Published in final edited form as:

Pharm Stat. 2018 September ; 17(5): 458–476. doi:10.1002/pst.1865.

Design and Inference for 3-Stage Bioequivalence Testing with Serial Sampling Data

Fangrong Yan¹, Huihong Zhu¹, Junlin Liu¹, Liyun Jiang¹, and Xuelin Huang^{2,*}

¹Research Center of Biostatistics and Computational Pharmacy, China Pharmaceutical University, Nanjing 210009, P.R. China

²Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, 77030, USA

Abstract

A bioequivalence test is to compare bioavailability parameters, such as the maximum observed concentration (C_{max}) or the area under the concentration-time curve (AUC), for a test drug and a reference drug. During the planning of a bioequivalence test, it requires an assumption about the variance of C_{max} or AUC for the estimation of sample size. Since the variance is unknown, current 2-stage designs use variance estimated from stage-1 data to determine the sample size for stage 2. However, the estimation of variance with the stage-1 data is unstable, and may result in too-large or too-small sample size for stage 2. This problem is magnified in bioequivalence tests with a serial sampling schedule, by which only one sample is collected from each individual and thus the correct assumption of variance becomes even more difficult. To solve this problem, we propose 3-stage designs. Our designs increase sample sizes over stages gradually, so that extremely large sample sizes will not happen. With one more stage of data, the power is increased. Moreover, the variance estimated using data from both stages 1 and 2 is more stable than that using data from stage 1 only in a 2-stage design. These features of the proposed designs are demonstrated by simulations. Testing significance levels are adjusted to control the overall type I errors at the same level for all the multi-stage designs.

Keywords

Sequential design; Bioequivalence testing; serial sampling data; sample size estimation; statistical power

1 Introduction

A bioequivalence clinical trial is to demonstrate that a new drug (test drug) is therapeutically interchangeable with a marketed drug (reference drug), both of which contain the same active substance. It ensures the pharmaceutical products' quality when replacing expensive originator drugs by cheap generic drugs. Such replacements have been greatly increasing during the past decades to reduce the huge health care cost. Consequently, a large number of

Address correspondence to: Xuelin Huang, Department of Biostatistics, University of Texas, MD Anderson Cancer Center, 1400 Pressler Street, Unit 1411, Houston, TX 77230-1402, Fax: 713-563-4243, Phone: 713-794-4172, xluang@mdanderson.org.

bioequivalence studies are conducted in pharmaceutical industries. Well-designed statistical procedures for these studies are important for ensuring their validity and efficiency, and thus of great interest not only to the industries, but also to consumers and government agencies.

In U.S.A., bioequivalence studies are performed on the basis of the requirements set forth in Section 21 of the Code of Federal Regulations and Guidance given by the Food and Drug Administration (FDA) [1]. A bioequivalence test is commonly constructed by comparing the bioavailability parameters, such as the maximum observed concentration (C_{max}) or the area under the concentration-time curve (AUC), for the test drug and the reference drug. Two one-sided t-tests are often used and bioequivalence is met if the 90% confidence interval of the ratio of the population means for the test/reference ratio lies within 0.80 to 1.25.

The estimation of C_{max} and AUC of a drug need multiple samples at each of a series of time points. Usually, multiple samples (one for each time point) are collected from each participant in the study. This is called a complete data design, which include the commonly seen crossover and parallel designs [2]. However, some times, due to ethical or scientific reasons, totally only one sample is collected from each individual, and samples at different time points must be collected from different individuals. For example, due to a limited quantity, aqueous humor can only be collected at one timepoint from an individual during the entire study, and data needed to be collected from multiple subjects sampled at distinct timepoints. This type of sampling is usually referred to as the serial sampling, which results in much more sparse data than the complete data design with the same number of participants (sample size). With a complete data design, bioavailability parameters such as AUC can be estimated for each individual separately. However, this is not applicable in serial sampling design where only one sample is taken from each individual. Therefore, estimation of the important bioavailability parameters such as AUC, calculation of the test statistics and degree of freedom should be modified. Hua [3] provided a comprehensive introduction of statistical issues in bioequivalence tests with serial sampling data. However, they considered only single-stage trials.

For a bioequivalence test, the sample size is usually chosen as the smallest number of subjects that will enable the power level to reach 80%. Here the power is defined as the probability of showing bioequivalence when the true value of the test/reference ratio is within 0.80 and 1.25. It depends on not only the sample size, but also the population variance and the effect size at which the power is estimated. These values are difficult to estimate correctly prior to the start of a clinical trial. If the variance used in the power calculation is too large, the sample size will be unnecessarily inflated. This will increase the cost of the study, but will not affect the conclusion of the test. On the other hand, if the variance used in the power calculation is too low, the sample size will be underestimated. This will result in a low power level and may cause false rejection of bioequivalence. In view of this issue, some regulatory departments permit “add-on” designs [4, 5]. If low power is the reason for the failure of a bioequivalence test, the investigators are allowed to add more subjects to the study and pool the data collected from the additional subjects with the original data for a second bioequivalence test. However, this method will inflate the type I error rate. In order to solve this problem, Potvin [6] proposed 2-stage sequential design approaches for bioequivalence studies. Their simulation results showed that these methods

can maintain adequate power and control the overall type I error rate under most levels of variability and initial sample sizes. Fuglsang [7, 8] applied these 2-stage methods to complete data designs and studied different futility rules for these methods. More information about 2-Stage design in bioequivalence trials can be found in [9]. To the best of our knowledge, these methods have not been applied to bioequivalence tests with serial sampling data, which are in great need. In this article, we apply and improve these methods for bioequivalence tests with serial sampling data.

Serial sampling designs result in much more sparse data than the complete data design with the same number of participants (sample size). In 2-stage designs, variance estimated from the stage-1 data is used to determine the sample size for the second stage. With sparse serial sampling data, however, the estimation of variance with the stage-1 data is unstable. It may give a severely underestimated variance leading to a sample size too small and thus an underpowered study, or an overestimated variance that asks for an unnecessarily large or unaffordable sample size for the second stage. To solve this problem, we propose 3-stage designs to replace the 2-stage ones. Our designs increase sample sizes over stages gradually, so that extremely large sample sizes will not happen. With one more stage of data, the power is increased. Moreover, the variance estimated from data from both stage 1 and stage 2 is used to determine the sample size for stage 3. This variance estimation is more stable than that using data from stage 1 only in a 2-stage design. Consequently, the proposed 3-stage designs increase the power and reduce the unnecessarily large sample size. Simulations demonstrate these properties by comparing the proposed 3-stage designs with their 2-stage counterparts. Testing significance levels are adjusted to control the type I errors at the same level for all multi-stage designs.

The rest of this paper is arranged as follows. In Section 2, based on four 2-stage designs currently in the literature, we propose four 3-stage designs respectively for bioequivalence testing with serial sampling data. In Section 3, we conduct simulations to evaluate and compare the operating characteristics of these eight sequential designs. We conclude with discussion in Section 4.

2 Methods

In this Section, we first give the assumptions in Subsection 2.1 for all the multistage designs described in this article. Then, in Subsection 2.2, we extend the four 2-stage designs of Potvin [6] to 3-stage. In Subsection 2.3, we provide the estimation and inference procedure for the proposed designs.

2.1 Assumptions

All multi-stage designs are constructed under the following six assumptions.

1. Only a single endpoint, the AUC or C_{max} is considered in the study. This choice is based on a policy of the Center for Drug Evaluation and Research (CDER) of the FDA [1], which requires that statistical tests for assessing bioequivalence preserve the type I error rate α at 0.05 for each individual endpoint.

2. A bioequivalence criterion at a specific level is that the confidence intervals for the ratio of the population means (test/reference) fall within 80%–125% at the $(1-2\alpha)\%$ level. The specific method for testing bioequivalence with serial sampling data will be introduced in Section 2.2.
3. Following the work of Potvin [6], all power calculations are performed at a specific α level using a ratio of the population means (test/reference) of 0.95. The value of 0.95 is chosen as a typical value for designing bioequivalence studies, allowing for some difference in the reference value.
4. The sample sizes for the test drug and reference drug are assumed to be equal in each stage. The sample size re-estimation will be introduced in Section 2.3.
5. No rules about pooling data are applied, which means that pooling data from different stages is always allowed, even if there is a statistically significant difference between the results from the two stages.
6. No overflow regarding interim analysis is existent. The number of recruited subjects at interim is equal to the number of subjects to be analyzed.

2.2 Sequential design methods for bioequivalence test

Potvin [6] proposed three categories of 2-stage sequential designs (methods B, C and D) for bioequivalence studies. These three methods all involved no stopping criteria or futility rules, but 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. If the trial continues to stage 2 due to insufficient power at stage 1, the sample size will be re-estimated based on the observed CV (within-subject coefficient of variation) and population means for the test/reference ratio using stage 1 data [10]. Method D is almost the same as method C with a little modification of the significance level, so we consider only methods B and C in this article, and denote them as methods S2-BE and S2-PW. Based on methods S2-BE and S2-PW, the modified methods S2-BE_c and S2-PW_c specify that the maximum sample size for stage 2 should not exceed four times of the sample size of stage 1. This was based on the observation that sometimes a unstable variance estimation from stage 1 results in a huge and unaffordable sample size for stage 2. In this article, for achieving a better performance in bioequivalence test with serial sampling data, we modify the methods S2-BE, S2-BE_c, S2-PW and S2-PW_c to give new 3-stage designs S3-BE, S3-BE_c, S3-PW and S3-PW_c respectively. The proposed 3-stage designs are illustrated in Figures 1 and 2 respectively. The 2-stage designs are described by the block diagrams in Appendix B.

Methods S3-BE & S3-PW: By methods S2-BE and S2-PW [6], the sample size is re-estimated to reach the power level of 80% for stage 2. This may result in a very large sample size that is unnecessary or unaffordable. In such a situation, one may try to conduct the trial with a smaller sample size and lower power. The bioequivalence conclusion may still be achieved. Inspired by this idea, we propose 3-stage sequential design methods S3-BE and S3-PW, in order to decrease the sample size without affecting the power of the whole bioequivalence test. The first stage remains the same as in Potvin's methods. At stage 2, the

sample size is re-estimated at the power level of 60% instead of 80% at stage 2. The threshold of 60% is determined by lots of pre-simulation studies in order to maintain the overall power above 80%. If bioequivalence is concluded, the trial can be stopped, which reduces the sample size. If the bioequivalence test fails, we move on to stage 3, calculate the sample size again to reach the power level of 80%, and then conduct the bioequivalence test at stage 3. Note that the sample size re-estimation procedure for stage 3 is based on the variance estimate using data from both stages 1 and 2. This means more subjects are included and the variance estimate is more stable using data from stage 1. This reduces the chance of unnecessarily large sample size due to over-estimated variance. Following Pocock [11], the significance level in each stage of the 3-stage classical group sequential design is 0.021. In this article, due to the permission of sample size re-estimation, we conduct large amounts of pre-simulations to adjust the α level in order to limit the overall type I error rate to a maximum of 5%. Finally, we chose the adjusted α in each stage in S3-BE and S3-PW to be 0.023 for simplicity. Denoting the initial sample size at stages 1, 2 and 3 as n_1 , n_2 and n_3 respectively.

Methods S3-BE_c & S3-PW_c: Methods S3-BE_c and S3-PW_c are natural extensions of methods S3-BE and S3-PW by adding stopping rules. The stopping rules specify that neither n_2 nor n_3 can exceed $4n_1$; otherwise, the trial is stopped and bioequivalence is not met. This is based on practical considerations, and is similar as the extension from method S2-BE to S2-BE_c, and S2-PW to S2-PW_c [6]. Fuglsang [7] found that S2-BE_c and S2-PW_c have a very low power level when the initial sample size at stage 1 is small or the co-efficient of variation is large. In such a situation, the sample size required to reach 80% power in stage 2 is too large and exceeds the limit. Then by the stopping rule, the trial fails and no bioequivalence is concluded. We believe our proposed S3-BE_c and S3-PW_c will not suffer so much from this problem because these designs require only 60% power at stage 2. The additional stage 3 will make up the power loss. The significance level for each stage is also chosen to be 0.023 in order to control the overall type I error rate at a maximum of 5%. The methods S3-BE and S3-BE_c are illustrated in Figure 1, and S3-PW and S3-PW_c are illustrated in Figure 2.

Although we focus on sequential design methods for bioequivalence testing with serial sampling data, the four proposed methods (S3-BE, S3-BE_c, S3-PW and S3-PW_c) can also be used for testing with complete data designs. The methods differ in parameter estimation, test statistic calculations and power estimation. These procedures are described for serial sampling data in the next Subsection.

2.3 Bioequivalence testing with serial sampling data

The sequential design methods by Potvin [6] were studied for bioequivalence tests with complete data designs, where multiple samples (one for each time point) are collected from each participant in the study. However, with serial sampling data, only one sample is collected from each subject, samples at different time points are collected from different participants. To conduct bioequivalence testing in this setting, we modify the testing method, power calculation and sample size re-estimation.

2.3.1 Estimation of AUC—In a bioequivalence test with serial sampling data, the AUC is commonly used. Its estimation forms the basis of the whole sequential method. Nedelman [12] proposed a method to estimate the AUC from 0 to the last time point. Wolfsegger [13] proposed a method to estimate the AUC from 0 to infinity in serial sampling data. For simplicity, in this article, we will use the AUC from 0 to the last time point.

The total number of time points is the same for all the stages, and it is denoted as K . For each drug (test or reference), the number of samples at each time point is m_s in stage s , for $s = 1, 2$ or 3 . The sample size for stage s is then $n_s = 2Km_s$, and the total sample size for the whole trial is $N = n_1 + n_2 + n_3$.

Let $j=R$ and $j=T$ represent the reference drug and test drug, respectively, the AUC can be estimated as

$$AUC_j = \sum_{k=1}^K w_k \mu_{jk} \quad (1)$$

$$w_1 = \frac{t_2 - t_1}{2}, w_K = \frac{t_K - t_{K-1}}{2}, \text{ and for } 1 < k < K, w_k = \frac{t_{k+1} - t_{k-1}}{2},$$

where μ_{jk} represents the mean concentration at time point k for drug j , w_k represents the trapezoidal weights and K is the total number of time points. Let y_{ijk} be the observed drug concentration of drug j for subject i at time point k , the AUC estimation can be given as

$$\hat{AUC}_j = \sum_{k=1}^K w_k \bar{Y}_{jk}, \text{ with } \bar{Y}_{jk} = \frac{\sum_{i=1}^{m_s} y_{ijk}}{m_s}, \quad (2)$$

where m_s is the sample size at time point k for drug j in stage s . The sample variance of the AUC for drug j is

$$\text{var}(\hat{AUC}_j) = S_j^2 = \frac{\sum_{k=1}^K w_k^2 S_{jk}^2}{m_s} \text{ with } S_{jk}^2 = \frac{\sum_{i=1}^{m_s} (y_{ijk} - \hat{Y}_{jk})^2}{(m_s - 1)}. \quad (3)$$

Whether the $\text{var}(AUC_T)$ and $\text{var}(AUC_R)$ are equal can be tested by using the F test. Here we describe only the case when $\text{var}(AUC_T) = \text{var}(AUC_R)$, and we construct the bioequivalence test with the assumption of equal variances. The pooled variance estimator is provided below. Modification for unequal variances is straightforward.

2.3.2 Bioequivalence testing with equal variances—Bioequivalence testing conducted at each stage of a sequential trial. The multiplicative testing hypothesis for a bioequivalence test with serial sampling data is as follows [14]:

$$H_0: \frac{AUC_T}{AUC_R} < \theta_1 \text{ or } \frac{AUC_T}{AUC_R} > \theta_2, \quad (4)$$

$$H_1: \theta_1 < \frac{AUC_T}{AUC_R} < \theta_2$$

where $\theta_1 = 0.8$ and $\theta_2 = 1.25$ according to the FDA [1] guidance. When the variances of the two AUCs are equal, the test statistic for the test hypothesis can be derived as

$$T_1 = \frac{\hat{AUC}_T - \theta_1 \hat{AUC}_R}{\sqrt{V_p(\hat{AUC})(1 + \theta_1^2)}}, T_2 = \frac{\hat{AUC}_T - \theta_2 \hat{AUC}_R}{\sqrt{V_p(\hat{AUC})(1 + \theta_2^2)}}, \quad (5)$$

where

$$V_p(\hat{AUC}) \approx \frac{\hat{v}_T S_T^2 + \hat{v}_R S_R^2}{\hat{v}_T + \hat{v}_R}, \quad (6)$$

$$\hat{v}_j = \frac{\left(\sum_{k=1}^K (w_k^2 S_{jk}^2 / m_s) \right)^2}{\sum_{k=1}^K w_k^4 S_{jk}^4 / \{m_s^2(m_s - 1)\}}, \quad j = R \text{ or } T. \quad (7)$$

The H_0 in equation (4) can be rejected and bioequivalence can be concluded if

$$T_1 > t_{1-\alpha}(d\hat{f}) \text{ and } T_2 < -t_{1-\alpha}(d\hat{f}), \quad (8)$$

where the degree of freedom, $d\hat{f}$, can be given as

$$d\hat{f} = \frac{\left[\sum_{k=1}^K w_k^2 (\lambda_T S_{Tk}^2 / m_s + \lambda_R S_{Rk}^2 / m_s) \right]^2}{\sum_{k=1}^K w_k^4 [\lambda_T^2 S_{Tk}^4 / \{m_s^2(m_s - 1)\} + \lambda_R^2 S_{Rk}^4 / \{m_s^2(m_s - 1)\}]}, \quad (9)$$

$$\text{and } \lambda_T = \frac{\hat{v}_T}{\hat{v}_T + \hat{v}_R}, \lambda_R = \frac{\hat{v}_R}{\hat{v}_T + \hat{v}_R}.$$

It should be noted that the df calculated here is different from that of the bioequivalence test with complete data designs where $df = n_T + n_R - 2$. The df calculated here relates to the variance estimation and the number of subjects at each time point, so it can be greatly influenced by subject allocation. In this paper, we assume that the total sample size N is equally allocated to two drugs (test and reference) and to every time point.

2.3.3 Power and sample size re-estimation—Power and sample size calculations are crucial steps in the sequential design. The power calculation method used by Potvin [6] assumed the two statistics T_1 and T_2 in equation (5) are correlated, they use an approximation to the exact Power using Owens [15] method. In this article, we will calculate the exact power. When $\text{var}(AUC_T) = \text{var}(AUC_R)$, the power function based on test hypothesis (4) is given as

$$P\left(T_1 \geq t_{1-\alpha}(df) \text{ and } T_2 \leq -t_{1-\alpha}(df) \middle| \theta_1 < \frac{AUC_T}{AUC_R} < \theta_2, V_p(\hat{AUC}), \rho\right) = 1 - \beta, \quad (10)$$

where β is the type II error rate and $1 - \beta$ is the power, which is the probability of correctly rejecting H_0 in equation (4). According to Owen [15], T_1 and T_2 are correlated and the random vector (T_1, T_2) follows a bivariate, non-central t-distribution with non-centrality parameters ϕ_1 and ϕ_2 and correlation coefficient ϕ , which are given as follows:

$$\begin{aligned} \phi_1 &= \frac{AUC_T/AUC_R - \theta_1}{[\sqrt{V_p(\hat{AUC}/AUC_R)} * \sqrt{(1 + \theta_1^2)}]}, \quad \phi_2 = \frac{AUC_T/AUC_R - \theta_2}{[\sqrt{V_p(\hat{AUC}/AUC_R)} * \sqrt{(1 + \theta_2^2)}]} \\ \rho &= \frac{\text{cov}[(AUC_T - \theta_1 AUC_R), (AUC_T - \theta_2 AUC_R)]}{\sqrt{[V(AUC_T - \theta_1 AUC_R)][V(AUC_T - \theta_2 AUC_R)]}} = \frac{1 + \theta_1 \theta_2}{\sqrt{(1 + \theta_1^2)}\sqrt{(1 + \theta_2^2)}} \end{aligned} \quad (11)$$

The power function (10) can be calculated using Owen's Q method. The specific process is described in Appendix A. The sample size re-estimation procedure used in all the sequential design methods in this article is based on this power calculation.

3 Simulation

3.1 Simulation settings

The performances of all eight sequential design methods for a bioequivalence test with serial sampling data are studied through simulations. We assume that the drug concentration data come from a one-compartment model, with first-order absorption and first-order elimination:

$$f(t, \varphi) = \frac{Dk_a}{CL - V k_a} \left\{ \exp(-k_a t) - \exp\left(-\frac{CL}{V} t\right) \right\}, \quad \text{with } \varphi = (k_a, CL, V). \quad (12)$$

where D is the dose, k_a is the absorption rate constant, CL is the clearance of the drug and V is the volume of distribution. For serial sampling data, each subject is allocated to only one time point. A parallel design is used for bioequivalence testing, which indicates that if the total sample size is N , then $N/2$ subjects receive the test drug and the other $N/2$ subjects receive the reference drug. An additive error model is used to generate the simulation data. Let y_{ijk} be the simulated drug concentration of drug j for subject i at time point k :

$$y_{ijk} = f(t_k, \varphi_j) + f(t_k, \varphi_j)\varepsilon_i + \eta_{ik}, \quad \varepsilon_i \sim N(0, \sigma^2), \quad \eta_{ik} \sim N(0, \sigma_1^2) \quad (13)$$

$$i = 1, \dots, N/2, \quad k = 1, \dots, K, \quad j = T \text{ or } R$$

For the simulation study, we fix the dose at 50mg. In order to study the power and type I error rate of the proposed method, we use two parameter settings. The power level is studied when the true ratio of the means is 0.95 and the parameters are set as follows,

$$\varphi_T = (k_a, CL, V) = (0.3, 2, 2.7), \varphi_R = (k_a, CL, V) = (0.3, 1.9, 2.7).$$

From Dubois [16], we know that the $AUC = D/CL$, so the $AUC_T/AUC_R = 1.9/2 = 0.95$. The type I error rate is studied when the true ratio of the means is 1.25 and the parameters are:

$$\varphi_T = (k_a, CL, V) = (0.3, 1.52, 2.7), \varphi_R = (k_a, CL, V) = (0.3, 1.9, 2.7),$$

where the $AUC_T/AUC_R = 1.9/1.52 = 1.25$.

We fix the measurement error variation parameter σ_1 at 0.05, and set the between-subjects variation parameter σ to be 0.1, 0.15, 0.22, 0.26, 0.3 and 0.4. The above mentioned two settings are typical in pharmacokinetics studies [16, 17, 18, 19]. Moreover, the critical parameter that affects the performance of our design is σ , the variation between-subjects. This parameter is the most important consideration for our design. It would be too many combinations to present if we let all parameters to vary in a range. We use the same sampling schedule for the test and reference drugs. We simulate a sparse sampling situation, where the drug concentration is collected at only four time points (0.5h, 4h, 8.5h, and 14h after dosing) and the number of subjects allocated to each time point is set at 3, 4, 5 or 6. We also simulate a dense sampling situation, where the drug concentration is collected at eight time points (0.5h, 1.5h, 2.5h, 4h, 6h, 8h, 10h and 14h after dosing) and the number of subjects allocated to each time point is set to 2, 3, 4 or 6.

3.2 Results and Conclusion

For each of the above scenarios with the variation parameter σ equal to 0.1, 0.15, 0.22, 0.26, 0.3 and 0.4 respectively, we conducted 10,000 simulations to study the power, and 100,000 simulations to study the type I error rate. For the scenario with a relatively large $\sigma = 0.26$, the power levels at $AUC_T/AUC_R = 0.95$ (which is within the range of bioequivalence) for all the eight designs are shown in Figure 3. For easy comparison, each one of the 3-stage designs is put in the same panel with its 2-stage counterpart. It can be easily seen that our

proposed 3-stage designs S3-BE, S3-BE_c, S3-PW and S3-PW_c have higher power levels than the current 2-stage designs in the literature, namely S2-BE, S2-BE_c, S2-PW and S2-PW_c respectively. That is to say, when the variance is large, comparing with each of the 2-stage designs, the proposed 3-stage designs have higher power levels.

For the same simulations reported in Figure 3, the average half final sample sizes ($N/2$, the number of samples for the reference or test drug) are shown in Figure 4. Surprisingly, the average sample sizes used by each of the 3-stage designs are less than or close to its 2-stage counterpart. How can smaller sample sizes result in higher power levels? To understand this phenomenon, we first make it clear that the above are results on average, not for each individual simulated trial. Then we describe two possible situations as an explanation. In the first situation, the data from stage-1 result in a very large variance estimate. By a 2-stage design, it demands a very large sample size for stage 2 to reach 80% power. However, by our proposed corresponding 3-stage design, we use a smaller sample size to reach only 60% power. This smaller sample size may still conclude bioequivalence as done by the 2-stage design, especially when the large variance in stage-1 is the result of over-estimation due to sparse serial sampling data. In this situation, the 3-stage design achieves the same conclusion with a smaller sample size. The 2-stage design overestimates and thus wastes some sample size. In another situation, suppose bioequivalence is rejected at stage 2 in a 2-stage study. However, in a 3-stage study, it has one more chance. In stage 3, the conclusion of bioequivalence may be achieved with a slightly larger total sample size. That is to say, the 3-stage design has a higher power level in this situation. Combine these two situations, it is possible that, comparing with a 2-stage design, an efficient 3-stage design achieves a higher power level with a smaller sample size on average.

The power results for scenarios with other σ values are reported in Table 3. In general, for each design, the power levels increase with sample sizes enrolled in stage 1, because more subjects will make the estimation of the variance more stable. However, the power levels decrease with the increasing variation of AUC. When the variation is small, such as $\sigma = 0.15$, the power levels achieved by the eight methods are similar and are all above 80%. When σ increases to 0.4, the power levels achieved by some methods are much less than 80%. The type I error rates at $AUC_T/AUC_R = 1.25$ are all well-controlled to be less than 0.05 (Table 4). What's more, when the variation is very large, such as $\sigma = 0.4$, we find that the type I error is smaller than that of a small variation and much less than 0.05, especially when a stopping rule is applied. Because a large variation can result in a large re-estimation sample size to reach a power level of 60% or 80% and sometimes no such a new sample size meeting the power level, then the tests stop and conclude bioinequivalence correctly. The average final half sample size are reported in Tables 5, 6, 7, and 8. Although the 3-stage designs may require slightly larger average sample sizes than the 2-stage designs when the true ratio of the means is 1.25, however, in some situations, those 2-stage designs may end up with extremely large sample sizes that are not affordable in practice. And we also can see from these four tables that the 95th percentile of sample size of 2-stage designs is large than that of 3-stage designs in most scenarios.

The designs S2-BE_c, S2-PW_c, S3-BE_c and S3-PW_c restrict that stage 2 (and stage 3) sample sizes cannot exceed four times of the stage-1 sample size n_1 . If the initial sample size

in stage 1 is small, and the variance is large, these designs have low power levels. These are the situations in the second and fourth panels in the top row of Figure 3. Still, in these situations, the power levels of the proposed 3-stage sequential designs (S3-BEc and S3-PWc) are larger than the corresponding 2-stage designs (S2-BEc and S2-PWc) respectively.

3.3 Comparison with other two-stage designs

We study the 3-stage designs under $\sigma=0.26$, $K=4$ and $m_1=6$, then compare their performance with the fixed designs and other two-stage designs, such as classical group sequential designs, promising zone method[20] and inverse normal method[21]. The number of simulation of each method is the same as that specified in Section 3.2. The promising zone two-stage design allows an adaption of the sample size which based on the estimated treatment effect and used the conditional power for the re-calculation of the sample size. As compared with the promising zone approach, recalculation of inverse normal method is not restricted to conditional power arguments.

Table 1 shows the simulated type I error rate or power, respectively, the average sample size (ASN) of different designs. From the results, we can see that GSD_56 has a similar performance with S3-BEc and S3-PWc. GSD_76 and promising zone have smaller ASNs compared with S3-BE and S3-PW when the true ratio of the means is 1.25, however, larger ASNs when the true ratio of the means is 0.95. So on the whole, each has advantages and disadvantages. What's more, the Inverse normal method need a larger sample size compared with S3-BE and S3-PW.

3.4 A simulated example

We used the simulated data to carry out a case analysis for method S3-BEc and generated the data under the setting of $\sigma=0.26$, $K=8$ and $m_1=2$.

According to the design, 16 subjects were enrolled in stage 1 and two subjects were allocated in each timepoint. After the first stage, the two statistics T_1 and T_1 were 2.3732 and -2.6865 respectively (see Table 2), so H_0 can't be rejected and BE can't be concluded. Because the power=23.18% < 60%, the trial continued and needed to reestimate the sample size to reach the power level of 60%. Based on variance from stage 1 and $\alpha=0.023$, the new sample size n_2 became 32, smaller than the constraint on maximum sample size, so the stopping rules didn't satisfied and 16 new subjects were enrolled in stage 2. According to the data from both stages, the two statistics were 3.1192 and -4.8594, therefore, we rejected the H_0 and concluded BE.

4 Discussions

We have proposed four 3-stage sequential design methods for bioequivalence studies, to respectively improve over the four 2-stage designs in the literature. Simulation studies show that our proposed designs control type I error rates, save sample sizes and improve power. These new designs help planning and conducting of valid and efficient bioequivalence studies, which evaluate the therapeutic compatibility between expensive originator drugs and cheaper generic drugs. This plays an important role on reducing the dramatically increasing medical cost and safeguarding drug efficacy.

Note that, in general, sequential design methods are not a means to reduce sample size. Instead, they are remedial measures when a bioequivalence test fails because of insufficient power. Actually, the sample size needed for a bioequivalence test using a multi-stage sequential design to achieve 80% power is larger than that needed for a single-stage bioequivalence test because the significance level α is split between the multiple testing in sequential designs. Therefore, when the variance estimate is accessible and the sample size can be correctly estimated, sequential design methods should be avoided. However, in most bioequivalence studies, the variances of the bioavailability parameters are unknown, thus the required total sample sizes are unknown, but can only be determined along the way when data are cumulated. During such a process, comparing with previous 2-stage designs, the proposed 3-stage designs are more exible to adaptively and efficiently deal with various aspects of uncertainty along the way. It warrants future investigations to optimize the multi-stage sequential designs for bioequivalence testing, especially when considering the large number of bioequivalence studies, and the great potential of cost-saving brought by efficient statistical designs.

In this paper, we choose the adjustment significance level in each stage of the proposed four 3-stage design to be 0.023, which is exclusively determined by simulation studies. And the overall type I error rate control is only demonstrated by simulations, which is lacking of a theoretical justification. We will address this in a future work.

Acknowledgments

The research of F.Y. was supported by project National Social Science of China (No. 16BTJ021). The research of X.H. was supported in part by USA NIH grants U54 CA096300, U01 CA152958 and 5P50 CA100632.

Appendix A: Power Calculation

Owen's Q function is used to express the power function (10). The structure of Owen's Q function is shown below:

$$Q(t_1, t_2, \varphi_1, \varphi_2, \rho) = \frac{\sqrt{2\pi}}{\Gamma(v/2)2^{(v-2)/2}} \int_0^\infty \Phi_2\left(\frac{t_1x}{\sqrt{v}} - \varphi_1, \frac{t_2x}{\sqrt{v}} - \varphi_2, \rho\right) x^{v-1} \Phi'(x) dx, \quad (\text{A.1})$$

$$\Phi_2(x, y, \rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \int_{-\infty}^x \int_{-\infty}^y \exp\left(-\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)}\right) dudv,$$

$$\Phi'(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

where v represents the degrees of freedom df and equals to df in equation (9). Applying (A.1) to the power function (10), the power of a bioequivalence test for the ratio of the two AUCs at a desired power level of $1 - \beta$ under the condition of equal variances can be given as

$$\begin{aligned}
1 - \beta &= P\left(T_1 \geq t_{1-\alpha}(df) \text{ and } T_2 \leq -t_{1-\alpha}(df) \mid \theta_1 < \frac{AUC_T}{AUC_R} < \theta_2, V_p(\widehat{AUC}), \rho\right) \\
&= Q(\infty, t_{1-\alpha}, v, \varphi_1, \varphi_2, \rho) - Q(t_{1-\alpha}, v, -t_{1-\alpha}, \varphi_1, \varphi_2, \rho) \\
&= \frac{1}{\Gamma(v/2)2^{(v-2)/2}\pi\sqrt{1-\rho^2}} \times \\
&\quad \int_0^\infty \left[\int_{(t_{1-\alpha}, v^{z/\sqrt{v}}) - \varphi_1}^\infty \int_{-\infty}^{(-t_{1-\alpha}, v^{z/\sqrt{v}}) - \varphi_2} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) dx dy \right] z^{v-1} \exp\left(-\frac{z^2}{2}\right) dz
\end{aligned}
\tag{A.2}$$

where

$$\begin{aligned}
\varphi_1 &= \frac{AUC_T/AUC_R - \theta_1}{\left[\sqrt{V_p(\widehat{AUC}/AUC_R)} * \sqrt{(1+\theta_1^2)}\right]}, \varphi_2 = \frac{AUC_T/AUC_R - \theta_2}{\left[\sqrt{V_p(\widehat{AUC}/AUC_R)} * \sqrt{(1+\theta_2^2)}\right]} \\
\rho &= \frac{\text{cov}[(AUC_T - \theta_1 AUC_R), (AUC_T - \theta_2 AUC_R)]}{\sqrt{[V(AUC_T - \theta_1 AUC_R)][V(AUC_T - \theta_2 AUC_R)]}} = \frac{1 + \theta_1 \theta_2}{\sqrt{(1+\theta_1^2)}\sqrt{(1+\theta_2^2)}}
\end{aligned}
\tag{A.3}$$

Appendix B: Review of 2-stage Designs

Methods S2-BE & S2-PW: Methods S2-BE and S2-PW are the foundation of six other methods and they equal to the two-stage sequential design method B and C proposed by Potvin [6]. The detailed description of these two methods can be found in Potvin [6]. In order to control the overall type I error rate to a maximum of 5% for two-stage sequential designs, the significance level α is chosen to be 0.0294 according to Potvin [6] and Pocock [11].

Methods S2-BEc & S2-PWc: Stopping criteria or futility rules are not involved in methods S2-BE and S2-PW, but EMA [22] explicitly states that the two-stage trials to demonstrate bioequivalence is acceptable when the stopping criteria is clearly defined. Fuglsang's [7] work investigated stopping criteria defined as scenarios when the total sample size (n_2) cannot exceed $2\times$, $3\times$, $4\times$ the initial sample size (n_1) at stage 1 or the total sample size (n_2) is fixed to 60 and 80. These criteria have different levels of effects on the power of bioequivalence trials but the tendencies are similar, so we only studied the case when stopping criterion is defined as $n_2 = 4 \times n_1$. After applying this stopping criterion to methods S2-BE and S2-PW, we can get methods S2-BEc and S2-PWc. They are also two-stage sequential designs and is still chosen to be 0.0294 in order to control the overall type I error rate to a maximum of 5%. The detailed procedures of methods S2-BE, S2-BEc, S2-PW and S2-PWc are illustrated in decision trees shown in Figures 5 and 6.

Figure 5.

Sequential design method S2-BE. The method S2-BE was similar to S2-BEc, which is only lack of the branch in bold dashed box and continue to stage 2 directly

Figure 6.

Sequential design method S2-PW. The method S2-PW was similar to S2-PWc, which is only lack of the branch in bold dashed box and continue to stage 2 directly

Appendix C: Simulation Results

Table 3

The powers of the eight sequential design methods for the bioequivalence test.

σ	K	m_1	$n_1/2$	S2-BE	S2-BE _c	S3-BE	S3-BE _c	S2-PW	S2-PW _c	S3-PW	S3-PW _c
0.1	4	3	12	0.9690	0.9705	0.9678	0.9679	0.9704	0.9715	0.9709	0.9741
		4	16	0.9768	0.9740	0.9758	0.9747	0.9827	0.9833	0.9850	0.9821
		5	20	0.9883	0.9856	0.9822	0.9848	0.9925	0.9920	0.9932	0.9926
		6	24	0.9942	0.9958	0.9923	0.9926	0.9978	0.9970	0.9977	0.9980
	8	2	16	0.9764	0.9761	0.9739	0.9770	0.9831	0.9823	0.9849	0.9841
		3	24	0.9979	0.9982	0.9974	0.9965	0.9993	0.9992	0.9992	0.9995
		4	32	1.0000	1.0000	0.9999	0.9995	0.9999	0.9999	0.9999	0.9999
		6	48	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.15	4	3	12	0.9312	0.8618	0.9502	0.9189	0.9401	0.8675	0.9472	0.9165
		4	16	0.9395	0.9300	0.9372	0.9410	0.9479	0.9349	0.9449	0.9402
		5	20	0.9442	0.9457	0.9456	0.9480	0.9527	0.9484	0.9517	0.9538
		6	24	0.9509	0.9455	0.9538	0.9583	0.9603	0.9578	0.9629	0.9600
	8	2	16	0.9442	0.9446	0.9686	0.9645	0.9405	0.9353	0.9489	0.9504
		3	24	0.9515	0.9466	0.9740	0.9436	0.9612	0.9633	0.9629	0.9802
		4	32	0.9755	0.9806	0.9943	0.9948	0.9897	0.9867	0.9978	0.9976
		6	48	0.9974	0.9987	0.9996	1.0000	0.9995	0.9998	1.0000	0.9999
0.22	4	3	12	0.9455	0.5836	0.9184	0.6174	0.9369	0.5475	0.9147	0.6135
		4	16	0.8937	0.7246	0.9200	0.8192	0.9080	0.7245	0.9507	0.8739
		5	20	0.9135	0.8573	0.9171	0.8881	0.9211	0.8399	0.9326	0.8914
		6	24	0.9246	0.8969	0.9198	0.9145	0.9217	0.9033	0.9407	0.9136
	8	2	16	0.9046	0.7970	0.9199	0.8780	0.8996	0.8100	0.9065	0.8717
		3	24	0.9091	0.9019	0.9278	0.9240	0.9092	0.9053	0.9120	0.9075
		4	32	0.9156	0.9196	0.9236	0.9207	0.9281	0.9317	0.9787	0.9796
		6	48	0.9453	0.9452	0.9519	0.9484	0.9619	0.9643	0.9660	0.9608
0.26	4	3	12	0.8292	0.3323	0.9034	0.3970	0.8301	0.3315	0.9023	0.3949
		4	16	0.8800	0.5010	0.9022	0.6137	0.8940	0.5220	0.9040	0.6386
		5	20	0.9306	0.7135	0.9301	0.7985	0.9170	0.6889	0.9499	0.8180
		6	24	0.9563	0.8208	0.9662	0.8580	0.9573	0.8100	0.9657	0.8909
	8	2	16	0.8680	0.6333	0.8952	0.7806	0.9185	0.6402	0.9044	0.7621
		3	24	0.8974	0.8946	0.8957	0.9003	0.8950	0.9155	0.9094	0.8999
		4	32	0.9060	0.9077	0.9043	0.9002	0.9160	0.9105	0.9914	0.9774
		6	48	0.9025	0.9016	0.9930	0.9934	0.9158	0.9161	0.9912	0.9922
0.3	4	3	12	0.8360	0.2009	0.9173	0.2277	0.8265	0.1920	0.9206	0.2282
		4	16	0.8779	0.3943	0.9185	0.4059	0.9179	0.3869	0.9191	0.4055
		5	20	0.8905	0.5210	0.9289	0.5958	0.8938	0.4673	0.9203	0.5974
		6	24	0.9008	0.6189	0.9289	0.7571	0.9033	0.6164	0.9318	0.7624
	8	2	16	0.8308	0.4684	0.9589	0.6061	0.8406	0.4104	0.9432	0.5912
		3	24	0.8722	0.7504	0.9048	0.8885	0.8742	0.7568	0.9075	0.8869
		4	32	0.9302	0.9212	0.9848	0.9707	0.9319	0.9200	0.9853	0.9693

σ	K	m_1	$n_1/2$	S2-BE	S2-BEc	S3-BE	S3-BEc	S2-PW	S2-PWc	S3-PW	S3-PWc
0.4	4	6	48	0.8901	0.8914	0.9196	0.9909	0.9238	0.9310	0.9395	0.9899
		3	12	0.6972	0.0406	0.8395	0.0406	0.6980	0.0422	0.8768	0.0511
		4	16	0.7779	0.0715	0.8871	0.0916	0.7801	0.0731	0.8907	0.0888
		5	20	0.8121	0.1486	0.8860	0.1750	0.8074	0.1405	0.8997	0.1797
	8	6	24	0.8813	0.2793	0.9858	0.3151	0.8781	0.2870	0.9014	0.3120
		2	16	0.8753	0.1700	0.8860	0.1531	0.8711	0.1691	0.8972	0.1984
		3	24	0.9699	0.5078	0.9907	0.6291	0.9666	0.4714	0.9876	0.6287
		4	32	0.9070	0.7885	0.8918	0.8444	0.9047	0.7892	0.8885	0.8464
		6	48	0.9178	0.9249	0.8945	0.8906	0.8793	0.8790	0.9047	0.9084

The total number of time points K is the same for all the stages. For each drug (test or reference), the number of samples at each time point is m_s in stage s , for $s = 1; 2$ or 3 . The sample size for stage s is then $n_s = 2Km_s$, and the total sample size for the whole trial is $N = n_1 + n_2 + n_3$.

Table 4

The type I error rates of the eight sequential design methods for the bioequivalence test.

σ	K	m_1	$n_1/2$	S2-BE	S2-BEc	S3-BE	S3-BEc	S2-PW	S2-PWc	S3-PW	S3-PWc
0.1	4	3	12	0.00059	0.00932	0.00169	0.00024	0.00527	0.00762	0.01182	0.01344
		4	16	0.00067	0.00873	0.00176	0.00040	0.00564	0.00841	0.01406	0.01287
		5	20	0.00093	0.00877	0.00202	0.00033	0.00672	0.00843	0.01443	0.01389
		6	24	0.00079	0.00875	0.00554	0.00037	0.00648	0.00838	0.01511	0.01302
	8	2	16	0.00108	0.01215	0.00283	0.00049	0.00278	0.01313	0.02123	0.02027
		3	24	0.00157	0.01648	0.00377	0.00076	0.00164	0.01675	0.03091	0.02743
		4	32	0.00169	0.01825	0.00698	0.00074	0.01268	0.02884	0.00659	0.02944
		6	48	0.00207	0.01887	0.00781	0.00082	0.01808	0.01970	0.01702	0.03073
0.15	4	3	12	0.00469	0.01047	0.00291	0.00260	0.00125	0.01026	0.00484	0.00621
		4	16	0.00446	0.01104	0.00326	0.00209	0.00348	0.01121	0.00515	0.02392
		5	20	0.00420	0.01035	0.00324	0.00150	0.00550	0.01231	0.00586	0.00707
		6	24	0.00514	0.01088	0.00315	0.00300	0.00648	0.01391	0.00480	0.01332
	8	2	16	0.01023	0.01714	0.00405	0.00609	0.01011	0.01983	0.01184	0.01837
		3	24	0.01163	0.01848	0.00575	0.00568	0.01535	0.03332	0.01754	0.01960
		4	32	0.01329	0.01720	0.00484	0.01073	0.01364	0.03762	0.02181	0.01624
		6	48	0.01393	0.01691	0.00536	0.01196	0.02359	0.04042	0.02234	0.01719
0.22	4	3	12	0.00291	0.00890	0.00287	0.00085	0.00567	0.00914	0.00668	0.00584
		4	16	0.00377	0.01019	0.00245	0.00080	0.00500	0.01080	0.00623	0.00724
		5	20	0.00398	0.01021	0.00277	0.00117	0.00610	0.01156	0.00305	0.00715
		6	24	0.00338	0.01062	0.00255	0.00076	0.00617	0.01177	0.00292	0.00706
	8	2	16	0.00852	0.01655	0.00476	0.00089	0.00949	0.01043	0.00566	0.01737
		3	24	0.00803	0.02244	0.00674	0.00071	0.01142	0.02207	0.00375	0.01768
		4	32	0.01008	0.01848	0.00471	0.00103	0.01268	0.02543	0.00757	0.02127
		6	48	0.01443	0.01924	0.00649	0.01281	0.00995	0.02931	0.01301	0.02493
0.26	4	3	12	0.00654	0.00469	0.00787	0.00310	0.00529	0.00315	0.00530	0.00576
		4	16	0.00707	0.00629	0.00818	0.00439	0.00510	0.00502	0.00517	0.00777

σ	K	m_1	$n_1/2$	S2-BE	S2-BEc	S3-BE	S3-BEc	S2-PW	S2-PWc	S3-PW	S3-PWc	
0.3	8	5	20	0.00759	0.00764	0.00849	0.00491	0.00556	0.00590	0.00541	0.00780	
		6	24	0.00808	0.00837	0.00543	0.00492	0.00488	0.00688	0.00437	0.01031	
		2	16	0.01175	0.00893	0.00265	0.01232	0.00861	0.00709	0.00892	0.01271	
		3	24	0.01327	0.01220	0.00352	0.00820	0.01179	0.01200	0.00631	0.01847	
		4	32	0.01783	0.00282	0.00652	0.01571	0.00964	0.01295	0.00607	0.01833	
		6	48	0.01516	0.00383	0.00503	0.01527	0.01309	0.01622	0.01009	0.02466	
	4	3	12	0.00714	0.00113	0.00709	0.00282	0.00212	0.00400	0.00304	0.00336	
		4	16	0.00741	0.00204	0.00740	0.00400	0.00956	0.00597	0.00335	0.00458	
		5	20	0.00765	0.00214	0.00791	0.00521	0.01122	0.00786	0.00317	0.00614	
		6	24	0.00770	0.00281	0.00788	0.00594	0.01087	0.00967	0.00305	0.00647	
		8	2	16	0.00504	0.00582	0.00754	0.00886	0.01757	0.01181	0.01870	0.01077
			3	24	0.01321	0.01133	0.00884	0.01560	0.02006	0.02318	0.01849	0.01770
	0.4	8	4	32	0.01457	0.01671	0.00943	0.01670	0.02034	0.02958	0.02025	0.02020
			6	48	0.01115	0.02004	0.00968	0.01575	0.01980	0.02980	0.02143	0.01826
4			3	12	0.00419	0.00078	0.00784	0.00061	0.00327	0.00027	0.00462	0.00053
			4	16	0.00369	0.00083	0.00767	0.00106	0.00329	0.00021	0.00458	0.00123
			5	20	0.00371	0.00129	0.00746	0.00151	0.00319	0.00074	0.00671	0.00196
			6	24	0.00370	0.00209	0.00778	0.00187	0.00742	0.00154	0.00643	0.00200
		8	2	16	0.00904	0.00118	0.00921	0.00320	0.01513	0.00074	0.00241	0.00346
			3	24	0.00085	0.00278	0.00773	0.00801	0.00609	0.00214	0.00633	0.00930
4		4	32	0.00771	0.00616	0.00741	0.00679	0.01794	0.00520	0.00535	0.01302	
		6	48	0.00930	0.00873	0.00940	0.00873	0.01680	0.00493	0.00758	0.01631	

The total number of time points K is the same for all the stages. For each drug (test or reference), the number of samples at each time point is m_s in stage s , for $s = 1; 2$ or 3 . The sample size for stage s is then $n_s = 2Km_s$, and the total sample size for the whole trial is $N = n_1 + n_2 + n_3$.

Table 5

Estimated sample sizes for methods S2-BE, S2-BEc, S3-BE and S3-BEc when the true ratio of the means is 1.25.

σ	K	m_1	$n_1/2$	Average sample size(5th, 50th, 95th percentile)			
				S2-BE	S2-BEc	S3-BE	S3-BEc
0.1	4	3	12	18.65(12,16,36)	17.88(12,16,32)	19.84(12,20,36)	19.48(12,20,36)
		4	16	19.85(16,16,32)	19.45(16,16,32)	20.72(16,16,32)	20.83(16,16,32)
		5	20	21.90(20,20,32)	21.56(20,20,28)	22.61(20,20,32)	22.65(20,20,32)
		6	24	24.76(24,24,28)	24.58(24,24,28)	25.10(24,24,32)	25.30(24,24,32)
	8	2	16	19.91(16,16,32)	19.76(16,16,32)	23.52(16,16,32)	23.47(16,16,32)
		3	24	24.30(24,24,24)	24.16(24,24,24)	24.79(24,24,24)	24.89(24,24,40)
		4	32	32.00(32,32,32)	32.00(32,32,32)	32.02(32,32,32)	32.02(32,32,32)
		6	48	48(48,48,48)	48(48,48,48)	48(48,48,48)	48(48,48,48)
0.15	4	3	12	32.56(12,28,76)	23.49(12,20,44)	34.06(12,32,64)	25.61(12,24,48)
		4	16	32.77(16,28,68)	29.38(16,28,56)	33.58(16,32,60)	31.72(16,28,56)
		5	20	33.12(20,28,60)	32.13(20,28,56)	33.82(20,32,56)	33.67(20,32,56)

				Average sample size(5th, 50th, 95th percentile)					
σ	K	m_1	$n_1/2$	S2-BE	S2-BE _c	S3-BE	S3-BE _c		
0.22	6	24	24	33.72(24,28,60)	33.36(24,28,56)	34.58(24,32,56)	34.81(24,32,56)		
		8	2	30.18(16,24,56)	29.29(16,24,48)	34.27(16,32,56)	33.68(16,32,48)		
		3	24	30.10(24,24,48)	29.92(24,24,48)	33.82(24,40,48)	33.86(24,40,48)		
		4	32	33.51(32,32,40)	33.34(32,32,40)	35.67(32,32,48)	35.42(32,32,48)		
		6	48	48.01(48,48,48)	48.01(48,48,48)	48.09(48,48,48)	48.05(48,48,48)		
	4	3	12	61.14(16,48,148)	21.40(12,12,44)	67.27(20,64,128)	18.58(12,12,44)		
		4	16	62.52(20,56,136)	31.93(16,28,60)	65.58(24,64,116)	30.83(16,16,64)		
		5	20	62.93(24,56,128)	42.91(20,40,76)	64.76(28,64,112)	45.19(20,44,76)		
		6	24	63.13(24,56,120)	51.98(24,52,88)	64.22(32,64,104)	56.34(24,56,92)		
		8	2	16	53.63(24,48,112)	35.99(16,32,64)	57.92(32,56,96)	38.52(16,40,64)	
3	24		52.79(24,48,88)	50.52(24,48,80)	55.56(40,56,80)	54.79(24,56,80)			
4	32		52.82(32,48,80)	52.44(32,48,80)	55.70(32,56,80)	55.84(32,56,80)			
6	48		54.97(48,48,72)	54.69(48,48,72)	59.10(48,64,80)	58.96(48,64,80)			
0.26	4		3	12	77.75(20,68,164)	18.84(12,12,44)	90.92(32,88,156)	14.87(12,12,40)	
		4	16	80.64(24,72,160)	28.01(16,16,60)	89.39(36,88,148)	23.44(16,16,60)		
		5	20	82.36(28,76,156)	39.28(20,32,76)	88.54(36,88,140)	36.36(20,20,76)		
		6	24	83.20(32,80,148)	51.70(24,52,92)	88.04(40,88,136)	53.55(24,56,96)		
		8	2	16	70.54(24,64,144)	32.20(16,24,64)	77.14(32,72,128)	29.59(16,16,64)	
	3	24	24	70.07(32,64,120)	57.15(24,56,96)	73.84(40,72,112)	64.76(24,64,96)		
		4	32	69.79(44,64,112)	68.54(32,64,112)	72.76(48,72,104)	72.35(48,72,104)		
		6	48	70.11(48,72,104)	70.52(48,72,104)	74.22(48,72,104)	74.19(48,72,96)		
		0.3	4	3	12	92.26(24,88,172)	17.06(12,12,44)	113.03(48,116,172)	13.21(12,12,20)
				4	16	96.13(32,92,168)	24.40(16,16,60)	112.52(52,116,164)	18.81(16,16,48)
5	20			99.07(40,96,164)	33.91(20,20,76)	111.94(56,116,160)	26.71(20,20,72)		
6	24		24	101.05(44,100,160)	46.05(24,24,92)	111.43(56,116,156)	38.99(24,24,92)		
	8		2	16	87.88(32,80,168)	27.74(16,16,64)	96.61(48,96,152)	20.67(16,16,56)	
		3	24	87.69(40,88,144)	55.14(24,56,96)	92.49(56,96,128)	60.79(24,72,96)		
		4	32	87.56(48,88,136)	81.72(32,80,120)	90.85(56,88,120)	88.70(48,88,120)		
		6	48	87.58(56,88,120)	87.89(56,88,120)	90.49(64,88,120)	90.58(64,88,120)		
0.4		4	3	12	112.93(36,112,180)	13.92(12,12,32)	135.77(92,132,184)	12.16(12,12,12)	
	4		16	116.97(52,116,172)	18.73(16,16,48)	135.15(100,132,176)	16.29(16,16,16)		
	5		20	119.50(64,120,168)	25.48(20,20,72)	134.46(100,132,172)	20.64(20,20,20)		
	6		24	121.15(72,124,168)	36.30(24,24,92)	134.04(100,132,168)	26.34(24,24,24)		
	8		2	16	124.61(48,120,216)	20.72(16,16,56)	128.36(88,120,208)	16.34(16,16,16)	
	3	24	24	117.20(64,112,184)	41.05(24,24,96)	120.64(88,112,168)	36.88(24,24,96)		
		4	32	113.92(80,112,160)	88.40(32,96,128)	116.78(88,112,160)	92.70(32,104,128)		
		6	48	108.00(80,104,144)	107.76(80,104,144)	112.33(88,112,144)	112.19(88,112,144)		

The total number of time points K is the same for all the stages. For each drug (test or reference), the number of samples at each time point is m_s in stage s , for $s = 1; 2$ or 3 . The sample size for stage s is then $n_s = 2Km_s$, and the total sample size for the whole trial is $N = n_1 + n_2 + n_3$.

Table 6

Estimated sample sizes for methods S2-PW, S2-PWc, S3-PW and S3-PWc when the true ratio of the means is 1.25.

σ	K	m_1	$n_1/2$	Average sample size(5th, 50th, 95th percentile)			
				S2-PW	S2-PWc	S3-PW	S3-PWc
0.1	4	3	12	18.06(12,12,36)	17.35(12,12,36)	18.15(12,12,32)	17.92(12,12,32)
		4	16	19.21(16,16,32)	19.07(16,16,32)	18.90(16,16,32)	18.91(16,16,32)
		5	20	21.37(20,20,32)	21.29(20,20,32)	21.08(20,20,28)	21.11(20,20,28)
		6	24	24.42(24,24,24)	24.39(24,24,24)	24.30(24,24,24)	24.33(24,24,24)
	8	2	16	18.69(16,16,32)	18.44(16,16,32)	19.47(16,16,32)	19.53(16,16,32)
		3	24	24.09(24,24,24)	24.07(24,24,24)	24.06(24,24,24)	24.07(24,24,24)
		4	32	32(32,32,32)	32(32,32,32)	32(32,32,32)	32(32,32,32)
		6	48	48(48,48,48)	48(48,48,48)	48(48,48,48)	48(48,48,48)
0.15	4	3	12	32.39(12,28,76)	23.04(12,20,44)	32.65(12,32,64)	24.82(12,24,48)
		4	16	32.30(16,28,68)	28.95(16,28,56)	32.03(16,28,60)	30.56(16,28,56)
		5	20	32.53(20,28,60)	31.60(20,28,60)	32.08(20,28,56)	31.87(20,28,56)
		6	24	33.04(24,24,60)	32.62(24,24,56)	32.68(24,24,56)	32.36(24,24,52)
	8	2	16	29.28(16,24,56)	28.26(16,24,48)	31.50(16,32,56)	31.07(16,32,48)
		3	24	28.60(24,24,48)	28.07(24,24,40)	29.26(24,24,48)	29.24(24,24,48)
		4	32	32.06(32,32,32)	32.50(32,32,32)	32.82(32,32,32)	32.91(32,32,48)
		6	48	48.00(48,48,48)	48.00(48,48,48)	48(48,48,48)	48.00(48,48,48)
0.22	4	3	12	61.16(12,52,148)	21.30(12,12,44)	66.20(12,64,124)	18.16(12,12,44)
		4	16	62.22(16,56,136)	31.82(16,28,60)	64.76(16,64,116)	31.13(16,16,64)
		5	20	62.48(20,56,128)	42.71(20,40,76)	63.75(20,64,108)	45.97(20,48,80)
		6	24	62.60(24,56,120)	51.81(24,52,88)	63.27(24,64,104)	56.56(24,60,92)
	8	2	16	53.51(16,48,112)	34.97(16,32,64)	56.92(16,56,96)	40.36(16,48,64)
		3	24	52.34(24,48,88)	50.10(24,48,80)	53.88(24,56,88)	53.10(24,56,80)
		4	32	51.68(32,48,80)	51.36(32,48,80)	52.38(32,56,80)	52.24(32,56,80)
		6	48	53.20(48,48,80)	52.34(48,48,72)	53.25(48,48,80)	52.80(48,48,80)
0.26	4	3	12	78.12(20,68,168)	18.84(12,12,44)	90.24(28,88,156)	14.47(12,12,40)
		4	16	80.39(24,72,160)	27.90(16,16,60)	89.27(32,88,148)	22.95(16,16,60)
		5	20	82.29(28,76,156)	38.92(20,32,76)	88.33(36,88,144)	36.33(20,20,80)
		6	24	83.05(32,80,152)	51.24(24,52,92)	87.44(36,88,140)	54.10(24,56,96)
	8	2	16	70.38(24,64,144)	31.96(16,24,64)	76.15(32,72,120)	29.20(16,16,64)
		3	24	69.83(24,64,120)	56.91(24,56,96)	72.94(40,72,112)	65.24(24,72,96)
		4	32	69.63(32,64,112)	68.17(32,64,104)	71.38(32,72,104)	71.26(32,72,104)
		6	48	68.52(48,72,104)	68.50(48,72,104)	69.85(48,72,104)	69.72(48,72,96)
0.3	4	3	12	90.79(28,84,172)	16.58(12,12,44)	111.92(40,116,172)	13.03(12,12,12)
		4	16	96.81(32,96,168)	24.10(16,16,60)	111.86(48,116,164)	18.56(16,16,44)
		5	20	99.36(40,96,164)	33.68(20,20,76)	111.32(52,116,160)	26.54(20,20,72)
		6	24	101.23(44,100,160)	46.38(24,24,92)	110.90(52,116,156)	38.93(24,24,92)
	8	2	16	88.65(32,80,168)	27.62(16,16,64)	96.52(48,96,144)	20.42(16,16,56)

σ	K	m_1	$n_1/2$	Average sample size(5th, 50th, 95th percentile)			
				S2-PW	S2-PWc	S3-PW	S3-PWc
0.4	3	3	24	87.94(40,88,144)	56.24(24,64,96)	92.50(56,96,128)	61.07(24,72,96)
		4	32	87.59(48,88,128)	82.73(32,88,120)	90.60(56,96,120)	88.68(48,88,120)
		6	48	87.16(48,88,120)	87.41(48,88,120)	89.41(48,88,112)	89.30(48,88,112)
		4	3	112.38(36,112,176)	14.16(12,12,36)	134.81(92,132,180)	12.14(12,12,12)
		4	16	116.32(52,116,172)	18.87(16,16,48)	134.27(96,132,176)	16.26(16,16,16)
		5	20	119.28(64,120,168)	26.05(20,20,72)	134.31(100,132,172)	20.56(20,20,20)
	8	6	24	122.34(72,124,168)	37.11(24,24,92)	133.97(100,132,168)	26.39(24,24,24)
		2	16	126.41(48,120,216)	20.85(16,16,56)	127.15(88,120,208)	16.31(16,16,16)
		3	24	118.19(64,112,184)	41.42(24,24,96)	120.60(88,112,176)	36.69(24,24,96)
		4	32	115.25(80,112,160)	88.05(32,96,128)	116.04(88,112,160)	93.24(32,104,128)
		6	48	108.02(80,104,144)	107.37(80,104,144)	112.37(88,112,144)	112.20(88,112,144)

The total number of time points K is the same for all the stages. For each drug (test or reference), the number of samples at each time point is m_s in stage s , for $s = 1; 2$ or 3 . The sample size for stage s is then $n_s = 2Km_s$, and the total sample size for the whole trial is $N = n_1 + n_2 + n_3$.

Table 7

Estimated sample sizes for methods S2-BE, S2-BEc, S3-BE and S3-BEc when the true ratio of the means is 0.95.

σ	K	m_1	$n_1/2$	Average sample size(5th, 50th, 95th percentile)			
				S2-BE	S2-BEc	S3-BE	S3-BEc
0.1	4	3	12	14.25(12,12,28)	14.29(12,12,28)	13.81(12,12,24)	13.90(12,12,24)
		4	16	16.41(16,16,16)	16.50(16,16,16)	16.46(16,16,20)	16.47(16,16,20)
		5	20	20.05(20,20,20)	20.05(20,20,20)	20.07(20,20,20)	20.06(20,20,20)
		6	24	24.00(24,24,24)	24.00(24,24,24)	24.00(24,24,24)	24.00(24,24,24)
		8	2	16.60(16,16,24)	16.60(16,16,24)	16.78(16,16,24)	16.79(16,16,24)
		3	24	24.00(24,24,24)	24.00(24,24,24)	24.00(24,24,24)	24.00(24,24,24)
	8	4	32	32.00(32,32,32)	32.00(32,32,32)	32.00(32,32,32)	32.00(32,32,32)
		6	48	48.00(48,48,48)	48.00(48,48,48)	48.00(48,48,48)	48.00(48,48,48)
		4	3	23.69(12,16,56)	20.17(12,12,40)	21.44(12,20,44)	20.02(12,16,40)
		4	16	22.89(16,16,48)	22.07(16,16,48)	21.53(16,16,40)	21.46(16,16,40)
		5	20	23.20(20,20,44)	23.21(20,20,44)	22.63(20,20,36)	22.70(20,20,36)
		6	24	25.22(24,24,32)	25.26(24,24,32)	25.12(24,24,32)	25.18(24,24,36)
0.15	8	2	16	22.13(16,16,40)	22.07(16,16,40)	21.62(16,16,32)	21.46(16,16,32)
		3	24	24.37(24,24,24)	24.39(24,24,24)	24.46(24,24,24)	24.55(24,24,32)
		4	32	32.00(32,32,32)	32.00(32,32,32)	32.00(32,32,32)	32.00(32,32,32)
		6	48	48.00(48,48,48)	48.00(48,48,48)	48.00(48,48,48)	48.00(48,48,48)
		4	3	46.98(12,40,120)	21.04(12,12,44)	42.27(12,40,88)	23.59(12,20,48)
		4	16	44.78(16,40,96)	30.85(16,28,60)	40.08(16,36,80)	32.37(16,32,60)
	8	5	20	43.45(20,40,92)	36.86(20,28,72)	38.73(20,36,72)	36.53(20,36,68)
		6	24	42.03(24,24,88)	39.35(24,24,80)	38.03(24,36,68)	37.72(24,36,68)
		2	16	40.57(16,40,80)	33.30(16,32,64)	36.45(16,32,64)	34.51(16,32,56)

σ	K	m_1	$n_1/2$	Average sample size(5th, 50th, 95th percentile)			
				S2-BE	S2-BEc	S3-BE	S3-BEc
0.26	3	3	24	34.34(24,24,64)	34.26(24,24,64)	33.13(24,24,56)	33.07(24,24,56)
		4	32	34.92(32,32,56)	34.67(32,32,56)	35.20(32,32,48)	35.12(32,32,48)
		6	48	48.06(48,48,48)	48.04(48,48,48)	48.11(48,48,48)	48.10(48,48,48)
		8	64	61.12(12,52,144)	20.69(12,12,44)	58.61(12,56,120)	20.43(12,12,48)
	4	3	12	61.51(16,56,132)	30.24(16,16,60)	55.74(16,52,108)	32.51(16,28,60)
		4	16	61.31(20,56,128)	38.87(20,20,76)	53.79(20,52,104)	41.77(20,40,76)
		6	24	60.58(24,60,124)	47.67(24,48,88)	50.65(24,48,96)	48.04(24,48,84)
		8	32	54.57(16,48,112)	34.06(16,32,64)	48.82(16,48,88)	35.83(16,32,64)
0.3	3	3	24	48.33(24,48,88)	45.53(24,48,88)	44.44(24,48,72)	43.66(24,40,72)
		4	32	43.69(32,32,80)	43.77(32,32,88)	42.20(32,32,72)	42.32(32,32,72)
		6	48	49.27(48,48,56)	49.44(48,48,56)	48.80(48,48,48)	48.74(48,48,48)
		8	64	75.56(12,68,160)	18.68(12,12,44)	75.83(12,72,148)	16.83(12,12,44)
	4	3	12	77.69(16,72,152)	27.36(16,16,60)	73.99(16,72,140)	27.79(16,16,60)
		4	16	79.16(20,76,152)	37.28(20,20,76)	71.68(20,68,132)	40.30(20,36,76)
		6	24	78.24(24,76,144)	50.15(24,52,92)	70.19(24,68,128)	52.47(24,52,92)
		8	32	68.97(16,64,136)	30.94(16,16,64)	61.91(16,56,112)	32.98(16,32,64)
0.4	3	3	24	66.02(24,64,120)	53.80(24,56,96)	58.38(24,56,96)	53.28(24,56,88)
		4	32	59.71(32,56,112)	58.72(32,56,104)	52.62(32,48,88)	53.50(32,56,88)
		6	48	54.99(48,48,48)	55.37(48,48,96)	54.94(48,48,80)	53.23(48,48,80)
		8	64	102.56(30,100,172)	14.85(12,12,40)	114.98(52,120,168)	12.81(12,12,12)
	4	3	12	105.43(40,104,164)	20.97(16,16,56)	112.13(52,116,160)	18.82(16,16,48)
		4	16	108.10(48,108,160)	30.06(20,20,76)	110.68(48,116,156)	27.33(20,20,76)
		6	24	108.78(52,112,156)	41.22(24,24,92)	105.95(44,112,152)	40.31(24,24,92)
		8	32	104.86(32,96,200)	23.96(16,16,54)	102.15(48,96,168)	21.31(16,16,64)

The total number of time points K is the same for all the stages. For each drug (test or reference), the number of samples at each time point is m_s in stage s , for $s = 1; 2$ or 3 . The sample size for stage s is then $n_s = 2Km_s$, and the total sample size for the whole trial is $N = n_1 + n_2 + n_3$.

Table 8

Estimated sample sizes for methods S2-PW, S2-PWc, S3-PW and S3-PWc when the true ratio of the means is 0.95.

σ	K	m_1	$n_1/2$	Average sample size(5th, 50th, 95th percentile)			
				S2-PW	S2-PWc	S3-PW	S3-PWc
0.1	4	3	12	14.25(12,12,28)	14.14(12,12,28)	13.75(12,12,24)	13.71(12,12,24)
		4	16	16.48(16,16,16)	16.48(16,16,16)	16.33(16,16,16)	16.34(16,16,16)
		5	20	20.04(20,20,20)	20.05(20,20,20)	20.04(20,20,20)	20.03(20,20,20)
		6	24	24.00(20,20,20)	24.01(20,20,20)	24.00(24,24,24)	24.00(24,24,24)
	8	2	16	16.45(16,16,16)	16.47(16,16,16)	16.46(16,16,24)	16.53(16,16,24)
		4	32	16.45(16,16,16)	16.47(16,16,16)	16.46(16,16,24)	16.53(16,16,24)

σ	K	m_1	$n/2$	Average sample size(5th, 50th, 95th percentile)			
				S2-PW	S2-PWc	S3-PW	S3-PWc
0.15		3	24	24.00(24,24,24)	24.00(24,24,24)	24.00(24,24,24)	24.00(24,24,24)
		4	32	32.00(32,32,32)	32.00(32,32,32)	32.00(32,32,32)	32.00(32,32,32)
		6	48	48.00(48,48,48)	48.00(48,48,48)	48.00(48,48,48)	48.00(48,48,48)
		8	64	64.00(64,64,64)	64.00(64,64,64)	64.00(64,64,64)	64.00(64,64,64)
	4	3	12	23.54(12,12,52)	20.42(12,12,44)	21.14(12,16,44)	19.91(12,16,40)
		4	16	22.92(16,16,48)	22.06(16,16,48)	21.16(16,16,40)	21.03(16,16,40)
		5	20	23.12(20,20,44)	23.26(20,20,44)	22.39(20,20,36)	22.52(20,20,36)
		6	24	25.20(24,24,32)	25.10(24,24,24)	24.96(24,24,32)	24.95(24,24,32)
0.22		3	24	24.25(24,24,24)	24.21(24,24,24)	24.23(24,24,24)	24.33(24,24,24)
		4	32	32.00(32,32,32)	32.00(32,32,32)	32.00(32,32,32)	32.00(32,32,32)
		6	48	48.00(48,48,48)	48.00(48,48,48)	48.00(48,48,48)	48.00(48,48,48)
		8	64	64.00(64,64,64)	64.00(64,64,64)	64.00(64,64,64)	64.00(64,64,64)
	4	3	12	46.80(12,40,120)	21.82(12,12,44)	42.43(12,40,92)	23.25(12,20,48)
		4	16	45.44(16,40,100)	30.92(16,28,60)	38.78(16,36,80)	30.38(16,28,60)
		5	20	43.73(20,40,92)	36.63(20,28,72)	38.10(20,36,76)	36.32(20,36,68)
		6	24	41.74(24,24,880)	39.39(24,24,80)	37.70(24,32,72)	37.29(24,32,68)
0.26		3	24	33.98(24,24,64)	34.07(24,24,64)	32.62(24,24,56)	32.42(24,24,56)
		4	32	34.40(32,32,56)	34.39(32,32,56)	34.30(32,32,48)	34.23(32,32,48)
		6	48	48.02(48,48,48)	48.02(48,48,48)	48.01(48,48,48)	48.02(48,48,48)
		8	64	64.00(64,64,64)	64.00(64,64,64)	64.00(64,64,64)	64.00(64,64,64)
	4	3	12	61.52(12,52,144)	20.55(12,12,44)	59.04(12,56,124)	20.26(12,12,48)
		4	16	62.00(16,56,132)	29.84(16,16,60)	56.38(16,52,112)	31.35(16,28,60)
		5	20	60.93(20,56,128)	39.76(20,40,76)	52.43(20,48,104)	40.72(20,40,76)
		6	24	60.02(24,60,120)	47.42(24,48,88)	50.84(24,48,96)	47.20(24,44,84)
0.3		3	24	54.31(16,48,112)	33.94(16,32,64)	48.85(16,48,88)	36.56(16,40,64)
		4	32	48.27(24,48,96)	45.20(24,40,88)	43.99(24,40,72)	43.36(24,40,72)
		6	48	43.71(32,32,88)	43.77(32,32,88)	41.19(32,32,72)	41.05(32,32,72)
		8	64	48.79(48,48,48)	48.86(48,48,48)	48.74(48,48,48)	48.60(48,48,48)
	4	3	12	74.86(12,68,160)	18.45(12,12,44)	76.36(12,72,148)	17.09(12,12,44)
		4	16	77.72(16,72,156)	27.26(16,16,60)	74.16(16,72,140)	27.84(16,16,60)
		5	20	78.80(20,76,148)	38.16(20,20,76)	71.70(20,68,132)	40.69(20,36,76)
		6	24	78.92(24,76,144)	49.97(24,52,92)	69.70(24,68,128)	52.42(24,52,92)
0.4		3	24	70.70(16,64,136)	31.66(16,16,64)	61.99(16,56,112)	33.18(16,32,64)
		4	32	66.06(24,64,120)	54.48(24,56,96)	58.08(24,56,96)	53.55(24,56,88)
		6	48	59.89(32,56,112)	59.17(32,56,112)	52.96(32,48,96)	53.26(32,56,88)
		8	64	54.07(48,48,96)	53.87(48,48,96)	53.78(48,48,80)	53.14(48,48,80)
	4	3	12	102.53(32,100,172)	14.99(12,12,40)	113.11(48,116,168)	12.90(12,12,12)
		4	16	105.99(40,108,168)	21.00(16,16,56)	111.86(52,116,160)	18.65(16,16,48)
		5	20	107.80(48,108,160)	29.72(20,20,76)	110.44(48,116,156)	27.42(20,20,72)
		6	24	108.92(52,112,156)	42.17(24,24,92)	109.09(48,116,156)	39.60(24,24,92)
	8	2	16	104.23(32,96,200)	24.03(16,16,64)	102.42(48,96,168)	22.22(16,16,64)
		3	24	102.26(24,104,168)	48.76(24,24,96)	90.27(40,88,144)	54.56(24,56,96)

σ	K	m_1	$n_{1/2}$	Average sample size(5th, 50th, 95th percentile)			
				S2-PW	S2-PWc	S3-PW	S3-PWc
	4	32		100.85(32,104,144)	85.57(32,96,128)	91.00(32,96,136)	85.47(32,88,120)
	6	48		87.99(48,96,136)	88.35(48,96,128)	85.57(48,88,120)	85.42(48,88,120)

The total number of time points K is the same for all the stages. For each drug (test or reference), the number of samples at each time point is m_s in stage s , for $s = 1; 2$ or 3 . The sample size for stage s is then $n_s = 2Km_s$, and the total sample size for the whole trial is $N = n_1 + n_2 + n_3$.

References

1. FDA. Guidance for industry: bioavailability and bioequivalence studies for orally administered drug products general considerations. Food and Drug Administration; Washington, DC: 2003.
2. Chow SC, Wang H. On sample size calculation in bioequivalence trials. *Journal of Pharmacokinetics and Pharmacodynamics*. 2001; 28:155–169. [PubMed: 11381568]
3. Hua SY, Hawkins D, Zhou J. Statistical considerations in bioequivalence of two area under the concentration–time curves obtained from serial sampling data. *Journal of Applied Statistics*. 2013; 40(5):1140–1154.
4. HPFB. Guidance for industry: conduct and analysis of bioavailability and bioequivalence studies—part a: oral dosage formulations used for systemic effects. Health Canada, Ministry of Health-Health Products and Food Branch; 1992.
5. WHO. WHO expert committee on specifications for pharmaceutical preparations. fortieth report. World Health Organization technical report series. 2006; 937:1. [PubMed: 16836287]
6. Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ. Sequential design approaches for bioequivalence studies with crossover designs. *Pharmaceut Statist*. 2008; 7(4):245–62.
7. Fuglsang A. Futility rules in bioequivalence trials with sequential designs. *The AAPS journal*. 2014; 16(1):79–82. [PubMed: 24218038]
8. Fuglsang A. Sequential bioequivalence approaches for parallel designs. *The AAPS journal*. 2014; 16(3):373–378. [PubMed: 24526610]
9. Schütz H. Two-stage designs in bioequivalence trials. *European journal of clinical pharmacology*. 2015; 71(3):271–281. [PubMed: 25604509]
10. Xu J, Audet C, Diliberti CE, Hauck WW, Montague TH, Parr AF, Potvin D, Schuirmann DJ. Optimal adaptive sequential designs for crossover bioequivalence studies. *Pharmaceutical Statistics*. 2016; 15(1):15–27. [PubMed: 26538182]
11. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977; 64(2):191–199.
12. Nedelman JR, Gibiansky E, Lau DTW. Applying bailer's method for auc confidence intervals to sparse sampling. *Pharmaceutical research*. 1995; 12(1):124–128. [PubMed: 7724473]
13. Wolfsegger MJ, Jaki T. Estimation of AUC from 0 to infinity in serial sacrifice designs. *Journal of pharmacokinetics and pharmacodynamics*. 2005; 32(5-6):757–766. [PubMed: 16328101]
14. Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*. 1996; 11(4):283–319.
15. Owen D. A special case of a bivariate non-central t-distribution. *Biometrika*. 1965; 52(3–4):437–446.
16. Dubois A, Lavielle M, Gsteiger S, Pigeolet E, Mentre F. Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm. *Statistics in medicine*. 2011; 30(21):2582–2600. [PubMed: 21793036]
17. Jaki T, Wolfsegger MJ, Lawo JP. Establishing bioequivalence in complete and incomplete data designs using aucs. *Journal of Biopharmaceutical Statistics*. 2010; 20(4):803–820. [PubMed: 20496207]

18. Panhard X, Mentré F. Evaluation by simulation of tests based on nonlinear mixed-effects models in pharmacokinetic interaction and bioequivalence cross-over trials. *Statistics in medicine*. 2005; 24(10):1509–1524. [PubMed: 15761916]
19. Dubois A, Gsteiger S, Pigeolet E, Mentré F. Bioequivalence tests based on individual estimates using non-compartmental or model-based analyses: evaluation of estimates of sample means and type i error for different designs. *Pharmaceutical research*. 2010; 27(1):92–104. [PubMed: 19876723]
20. Kieser M, Rauch G. Two-stage designs for cross-over bioequivalence trials. *Statistics in Medicine*. 2015; 34(16):2403. [PubMed: 25809815]
21. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999; 55(4):1286–1290. [PubMed: 11315085]
22. EMA. Guideline on the investigation of bioequivalence. European Medicines Agency (EMA); London: 2010. 27

Figure 1.

Proposed 3-stage sequential design method S3-BE, aiming to improve over 2-stage design S2-BE. The method S3-BE was similar to S3-BE, which is only lack of the branches in bold dashed boxes and continue to the next stage directly, aiming to improve over 2-stage design S2-BE.

Figure 2.

Proposed 3-stage sequential design method S3-PWc, aiming to improve over 2-stage design S2-PWc. The method S3-PW was similar to S3-PWc, which is only lack of the branches in bold dashed boxes and continue to the next stage directly, aiming to improve over 2-stage design S2-PW.

Figure 3.

Estimated by 10,000 simulations, the power levels of the eight sequential design methods for the bioequivalence test when $\sigma = 0.26$, and the true ratio of mean AUCs is 0.95. All designs have the same number of time points, K , in each stage. Top panel: $K=4$; bottom panel: $K=8$. In stage 1, all designs have the same number of samples, m , at each time point.

Figure 4.

Average half final sample sizes ($N/2$) used by the eight sequential design methods for the bioequivalence test when $\sigma = 0.26$, and the true ratio of mean AUCs is 0.95. All designs have the same number of time points, K , in each stage. Top panel: $K=4$; bottom panel: $K=8$. In stage 1, all designs have the same number of samples, m , at each time point.

Table 1

Performance of 3-stage designs under $\sigma=0.26$, $K=4$ and $m_1=6$ with other 2-stage designs.

Design	power		ASN	
	0.95	1.25	0.95	1.25
Fixed_48	0.8789	0.01988	48	48
Fixed_60	0.9566	0.01270	60	60
GSD_56 (24)	0.8660	0.00989	47.03	55.72
GSD_76 (24)	0.9612	0.00838	64.26	75.69
Promising zone	0.9576	0.00372	61.95	77.98
Inverse normal	0.9682	0.00876	80.60	95.61
S2-BEc	0.8208	0.00837	47.67	51.70
S2-PWc	0.8100	0.00688	47.42	51.24
S2-BE	0.9563	0.00808	60.58	83.20
S2-PW	0.9573	0.00488	60.02	83.05
S3-BEc	0.8580	0.00492	48.04	53.55
S3-PWc	0.8909	0.01031	47.20	54.10
S3-BE	0.9662	0.00543	50.65	88.04
S3-PW	0.9657	0.00437	50.84	87.32

Table 2

Simulated concentrations and statistics at each stage.

stage	t1	t2	t3	t4	t5	t6	t7	t8
R	2.4464	3.0380	3.6216	3.6589	2.0583	1.4660	0.5153	0.2048
	3.5078	4.5959	4.7424	4.0566	2.3326	0.5236	0.7374	0.3308
	1.7970	4.5758	4.4836	4.2461	1.5042	1.7135	0.5700	0.2561
T	1.4887	4.5122	4.5031	2.6143	1.9624	0.9653	0.7900	0.2964
	$T1=2.3732, T2=-2.6865, df=5.777, t_{0.977} (df)=2.533, \text{power}=23.18\%$							
	1.3157	4.2358	4.9457	4.6564	2.5012	1.5603	0.6209	0.2132
2	2.1767	5.8431	4.1943	3.6932	1.6449	1.3130	0.7408	0.2163
	3.4120	3.0948	5.4394	3.6082	1.6167	1.7861	0.6847	0.1840
	2.5212	3.3374	3.9434	4.2023	2.1023	1.0366	0.5229	0.2647
$T1=3.1192, T2=-4.8594, df=27.707, t_{0.977} (df)=2.089$								