaces

As described earlier, methods E and F were optimized separately for each of two ISCV ranges: a low ISCV range of 10–30% and a high ISCV range of 30–55%. This yielded a total of four different design spaces, viz, method E low ISCV, method E high ISCV, method F low ISCV, and method F high ISCV. In each case, optimization was done over a range of GMR values (0.70–1.00), which further defined the design spaces, and which were intended to represent realistic streams of formulations that may or may not match the reference product well.

Within each design space, each specific design variant was defined by four parameters (α_1, α_2, n_1, and f). In the simulations for each design space, these parameters were allowed to vary to achieve optimum performance over that design space. In order to test the performance of each sequential design variant over its design space, it was evaluated at a 2 dimensional grid of test points corresponding to different values of GMR and ISCV spaced at 5% intervals within that design space. Using the low ISCV case as an example, design variant performance was tested at population GMR values of 70%, 75%, 80%, 85%, 90%, 95%, and 100% and population ISCV values of 10%, 15%, 20%, 25%, and 30% for a 7 × 5 grid with a total of 35 points.

The first step in evaluating a specific design variant (i.e., combination of α_1, α_2, n_1, and f) was to test whether it met the design criteria of providing a type I error rate of ≤ 0.05 and a power of ≥ 0.80 for all ISCV values within the design space. Type I error rate and power were determined by the fraction of studies meeting the BE criteria at population GMR values of 0.80 and 0.95, respectively. Using the low ISCV case as an example, type I error rate was assessed at population ISCV values of 10%, 15%, 20%, 25%, and 30%, while holding the population GMR constant at 0.80. Similarly, power was assessed at population ISCV values of 10%, 15%, 20%, 25%, and 30%, while holding the population GMR constant at 0.95. In other words, these tests for type I error rate and power were done on two different subsets of the overall 2D grid of GMR and ISCV combinations. If, for any of the ISCVs tested within the design space, either the type I error rate was > 0.05 or the power was < 0.80, the design variant was rejected as unacceptable, and no simulations were done for the remaining points in the 2D grid. This allowed for considerable saving of computational time.

If a particular design variant met the type I error rate and power criteria for all ISCV values within the design space, simulations were done at the remaining points in the 2D test grid. For each point in the 2D test grid, the performance of the design variant was recorded (including the percentages of studies passing overall, and at each stage, as well as sample size statistics). Based on the sample sizes at each point in the 2D test grid a cost function (described below) was calculated as a global measure of sample size within the design space. The values of the design parameters α_1, α_2, n_1, and f were then varied, as described below (Optimal design search algorithm section), to minimize this cost function.

As a comparator, both as a component of the cost function, and for the sample size statistics reported for the optimized designs, the smallest sample size for a conventional, single stage, two-way crossover design providing power ≥ 0.80 for a population GMR of 0.95 and a given population ISCV was calculated using Equation (1) below.

Constraints on design parameters

Constraints for the four design parameters were set to ensure that their optimized values would be reasonable. The minimum size of 12 for stage 1 (n_1) was selected to provide the basis for a reasonable variance estimate and BE/futility decision and because this is the smallest study that FDA will accept for the pivotal demonstration of bioequivalence. The maximum size for stage 1 was set as 30 or 60, for ISCVs $\leq 30\%$ or ISCVs $> 30\%$, respectively. The maximum total sample size from both stages was set as 42 or 180 for ISCVs $\leq 30\%$ or ISCVs $> 30\%$, respectively, based on practical considerations explained below.

Tothfalusi and Endrenyi [11] reported that the smallest SABE study that would provide at least 80% power for a population GMR of 0.95 and a population ISCV of 30% would be a 24-subject, 3-period, 3-sequence partial replicate design. In order to account for the possibility of at least a small percentage of dropouts, the smallest balanced 3-period SABE study size providing at least 80% power for a population GMR of 0.95 and a population ISCV of 30% would then start with 27 subjects. Such a design would have $27*3 = 81$ administrations of drug. The smallest balanced two-way crossover design with at least this number of drug administrations, and, therefore, a cost similar to that of the smallest practical SABE study would use 42 subjects.

The largest study sizes for the low ISCV range of 10–30% would be expected to occur at an ISCV of 30%. If the ISCV is expected to be 30%, and if the option of conducting a SABE study were available, the SABE design may be a better option because of its superior power over a sequential design at an ISCV of 30%. If the ISCV was uncertain but as high as 30% a sequential design study could be attractive as long as the cost would not exceed that of the possible alternative (SABE).[2] However, if the sequential design could potentially cost more than a SABE design, then the SABE design might still be a better option even if the population ISCV might be a little lower than 30%, as the SABE design would provide more power, especially for even higher observed ISCV values. Hence, a maximum total sample size of 42 was selected for simulations of the low ISCV range to ensure that it would provide an attractive alternative to SABE designs for ISCVs in the vicinity of 30%.

The maximum total sample size limit for the high ISCV case (180) was chosen based on practical considerations–only rarely would studies larger than this be conducted. The maximum n_1 values for the low and high ISCV cases (30 and 60, respectively) were chosen to ensure that the final optimized designs

[2] It should be noted that a partial- or fully-replicate SABE design study for which the observed ISCV is $< 30\%$ is still valid, but will not benefit from scaling, and thus becomes less desirable for ISCV values $< 30\%$.

would not start with such large stage 1 groups as to be unattractive to sponsors. The final optimized designs were found to use values of n_1 that were substantially lower than these upper design limits for n_1, and so were not constrained by these design limit selections.

Simulation of individual BE studies

For each grid point (GMR, ISCV) within each of the four design spaces, two-way crossover BE studies with up to two stages were simulated. The individual ln(Test)-ln(Reference) differences from a two-way crossover BE design were simulated as normally distributed values with population mean $\ln(\theta)$ and population variance $2\sigma^2$, where θ was the ratio of the Test to the Reference population geometric means(GMR) and σ^2 was the population intra-subject variance (ln-scale) of the drug. The population variance (ln-scale) and population ISCV (untransformed scale), based on an assumed lognormal distribution, are related by:

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

Power calculations for each combination of GMR and ISCV were performed iteratively to find the smallest even n that was needed to attain at least the desired power, $1 - \beta$:

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

where β represents the probability of a type II error, 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 probability function of Student's t distribution with DF degrees of freedom and $t_{1-\alpha,DF}$ is the $(1-\alpha)$th percentile of the Student t density function. This formula was derived by one of us (Potvin) from the work by Hauschke et al. [10].

Simulation and evaluation at each grid point within design space

Because this entire simulation process had to be repeated at multiple grid points for each change to one or more of the four parameters being optimized, the number of simulated studies involved was enormous. Therefore, in order to make the process more efficient computationally, importance sampling was used [12,13,14]. This drastically reduced the number of simulated studies required to achieve a given precision in the estimates of the performance metrics of a given sequential design variant (type I error rate, power, sample size.) Importance sampling suggests that the probabilities of claiming BE under the null hypothesis (i.e.; GMR = 1.25 or 0.8) can be estimated as:

$$\textit{Prob}\ [\textit{Claim BE}|H_0 : GMR = 1.25] = \frac{1}{m} \sum L \cdot I_R \tag{2}$$

where $I_R = 1$ when the event Claim BE occurs and L is the ratio of the densities of observations under the null and alternative hypotheses. The variance of the importance sampling estimate is always less than or equal to the variance of the direct Monte Carlo estimates for the same number of simulations. By using importance sampling, not only was the number of simulations reduced ($n = 40,000$), but both the type I error rate and power were estimated using the same set of simulations run under the alternative hypothesis (i.e.; GMR = 0.95). However, for assumed ISCV's less than 15%, the importance sampling estimate of the type I error rate failed to converge properly. As such, for those cases, the type I error was estimated using Monte Carlo sampling ($n = 100,000$).

Cost function

To evaluate and determine the optimal design, a cost function was developed to measure the deviation in the average sample size (see average sample size section below) of the adaptive sequential two-stage designs from the average sample size of the single stage design across the 2D grid of ISCVs and GMRs. As such, a "cost" value was calculated for each unique combination of ISCV and GMR, based on the difference between the average sample size of the two-stage sequential design and the sample size of the smallest single stage design providing power ≥ 0.80 at a population GMR of 0.95 and the same population ISCV using Equation (3).

The cost function penalized two-stage sequential designs that had an average sample size that was greater than the sample size of the single stage design more heavily (quadratic cost) than it rewarded those that had an average sample size less than that of the single stage design (linear reward). This was done because it was felt that reducing sample sizes below those of the corresponding single stage n was a less important goal than was preventing inflation of sequential design sample sizes beyond the corresponding single stage n. In other words, we did not want a large reward for some instances (within the 2D grid of ISCV and GMR values) of sequential design sample sizes substantially below those of the corresponding single stage n to counterbalance, and thus permit other instances of sequential design sample sizes (i.e., other points in the same 2D grid) to substantially exceed those of the corresponding single stage n.

Thus, for the ith GMR and the jth ISCV, the cost value (at a particular grid point) is defined as:

$$Cost_{\{GMR_i,CV_j\}} = \begin{cases} \left(\tilde{n}^S_{\{GMR_i,CV_j\}} - \tilde{n}^1_{\{0.95,CV_j\}}\right)^2 & \text{if } \tilde{n}^S_{\{GMR_i,CV_j\}} - \tilde{n}^1_{\{0.95,CV_j\}} > 1 \\ \left(\tilde{n}^S_{\{GMR_i,CV_j\}} - \tilde{n}^1_{\{0.95,CV_j\}}\right) & \text{if } \tilde{n}^S_{\{GMR_i,CV_j\}} - \tilde{n}^1_{\{0.95,CV_j\}} \leq 1 \end{cases} \tag{3}$$

Where $\tilde{n}^S_{\{GMR_i,CV_j\}}$ is the average sample size of the two-stage sequential design for the ith GMR and the jth ISCV and $\tilde{n}^1_{\{0.95,CV_j\}}$ is the sample size of the single stage design for the jth ISCV, which assumes GMR of 0.95. The overall cost function for the two-stage sequential design for the $n \times m$ matrix of GMRs and CVs is then defined as:

$$\text{Cost function } = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} Cost_{\{GMR_i,CV_j\}}$$

An optimal design is defined as the design that has at least 80% power and at most 5% overall type I error rate for all ISCV values within the specified ISCV range, and achieves the smallest cost function value described above.

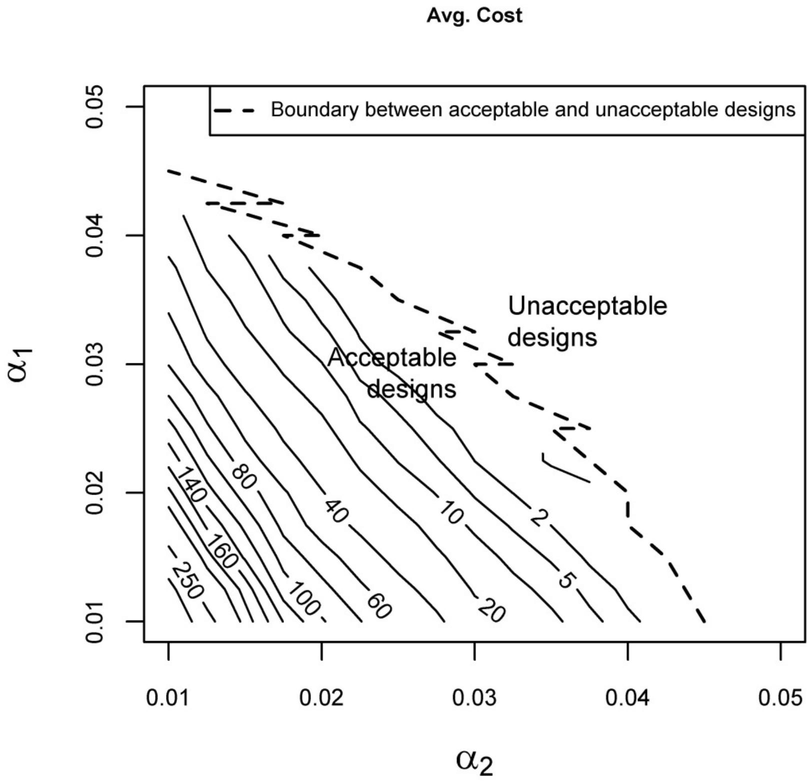
Optimal design search algorithm

The following four parameters were varied to optimize study designs using method E or F: two significance levels, α_1, and α_2, the stage 1 sample size (n_1), and the futility criterion (f). For each ISCV range, $n_{\max}$ was not optimized, but rather was fixed based on practical considerations (42 for the ISCV range 10–30%, and 180 for the ISCV range 30–55%).

The NOMAD implementation [15] of the MADS algorithm [16] for nonsmooth constrained optimization was used to search for the optimal study design. MADS is designed for blackbox optimization problems [17] in which the cost function and constraints are evaluated by a time-consuming simulation code. The optimization problem had four bound-constrained variables, three of them being continuous, and the fourth one (n_1) discrete. The cost function and the eight constraints on power and on type I error rate were computed by the simulations described above.

The cost function surface consisted of a steep slope from bottom left to top right, and an irregular edge that separates the acceptable (bottom left) and unacceptable (top right) design regions as illustrated in Figure 3. The top right region represents those designs whose power is less than 80% or type I error rate is greater than 5%. Unacceptable designs were assigned an extremely high cost value in order to keep the search within the acceptable region of the design space. The optimal design will be located on the edge, as otherwise, it can be improved with a larger alpha level or smaller power. Unlike conventional optimization algorithms, such as Nelder Mead, NOMAD's ability to handle constraints allows it to search effectively along this edge.

An initial study design with at least 80% power and at most 5% type I error rate on all range of CVs is given to the algorithm. From this initial study design, NOMAD evaluated the cost function, and automatically searched for an optimal design within the design space until convergence to the feasible design variant with the smallest cost function.



The cost function was plotted against alpha allocation in stages 1 and 2 as an illustration of the cost function surface. The cost function decreases as both α_1 and α_2 increases as shown in the contour. Acceptable designs (type I error rate = 0.05 and power = 80%) are on the lower left side of the irregular boundary illustrated as the dashed diagonal sawtooth curve.

Figure 3: The contour plot of cost function versus α_1 and α_2.

Simulations and optimizations were run on a parallel computing Linux cluster containing 5–6 individual PCs. With this setup, a complete search for an optimal design only took a few hours, rather than days that would have been required using a single PC. For example, it typically took about 5 hours to find an optimal design using the cluster, where about 100 design variants were tested before achieving convergence, taking an average of about 3 minutes per design variant.

Average sample size

Once the optimal designs were found for each of the four design spaces, the average sample size differences relative to the conventional, single stage designs were calculated to compare global method performance across the design space. In each case, averaging over the corresponding design space was done as follows.

For each of the four optimal designs the average sample size difference for a range of population GMRs and CVs, Δ, is defined as the arithmetic mean of the sample size difference between sequential design and one-stage design for each combination of GMRs and CVs as follows.

$$\Delta = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\tilde{n}^{S}_{\{GMR_i,CV_j\}} - \tilde{n}^{1}_{\{0.95,CR_j\}}\right) \tag{4}$$

The average sample size difference measures how many more subjects on average a sequential design requires than a one-stage design. A positive (negative) average sample size difference means a sequential design requires more (fewer) subjects than a one-stage design. Thus, a negative average sample size difference represents an improvement of the sequential design over the one-stage design.

Results and discussion

The optimized study designs based on methods E and F for the two desired ISCV ranges are presented in Table 1. Within each ISCV range, the optimized values for α_1, α_2, futility region $[\exp(-f), \exp(f)]$, and n_1 were very similar for methods E and F. Each of the optimized methods E and F in each of the ISCV ranges exhibited a type I error rate of ≤ 0.05 and a power of at least approximately 80%, and thus met the design criteria.

For the low ISCV range (10–30%), the average sample sizes for both methods E and F were only modestly larger than those of the corresponding conventional single stage designs. For the high ISCV range (30–55%), both optimized methods E and F yielded average sample sizes that were less than the average sample sizes required for a conventional single stage design at the same ISCV levels, and yielding the same power. This is particularly noteworthy, because the sequential design methods assumed no *a priori* knowledge of the ISCV, other than that it fell within the range of 30 to 55%, whereas the conventional single stage sample size calculation assumed that the population ISCV was known exactly and that the population GMR was exactly 0.95, rather unreasonable, though widely used assumptions.

For all four optimized methods, the futility regions seemed reasonable, in that they determined that there would be little point in proceeding to stage 2 if the 90% confidence interval about the GMR for the stage 1 results fell entirely outside the interval 100% ± about 5%–7%.

Table 2 shows the distribution of sample sizes for both methods E and F in both ISCV ranges as well as the average percentage of studies proceeding to stage 2. This was done for two different scenarios; a population GMR of 0.80, corresponding to a test formulation that did not match the reference formulation well, and a

Table 1: Group sequential design performance on a range of CVs.

ISCV%	Method	α_1	α_2	Futility region[1]	n_1	Maximum estimated type I error rate[2]	Minimum estimated power[2]	Average sample size difference[3]	True ISCV% (sample size of one stage design)	Average sample size difference[3]
10–30	E	0.0249	0.0363	93.74; 106.67	18	0.050	0.80	1.9	10 (7)	11
									15 (12)	6.1
									20 (19)	0.7
									25 (28)	-3.1
									30 (39)	-5.2
	F	0.0248	0.0364	94.92; 105.35	18	0.050	0.80	1.9	10 (7)	11
									15 (12)	6.0
									20 (19)	0.6
									25 (28)	-3.0
									30 (39)	-4.9
30–55	E	0.0254	0.0357	93.05; 107.47	48	0.050	0.80	-6.1	30 (39)	9.5
									35 (52)	-0.1
									40 (66)	-6.9
									45 (81)	-11
									50 (98)	-13.6
									55 (116)	-14.6
	F	0.0259	0.0349	93.50; 106.95	48	0.050	0.80	-6.7	30 (39)	8.4
									35 (52)	-0.4
									40 (66)	-7.2
									45 (81)	-12.2
									50 (98)	-13.9
									55 (116)	-15.0

The optimal design parameters of methods E and F are listed for each of the two ranges of ISCV% (10–30% and 30–55%). The maximum estimated type I error rate and minimum estimated power were calculated from the corresponding ISCV range to represent the worst case scenario in type I error rate and power for the optimal designs. Average sample size differences between optimal designs and single stage designs across the ISCV range and in each specific ISCV were calculated by averaging across GMRs.

[1] Stop for futility if the 90% confidence interval around the GMR after stage 1 is outside the futility region of $[\exp(-f), \exp(f)]$.

[2] Type I error rate is estimated at the true ratio of geometric means of 0.80 and power is estimated at the true ratio of geometric means of 0.95.

The standard errors of the estimated type I error rates and estimated powers are no more than 0.003 and 0.002, respectively.

[3] Average sample size differences as compared to a one-stage crossover design: a negative number represents a decrease in the number of subjects as compared to a one-stage design. A positive number represents an increase.

Table 2: Total sample size and proportions of studies requiring a second stage.

ISCV% range	Mean n total (5th, 50th, 95th) % of studies in stage 2	Ratio = 0.80 method		Ratio = 0.95 method	
	True ISCV%	E	F	E	F
10–30	10	18(18,18,18) 0%	18(18,18,18) 0%	18(18,18,18) 0%	18(18,18,18) 0%
	15	18(18,18,18) 0.9%	18(18,18,18) 0.5%	18.1(18,18,18) 2.4%	18.1(18,18,18) 1.3%
	20	19.3 (18,18,28) 13.7%	19.3(18,18,28) 12.7%	20.3(18,18,32) 24.1%	20.3(18,18,32) 21.8%
	25	23.8 (18,18,42) 31.4%	24.0(18,18,42) 32.0%	28.1 (18,24,42) 54.2%	28.2(18,24,42) 53.7%
	30	31.7(18,18,42) 44.9%	32.1(18,18,42) 46.3%	40.7 (18,42,42) 75.8%	41.0(18,42,42) 76.9%
30–55	30	48.5(48,48,52) 6.5%	48.4(48,48,48) 2.8%	48.6(48,48,52) 7.6%	48.5(48,48,48) 3.6%
	35	52(48,48,72) 24.8%	51.6(48,48,72) 18.1%	52.7(48,48,74) 28.2%	52.5(48,48,74) 22.8%
	40	59.3(48,48,94) 37.1%	58.6(48,48,94) 33.4%	62.2(48,48,98) 46.2%	62.2(48,48,98) 44.0%
	45	69.4(48,48,118) 45.3%	68.6(48,48,118) 42.8%	77.6(48,80,124) 61.3%	77.8(48,80,124) 60.5%
	50	82.1(48,76,144) 51.4%	81.4(48,48,146) 49.4%	97.7(48,104,150) 74.3%	98.1(48,104,152) 73.6%
	55	96.6(48,104,172) 56.4%	96.4(48,102,172) 55.3%	121.3(48,128,176) 85.2%	121.3(48,128,180) 84.3%

Average total sample size, (5th, 50th, 95th) percentile and percent of studies continuing to stage 2 were listed by ISCV% in the increment of 5% and GMR (80% for non-BE and 95% for BE) for optimal designs (α_1, α_2, futility region and n_1) presented in Table 1.
Mean n total (5th, 50th, 95th) is the average, 5 percentile, median and 95 percentile of total n in each ISCV for a certain true GMR ratio. And the % of studies in stage 2 is the percentage of decisions made to go in stage 2 in each ISCV.

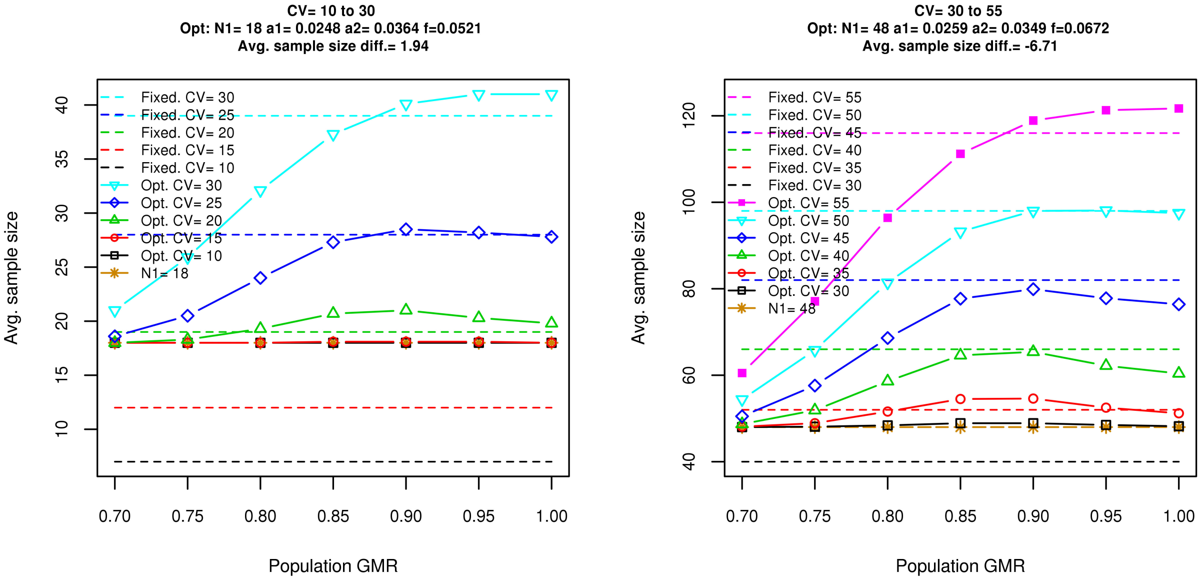
population GMR of 0.95, corresponding to a test formulation that did match the reference formulation well. As was the case with Table 1, these results show little difference between the performance characteristics of methods E or F. However, for the high ISCV range, method F may offer a slight performance advantage in terms of smaller median sample size, and less likelihood of proceeding to stage 2, particularly for the GMR = 0.80 case. The smaller sample sizes and frequencies of proceeding to stage 2 for the GMR = 0.80 case as compared to the GMR = 0.95 case reflect the beneficial impact of the futility rules embedded in these designs.

Figure 4 shows plots of average total sample size as a function of population GMR for optimized method F[3] as well as for the corresponding conventional single stage design at each specific ISCV level tested. These plots show that method F may sometimes require somewhat larger average sample sizes than the corresponding single stage design, particularly at the lower end of each optimized ISCV range (e.g. ISCV=10% on the left and ISCV = 30% on the right part of the figure) and when the population GMR is in the range 0.90–1.00 for range of ISCV of 10–30% and ISCV = 55%. However, when the population GMR is far from 1.00, method F may exhibit substantially smaller average sample sizes than the corresponding single stage design, particularly towards the upper end of the optimized ISCV range. In this sense, the optimized designs exhibit some of the benefits of pilot studies (i.e., typically low overall cost for a formulation that turn out to match the reference product poorly), but, unlike pilot studies, without the need to discard the initial data collected (i.e., for formulations that match the reference product reasonably well). In an effort to be conservative the comparisons made between the conventional single stage design and the sequential design put the sequential design at a disadvantage, just because of the constraints we imposed on the sequential design. For the sequential design, we imposed a minimum for n of 12, whereas for the conventional single stage design, neither a minimum for n nor a requirement for an even sample size n was imposed for the comparison.

It is interesting to note that the optimized stage 1 sample size (n_1) for the low end of each ISCV range resulted in overpowering, but much more so for the low ISCV range than for the high ISCV range (viz. single stage power for $n = 18$, ISCV = 10%, and GMR = 0.95 is 0.99, whereas for $n = 48$, ISCV = 30%, and GMR = 0.95 is only 0.88). This appears to reflect a fundamental limitation in the ability of a sequential design method to effectively span a large range of possible ISCV values, which can be seen as follows.

Assuming a population GMR of 0.95, the single stage n required for ISCVs of 10%, 30%, and 55% are 7, 39, and 115 subjects, respectively. Therefore, a sequential design attempting to cover the low ISCV range of 10–30% would be attempting to match the performance of single stage designs employing a range of 7 to 39

[3] Plots of method performance characteristics are provided only for method F, because its performance was slightly better than that of method E, and the plots for method E were similar.



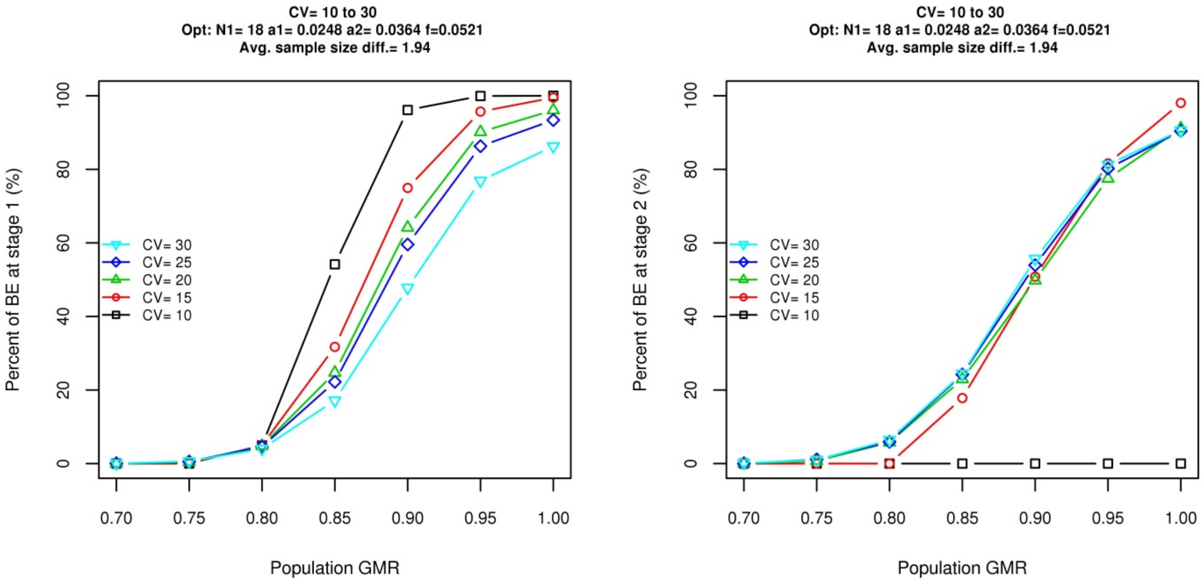
Average sample sizes were plotted against population GMR by ISCV% (in the increment of 5%), for optimal design method F, when ISCV% is in each of two ranges (10–30% and 30–55%) using solid line. The horizontal dashed lines illustrated sample size from single stage design for comparison.

Figure 4: Average Sample Size per range of CVs and per true ratio of geometric means for Optimal design method F.

subjects, which amounts to a 5.57-fold range in sample sizes. In contrast, a sequential design attempting to cover the high ISCV range of 30–55% would be attempting to match the performance of single stage designs employing a range of 39–115 subjects, which amounts to only a 2.95-fold range in sample sizes. It appears that the greater overpowering of n_1 for the low ISCV range simply reflects the inability of a sequential design to meet the design criteria of type I error rate and power over what amounts to the larger span of corresponding single stage sample sizes (5.57-fold) that would be required for the low ISCV range, while still keeping n_1 low.

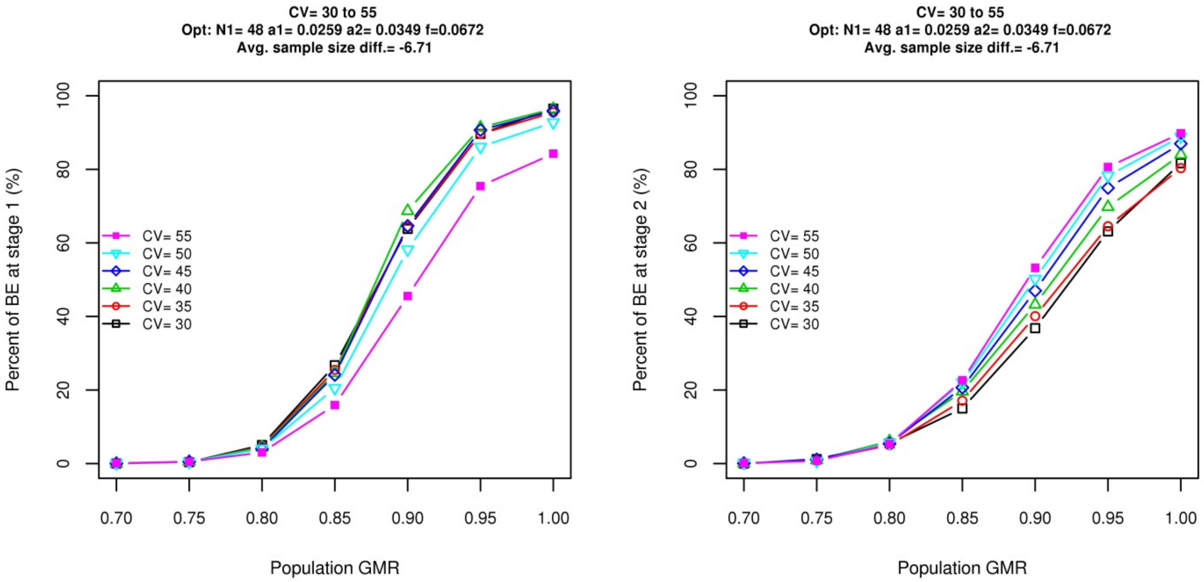
Figure 5 shows the probability of passing (declaring BE) as a function of population GMR for method F optimized over the ISCV range of 10–30%, for each individual ISCV level tested at stage 1 (left panel), and at stage 2 (right panel). These plots show that the likelihood of passing at stage 1 increases substantially as the population ISCV decreases, particularly for borderline formulations whose population GMR's are between about 0.85 and 0.90. This is entirely expected, considering that method F employs a fixed initial sample size (n_1). The right panel, however, shows that the likelihood of passing at stage 2 (i.e., the probability of passing based on the combined data from stages 1 and 2 for that subset of studies that actually proceeds to stage 2) has little dependence on the population ISCV.

Figure 6, which shows the results for method F optimized over the ISCV range of 30–55%, is similar to Figure 5 except that it shows less change in stage 1 power as ISCV changes, but more change in stage 2 power as ISCV changes. This is probably due to the greater overpowering of stage 1 for low ISCVs in the low ISCV range than in the high ISCV range discussed above. In other words, for the low end of the high ISCV range, stage 1 never achieves the extremely high power seen for the low end of the low ISCV range. This results in more studies passing at stage 1 for the low ISCV range, which more effectively screens the stream of studies proceeding to stage 2 than is the case for the high ISCV range. In other words, the optimized sequential design is more similar to a single stage design over the low ISCV range than it is over the high ISCV range. This can be seen in Figure 7, which shows the percentage of decisions made in stage 1 for the low and high ISCV ranges (left and right panels, respectively). Overall, there is a tendency for more decisions to be made at stage 1 for the low ISCV range than for the high ISCV range.



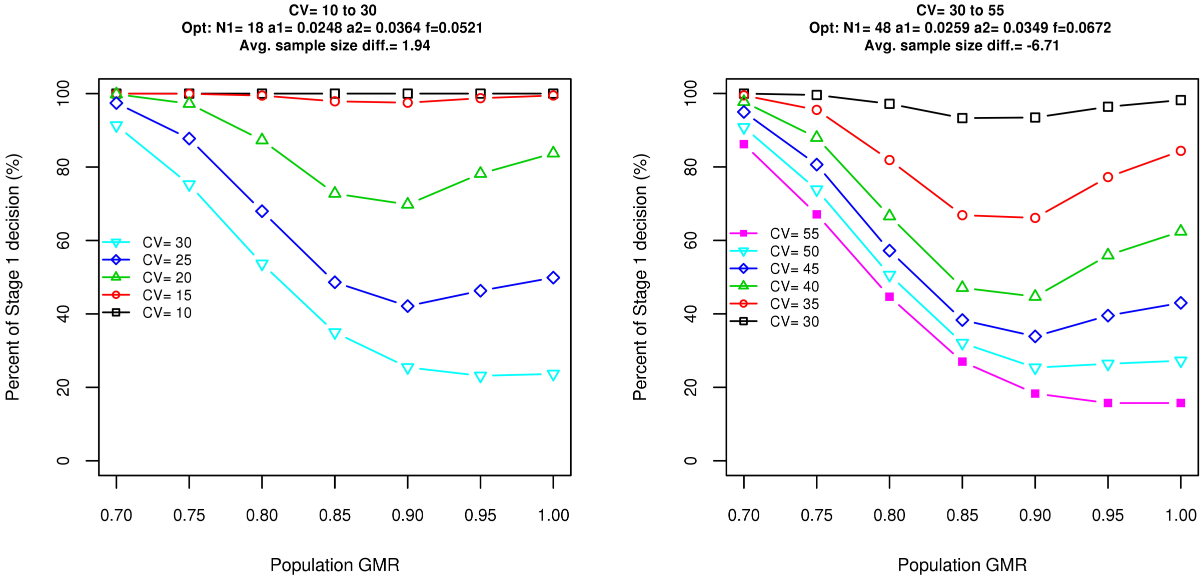
The percent of BE decisions at stage 1 (left) and stage 2 (right) were plotted against population GMR by ISCV% (in the increment of 5%) for optimal design method F, when ISCV% is in the range of 10–30%. Here percent of BE decisions in stage 1 (stage 2), was calculated as the proportion of BE decisions out of decisions made at stage 1 (stage 2).

Figure 5: Percentage of BE decisions made at stages 1 and 2 in the optimal design of method F for ISCV ranging between 10% and 30%.



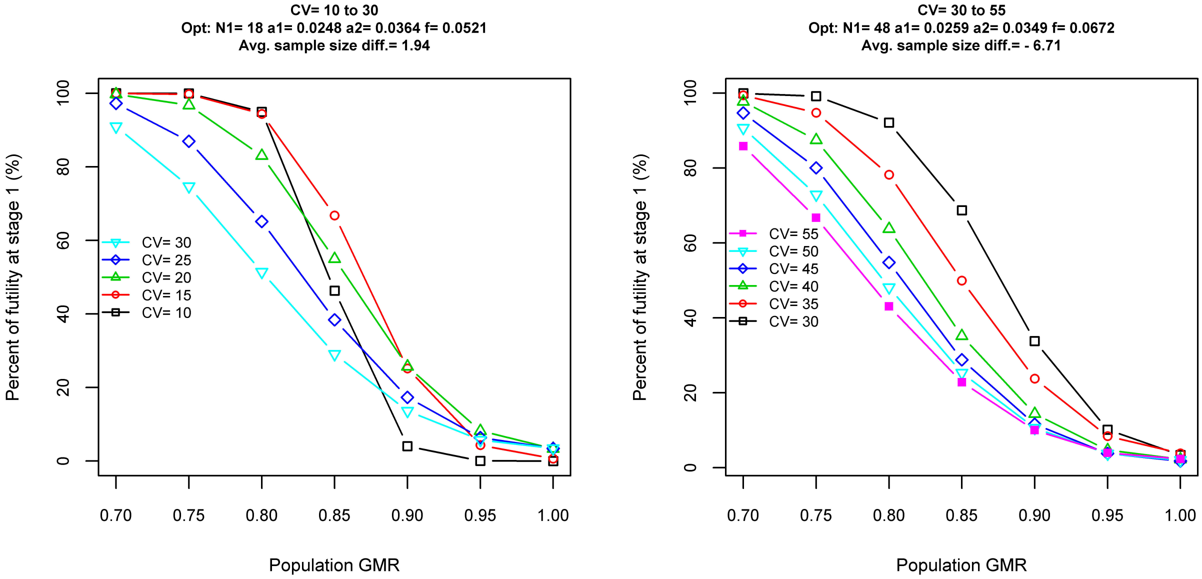
The percent of BE decisions made at stages 1 (left) and 2 (right) were plotted against true population GMR by ISCV% (in the increment of 5%) for optimal design method F, when ISCV% is in the range of 30–55%. Here the percent of BE decisions at stage 1 (stage 2), was calculated as the proportion of BE decisions out of decisions made at stage 1 (stage 2).

Figure 6: Percentage of BE decisions made at stages 1 and 2 in the optimal design of method F for ISCV ranging between 30% and 55%.



The percent of decisions made at stage 1 were plotted against true population GMR by ISCV%(in the increment of 5%) for optimal design method F, when ISCV% is in the range of 10–30%(left) and 30–55%(right). Here the percent of decisions at stage 1 was calculated as the proportion of decisions at stage 1 out of all decisions made in both stages.

Figure 7: Percentage of decisions made in stage 1 in the optimal design method F for ISCV ranging between 10% and 30% and between 30% and 55%.



The percent of futility made at stage 1 were plotted against population GMR by ISCV% (in the increment of 5%) for optimal design method F, when ISCV% is in the range of 10–30% (left) and 30–55% (right). Here the percent of futility was calculated as the proportion of futility decisions out of decisions at stage 1.

Figure 8: Figure 8. Percentage of futility in stage 1 in the optimal design method F for ISCV ranging between 10% and 30% and between 30% and 55%.

Figures 7 and 8 also illustrate the success of the futility rule used as reflected in the high decision rates (non BE) at stage 1 for extreme (low) population GMRs. The minima in stage 1 decision rates seen in many curves around population GMRs of 0.85–0.90 are entirely expected, and reflect the fact that, in this range of GMRs, the greatest uncertainty regarding formulation performance based on stage 1 data occurs. In effect, the decision is easier for very good (GMR close to 1) or very poor (GMR far from 1) formulations. At the highest ISCV levels within each ISCV range, the minima in the stage 1 decision rates do not appear in the GMR range of 0.85–0.90 because of the very low power at stage 1, even for GMRs close to 1. Nevertheless, even though, at the highest ISCV within each of the two ISCV ranges, stage 1 power is low (resulting in low likelihood of passing at stage 1), there is still excellent rejection (∼85–90%) of poor formulations, which further underscores the good performance of the futility rule employed.

Table 3 shows the performance characteristics of the optimized sequential designs E and F relative to the earlier sequential designs B, C, and D for the low ISCV range. Relative to the earlier designs B, C, and D, the optimized designs show substantially lower, but still slight sample size penalties (as compared to single stage designs). Table 4 shows similar comparative performance data for the high ISCV range. Several design features of the new sequential designs, including the maximum sample size constraint, the futility rule, and optimizing over a narrower range of ISCV values allowed for a slight reduction in α_1 and but a larger increase in α_2 in the optimized designs E and F relative to the earlier methods B, C, and D, while still controlling the overall type I error rate to ≤ 0.05. This, in turn, allowed for a substantial reduction in sample size relative to methods B, C, and D, as well as little or no statistical "penalty" relative to single stage designs.

Table 5 shows the performance characteristics of optimized methods E and F when they are used for products whose population ISCV values are outside the design ranges of either 10–30% or 30–55% for the low and high ISCV versions. These results show that, even if the population ISCV is outside the design limits of the optimized designs, the type I error rate is not inflated. As expected, however, there is a modest loss of power below the desired 0.80 when the population ISCV is higher than the design ranges for the optimized sequential methods E and F. Thus, optimized sequential methods E and F are suitable for use even in cases where the population ISCV is outside the design ranges of 10–30% or 30–55% for the low and high ISCV optimized forms, although power may be compromised for ISCV values above the upper limits of the design ranges.

Table 3: Comparison of sample size reduction from Potvin et al. vs. optimal designs (ISCV range = 10–30%).

Method (Potvin et al.)	Method (optimal design)	α_1	α_2	Futility region	n_1	Max α	Min power	Overall average alpha	Overall average power	Average sample size difference[1]
B		0.0294	0.0294	None	12	0.048	0.79	0.041	0.86	4.9
					24	0.048	0.83	0.036	0.91	8.5[†]
					30*	0.044	0.83	0.033	0.92	12.2[†]
C		0.0294	0.0294	None	12	0.051	0.79	0.049	0.87	4.8
					24	0.050	0.83	0.049	0.92	8.3[†]
					30*	0.051	0.84	0.049	0.93	12.0
D		0.0280	0.0280	None	12	0.050	0.78	0.047	0.87	5.1
					24	0.050	0.83	0.049	0.92	8.5
					30*	0.051	0.84	0.049	0.93	12.2
	E	0.0249	0.0363	93.74; 106.67	18	0.050	0.80	0.043	0.88	1.9[†]
	F	0.0248	0.0364	94.92; 105.35	18	0.050	0.80	0.049	0.89	1.9[†]

The optimal designs are compared with methods B, C, and D reported previously in Potvin et al. for the ISCV% range of 10–30%. Maximum type I error rate estimates and minimum power estimates are listed to illustrate the worst case scenario in the ISCV% range. Average alphas, average powers, and average sample size differences from single stage design were provided for comparison.

* n_1 is reset to 30 if greater than 30 to match the method used in optimal design.

[†] Designs that meet the type I error and power requirement.

[1] Average sample size difference as presented in Equation (4).

Table 4: Comparison of sample size reduction from Potvin et al. vs. optimal designs (ISCV range = 30–55%).

Method (Potvin et al.)	Method (optimal design)	α_1	α_2	Futility region	n_1	Max α	Min power	Overall average alpha	Overall average power	Average sample size difference
B		0.0294	0.0294	None	12	0.044	0.73	0.035	0.75	14.9
					24	0.048	0.78	0.040	0.80	13.5
					36	0.049	0.79	0.043	0.82	11.2
					48	0.048	0.81	0.043	0.83	9.9†
					60	0.049	0.82	0.041	0.84	11.2†
C		0.0294	0.0294	None	12	0.045	0.73	0.035	0.75	14.9
					24	0.050	0.78	0.041	0.80	13.5
					36	0.049	0.79	0.045	0.82	11.0
					48	0.051	0.81	0.048	0.83	9.5
					60	0.050	0.82	0.048	0.85	10.7
D		0.0280	0.0280	None	12	0.042	0.73	0.034	0.75	16.3
					24	0.048	0.77	0.039	0.80	14.9
					36	0.047	0.79	0.043	0.82	12.5
					48	0.050	0.81	0.046	0.83	10.8
					60	0.050	0.82	0.047	0.85	11.8†
	E	0.0254	0.0357	93.05; 107.47	48	0.050	0.80	0.045	0.82	-6.1†
	F	0.0259	0.0349	93.50; 106.95	48	0.050	0.80	0.048	0.83	-6.7†

The optimal designs were compared with methods B, C, and D reported previously in Potvin et al. for the ISCV% range of 30–55%. Maximum type I error rate estimates and minimum power estimates are listed to illustrate the worst case scenario in the ICSV% range. Average alphas, average powers, and average sample size differences from the single stage design were provided for comparison.

† Designs that meet the type I error and power requirement.

Table 5: Type I error rate and power outside of optimal ISCV range (methods E and F).

ISCV%	Method	α_1	α_2	Futility region	N_1	Estimated type I error		Estimated power	
						ISCV = 5%	ISCV = 35%	ISCV = 5%	ISCV = 35%
10–30	E	0.0249	0.0363	93.74; 106.67	18	0.037	0.041	0.999	0.777
	F	0.0248	0.0364	94.92; 105.35	18	0.050	0.459	0.999	0.780
						ISCV = 25%	ISCV = 60%	ISCV = 25%	ISCV = 60%
30–55	E	0.0254	0.0357	93.05; 107.47	48	0.039	0.036	0.852	0.764
	F	0.0259	0.0349	93.50; 106.95	48	0.050	0.036	0.869	0.762

Estimated type I error rates and powers at ISCV outside the target range are shown to illustrate the impact from out-of-range ISCV. For example, the optimal design in method E in the first row was destined for ISCV% range of 10–30%. The estimated type I error rates and powers are listed for ISCV = 5% and 35%.

Example of use

Following is a hypothetical example to demonstrate how decision rules are applied based on method E and method F when using the optimal design on an unknown ISCV in the range of 30–55%.

Method E

Assume an adaptive two-stage design for two-period crossover bioequivalence studies targeting on an unknown ISCV in the range of 30–55% was conducted. Based on Table 1, optimal design parameters ($\alpha_1 = 0.0254$, $\alpha_2 = 0.0357$, $N_1 = 48$, Futility region = [93.05%; 107.47%]) were determined if method E was chosen. At stage 1 with 48 completed subjects, BE was evaluated at alpha level of $\alpha_1 = 0.0254$. The observed 94.92% confidence interval of GMR in original scale was (0.78, 1.14) with an observed ISCV = 48.3%. As BE was not met, the power was estimated as 35.8% with alpha level at $\alpha_2 = 0.0357$. Since power was less than 80%, futility was then checked. Given the observed 90% confidence interval (0.81, 1.11) was not completely out of futility region, futility was not met and overall sample size was re-estimated as 104. Thereafter, stage 2 was initiated and additional 56 subjects were enrolled. BE was evaluated at stage 2 using data from both stages at alpha level of $\alpha_2 = 0.0357$. Given the observed 92.86% confidence interval (0.91, 1.12) was within (0.80, 1.25), BE was concluded.

Method F

If method F was instead chosen for the same study, optimal design parameters ($\alpha_1 = 0.0259$, $\alpha_2 = 0.0349$, $N_1 = 48$, Futility region = [93.50%; 106.95%]) were determined from Table 1. At stage 1, the same 48 subjects completed the study. The power was estimated as 48.7% with alpha level at 0.05. Since power $< 80\%$, BE was evaluated at alpha level $\alpha_1 = 0.0259$. The observed 94.82% confidence interval of GMR in original scale was (0.78, 1.14) with an observed ISCV = 48.3%. As BE was not met at stage 1, futility was then checked. Given the observed 90% confidence interval (0.81, 1.11) was not completely out of futility region, futility was not met and overall sample size was re-estimated as 104. Thereafter, stage 2 was initiated and additional 56 subjects were enrolled. BE was evaluated at stage 2 using data from both stages at alpha level of $\alpha_2 = 0.0349$. Given the observed 93.02% confidence interval (0.91, 1.12) was within (0.80, 1.25), BE was concluded.

Conclusion

The four optimized designs (methods E and F, each optimized for low and high ISCV) offer the benefits of adapting automatically to substantial uncertainty in the anticipated formulation GMR as well as moderate uncertainty in the ISCV, with only a modest sample size penalty for the low ISCV range, and a modest sample size benefit (reduction) for the high ISCV range, all while preserving type I error rates ≤ 0.05 and power ≥ 0.80. The inclusion of a futility rule provides for the early abandonment of poor formulations and thereby effectively controls cost. Unlike our prior methods, the upper limits imposed on overall sample size by the optimized designs provide sponsors with upper limits on overall study cost. The performance difference between the two optimized algorithms based on methods E and F is negligible, and these two classes of sequential algorithms work equally well. Overall, the optimized designs described here provide attractive approaches to addressing uncertainties in GMR and ISCV, whether or not scaled average bioequivalence methods are available as alternatives.

References

[1] Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ, Smith, RA. Sequential Design approaches for bioequivalence studies with crossover designs. Pharmaceutical Statistics 2008; 7:245–262.

[2] Montague TH, Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ. Additional results for “Sequential design approaches for bioequivalence studies with crossover designs”. Pharmaceutical Statistics 2012; 11:8–13.

[3] Pocock SJ. Group Sequential Methods in the Design and Analysis of Clinical Trials. Biometrika 1977; 64:191–199.

[4] Bandyopadhyay N, Dragalin V. Implementation of an adaptive group sequential design in a bioequivalence study. Pharmaceutical Statatistics 2007; 6:115–122.

[5] Fuglsang A. Futility rules in bioequivalence trials with sequential designs. AAPS J 2014; 16:79–82

[6] Fuglsang A. A sequential bioequivalence design with a potential ethical advantage. AAPS J 2014; 16:843–846.

[7] US Food and Drug Administration. Bioequivalence Recommendations for Specific Products: Draft Guidance on Progesterone Capsules, December 12, 2012. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM209294.pdf (accessed 29 June 2015).

[8] European Medicines Agency – Committee For Medicinal Products For Human Use (CHMP) Guideline On The Investigation Of Bioequivalence, London, 20 January 2010. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf (accessed 29 June 2015).

[9] 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.

[10] 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.

[11] Tothfalusi L, Endrenyi L. Sample Sizes for Designing Bioequivalence Studies for Highly Variable Drugs. J Pharm Pharmaceut Sci 2012; 15:73-84.

[12] Siegmund D. Importance Sampling in the Monte Carlo Study of Sequential Tests. The Annuals of Statistics 1976; 4:673–684.

[13] Rubinstein RY. Simulations and the Monte Carlo Method. Wiley and Sons: New York, NY, 1981.

[14] Fishman GA. Monte Carlo: Concepts, Algorithms and Applications. Springer: 1996.

[15] Le Digabel S. Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. ACM Transactions on Mathematical Software 2011; 37:44:1–44:15.

[16] Audet C and Dennis, Jr. JE. A progressive barrier for derivative-free nonlinear programming. SIAM Journal on Optimization 2009; 20:445–472.

[17] Audet C. A survey on direct search methods for blackbox optimization and their applications. Chapter 2 in Mathematics without boundaries: Surveys in interdisciplinary research. Pardalos PM, Rassias TM (Eds.). Springer, 2014, 31–56.