Pharm Res (2013) 30:1824–1835
DOI 10.1007/s11095-013-1026-3

RESEARCH PAPER

An Insight into the Properties of a Two-Stage Design in Bioequivalence Studies

Vangelis Karalis • Panos Macheras

Received: 17 October 2012 / Accepted: 12 March 2013 / Published online: 9 April 2013

ABSTRACT

Purpose Unveil the properties of a two-stage design (TSD) for bioequivalence (BE) studies.

Methods A TSD with an upper sample size limit (UL) is described and analyzed under different conditions using Monte Carlo simulations. TSD was split into three branches: A, B1, and B2. The first stage included branches A and B1, while stage two referred to branch B2. Sample size re-estimation at B2 relies on the observed GMR and variability of stage 1. The properties studied were % BE acceptance, % uses and % efficiency of each branch, as well as the reason of BE failure.

Results No inflation of type I error was observed. Each TSD branch exhibits different performance. Stage two exhibits the greatest % BE acceptances when highly variable drugs are assessed with a low starting number of subjects (N_1) or when formulations differ significantly. Branch A is more frequently used when variability is low, drug products are similar, and a large N_1 is included. BE assessment at branch A is very efficient.

Conclusions The overall acceptance profile of TSD resembles the typical pattern observed in single-stage studies, but it is actually different. Inclusion of a UL is necessary to avoid inflation of type I error.

KEY WORDS adaptive design · bioequivalence · Monte Carlo simulations · sample size · two-stage design

ABBREVIATIONS

A	Branch of the two-stage design selected for bioequivalence assessment when 'power' of the study using the starting population is higher or equal to 80%.
ANOVA	Analysis of variance
AUC	Area under the concentration-time curve
B	Branch of the two-stage design selected for bio-equivalence assessment when 'power' of the study using the starting population is less than 80%
B1	First segment of branch B which belongs to stage 1 of the design
B2	The second segment of branch B which belongs to stage 2 of the design and where sample re-estimation takes place
BE	Bioequivalence
CI	Confidence interval
Cmax	Maximum plasma concentration value
CVw	Within-subject coefficient of variation
EMA	European Medicines Agency
F_A	Type of failure because the 90% confidence interval around GMR of branch A lies outside the BE limits
F_{B1}	Type of failure because the 94.12% confidence interval around GMR of branch B1 lies outside the BE limits
F_{B2}	Type of failure because the 94.12% confidence interval around GMR of branch B2 lies outside the BE limits
FDA	The US Food and Drug Administration
F_N	Type of failure because the required number of subjects, after sample size re-estimation, leads to a total sample size that is greater than the pre-set maximum allowable level
GMR	Geometric mean ratio
N	Total number of subjects participating in the study
N_1	Starting number of subjects enrolled in the study
N_2	Additional number of subjects recruited at the second stage
PK	Pharmacokinetic
R	Reference formulation
T	Test formulation

V. Karalis (✉) · P. Macheras
Laboratory of Biopharmaceutics-Pharmacokinetics Faculty of Pharmacy
National and Kapodistrian University of Athens, Panepistimiopolis
Athens 15771, Greece
e-mail: vkaralis@pharm.uoa.gr

TSD	The two-stage design introduced and studied
UL	Upper sample size limit
α	Type I error of the nominal statistical hypothesis

INTRODUCTION

The aim of bioequivalence (BE) studies is to assess the *in vivo* equivalence between two drug products (1,2). Classically, if no other reason for biowaiver can be granted, BE can be proved through specifically designed clinical studies. BE studies are actually clinical trials and their planning follows the same principles as in clinical studies. Therefore, among others, estimating sample size requires a prior knowledge of the *variability*, of the active moiety under investigation, as well as an estimate for the *difference* in the mean values of the BE measures (e.g., AUC, Cmax) between the two products, i.e., test (T) and reference (R). Since, BE studies are usually conducted using crossover designs, the estimate of variability, actually, refers to within-subject variability of the drug. The mean difference between the pharmacokinetic (PK) measures of the two products is expressed by their geometric mean ratio (GMR). Thus, if an estimate for within-subject coefficient of variation (CVw) and/or GMR cannot be pre-established accurately, the study might be underpowered if few subjects are included or overpowered when the assumed sample size is higher than actually required. Both situations lead to severe risks such as aimless costs and unnecessary human exposure to drugs. Therefore, in order to face these problems, alternate designs can be considered instead of a typical single-stage study.

Alternate design methods may refer to a wide variety of approaches, such as add-on, group sequential, and adaptive designs. These methods are not new in clinical research, and introduced since the 1970s (3–11). Even though alternate methods offer many advantages in clinical research, some difficulties are also present (12,13). These problems arise from the fact that many adaptations of the study may lead to a significantly different trial, while the type I error (i.e., the significance level α) of the statistical hypothesis may be inflated. The greater the number of interim analyses carried out, the greater the chance of Type I errors. Thus, a major concern is to preserve the overall type I error rate at a specific level (e.g., 5%) for the pharmacokinetic endpoint (e.g. AUC, Cmax) (14,15). In order to resolve this problem, several methods have been proposed such as Bonferroni correction and the approaches introduced by Pocock, O'Brien–Fleming, and Lan–DeMets (15–19).

Alternate design methods have also attracted attention in BE assessment (20–24). In addition, regulatory authorities like World Health Organization and the Japanese Pharmaceuticals and Medical Devices Evaluation Agency allow the use of add-on designs in BE assessment (25,26). In case of Health Canada, the latest guidance on BE studies posted in 2012 allows the application of group-sequential and adaptive designs (27). The Korean Food and Drug Administration currently allows the conduction of an additional trial (28). Besides, according to the Australian regulatory guidelines add-on or sequential designs are not addressed at this time, but it is anticipated that group sequential BE studies using the Bonferroni method will be accepted (29). Finally, the US Food and Drug Administration (FDA) also recommends the use of two-stage group-sequential design approaches (30,31).

The European Medicines Agency (EMA) suggested the use of two-stage designs for BE purposes (1). According to EMA's approach, an initial group of subjects can be treated and their data analyzed; if BE is not demonstrated then an additional group can be included and the results of both studies should be combined in final analysis. The EMA 2010 guideline generally mentions that appropriate steps should be followed to preserve the overall type I error at the nominal level of 5%. The latter can be accomplished using 94.12% confidence intervals (CI) for the analysis of both the first stage as well as of the entire set of data after completing the second stage of the study (1). However, specific criteria and stopping rules explaining how to perform a two-stage BE study are not defined in the guideline.

In this context, Potvin *et al.* (32) and Montague *et al.* (33) published two very interesting studies towards the validation of methods on two-stage cross-over BE studies. The authors evaluated four methods for sample size re-estimation in BE trials: a simple naive approach and three variations of adaptive methods assuming GMR values equal to 0.95 or 0.90 (32,33). Finally, the authors ended-up with recommendations, for the regulatory agencies or the sponsors, on the appropriateness of each method regarding their application at specific conditions (e.g., an assumed population GMR value equal to 0.90 or 0.95).

The aim of this study is to present a two-stage design (TSD) for BE studies and to unveil its underlying properties. The TSD approach used in this analysis originates from the 'C' method quoted in the Potvin/Montague articles (32,33), but differs in the following two points: a) sample size re-estimation in our TSD is based on the actual GMR and CVw observed at stage 1 and b) in the presented TSD an upper sample size limit is introduced. The entire task was accomplished by using Monte Carlo simulations. The performance of the entire design as well as the individual contribution of each branch of the TSD were investigated. For this reason, several, already established or newly proposed in this study, methodological tools were used. The properties examined for each one of the branches was the % BE acceptance, % uses, % efficiency, and the % individual branch failure.

MATERIALS AND METHODS

Two-Stage Design

In this analysis, a two-stage design for BE studies is introduced, which originates from the so quoted 'C' adaptive sample size sequential method by Potvin *et al.* and Montague *et al.* (32,33). A schematic representation of this TSD is depicted in Fig. 1. Each stage of this TSD consists of a two-sequence, two-period (2×2) crossover design.

The first step of the analysis is the evaluation of the power at stage 1 using the estimated CVw, GMR, and $\alpha=5\%$. If this power estimate is higher than or equal to 80%, then BE should be assessed at the 5% level of significance (branch A in Fig. 1). Irrespectively of the outcome of BE assessment, BE or failure, the analysis should be ended afterwards. A common feature of this TSD, which is also noted in the Potvin/Montague C method, is that the predefined significance level at the first stage is 5%, namely, equal to the one applied to a typical single-stage BE study.

If the power estimate of the study is lower than 80%, branch B of the TSD method should be followed (Fig. 1). In this case, BE should be assessed by setting type I error equal to 2.94%. Again, if BE is proved then assessment should be stopped (branch B1 in Fig. 1). Otherwise, the TSD method will proceed into branch B2 (i.e., the second stage of TSD) where sample size re-estimation takes place. The latter will be based on the variability and the GMR estimates derived from stage 1 and setting $\alpha=2.94\%$. Finally, assessment of BE will be made using the entire data set, from both stages 1 and 2, on the basis of a stringent type I error equal to 2.94%.

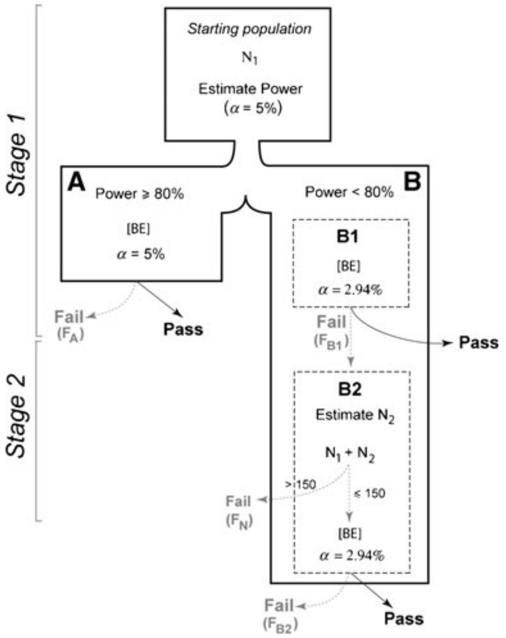


Fig. 1 Schematic representation of the two-stage design (TSD) investigated in this study. Based on the power estimate of the starting sample the entire TSD is initially divided into two main branches A and B. Passage through sub-branch B1 is a premise for the assessment of bioequivalence at B2 of the second stage. *Key*: N_1, starting sample size; N_2, additional number of subjects recruited at the second stage; 150, the maximum allowed number of subjects from stage 1 and 2; α, type I error of the nominal hypothesis of bio-in-equivalence; [BE], bioequivalence assessment. The terms F_A, F_{B1}, F_{B2}, and F_N refer to the possible types of failure (see Table I).

It should be highlighted that in this study sample size re-estimation is based on the actual CVw and the GMR estimated of stage 1 rather than an assumed population GMR of 0.90 or 0.95. In addition, an upper limit (UL) to the total sample size (for both stages) is introduced. The inclusion of UL was found to be necessary in order to avoid inflation of type I error. It has been reported that sample size increases, based on interim results of the treatments difference, may inflate type I error (34). The value of 150 was arbitrarily chosen, as it refers to a rational maximum sample size value for a BE study.

Bioequivalence Assessment

Wherever BE assessment was made, it was based on the concept of average BE (1,2). According to this approach, two drug products are declared bioequivalent if the calculated confidence interval (typically, 90% CI for single-stage studies) around the difference of the mean measures of bioavailability (in the *ln*-transformed scale) lies within predefined limits imposed by the regulatory authorities (1,2,35). These BE limits are usually set equal to 80.00% and 125.00%. Alternatively, this approach is equivalent to the two one-sided test procedure (36).

In this study, a general linear model (ANOVA) was applied to the *ln*-transformed values of the PK metric. For stage 1, the terms used in the ANOVA model were Sequence, Period, Treatment, and Subject-within-Sequence (1). In case of the analysis using the combined data, from stages 1 and 2, the factors considered for the model were Sequence, Treatment, Stage, as well as the nested terms Period-within-Stage and Subject-within-(Sequence × Stage) (1,32,33). All these effects were treated as 'fixed' factors (1). Since, cross-over design is used, the residual variability derived from the linear models was considered to reflect the within-subject variability of the drug under study.

The so derived within-subject variability was used to construct (1-2α)% confidence intervals around the GMR (T/R) of the PK parameter. In particular, for the determination of BE at branch A it was used a 90% CI, while a 94.12% CI was applied to branch B1 or to the combined dataset from stages 1 and 2.

Construction of the Design Matrix

For the purposes of this study, only a single PK endpoint (e.g., AUC or Cmax) was assumed. Simulated values for the PK parameter were generated assuming *log*-normal distribution. In case of the reference formulation the value of the PK endpoint was assumed to be equal to 100 units These PK values for each product (T or R) were appropriately assigned to the two Sequences and the two Periods of stage 1 of the study in a way that ensured randomness and balance with respect to sequence, period, and treatment.

If BE assessment proceeded into sample size re-estimation, then the simulated data at stage 2 were added to the previously created design matrix in order to construct the entire data matrix of the study. Data for stage 2 were generated similarly, so as to ensure balance with respect to sequence, period, and treatment. No tests or criteria for the poolability of the data from stages 1 and 2 were applied.

Sample Size Re-estimation

The basic advantage of two-stage designs in BE studies is the possibility for sample size re-estimation if BE was not proved at stage 1. In case of the TSD investigated in this study, an initial number of subjects (N_1) are assumed at stage 1, while sample size re-estimation takes place at branch B2 (Fig. 1). The additionally required number of subjects (N_2) is calculated based on the within-subject variability and the GMR estimate obtained at stage 1 assuming power equal to 80% and $\alpha = 2.94\%$.

In our case, sample size re-estimation was achieved through an automated iterative algorithm developed for the purposes of this study. In more details, this algorithm was composed from two main elements:

a) Firstly, an approximate method was used to obtain an initial rough estimate of the required sample size. The mathematical formula included also a ‘correction term’ in order to account for the cases of low number of subjects (37).
b) After an initial estimate of N_2 was obtained, the iterative method was applied to calculate accurately the number of subjects. Based on the degrees of freedom of the initially estimated N_2, the inverse of Student’s t cumulative distribution function was computed, which was inserted again in mathematical formulas to re-estimate N_2. Convergence to the final N_2 value was considered when either the difference in the N_2 values between two consecutive estimations was less than 0.5, or the maximum number of iterations (set equal to 100) was exceeded. Depending on the observed GMR, the algorithm was able to select between different formulas regarding difference or no treatment difference.

The value of N_2 obtained from the above-mentioned algorithm was then rounded to the nearest integer. If the so-derived number was odd, it was converted to even by adding 1.

In addition, a UL was set to the highest value of N_1+N_2 at stage 2. In order to be realistic, no more than a total number of N=N_1+N_2=150 subjects could be enrolled in the BE study. Thus, the number of N_2 could range between 2 and 150 – N_1. It should be mentioned that N_2 is estimated whether or not BE is going to be declared at the final step of BE assessment (i.e., B2).

After the construction of the final matrix, BE was assessed using the entire set of data as described earlier (e.g., ANOVA on the *ln*-transformed values *etc.*). It is worth mentioning that BE assessment was based on Pocock’s method (15) even though the sample sizes N_1 and N_2 might be different.

Simulations

In order to study the properties of TSD, two levels of theoretical CVw values of the initial population (Fig. 1) were considered in the simulations: 20% and 40%. These CVw values were selected to reflect the conditions of medium and high within-subject variability. In addition, three levels of starting values of sample size (i.e., N_1) were considered: 18, 30, and 60 subjects, which refer to low, medium, and relatively high number of subjects, respectively. It is admitted that some combinations of CVw and N_1 (e.g., CVw=20% and N_1=60) may not be realistic. Nevertheless, these extreme cases were included in the simulations not only for reasons of completeness, but as an effort to unveil any possible trends in the performance of the two-stage BE design that might have not been identified otherwise. The theoretical GMR value was gradually changed, from 1.00 to 1.25 using a step of 0.025. Under each condition, a number of 100,000 studies according to the TSD scheme (Fig. 1) were simulated. In each study, BE was declared if the (1-2α)% confidence interval around point GMR for the two T and R products was between the BE limits (36).

In case of the estimation of the type I error rate values 1,000,000 studies were simulated. When a UL of 150 was assumed, six levels of CVw (10%, 20%, 30%, 40%, 50%, and 60%) and seven different starting sample sizes (12, 18, 24, 30, 48, 60, and 96) were considered. For UL of 100 and 1,000 subjects, two CVw (20%, 40%) and four different N_1 (18, 30, 60, 96) values were assumed.

The entire programming work was implemented by developing all necessary functions in MATLAB® (The MathWorks, Inc). All functions were validated prior to their use, while the Monte Carlo simulation approach was in agreement with other published studies and our previous works (38–41).

Investigational Methods

In order to assess the performance and the underlying properties of the two-stage design of Fig. 1, several methodological tools were applied. In all cases, the calculated metric was plotted *versus* the theoretical GMR of the study.

Percentage of Acceptance: Total and Relative Contribution of Each Branch

The percentage of simulated studies showing BE was recorded and allowed the construction of *power curves* by plotting the % acceptance values as a function of GMR. For the purposes of this study, not only the total power (%) was recorded, but also the individual % BE acceptances of each branch of TSD (Fig. 1). The % BE acceptance values at each GMR value refer to:

$$\%\ \text{Acceptance} = 100 \cdot \frac{\text{Number of studies showing bioequivalence}}{\text{Total number of simulated studies}} \tag{1}$$

Percent Uses of Each Branch

It would be informative to know how many times BE was assessed (successfully or not) at each segment of TSD. To answer this demand, a new type of plots was constructed where the percentage of usage of each branch was recorded *versus* the GMR of the study, Eq. 2:

$$\%\ \text{Uses} = 100 \cdot \frac{\text{Number of branch accesses}}{\text{Total number of simulated studies}} \tag{2}$$

Efficiency of Each Branch

The % relative "efficiency" of each TSD branch expresses the number of studies declared bioequivalent by each branch divided by the number of times this branch was utilized, Eq. 3:

$$\begin{aligned}\%\ \text{Efficiency} &= 100 \cdot \frac{\%\ \text{Acceptance}}{\%\ \text{Uses}} \\ &= 100 \cdot \frac{\text{Number of studies showing BE}}{\text{Number of branch accesses}}\end{aligned} \tag{3}$$

Origin of Failure to Demonstrate Bioequivalence

In addition, the origin of "failure", namely, the inability to declare BE was explored. This task was undertaken since the reason of rejecting BE (i.e., failure) in the TSD can be a result of the inability to demonstrate BE at branches A, B1, and B2, as well as due to the fact that the total sample size (i.e., N_1+N_2) after sample size re-estimation at B2, is greater than the pre-set maximum allowable level (UL=150 in this study). These failures were termed as F_A, F_{B1}, F_{B2}, and F_N, respectively (Table I). Alternatively, the F_N type of failure could have been considered as belonging to F_{B2}; however, F_N is evaluated separately from F_{B2} in order to provide a better insight into the properties of the TSD.

The total number of failures at each GMR step of the study can easily be computed by summarizing all separate failures. The latter further allows the estimation of the % frequency of each failure by dividing the number of failures due to any reason by the total number of failures at each GMR.

$$\%\ \text{Frequency of failure}\ i = 100 \cdot \frac{\text{Number of failures}\ i}{\text{Total number of failures}} \tag{4}$$

where the symbol "i" refers to the specific type of failure of the TSD.

RESULTS

Table II lists the type I error values for the TSD used in this study. In all cases, the estimated values are lower than 5% and therefore no inflation of type I error beyond 5% becomes apparent. In contrast to the Potvin/Montague methods, where estimation of N_2 is based on a pre-defined population GMR of 0.90 or 0.95, the TSD under study uses the observed GMR of the stage 1 study. In the past, it has been shown that sample size increases, based on interim results of the treatments difference, may inflate type I error (34). However, the presented TSD approach further includes an upper limit of 150 and only two stages. If our design either included more than one interim analyses (apart from the final analysis), or allowed the recruitment of any number of subjects then possibly type I error would be inflated. Nevertheless, it should not be disregarded that this TSD applies to BE studies where only two stages are allowed (1) and in most of the cases less than 150 subjects are enrolled.

In order to verify this postulation, further results for type I error rate are listed in Table III. In this case, two different UL values were considered: a) a low UL equal to 100 and b) a high total number of subjects (up to 1,000) is allowed to be recruited in the BE study. The results quoted in Table III are in accordance with our theoretical expectations. As UL decreases, the TSD approach becomes rather conservative, whereas the expansion of UL leads to an increase of the type I error. For a UL equal to 1,000, type I error values do not exceed 5% for the conditions studied. Nevertheless, if no UL was considered, the increase would probably exceed 5%. In actual practice it is very rare for a BE study to enroll more than 150 subjects and for this reason a UL of 150 was considered as a rational limit.

Table I Origin of Possible Failure to Declare Bioequivalence (BE) in the Proposed Two-Stage Design

Symbol	Description	Branch
F_A	The 90% confidence interval (CI), assessed at branch A, lies outside the BE limits 80.00–125.00%.	A
F_{B1}	The 94.12% CI, assessed at branch B1, lies outside the BE limits 80.00–125.00%.	B1
F_{B2}	The 94.12% CI, assessed at branch B2, lies outside the BE limits 80.00–125.00%.	B2
F_N	The required number of subjects, after sample size re-estimation, leads to a total sample size that is greater than the pre-set maximum allowable level (150 in this study).	B2

Percentage of Bioequivalence Acceptance

Figure 2 depicts the percentage of simulated studies in which BE is accepted *versus* the GMR of the study. As it anticipated, the increase of variability, from 20% to 40%, leads to lower total and individual % acceptance values. In the same vein, as the number of subjects increases, the total % acceptance level is raised. In case of low variability values (Fig. 2a, c, and e), branch A exerts a predominant role. This behavior becomes more evident as sample size increases (Fig. 2c and e). As GMR rises, the % acceptances of B1 and B2 increase, but up to a maximum GMR value, after which both B1 and B2 lose their ability to declare BE In these cases, the relative contribution of B2 is more prominent at low sample sizes (Fig. 2a), while branch B1 dominates when a high number of subjects is used (Fig. 2e).

A quite different performance, for the individual role of each branch, is observed when the drug exerts high within-subject variability (right column of Fig. 2). Branch B2 appears to be the sole reason for success when a highly variable drug is assessed in a BE study with few subjects (Fig. 2b). However, as the number of subjects recruited in the BE study increases, the importance of both A and B1 becomes more potent (Fig. 2d). A further enlargement of sample size, makes BE assessment at branch A the most prominent method for addressing BE (Fig. 2f).

Table II Type I Error Rate for the Two-Stage Design Under Study

CVw	N_1	Type I error rate (%)	CVw	N_1	Type I error rate (%)
10	12	4.49	40	12	1.86
	18	4.20		18	2.35
	24	4.06		24	2.99
	30	3.88		30	3.46
	48	3.47		48	3.46
	60	3.18		60	3.18
	96	2.95		96	2.95
20	12	4.46	50	12	1.07
	18	4.20		18	1.29
	24	4.06		24	1.60
	30	3.88		30	2.03
	48	3.47		48	3.19
	60	3.18		60	3.12
	96	2.95		96	2.95
30	12	3.31	60	12	0.60
	18	3.86		18	0.76
	24	4.03		24	0.86
	30	3.87		30	1.00
	48	3.47		48	1.96
	60	3.18		60	2.57
	96	2.95		96	2.93

Six levels of within-subject variability (CVw) and seven levels of starting sample (N_1) size are quoted. An upper sample size limit of 150 is assumed

Percent Uses of Each Branch

Passage through a branch of TSD simply reflects the number of times BE was assessed (positively or negatively) within this branch. In order to accomplish this task Fig. 3 was constructed which shows the % uses of each of the three TSD branches *versus* the GMR of the study. In all cases, BE assessment at segment A becomes less frequently visited as

Table III Type I Error Rate for the Two-Stage Design Under Study if the Upper Limit (UL) is Set Equal to 100 or 1,000

CVw	N_1	Type I error rate (%)	
		UL = 100	UL = 1000
20	18	3.98	4.95
	30	3.50	4.78
	60	2.95	4.52
	96	2.95	4.26
40	18	2.05	3.13
	30	3.06	4.38
	60	2.95	4.52
	96	2.95	4.26

Two levels of within-subject variability (CVw) and four levels of starting sample (N_1) size are quoted

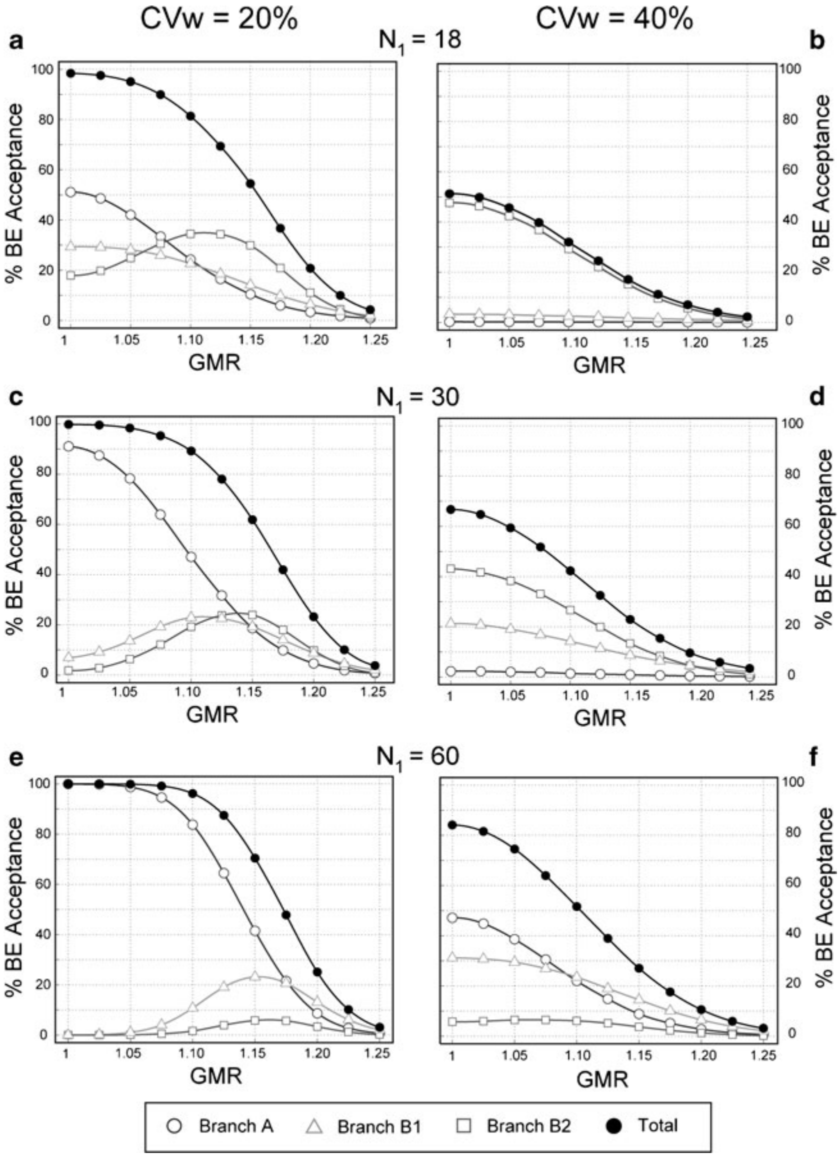


Fig. 2 Percentage of bioequivalence (BE) studies accepted *versus* GMR. The individual performance of each of the three branches (A, B1, B2) of the two-stage design are shown. The coefficient of variation of the within-subject variability (CVw) was equal to 20% (*left column*) and 40% (*right column*) for the initial population. Three different starting sample sizes, N_1, were assumed: 18, 30, and 60.

GMR differentiates from unity. The latter, in turn, results in a more frequent use of branch B (either B1, or B2). This attribute becomes more apparent as CVw increases and N_1 diminishes (Fig. 3).

An increase of within-subject variability makes the role of branch B more important. On the contrary, as the starting sample size gets larger the necessity of branch B fades and becomes only essential for two drugs much different to each other. In other words, as the CVw/N_1 ratio increases the role of sample size re-estimation at stage 2 becomes absolutely necessary in BE assessment.

Efficiency of Each Branch

The performances depicted in Figs. 2 and 3 represent the % acceptances and the % uses of each branch of the two-stage method examined. Combining these two properties into one criterion allows us to investigate the % efficiency of each branch. The term % efficiency expresses the relative ability of the A, B1, and B2 segments of the TSD to declare BE (Fig. 4).

Percent efficiency values for all three branches (A, B1, and B2) decline as GMR increases. In particular, branch A exerts almost excellent % efficiency at low variabilities (Fig. 4a, c, and e). The enlargement of the starting sample size further improves A's efficiency, while it drops to very low levels as within-subject variability increases. However, the inclusion of more subjects at stage 1 counterbalances the effect of high CVw (Fig. 4f). The % efficiency of branch A starts to diminish abruptly when the two drug products differ enough in their mean PK values. The % efficiency of branch B is relatively high at GMR values close to unity, but it is vanishing as GMR and/or within-subject variability increase (Fig. 4). Besides, as N_1 increases the % efficiency of B2 becomes worse, whereas it improves for branch B1.

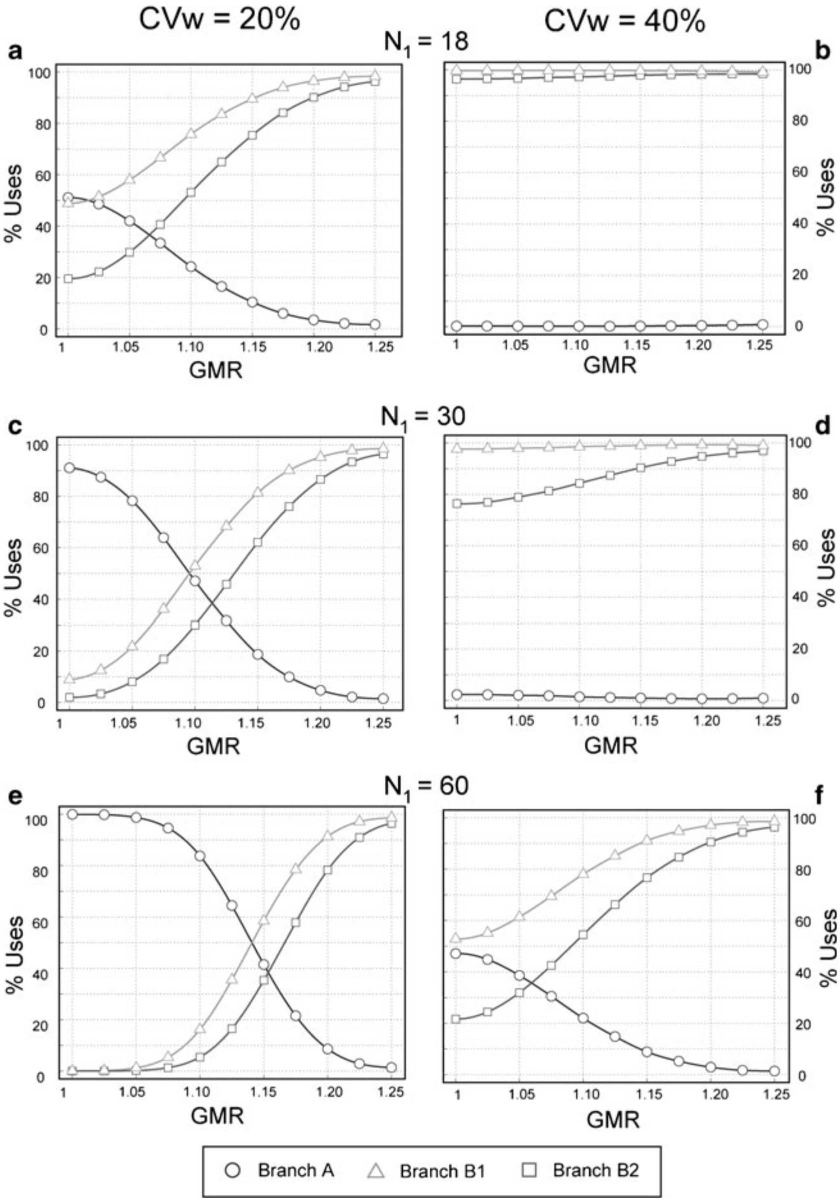


Fig. 3 Percentage of use of each branch (A, B1, B2) of the two-stage design *versus* GMR. The coefficient of variation of the within-subject variability (CVw) was equal to 20% (*left column*) and 40% (*right column*) for the initial population. Three different starting sample sizes, N$_1$, were assumed: 18, 30, and 60.

Percentage of Frequency of Each Failure

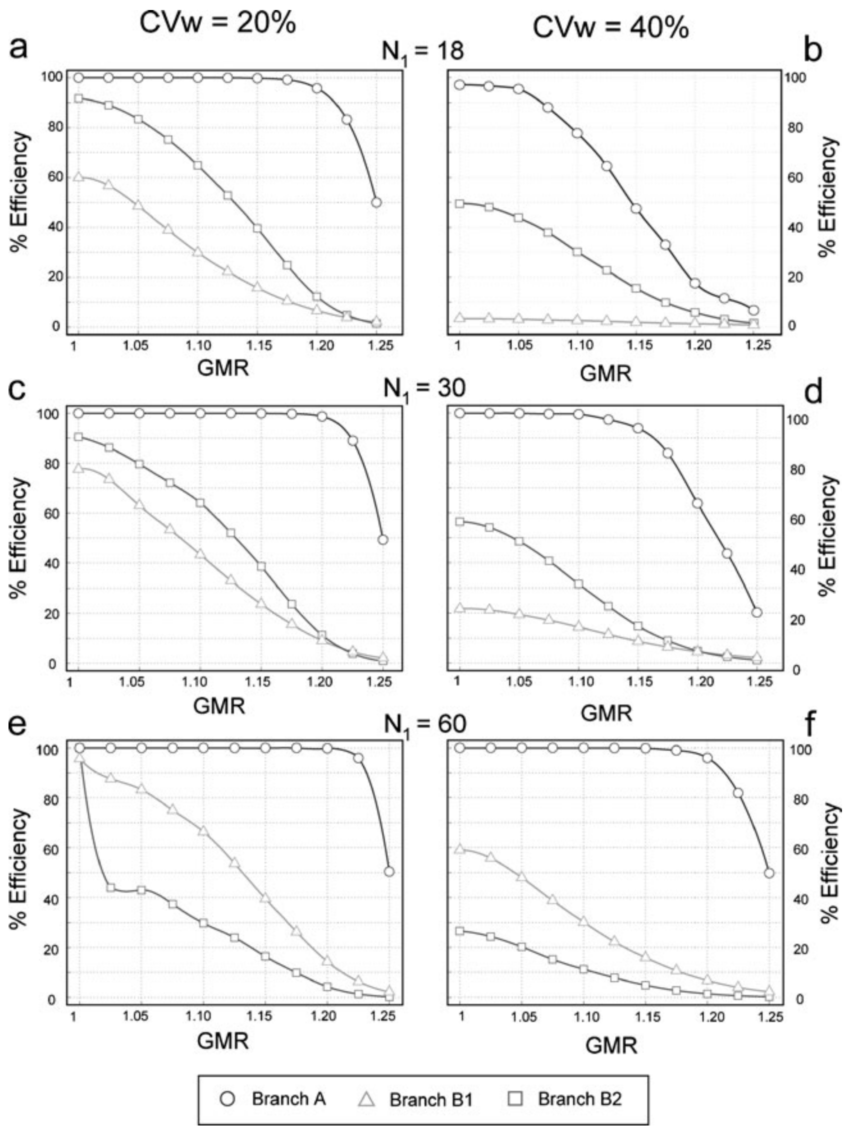
In Fig. 5 six different scenarios are shown which refer to the cases when medium and highly variable drugs are assessed with a low, moderate or a relatively high starting number of subjects. Visual inspection of Fig. 5 reveals that the causes of BE failure obey the following general ranking in terms of decreasing frequency of occurrence: $F_{B1} > F_N > F_{B2} > F_A$. In other words, F_A represents the least common reason of failure, whereas BE failure at B1 represents the most possible outcome at all GMR values. The existence of a too large sample size at stage 2 is a reason that turns out to be more important, as primarily N$_1$, and, to a lesser extent, CVw increase.

DISCUSSION

The aim of this analysis was to explore the underlying properties of the proposed two-stage design (Fig. 1). The TSD explored in this study was found to result in no inflation of type I error beyond 5% (Tables II and III). The underlying reason for this attribute is considered to be two-fold: firstly, the presented TSD includes only two stages and secondly, a UL of 150 was set. The latter apart from rational for use in BE studies was also found as capable of restricting inflation of type I error.

The percentage of simulated studies where BE is accepted *versus* the GMR of the study is shown in Fig. 2. Each branch of the TSD exhibits a different ability to declare BE, which further changes according to variability and sample

Fig. 4 Percentage of efficiency of each branch (A, B1, B2) of the two-stage design *versus* GMR. The coefficient of variation of the within-subject variability (CVw) was equal to 20% (*left column*) and 40% (*right column*) for the initial population. Three different starting sample sizes, N_1, were assumed: 18, 30, and 60.



size (Fig. 2). Among the three parts of TSD, branch A exerts the highest % BE acceptance values either when variability is low or moderate, or an adequately high number of subjects is recruited. The usefulness of branch B2 becomes mainly apparent when highly variable drugs are assessed with a low number of subjects and when the two drug products differ significantly. These two reasons partially reflect the basis for using two-stage designs, namely, in cases when the prior estimates for variability and GMR lead to sample size under-estimation.

An attribute strongly related to the percentages of acceptance is the % uses of each branch of TSD. The visit at each branch of TSD depends on the three key parameters: CVw, N_1, and GMR. Generally speaking, branch A is more frequently accessed in cases where the following factors co-exist: a) variability is low, b) the two drug preparations do not differ significantly, and c) an initial large number of subjects are used. However, when the difference between the two drug products increases or the drug exerts high variability, second stage is used more often (Fig. 3b and d).

Visual inspection and comparison of Figs. 2, 3, and 4 reveals that branch A exhibits a high % efficiency of declaring BE either when is assessed many, or few times. Plausibly, the underlying reason of this behavior relies on the design of the TSD (Fig. 1). According to this design, BE assessment at stage 1 is only allowed when the power of the study is greater than 80%. Thus, only adequately powered studies are assessed at A, which subsequently lead to high % efficiencies. However, as variability increases, the % acceptances of branch A fall to very low levels. Therefore, it appears that the initial 80% power criterion actually imposes the use of branch B when there is *a priori* low probability of declaring BE using the initial group of subjects.

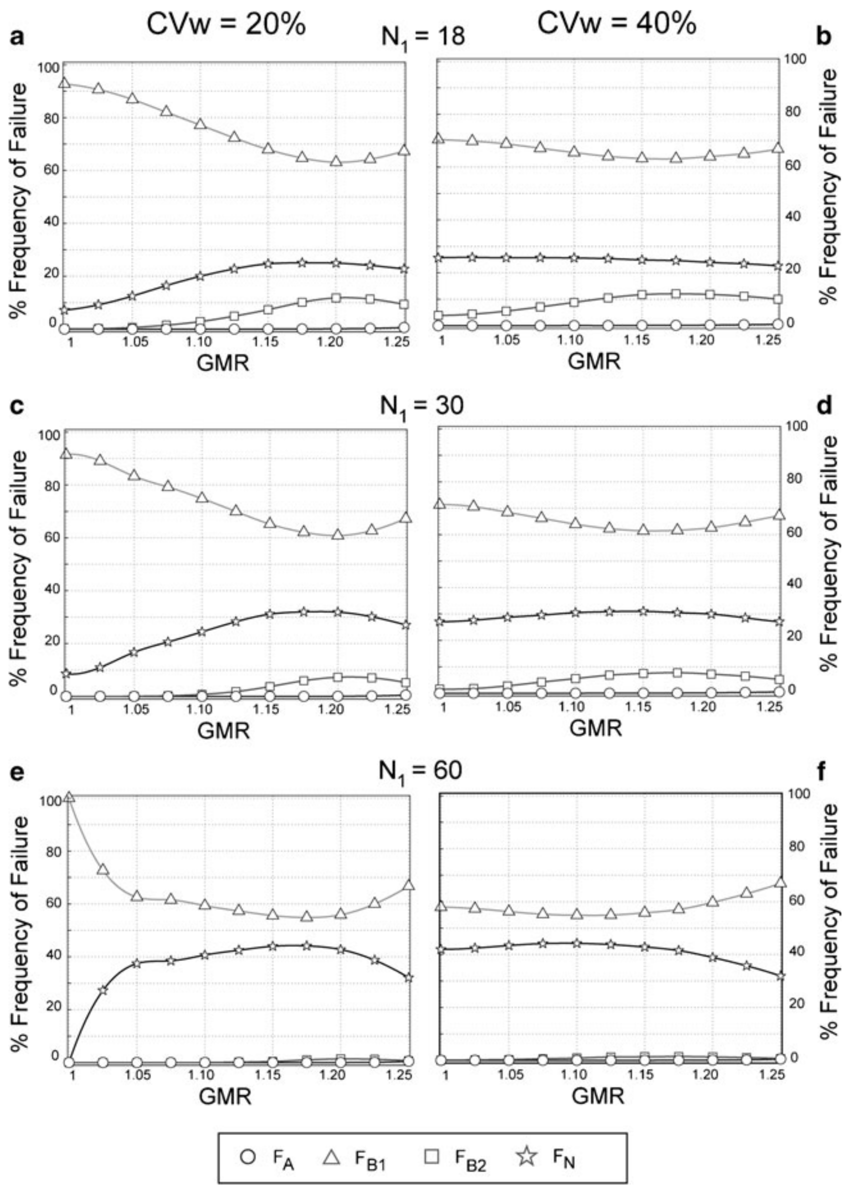


Fig. 5 Frequency (in % values) of each failure to declare bioequivalence *versus* GMR of the study. The coefficient of variation of the within-subject variability (CVw) was equal to 20% (*left column*) and 40% (*right column*) for the initial population. Three different starting sample sizes, N_1, were assumed: 18, 30, and 60. The terms referring to types of failure are listed in Table 1.

It is anticipated that if no UL was set to the re-estimated sample size value, N_2, then the % acceptances and therefore the % efficiency of branch B2 would have been increased. Nevertheless, the absence of a UL seems to be an unrealistic condition, since nobody would plan a BE study allowing the addition of a very large number of subjects at the second stage. In this study, the allowable N_2 number could range from 90 to 132 (namely, N_2=N-N_1 which results in 150−18=132 subjects and 150−60=90, for N_1=18 and N_1=60, respectively). Since these N_2 values refer to the second stage of a BE study and not to a phase III clinical trial, they were considered sufficient for our purposes.

Four types of failures were identified for the two-stage design and were further investigated (Fig. 5). The most common origin of BE failure is encountered at B1. This is a logical finding and refers to the inherent properties of the two-stage approach under study. It should not be disregarded that branch B1 is accessed when the power of the study at stage 1 is less than 80%. In fact, this power estimate is calculated assuming a higher value of type I error than the one applied to branch B1 (5% *versus* 2.94%). Therefore, the high % frequency of F_{B1} failure results from the stricter confidence interval criterion applied to B1 compared to the precedent "power" condition. The second more frequent type of failure is due to the fact that the required number of subjects, after sample size re-estimation, leads to a total sample size that is greater than UL. As more subjects are recruited in the initial phase of the study, the frequency of this failure increases. This finding is actually expected since a UL value was set to the total sample size value.

Finally, it should be mentioned that this analysis also included conditions where CVw was higher than 30%, namely, BE assessment of highly variable drugs. In our case, these studies were only treated in the light of two-stage approaches. Another possibility, for BE assessment, would be the use of reference-scaled BE limits, as it is currently recommended by regulatory authorities (1,42). However, the latter would require a replicate or a semi-replicate design in order to estimate CVw of the reference product. Nevertheless, this type of analysis exceeds the scope of the current study.

CONCLUSION

The aim of this study was to present a two-stage design for BE studies and unveil its properties under several conditions. Basic conclusions derived from our analysis include the following:

a) The TSD under study leads to no inflation of type I error rate beyond 5%. The underlying reason could be ascribed to the inclusion of an upper sample size limit and the fact that only two stages are considered.

b) Each branch of the TSD exhibits different % BE acceptances which are dependent on CVw and N_1. In particular:

  b1) Branch A exerts the highest ability to declare BE either when variability is low to moderate, or an adequately high number of subjects is recruited.

  b2) Second stage becomes mainly useful when highly variable drugs are assessed with a low number of subjects and/or the two drug products differ significantly.

c) BE assessment at branch A is more frequently encountered when: variability is low, the two drugs do not differ significantly, and a large starting number of subjects are used.

d) The % uses of the second stage increases when the two drug products are significantly different or they exert high within-subject variability.

e) BE assessment at branch A exhibits high efficiency to declare BE. On the contrary, branches B1 and B2 are usually less efficient in declaring BE.

f) The initial power criterion of 80% imposes the use of a second stage when there is an *a priori* low probability to conclude BE using the starting group of subjects.

REFERENCES

1. EMA (European Medicines Agency). Committee for medicinal products for human use, CHMP. Guideline on the Investigation of Bioequivalence, London; 2010.
2. FDA (Food and Drug Administration). Center for Drug Evaluation and Research (CDER), bioavailability and bioequivalence studies for orally administered drug products. General Considerations, Rockville, MD; 2003.
3. Efron B. Forcing a sequential experiment to be balanced. Biometrika. 1971;58(3):403–17.
4. Wei LJ. The adaptive biased coin design for sequential experiments. Ann Stat. 1978;6(1):92–100.
5. Lachin JM. Statistical properties of randomization in clinical trials. Control Clin Trials. 1988;9(4):289–311.
6. Demets DL. Group sequential procedures: calendar *versus* information time. Stat Med. 1987;8(10):1191–8.
7. Posch M, Bauer P. Adaptive two-stage designs and the conditional error function. Biom J. 1999;41(6):689–96.
8. Rosenberger WF, Stallard N, Ivanova A, Harper CN, Ricks ML. Optimal adaptive designs for binary response trials. Biometrics. 2001;57(3):909–13.
9. Shih WJ. Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: a comparison. Stat Med. 2006;25(6):933–41.
10. Chung-Stein C, Anderson K, Gallo P, Collins S. Sample size re-estimation: a review and recommendations. Drug Inf J. 2006;40(4):475–84.
11. Lu Q, Tse SK, Chow SC. Analysis of time-to-event data under a two-stage survival adaptive design in clinical trials. J Biopharm Stat. 2010;20(4):705–19.
12. Emerson SS, Fleming TR. Adaptive methods: telling "the rest of the story". J Biopharm Stat. 2010;20(6):1150–65.
13. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. Stat Med. 2011;30(28):3267–84.
14. Pocock SJ. Size of cancer clinical trials and stopping rules. Br J Cancer. 1978;38(6):757–66.
15. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977;64(2):191–9.
16. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilita. Instituto Superiore die Scienze Economiche e Commerciali di Firenze. 1936;8:3–62.
17. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979;35(3):549–56.
18. Lan KG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70(3):659–63.
19. DeMets DL, Hardy R, Friedman LM, Lan KK. Statistical aspects of early termination in the beta-blocker heart attack trial. Control Clin Trials. 1984;5(4):362–72.
20. Srinivasan R, Lanqenbebg P. A two-stage procedure with controlled error probabilities for testing bioequivalence. Biom J. 1986;28(7):826–33.
21. Racine-Poon A, Grieve AP, Fluhler H, Smith AF. A two-stage procedure for bioequivalence studies. Biometrics. 1987;43(4):847–56.
22. Gould AL. Group sequential extensions of a standard bioequivalence testing procedure. J Pharmacokinet Biopharm. 1995;23(1):57–86.
23. Hauck WW, Preston PE, Bois FY. A group sequential approach to crossover trials for average bioequivalence. J Biopharm Stat. 1997;7(1):87–96.
24. Bandyopadhyay N, Dragalin V. Implementation of an adaptive group sequential design in a bioequivalence study. Pharm Stat. 2007;6(2):115–22.

25. World Health Organization Expert Committee on Specifications for Pharmaceutical Preparations. 40th report, Annex 7, Regulatory guidance on interchangeability for multisource (generic) pharmaceutical products. WHO Technical Report 937, Geneva; 2006. p. 347–90.
26. National Institute of Public Health of Japan, Division of Drugs. Guideline for bioequivalence studies of generic products; 2012.
27. Health Canada, Ministry of Health, Health Products and Food Branch. Guidance document conduct and analysis of comparative bioavailability studies. 2012.
28. Korean Food and Drug Administration (KFDA). Guidance document for bioequivalence study. 2008.
29. Therapeutic Goods Administration (TGA). Australian regulatory guidelines for prescription medicines Appendix 15: Biopharmaceutic studies. 2004.
30. FDA (Food and Drug Administration). Draft guidance on Dexamethasone - Tobramycin (Rev. June); 2012.
31. Polli J, Cook J, Davit B, Dickinson P, Argenti D, Barbour N, *et al.* Summary workshop report: facilitating oral product development and reducing regulatory burden through novel approaches to assess bioavailability/bioequivalence. AAPS J. 2012;14(3):627–38.
32. Potvin D, DiLiberti C, Hauck W, Parr A, Schuirmann D, Smith R. Sequential design approaches for bioequivalence studies with crossover designs. Pharm Stat. 2008;7(4):245–62.
33. Montague T, Potvin D, DiLiberti C, Hauck W, Parr A, Schuirmann D. Additional results for 'Sequential design approaches for bioequivalence studies with crossover designs'. Pharm Stat. 2012;11(1):8–13.
34. Cui L, Hung J, Wang SJ. Modification of sample size in group sequential clinical trials. Biometrics. 1999;55(3):853–7.
35. FDA (Food and Drug Administration). Center for Drug Evaluation and Research (CDER), statistical approaches to establishing bioequivalence. Rockville, MD; 2001.
36. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987;15(6):657–80.
37. Julious S. Sample sizes for clinical trials with normal data. Stat Med. 2004;23(12):1921–86.
38. Tothfalusi L, Endrenyi L, Midha KK, Rawson MJ, Hubbard JW. Evaluation of the bioequivalence of highly-variable drugs and drug products. Pharm Res. 2001;18(6):728–33.
39. Tothfalusi L, Endrenyi L. Limits for the scaled average bioequivalence of highly variable drugs and drug products. Pharm Res. 2003;20(3):382–9.
40. Karalis V, Symillides M, Macheras P. Novel scaled average bioequivalence limits based on GMR and variability considerations. Pharm Res. 2004;21(10):1933–42.
41. Karalis V, Symillides M, Macheras P. Bioequivalence of highly variable drugs: a comparison of the newly proposed regulatory approaches by FDA and EMA. Pharm Res. 2012;29(4):1066–77.
42. Haidar SH, Davit B, Chen ML, Conner D, Lee L, Li QH, *et al.* Bioequivalence approaches for highly variable drugs and drug products. Pharm Res. 2008;25(1):237–41.