

Author Query Form

Journal BIMJ

Article bimj201900388

Dear Author,

During the copyediting of your manuscript the following queries arose.

Please refer to the query reference callout numbers in the page proofs and respond to each by marking the necessary comments using the PDF annotation tools.

Please remember illegible or unclear comments and corrections may delay publication.

Many thanks for your assistance.

Query No.	Description	Remarks
Q1	Please confirm that forenames/given names (blue) and surnames/family names (vermillion) have been identified correctly.	
Q2	Please check authors' affiliations for correctness.	
Q3	Please check and suggest if the abbreviations <i>RT/TR</i> , <i>GMR</i> , <i>TRTR/RTRT</i> , <i>RSABE</i> , <i>EMA</i> , etc. are to be set in roman throughout the text.	
Q4	The abbreviations HVD and T1E have been set in roman as per style throughout the text. Please check for correctness.	
Q5	As per journal style, the leading zeros have been eliminated in alpha values throughout the text. Please check for correctness.	
Q6	Please provide missing Table 8.	

Please confirm that Funding Information has been identified correctly.

Please confirm that the funding sponsor list below was correctly extracted from your article: that it includes all funders and that the text has been matched to the correct FundRef Registry organization names. If a name was not found in the FundRef registry, it may not be the canonical name form, it may be a program name rather than an organization name, or it may be an organization not yet included in FundRef Registry. If you know of another name form or a parent organization name for a "not found" item on this list below, please share that information.

FundRef Name	FundRef Organization Name
MINECO/FEDER	
Ministerio de Economía y Competitividad	
Generalitat de Catalunya	Generalitat de Catalunya

RESEARCH PAPER

An iterative method to protect the type I error rate in bioequivalence studies under two-stage adaptive 2×2 crossover designs

Eduard Molins¹  | Detlew Labes² | Helmut Schütz³ | Erik Cobo¹ | Jordi Ocaña⁴

¹ Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Catalunya, Spain

² Consultant Pharmacy Services, Berlin, Germany

³ BEBAC, Vienna, Austria

⁴ Department of Genetics, Microbiology and Statistics - Statistics Section, Universitat de Barcelona, Barcelona, Catalunya, Spain

Correspondence

Eduard Molins, Doctoral Program in Statistics and Operational Research, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain.
Email: eduard.molins@upc.edu

Funding information

MINECO/FEDER, Grant/Award Number: MTM2015-64465-C2-1-R; Ministerio de Economía y Competitividad, Grant/Award Number: 2014 SGR 464; Generalitat de Catalunya



This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Bioequivalence studies are the pivotal clinical trials submitted to regulatory agencies to support the marketing applications of generic drug products. Average bioequivalence (ABE) is used to determine whether the mean values for the pharmacokinetic measures determined after administration of the test and reference products are comparable. Two-stage 2×2 crossover adaptive designs (TSDs) are becoming increasingly popular because they allow making assumptions on the clinically meaningful treatment effect and a reliable guess for the unknown within-subject variability. At an interim look, if ABE is not declared with an initial sample size, they allow to increase it depending on the estimated variability and to enroll additional subjects at a second stage, or to stop for futility in case of poor likelihood of bioequivalence. This is crucial because both parameters must clearly be prespecified in protocols, and the strategy agreed with regulatory agencies in advance with emphasis on controlling the overall type I error.

We present an iterative method to adjust the significance levels at each stage which preserves the overall type I error for a wide set of scenarios which should include the true unknown variability value. Simulations showed adjusted significance levels higher than 0.0300 in most cases with type I error always below 5%, and with a power of at least 80%. TSDs work particularly well for coefficients of variation below 0.3 which are especially useful due to the balance between the power and the percentage of studies proceeding to stage 2. Our approach might support discussions with regulatory agencies.

KEYWORDS

average bioequivalence (ABE), generic drug product, significance level adjustment, two-stage adaptive designs (TSD), type I error control (TIE)

1 | INTRODUCTION

Bioequivalence studies typically involve testing two products, test, T , and reference, R , against each other in a two-period, two-sequence 2×2 crossover RT/TR trial. Primary pharmacokinetic metrics are C_{max} (maximum observed plasma concentration) and the area under the concentration time curve, AUC_{0-t} and $AUC_{0-\infty}$ (EMA, 2010a; FDA, 2014).

To test for average bioequivalence (ABE), the null hypothesis of bioinequivalence is tested against an alternative of bioequivalence, as follows:

$$H_0 : \phi \leq -\delta \text{ or } \phi \geq +\delta$$

$$H_1 : -\delta < \phi < +\delta$$

Based on the “interval inclusion rule,” to declare ABE (i.e., to reject the null hypothesis of bioinequivalence) at a significance level $\alpha = .05$, based on a normal \ln -linear model, the two-sided $1 - 2\alpha = 0.9$ symmetric confidence intervals for $\mu_T - \mu_R$, ϕ , should lie fully within the constant ABE limits of ± 0.223 , or equivalently, the back exponentially transformed confidence interval for the geometric mean ratio, should $GMR = e^\phi$ lie fully within 0.80 to 1.25 ($= 1/0.80$) (EMA, 2010a; Schütz, 2015).

Regulatory agencies usually accept conducting studies based on RT/TR two-stage adaptive 2×2 crossover designs (TSD) (Bandyopadhyay & Dragalin, 2007; EMA, 2010a, 2015; FDA, 2018; Health Canada, 2018; Schütz, 2015), whose application is becoming increasingly popular (Mistry, Dunn, & Marshall, 2017). TSDs allow declaring ABE at an interim look (or stage 1) with a small number of N_1 subjects; and if ABE is not met due to insufficient power, the sample size can be increased in a stage 2 based on the estimation of the within-subject variability, calculated by means of the pooled coefficient of variation of R and T , considering $CV_W = \sqrt{e^{\sigma_W^2} - 1}$ where σ_W^2 is the estimated value of the residual variance obtained from an ANOVA model on \ln -transformed data. Then, ABE is tested again at stage 2 with cumulated $N = N_1 + N_2$ sample size.

Also, TSDs provide investigators with an attractive solution to address some of the uncertainty that exists when the trial is originally designed (Coffey et al., 2012), allowing stopping the study at stage 1 with a small N_1 , avoiding to unnecessarily soar N above what is reasonable to attain a desired power, for example, 80%. And they are especially useful in case of drugs with little evidence about the true within-subject variability, and for highly variable drugs (HVD), that is, with a $CV_W \geq 0.3$ (Knahl, Lang, Fleischer, & Kieser, 2018; Molins, Cobo, & Ocaña, 2017). This discussion is important because the precise model for analysis must be prespecified in the protocol including the sources of variation that reasonably influence primary metrics (FDA, 2018; Potvin, DiLiberti, Hauck, Parr, & Schuirman, 2008). However, little guidance exists yet on how investigators should proceed when designing and planning an adaptive clinical trial (Thorlund, Haggstrom, Park, & Mills, 2018).

The critical point about using TSDs is the difficulty to preserve the type I error rate (T1E) (EMA, 2010a; Fuglsang, 2011; Kieser & Rauch, 2015; Maurer, Jones, & Chen, 2018). Significance level boundaries can be adjusted in various ways that are not fully specified in the regulations (EMA, 2010a). Using an a priori fixed sample size split at equal sequential groups, Pocock (1977) decision to stop the trial or continue was based on repeated significance tests of the accumulated data after each group was evaluated. Based on Pocock’s method but using sample size reassessment, that is, TSDs, Potvin et al. (2008) and Montague et al. (2012) proposed two methods to control the overall T1E rate: Their “Type 1” method consists on using the same adjusted significance level at stages 1 and 2, that is, $\alpha_{adj} = \alpha_1 = \alpha_2$; and “Type 2” method consists on using an unadjusted $\alpha = .05$ in the stage 1, if the interim power is of at least of 80% at stage 1, or else an adjusted α_1 and α_2 at stages 1 and 2, respectively.

Using simulations, Xu et al. (2016) implemented two methods (called E and F) to find optimal solutions of α_1, α_2, N_1 (and a futility parameter) by means of average cost functions of GMR and CV_W combination values. They presented optimal solutions for CV_W ranging from 10% to 30%, and for 30–55%. Maurer et al. (2018) used a principled approach using a standard inverse-normal p -value combination test, in conjunction with standard group sequential techniques (called it maximum combination test) to guarantee the control of T1E rate.

We present an iterative method, which is based on simulations, to adjust the significance levels at each stage, α_1 and α_2 , which preserves the overall T1E (usually at 5%) for a wide set of scenarios which should include the true unknown variability value. In addition, we propose an extended feature by allowing α_1 being different than α_2 . This method has been implemented in an R package called *betsd*, which is hosted on *GitHub*, which includes the function “t1e.tsd” to help to calculate both significance levels.

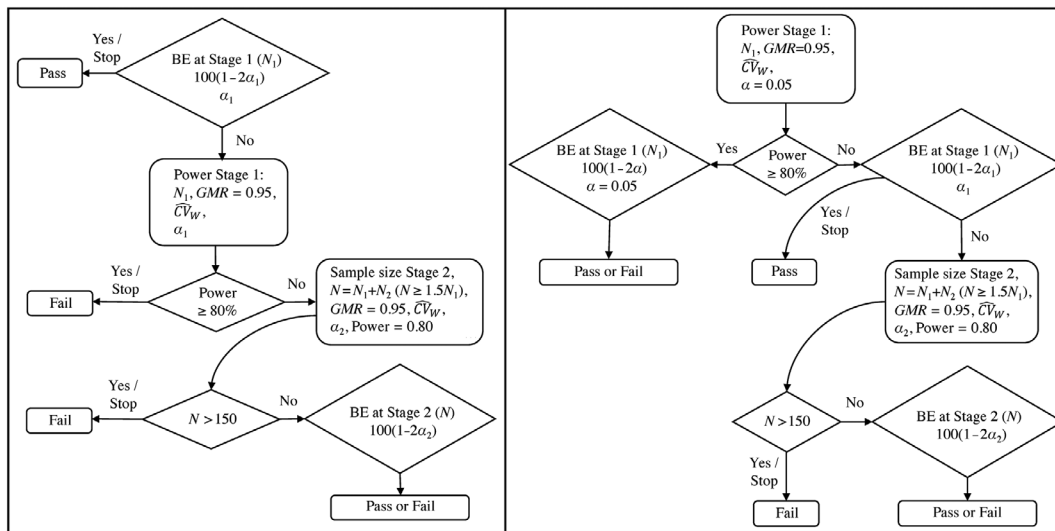


FIGURE 1 Testing ABE using TSD by means of type 1 (on the left) and type 2 (on the right) methodologies, with significance levels α_1 and α_2 at each stage

Note. Adapted from the figure depicted in detail by Montague et al. (2012) with the restriction of Karalis and Macheras (2014) of not including more than N_{max} subjects (150 by default), and $\min(N_2)$ ($N_2 \geq 0.5N_1$ by default); α_1 and α_2 , adjusted significance levels at stages 1 and 2 (α_1 may be different than α_2); TSD, two-stage adaptive designs; ABE, average bioequivalence; N_1 , initial fixed sample size; N_2 , additional number of subjects recruited at stage 2; GMR , geometric mean ratio; \widehat{CV}_w , simulation-based estimated within-subject coefficient of variation at stage 1

In Section 2, we present the methodology to obtain the adjusted significance levels using simulated samples. Then, we present the simulation results where we provide comparisons of our method with the most recent articles released by Xu et al. (2016), and Maurer et al. (2018), and we finalize with a discussion.

2 | METHODOLOGY TO OBTAIN THE ADJUSTED SIGNIFICANCE LEVELS

Figure 1 shows two algorithms to test ABE using TSD by means of the type 1 and 2 methodologies. They include two constraints, first on the minimum sample size at stage 2 of at least $N_2 \geq 0.5N_1$, and second, as previously discussed in Molins et al. (2017), Xu et al. (2016), and Karalis and Macheras (2014), with a futility criterion to stop the study at stage 1 based on a total study size upper limit, in our case of 150 subjects maximum. In contrast to the algorithms proposed by Potvin et al. (2008), we allow α_1 and α_2 being different.

Figure 2 shows the iterative method used to find an optimal significance level adjustment at stages 1 and 2, $\alpha_{adj} = (\alpha_1, \alpha_2)$, granting a global significance level below α (usually $\alpha = 5\%$).

These are the main inputs provided to the algorithm to obtain the adjusted α_1 and α_2 :

1. Arbitrary starting initial significance levels at each stage, for example, $(\alpha_1, \alpha_2) = (.0294, .0294)$ at stages 1 and 2, respectively (based on Potvin et al., 2008, constant).
2. An initial fixed sample size N_1 . A minimum of 12 subjects are required (European Generic Medicines Association, 2010; FDA, 2014).
3. A meaningful set of CV_w values trying to cover the true unknown variability value, a scalar or vector (larger set in case of higher uncertainty), for example, $CV_w = 0.2$.
4. An expected GMR for power calculation, for example, 0.95.
5. A true GMR for type I error assessment, let us say θ_0 , fixed at 1.25.
6. Type 1 or type 2 methodology (as shown in Figure 1).
7. A global significance level, for example, $\alpha = .05$.

By means of a “current” arbitrary significance level $\alpha_{adj} = (.0294, .0294)$ at stages 1 and 2, respectively, Figure 2 shows the algorithm which starts with a warm up period assessing the empiric TIE with 30,000 simulations for each test point

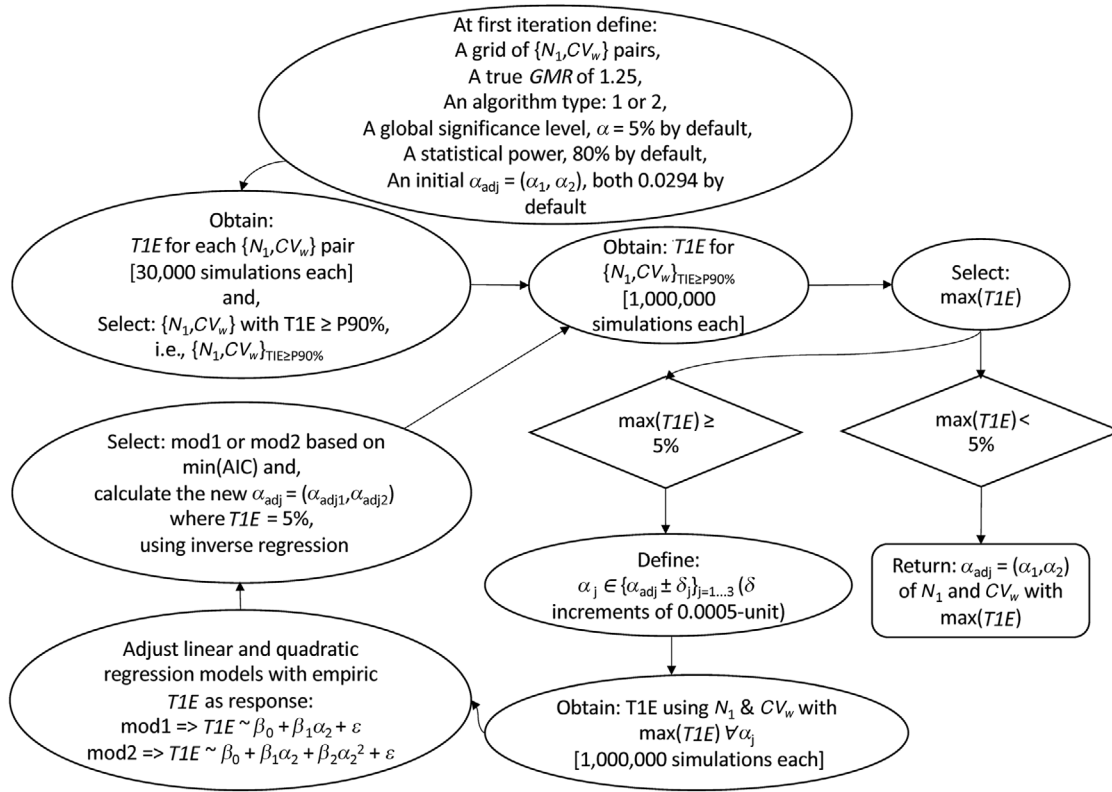


FIGURE 2 Iterative method to obtain adjusted α_1 and α_2 at each stage to grant a global TIE below α

Note. α , desired global significance level (be default, 5%); $\alpha_{\text{adj}} = (\alpha_1, \alpha_2)$ adjusted significance levels at each stage; N_1 , sample size at stage 1; CV_W , within-subject coefficient of variation; TIE, Type I error rate, assessed by means of R function “power.tsd” of Labes and Schütz (2016) and following Figure 1; P90%, percentile 90% of TIE; AIC, Akaike information criterion

at a grid of predefined $\{N_1, CV_W\}$ combinations (corresponding to $N_1 \times CV_W$ Cartesian product), and selecting those pairs exceeding the percentile 90%. For more accuracy, simulations are repeated for this subgroup 1,000,000 times each. The N_1 and CV_W pair with the maximum empiric TIE is selected. Six new significance levels of adjustment are then defined at ± 0.0005 distance from the “current” significance level, and the empiric TIE rate is assessed for each one (with 1,000,000 simulations each time), using the previous maximum N_1 and CV_W pair. This is the base to find the new adjusted significance level, $\alpha_{\text{adj}} = (\alpha_1, \alpha_2)$. To do so, regression models are adjusted with “empiric TIE” as response and “significant level” as covariate (linear and quadratic) as shown in the Figure 2. The model with the minimum Akaike Information Criterion (AIC) is selected, and the “adjusted” $\alpha_{\text{adj}} = (\alpha_1, \alpha_2)$ is established by isolating α_2 using the estimated parameters at a fixed TIE equals to 0.05. In summary, we output the solution if the “current” significance level protects the TIE below α , otherwise, the algorithm starts again from the beginning with the assignment of “current” = “adjusted” significance level.

To obtain the adjusted significance levels to preserve the overall TIE below α , simulations were performed with a true effect ratio θ_0 of 1.25 (i.e., considering the null hypothesis of bioequivalence true), where the treatment effect is just on the ABE frontier so that the likelihood of leading to a false positive result is highest. Under θ_0 equals to 1.25, we used an expected GMR at 0.95 to test for ABE following both TDSs algorithms shown in Figure 1. Once the adjusted significance levels were obtained and fixed, we conducted new simulations with θ_0 and GMR at 0.95 to predict the power at first stage and overall (stage 1 plus stage 2), the percentage of studies switching to stage 2, and the percentiles 5, 50, and 95 of $N = N_1 + N_2$ subjects.

Parameters N_1 and CV_W can be scalars or vectors. If they are vectors, for example, $N_1 = (12, 24)$ and $CV_W = (0.1, 0.15, 0.2, 0.25)$, then the (N_1, CV_W) combination of all $\{N_1, CV_W\}$ combinations with maximum TIE is selected for α_{adj} adjustment (see Figure 2). In addition, we propose an extended feature by allowing α_1 being different than α_2 . When this occurs, α_1 is considered fixed, and the adjustment is only based on α_2 . Since the true CV_W is unknown at the time that the simulations are conducted (before the study starts), and to avoid imprecise specifications for simulations based on tight ranges of CV_W

TABLE 1 Adjusted α_1 and α_2 in both stages preserving the overall TIE below 5%

N_1	CV_W LB – UB	Adjusted $\alpha_1 = \alpha_2$	TIE	% power Stg. 1	% to Stg. 2	% overall Power	P: 5, 50, 95
Type 1 methodology							
12	0.10–0.19	0.0299	0.046063	47.93	49.08	85.96	12, 12, 36
12	0.20–0.29	0.0307	0.049771	15.49	83.91	80.85	12, 34, 64
12	0.30–0.39	0.0303	0.044972	7.02	92.74	78.63	12, 44, 84
12	0.40–0.49	0.0377	0.044389	1.60	96.30	73.56	24, 66, 124
24	0.10–0.19	0.0381	0.039430	89.59	2.85	91.95	24, 24, 24
24	0.20–0.29	0.0306	0.048095	50.95	47.67	84.87	24, 24, 60
24	0.30–0.39	0.0302	0.049831	29.86	69.90	82.63	24, 50, 84
24	0.40–0.49	0.0306	0.045264	10.55	89.01	79.98	24, 76, 118
Type 2 methodology							
12	0.10–0.19	0.0280	0.049858	55.12	39.34	86.54	12, 12, 34
12	0.20–0.29	0.0280	0.049787	35.58	61.06	84.10	12, 22, 44
12	0.30–0.39	0.0295	0.044164	6.88	92.57	78.61	12, 44, 84
12	0.40–0.49	0.0377	0.044501	1.61	96.27	73.68	25, 66, 124
24	0.10–0.19	0.0314	0.049608	96.08	0.23	96.28	24, 24, 24
24	0.20–0.29	0.0301	0.049985	46.96	50.66	83.94	24, 36, 66
24	0.30–0.39	0.0303	0.049815	26.47	72.98	82.13	24, 54, 88
24	0.40–0.49	0.0306	0.044950	10.56	88.95	79.99	24, 76, 118

Note. Burn-in α_1 and α_2 values were initially set at .0294; N_1 , Initial fixed sample size; CV_W LB–UB, lower and upper bound (± 0.05) range of the within-subject coefficient of variation, analyzed at increments of 0.01 units; Adjusted $\alpha_1 = \alpha_2$, same adjusted significance levels at stages 1 and 2; TIE, empirical type I error; % power Stg. 1, power at stage 1; % to Stg. 2, percentage of studies which switch to stage 2; % overall power, overall power; P: 5, 50, 95, percentiles 5, 50, and 95 of $N = N_1 + N_2$.

(or a vague idea about the true/unknown CV_W) our methodology controls the TIE considering CV_W below and upper 0.05 from the values specified/considered.

By means of the function “power.tsd” included in the R package “Power2Stage,” developed by Labes and Schütz (2016), and hosted on CRAN, we developed an open R package called “betsd,” and hosted on GitHub <https://github.com/eduard-molins/betsd> to allow traceability of all versions. This package includes an accurate description of all functionalities of the “tle.tsd” function which serves to calculate the adjusted significance levels at stages 1 and 2. This function implements both methodologies shown in Figure 1, including the modifications proposed in Molins et al. (2017). Also, source code to reproduce the results is available as Supporting Information. In order to allow reproducibility of simulations, we used seed number 1234567.

In turn, this package follows the EMA Questions & Answers document (EMA, 2015), so that, in stage 1, the terms used in the ANOVA model are sequence, subject within sequence, period, and formulation. Fixed effects, rather than random effects, are used for all terms. In stage 2, the adjusted ANOVA model includes sequence, stage, sequence \times stage, subject within sequence \times stage, period within stage, and formulation. Note that models do not include carryover effects or treatment-by-period interactions.

3 | SIMULATION RESULTS

Using simulated samples, we found the adjusted significance levels when α_1 equals α_2 , with TIE rates always strictly below 5%. We assumed some credible scenarios for CV_W and N_1 . Table 1 shows the results for 16 scenarios corresponding to a preplanned fixed initial sample size N_1 of 12 and 24, and a priori true intrasubject CV_W in the following ranges: from 0.10 to 0.19 (a vector of discrete values analyzed at intervals of 0.01-units, i.e., 0.10, 0.11, 0.12, & 0.19), from 0.20 to 0.29, from 0.30 to 0.39, and from 0.40 to 0.49. We found the adjusted significance levels, TIE, % power at stage 1, % of studies jumping to stage 2, % overall power, and percentiles 5, 50 and 95 of N . 10E6 simulations were conducted per scenario.

TABLE 2 Type 1 method to adjust α_2 for a fixed α_1 preserving the overall T1E below 5%

N_1	CV_W LB-UB	Adjusted α_2	T1E	% power Stg. 1	% to Stg. 2	% overall Power	P: 5, 50, 95
$\alpha_1 = .0294 < \alpha_2$							
12	0.20–0.29	0.0310	0.049891	17.92	81.31	81.45	12, 32, 58
24	0.20–0.29	0.0318	0.048936	45.60	53.35	84.28	24, 36, 64
$\alpha_1 = .0320 > \alpha_2$							
12	0.20–0.29	0.0279	0.049767	27.96	71.02	83.25	12, 26, 52
24	0.20–0.29	0.0285	0.048875	47.88	51.30	84.54	24, 36, 66

Note. Burn-in α_2 value was set at .0300 for $\alpha_1 = .0294$, and at .0294 for $\alpha_1 = .0320$; N_1 , Initial fixed sample size; CV_W LB-UB, lower and upper bound (± 0.05) range of the within-subject coefficient of variation, analyzed at increments of 0.01 units; Adjusted α_2 , adjusted significance level at stage 2; T1E, empirical type I error; % power Stg. 1, power at stage 1; % to Stg. 2, percentage of studies which switch to stage 2; % overall power, overall power; P: 5, 50, 95, percentiles 5, 50, and 95 of $N = N_1 + N_2$.

Under the type 1 method, when N_1 equals 12, and considering CV_W from 0.1 to 0.19, the significance levels were adjusted at 0.0299 in both stages. This scenario provided 86% of power, with a likelihood of 49% of stepping up to stage 2, and with a percentile 95 of N equals to 36. When using the type 2 method, the adjusted significance levels were 0.0280, the power was 87%, and the likelihood of switching to stage 2 was 39% (bioequivalence was claimed at stage 1 frequently).

In all scenarios, significance levels were adjusted in at least 0.0299 and 0.0280 for type 1 and 2 methodologies, respectively, and bioequivalence met with a power of at least 80%, except for $N_1 = 12$ and 24 and true CV_W between 0.3 and 0.49, where the power was below 80% (and at stage 1 below 10%), and the likelihood of proceeding to stage 2 higher than 90%. In all cases, as CV_W increased, power at stage 1 decreased and the percentage of studies proceeding to stage 2 increased.

In Table 2, we found the adjusted α_2 , T1E, % power at stage 1, % of studies jumping to stage 2, % overall power, and percentiles 5, 50, and 95 of N , for four scenarios with initial sample sizes N_1 of 12 and 24, a priori assumption on the true intrasubject CV_W ranging from 0.20 to 0.29 (at intervals of 0.01-units) and given a fixed a priori α_1 . 10E6 simulations were conducted per scenario. Results of T1E rates were always below 5%. We considered both possibilities, to be more permissive at stage 1 with $\alpha_1 \leq \alpha_2$, or at stage 2 with $\alpha_1 \geq \alpha_2$. We can compare these results to the ones obtained in the Table 1 where $\alpha_1 = \alpha_2$.

For N_1 equals to 12, and a fixed $\alpha_1 = .0294 < \alpha_2$, the significance level at stage 2 was adjusted at 0.0310. These results contrast with the ones obtained in Table 1 with $\alpha_1 = \alpha_2 = .0307$, being the test less permissive at stage 1 and more permissive at the stage 2. In addition, a power of 81% was reached, with a likelihood of 81% of stepping up to stage 2, and with a percentile 95 of N equals to 58 subjects. Similarly, for N_1 equals to 12, and when $\alpha_1 = .0320 < \alpha_2$, α_2 was adjusted at .0279. This test is more permissive at stage 1 and less permissive at stage 2.

Given adjusted significance levels, Table 3 shows the empiric T1E rate and power for CV_W at 0.05 above and below the upper and lower CV_W bounds using the type 1 and 2 methodologies. Type 1 error and % overall power were calculated by means of 10E6 and 10E5 simulations per scenario, respectively. We can see that T1E never exceed the 5% global significance level and the power was around 80% or higher, except for CV_W values of 0.54 affected by the constraint of $\max(N = N_1 + N_2) = 150$.

Based on our method, protocols for ABE must include an initial N_1 , a method type (1 or 2) with constraint $\max(N = N_1 + N_2) = 150$ (if $N > 150$, ABE fails), and $N_2 \geq N_1/2$, a target power, and the significance levels to use, obtained by means of the function *tle.tsd()*. Figure 3 shows power contour plots, considering N_1 set to 12, the type 1 method, and a target power of 0.8. True unknown CV_W values range from 0.10 to 0.49 (y-axis), and true unknown GMR s between 0.80 and 1.25 (x-axis, extremes not included), both at increments of 0.05. Significance levels were taken from Tables 1 and 2: $\alpha_1 = \alpha_2 = .0299$; $\alpha_1 = \alpha_2 = .0307$; $\alpha_1 = .0294$ $\alpha_2 = .0310$; $\alpha_1 = .0320$ $\alpha_2 = .0279$. We tested 1,760 scenarios per graph (40 $CV_W \times 44$ GMR s) using the function *power.tsd()* with 10E5 simulations for scenario. We can see in all graphs that the constraint of a maximum of 150 subjects provokes a power decrease of at least 70% for CV_W values above 40%.

Xu et al. (2016) obtained α_1 , α_2 , N_1 , and a futility criterion (f) by means of average cost functions for GMR and CV_W combination values at increments of 5%. They varied (and fixed) the two significance levels α_1 and α_2 , N_1 , and a futility criterion (f), and checked whether the power was of at least of 80% (at a true GMR of 0.95) and the T1E rate (at a true GMR of 0.8 of bioequivalence) controlled for each GMR and CV_W combination value. They obtained optimal designs based on the lowest cost among valid combinations of α_1 , α_2 , N_1 , and f . We obtained type 1 and 2 α_1 and α_2 using the function *tle.tsd()*

TABLE 3 Empiric type 1 error and power for CV_W at 0.05 below and above LB and UB

N_1	CV_W LB – UB	Adjusted $\alpha_1 = \alpha_2$	Type 1 error		% overall power	
			CV_W LB – 0.05	CV_W UB + 0.05	CV_W LB – 0.05	CV_W UB + 0.05
Type 1 methodology						
12	0.10–0.19	0.0299	0.0299	0.0498	99.99	82.07
12	0.20–0.29	0.0307	0.0379	0.0411	90.09	76.61
12	0.30–0.39	0.0303	0.0498	0.0314	81.63	65.86
12	0.40–0.49	0.0377	0.0499	0.0297	77.01	48.68
24	0.10–0.19	0.0381	0.0378	0.0499	99.99	87.84
24	0.20–0.29	0.0306	0.0304	0.0499	97.48	82.14
24	0.30–0.39	0.0302	0.0436	0.0390	86.70	76.40
24	0.40–0.49	0.0306	0.0497	0.0253	81.97	52.04
Type 2 methodology						
12	0.10–0.19	0.0280	0.0499	0.0485	99.99	81.88
12	0.20–0.29	0.0280	0.0498	0.0370	90.29	76.17
12	0.30–0.39	0.0295	0.0499	0.0305	81.38	65.36
12	0.40–0.49	0.0377	0.0499	0.0297	76.91	48.92
24	0.10–0.19	0.0314	0.0496	0.0499	99.99	86.73
24	0.20–0.29	0.0301	0.0496	0.0498	98.63	82.18
24	0.30–0.39	0.0303	0.0492	0.0393	86.00	76.57
24	0.40–0.49	0.0306	0.0499	0.0254	81.75	52.04

Note. N_1 , initial fixed sample size; CV_W LB–UB, lower and upper bound values of the within-subject coefficient of variation; Adjusted $\alpha_1 = \alpha_2$, same adjusted significance levels at stages 1 and 2; Type 1 error, empirical type 1 error; % overall power, overall power.

TABLE 4 Xu et al. optimal TSD designs of methods E and F and our methodology (type 1 and 2 methods)

	CV_W range: 0.10–0.30 $N_1 = 18$	CV_W range: 0.30–0.55 $N_1 = 48$
Method E (Xu et al.)	$\alpha_1: .0249\alpha_2: .0363f: 93.74 - 106.67$	$\alpha_1: .0254\alpha_2: .0357f: 93.05 - 107.47$
Method F (Xu et al.)	$\alpha_1: .0248\alpha_2: .0364f: 94.92 - 105.35$	$\alpha_1: .0259\alpha_2: .0349f: 93.50 - 106.95$
Type 1 method	$\alpha_1 = \alpha_2 = .0303$	$\alpha_1 = \alpha_2 = .0305$
Type 2 method	$\alpha_1 = \alpha_2 = .0331$	$\alpha_1 = \alpha_2 = .0331$

Note. Type 1 and 2 based on N maximum of 150 subjects and $N_2 \geq 0.5N_1$. CV_W values were analyzed at increments of 0.05.

based on the N_1 and CV_W obtained by Xu et al. (Table 4). Due to design similarities, type 1 method (modified Potvin B) can be compared with Xu et al. Method E, and type 2 (modified Potvin C) to compare with Method F.

We used the *power.tsd()* function with 10E6 simulations per N_1 and CV_W pair with target power 80% and planned and true *GMR* 0.95 to calculate percentiles of $N = (N_1 + N_2)$ 5th, 50th, 95th, and % of studies in stage 2. Table 5 shows results which are comparable between type 1 and Method E, and type 2 and Method F. A power close to 80% was always obtained except for CV_W of 0.55 where maximum target of 150 subjects was reached (data not shown).

Maurer et al. (2018) used a standard inverse-normal *p*-value combination test, in conjunction with standard group sequential techniques (called it maximum combination test), to guarantee the control of TIE rate at any given significance level. The sample size N_2 at the second stage was based on comparing a “target conditional power,” with the power achieved at first stage, versus a “conditional power,” with the conditional errors for maximum combination test (using the CV_W estimation at interim), a formulation effect of 0.95, and N_2 . Starting on an initial N_2 set to 4, the “conditional power” was assessed at increments of two subjects until it exceeded the “target conditional power.”

Table 6 shows the power and sample size of different methods for HVD. Results from Potvin et al. (2008) ($\alpha_1 = \alpha_2 = .0294$) and Maurer et al. (2018) ($\alpha_1 = \alpha_2 = .0263$ for maximum combination test with (w, w^*): (0.5, 0.25)) were taken from Maurer et al. (2018) manuscript (table 8). Type 1 significance levels were obtained using the function *tle.tsd()*, considering $N_1 = (12, 24, 36)$, CV_W between 0.4 and 0.8 at increments of 0.01; and constraints $N \leq 4,000$ and $N_2 \geq 0.5N_1$. The result was $\alpha_1 = \alpha_2 = .0302$. Then, we used the *power.tsd()* function with 10E6 simulations per N_1 and CV_W pair with target power

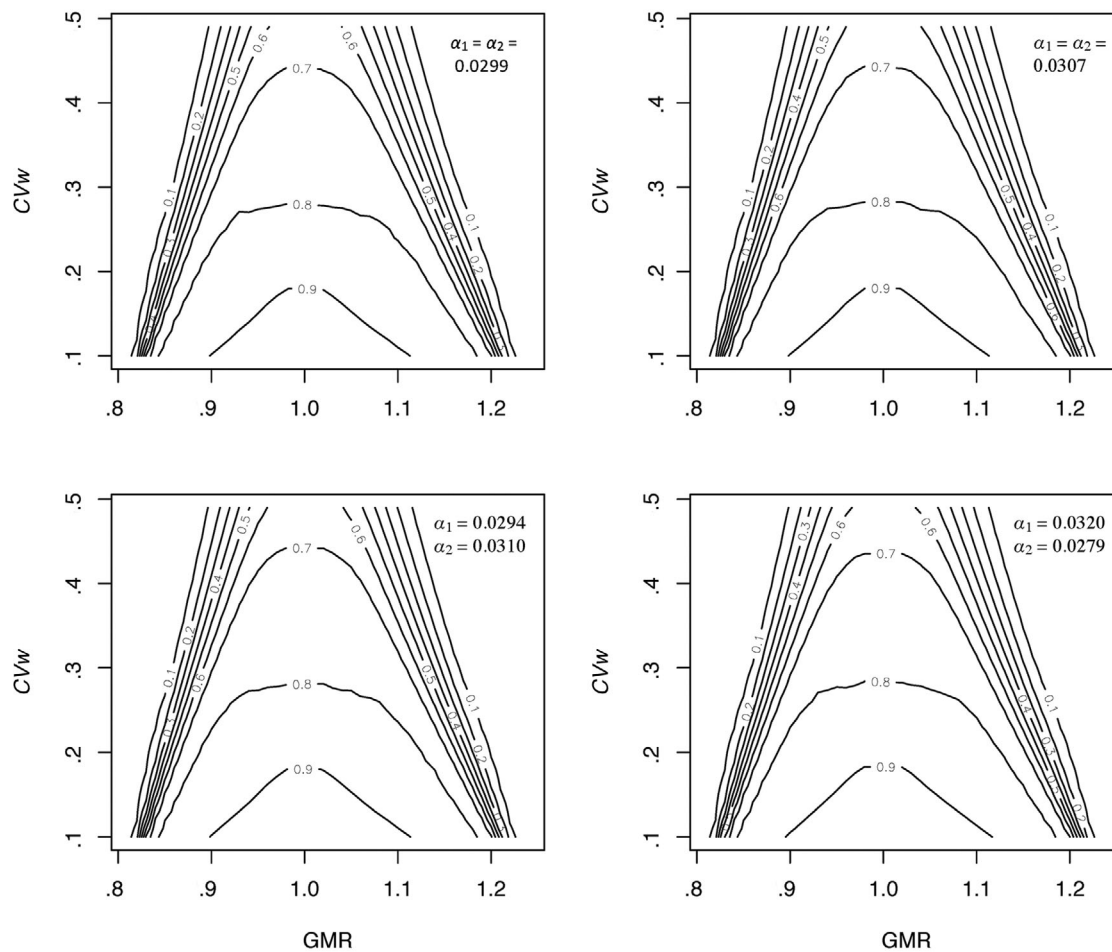


FIGURE 3 Power assessment based on true GMR and CV_W with $N_1 = 12$ and type 1 methodology
 Note. All combinations of GMR between 0.80 and 1.25 (extremes not included), and CV_W between 0.10 and 0.49, both defined as vectors of discrete values at intervals of 0.01-units, resulted on 1,760 scenarios which were simulated $10E5$ times each

TABLE 5 Percentiles of N (5th, 50th, 95th) and % of studies in stage 2

CV_W LB- UB, and N_1	CV_W	Xu et al.		Our method	
		Method E	Method F	Type 1 method	Type 2 method
0.10–0.30 $N_1 = 18$	0.10	(18,18,18) 0%	(18,18,18) 0%	(18,18,18) 0%	(18,18,18) 0%
	0.15	(18,18,18) 2.4%	(18,18,18) 1.3%	(18,18,18) 2.4%	(18,18,18) 0.9%
	0.20	(18,18,32) 24.1%	(18,18,32) 21.8%	(18,18,34) 24.9%	(18,18,32) 18.5%
	0.25	(18,24,42) 54.2%	(18,24,42) 53.7%	(18,28,54) 54.3%	(18,18,52) 49.5%
0.30–0.55 $N_1 = 48$	0.30	(18,42,42) 75.8%	(18,42,42) 76.9%	(18,44,74) 77.4%	(18,42,72) 74.4%
	0.35	(48,48,52) 7.6%	(48,48,48) 3.6%	(48,48,72) 8.7%	(48,48,48) 3.0%
	0.40	(48,48,74) 28.2%	(48,48,74) 22.8%	(48,48,76) 28.1%	(48,48,74) 20.4%
	0.45	(48,48,98) 46.2%	(48,48,98) 44.0%	(48,48,102) 45.0%	(48,48,98) 41.1%
	0.50	(48,80,124) 61.3%	(48,80,124) 60.5%	(48,80,128) 58.9%	(48,76,124) 56.6%
	0.55	(48,104,150) 74.3%	(48,104,152) 73.6%	(48,100,142) 65.3%	(48,98,140) 64.6%
		(48,128,176) 85.2%	(48,128,180) 84.3%	(48,102,146) 55.5%	(48,102,146) 57.7%

Note. CV_W LB-UB, lower and upper bound values of the within-subject coefficient of variation

Type 1 method: Modified Potvin B method with $\max(N = N_1 + N_2) = 150$, and $N_2 \geq N_1/2$

Type 2 method: Modified Potvin B method with $\max(N = N_1 + N_2) = 150$, and $N_2 \geq N_1/2$

Type 1 method is compared with Method E and type 2 with method F

Target power = 0.80 and planned and true $GMR = 0.95$.

TABLE 6 Power and mean sample size with constraint $N \leq 4,000$ for HVD

N_1	CV_W	Potvin et al.: Method B		Maurer, Jones, and Chen: MCT (w, w^*): (0.5, 0.25)		Type 1 method	
		Power (%)	Mean n	Power (%)	Mean n	Power (%)	Mean n
36	0.40	82	67	81	67	83	67
24	0.60	77	161	80	180	77	159
12	0.80	72	257	76	325	72	255

Note. Type 1 method: Modified Potvin B method with $\max(N = N_1 + N_2) = 4,000$, and $N_2 \geq N_1/2$

MCT, maximum combination test; HVD, highly variable drugs.

Target power = 0.80 and planned and true GMR = 0.95.

80% and planned and true GMR 0.95 to calculate the power achieved and mean N . Results show a power and sample size which are comparable across methods.

4 | DISCUSSION

ABE studies using adaptive designs (TSD) offer several advantages over conventional crossover trials. They provide an attractive solution to address some of the uncertainty that exists on the true variability value when the trial is originally designed, although they are typically more complex and exhaustive and require more efforts and time for planning and implementing (Thorlund et al., 2018). TSDs should be standardized and agreed between the pharmaceutical industries and the agencies, in particular, about the specific pathways to control the T1E rate, usually at 5%. We adapted two methodology types proposed initially by Potvin et al. (2008) to adjust the significance levels at each stage which controls the T1E. Adjusted significance levels were higher than 0.0300 in most cases with a power of at least 80%. We also adapt and compare our approach with Xu et al. (2016) and Maurer et al. (2018) to conclude that operating characteristics are comparable.

Our approach is implemented using our own function. In summary, given a grid of $\{N_1, CV_W\}$ and an initial warm-up α_1 and α_2 values, we found adjusted α_1 and α_2 and the (N_1, CV_W) pair with maximum empiric T1E (Tables 1 and 2). In the grid, we should cover an important range of CV_W values to ensure that the true/population CV_W is included. In Molins et al. (2017), we assessed a particular case assuming that the degree of uncertainty was encompassed by evaluating CV_W at 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, and 0.6. We have now improved this feature sweeping CV_W range values at intervals of 0.01-units. In addition, we have considered the case of an applicant/sponsor who assumes CV_W values which unfortunately do not contain the true unknown CV_W value. So, the T1E is controlled, by default, at an overall significance level considering the CV_W assumed ± 0.05 . We admit that though it is sometimes necessary to cover such a range of CV_W values, there is always a risk of losing some power.

We provide a methodology which usually adjusts significance levels above 0.0294 and strictly controls the T1E. The significance levels of 0.0294 at both stages (Pocock, 1977; Potvin et al., 2008) are not some kind of a “natural constant,” because they depend on the design, treatment effect, variability, target power, or sample size, and so they are entirely empiric and must be estimated in simulations. In addition, they did not always control the overall T1E rate at a maximum 5% (Karalis & Macheras, 2014; Montague et al., 2012). For example, the original “Potvin D” method only grants the maintenance of the T1E rate below 5.2%. And, by using the modified “Potvin C” method, with GMR = 0.95, $\alpha_1 = \alpha_2 = 0.0294$, $N_1 = 12$, and a true $CV_W = 0.2$, the T1E is assessed at 5.3% (5.5% in case of GMR = 0.90).

Other methods to adjust significance levels are discussed by some authors and regulatory instances (EMA, 2010b; FDA, 2018; Fuglsang, 2011; Health Canada, 2018; Kieser & Rauch, 2015; Maurer et al., 2018; Xu et al., 2016). In order to see how some operating characteristics compare to each other, we followed the frameworks (N_1 and CV_W range) used by Xu et al. (2016) and Maurer et al. (2018) to calculate the significance levels. We saw comparable results on the overall sample size, the percentage of studies jumping to stage 2, or the overall power (Tables 5 and 6). We highlight that our method is very flexible because it is customizable in many different ways.

We also allow α_1 and α_2 being different from each other. O’Brien and Fleming (1979) proposed a group sequential procedure with boundary values that decreased over the stages to make early stopping less likely. Xu et al. (2016) also found significance levels where $\alpha_1 < \alpha_2$ with α_1 at first stage close to .025. Adaptive strategies are persuasive because they allow stopping the trial at first stage and declare ABE with a low number of N_1 subjects. However, it will be difficult to declare ABE at first stage if α_1 is very conservative. Though $\alpha_1 < \alpha_2$ seems the most natural way of proceeding, an applicant

may be interested in being more permissive at stage 1, for example, Lan and DeMets α -spending function. We allowed both $\alpha_1 < \alpha_2$: .0294, .0310, and $\alpha_1 > \alpha_2$: .0320, .0279 (Table 2 and Figure 3).

Maurer et al. (2018) provided an attractive principled solution based on a maximum combination test to control the TIE inflation. While simulation-based approaches are criticized because require the investigation of many scenarios (in our case, CV_W range values should be large enough) to ensure the control of this error, this principled method also relies on specifying two weights w and w^* which need to be predefined a priori, and an initial guess on the CV_W . In addition, there is no a simple formula of obtaining the power which is desirable to compare the different settings (N_2 , weights, futility criteria). In analogy to the Potvin et al. (2008) methods, it is needed to undertake simulations to gain those values.

Some other differences between methodologies lie on the specifics of futility rules to stop the trial at first stage. Xu et al. (2016) and Maurer et al. (2018) specified futility rules based on 90% CI of the formulation effect completely outside of some margins. Also, Xu et al. (2016) and Karalis and Macheras (2014), included a futility criterion to stop the study at first stage based on a total study size upper limit. We included an upper limit for N of 150 subjects.

We consider HVD a special case under investigation (Endrenyi & Tothfalusi, 2009; Karalis, Symillides, & Macheras, 2012; Knahl et al., 2018; Muñoz, Alcaide, & Ocaña, 2016; Tothfalusi & Endrenyi, 2011; Tothfalusi, Endrenyi, Midha, Rawson, & Hubbard, 2001). We compared EMA Reference Scaled Average Bioequivalence (RSABE) based on replicate TRTR/RTTR designs and TSD methods (Molins et al., 2017). In terms of power, we saw that both approaches perform similarly despite adaptive methods usually requires a higher mean sample size to reach the same power, especially for clearly HVD. Nevertheless, we demonstrated suitable power at the first stage in some cases. But for true CV_W values above 0.29, the power at first stage is low and the proportion of studies switching to stage 2 high. In addition, assertion of ABE becomes difficult for CV_W greater than 0.5 (data shown in Tables 5 and 6), as ABE seldom can be declared at stage 2. It is arguable launching a drug into the market with such a within-subject variability, or even starting a study with such a low expected power (Fuglsang, 2014).

We calculated CV_W by means of the coefficient of variation under homoskedasticity assumption $CV_{WR} = CV_{WT} = CV_W$. Kang and Vahl (2017) showed that ABE testing with heterogeneous residual variances gives similar performance for CV_W lower than 0.4. In fact, power curves (Figure 3) show that the constraint that we are using of a maximum of 150 subjects provokes a power decrease for CV_W values above 0.4.

In conclusion, TSDs can be applied to bioequivalence studies more widely. We provide a function to adjust the significance levels at each stage which strictly grant the control of the type I error for different assumptions on the GMR, N_1 , and CV_W . With this paper, we would like to contribute toward a global harmonization and convergence of generic drug developments.


ACKNOWLEDGMENTS

This research is partially supported by the grant MTM2015-64465-C2-1-R (MINECO/FEDER) from the Ministerio de Economía y Competitividad (Spain) and by the grant 2014 SGR 464 from the Generalitat de Catalunya.

CONFLICT OF INTEREST

The authors have declared no conflict of interest

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Eduard Molins  <https://orcid.org/0000-0003-3895-6406>

REFERENCES

- Bandyopadhyay, N., & Dragalin, V. (2007). Implementation of an adaptive group sequential design in a bioequivalence study. *Pharmaceutical Statistics*, 6, 115–122.
- Coffey, C. S., Levin, B., Clark, C., Timmerman, C., Wittes, J., Gilbert, P., & Harris, S. (2012). Overview, hurdles, and future work in adaptive designs: Perspectives from a National Institutes of Health-funded workshop. *Clinical Trials*, 9, 671–680.

- Endrenyi, L., & Tothfalusi, L. (2009). Regulatory conditions for the determination of bioequivalence of highly variable drugs. *Journal of Pharmacy and Pharmaceutical Sciences*, *12*, 138–149.
- European Generic Medicines Association. (2010). Revised EMA Bioequivalence Guideline questions & answers. Summary of the discussions held at the 3rd EGA Symposium on Bioequivalence. Retrieved from https://www.medicinesforeurope.com/wp-content/uploads/2016/03/EGA_BEQ_QA_WEB_QA_1_32.pdf
- European Medicines Agency. (2010a). Guideline on the investigation of bioequivalence. CPMP/EWP/QWP/1401/98 Rev. 1. Retrieved from https://www.ema.europa.eu/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf
- European Medicines Agency. (2010b). Overview of comments received on draft guideline on the investigation of bioequivalence. EMA/CHMP/EWP/26817/2010. Retrieved from https://www.ema.europa.eu/en/documents/other/overview-comments-received-draft-guideline-investigation-bioequivalence-cpmp/ewp/qwp/1401/98-rev-1_en.pdf
- European Medicines Agency. (2015). Questions & Answers: Positions on specific questions addressed to the pharmacokinetics working party. EMA/618604/2008 Rev. 13. Retrieved from https://www.ema.europa.eu/en/documents/scientific-guideline/questions-answers-positions-specific-questions-addressed-pharmacokinetics-working-party_en.pdf
- Food and Drug Administration. (2014). Draft Guidance for Industry. Bioavailability and Bioequivalence Studies Submitted in NDAs or INDs - General Considerations. Retrieved from <https://www.fda.gov/media/88254/download>
- Food and Drug Administration. (2018). Draft Guidance for Industry. Adaptive Design Clinical Trials of Drugs and Biologics. Retrieved from <https://www.fda.gov/media/78495/download>
- Fuglsang, A. (2011). Controlling type I errors for two-stage bioequivalence study designs. *Clinical Research and Regulatory Affairs*, *28*, 100–105.
- Fuglsang, A. (2014). Futility rules in bioequivalence trials with sequential designs. *AAPS Journal*, *16*, 79–82.
- Health Canada. Guidance Document. (2018). *Conduct and analysis of comparative bioavailability studies*. Retrieved from <https://www.canada.ca/content/dam/hc-sc/documents/services/drugs-health-products/drug-products/applications-submissions/guidance-documents/bioavailability-bioequivalence/conduct-analysis-comparative.pdf>
- Kang, Q., & Vahl, C. I. (2017). Testing for bioequivalence of highly variable drugs from TR-RT crossover designs with heterogeneous residual variances. *Pharmaceutical Statistics*, *16*, 361–377.
- Karalis, V., & Macheras, P. (2014). On the statistical model of the two-stage designs in bioequivalence assessment. *Journal of Pharmacy and Pharmacology*, *66*, 48–52.
- Karalis, V., Symillides, M., & Macheras, P. (2012). Bioequivalence of highly variable drugs: A comparison of the newly proposed regulatory approaches by FDA and EMA. *Pharmaceutical Research*, *29*, 1066–1077.
- Kieser, M., & Rauch, G. (2015). Two-stage designs for cross-over bioequivalence trials. *Statistics in Medicine*, *34*, 2403–2416.
- Knahl, S. I. E., Lang, B., Fleischer, F., & Kieser, M. (2018). A comparison of group sequential and fixed sample size designs for bioequivalence trials with highly variable drugs. *European Journal of Clinical Pharmacology*, *74*, 549–559.
- Labes, D., & Schütz, H. (2016). Inflation of type I error in the evaluation of scaled average bioequivalence, and a method for its control. *Pharmaceutical Research*, *33*, 2805–2814.
- Maurer, W., Jones, B., & Chen, Y. (2018). Controlling the type I error rate in two-stage sequential designs when testing for average bioequivalence. *Statistics in Medicine*, *37*, 1587–1607.
- Mistry, P., Dunn, J. A., & Marshall, A. (2017). A literature review of applied adaptive design methodology within the field of oncology in randomised controlled trials and a proposed extension to the CONSORT guidelines. *BMC Medical Research Methodology*, *17*, Article No. 108.
- Molins, E., Cobo, E., & Ocaña, J. (2017). Two-stage designs versus European scaled average designs in bioequivalence studies for highly variable drugs: Which to choose? *Statistics in Medicine*, *36*, 4777–4788.
- Montague, T. H., Potvin, D., DiLiberti, C. E., Hauck, W. W., Parr, A. F., & Schuirmann, D. J. (2012). Additional results for “Sequential design approaches for bioequivalence studies with crossover designs”. *Pharmaceutical Statistics*, *11*, 8–13.
- Muñoz, J., Alcaide, D., & Ocaña, J. (2016). Consumer’s risk in the EMA and FDA regulatory approaches for bioequivalence in highly variable drugs. *Statistics in Medicine*, *35*, 1933–1943.
- O’Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*, 549–556.
- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*, 191–199.
- Potvin, D., DiLiberti, C. E., Hauck, W. W., Parr, A. F., & Schuirmann, D. J. (2008). Sequential design approaches for bioequivalence studies with crossover designs. *Pharmaceutical Statistics*, *7*, 245–262.
- Schütz, H. (2015). Two-stage designs in bioequivalence trials. *European Journal of Clinical Pharmacology*, *71*, 271–281.
- Thorlund, K., Haggstrom, J., Park, J. J. H., & Mills, E. J. (2018). Key design considerations for adaptive clinical trials: A primer for clinicians. *British Medical Journal*, *360*, k698.
- Tothfalusi, L., & Endrenyi, L. (2011). Sample sizes for designing bioequivalence studies for highly variable drugs. *Journal of Pharmacy and Pharmaceutical Sciences*, *15*, 73–84.
- Tothfalusi, L., Endrenyi, L., Midha, K. K., Rawson, M. J., & Hubbard, J. W. (2001). Evaluation of the bioequivalence of highly-variable drugs and drug products. *Pharmaceutical Research*, *18*, 728–733.
- Xu, J., Audet, C., DiLiberti, C. E., Hauck, W. W., Montague, T. H., Parr, A. F., & Schuirmann, D. J. (2016). Optimal adaptive sequential designs for crossover bioequivalence studies. *Pharmaceutical Statistics*, *15*, 15–27.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Molins E, Labes D, Schütz H, Cobo E, Ocaña J. An iterative method to protect the type I error rate in bioequivalence studies under two-stage adaptive 2×2 crossover designs. *Biometrical Journal*. 2020;1–12. <https://doi.org/10.1002/bimj.201900388>

UNCORRECTED PROOFS