Scaled Average Bioequivalence or Two-Stage Designs in Bioequivalence Studies for Highly Variable Drugs: Which to Choose?

Journal:	*Statistics in Medicine*
Manuscript ID	SIM-16-0869.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	28-May-2017
Complete List of Authors:	Molins Lleonart, Eduard; Universitat Politecnica de Catalunya, Department of Statistics and Operations Research Cobo, Erik; Technical University of Catalonia, Statistics and Operational Research Ocaña, Jordi; Universitat de Barcelona, Departamento de Estadística
Keywords:	Average Bioequivalence (ABE), Reference Scaled Average Bioequivalence (RSABE), Two-Stage Designs (TSD), Highly Variable Drugs (HVD), Significance Level Adjustment

POINT-BY-POINT RESPONSE TO REVIEWERS

We would like to thank the reviewers for their detailed comments and suggestions for the manuscript. We believe that the comments have identified important areas which required improvement. Below, you will find a point-by-point description of how each comment is addressed in the manuscript. Original reviewer comments are in boldface, responses in regular typeface. The new text is underlined while the crossed-out text refers to the deleted original text. Two appendixes with R code are included at the end of this document to support the answer to some of Reviewer 1's questions.

We have included a statement in the Acknowledgments:

We would like to thank the reviewers who identified areas of the manuscript that needed corrections or modifications.

Reviewer Comments:

Reviewer: 1

Comments to the Author

This is an excellent manuscript and it was a pleasure to read. However, some important issues require clarification.

A. Comments concerning RSABE

Page 2, lines 45-48: 'The main objective of this paper is to critically compare the EMA's original scaled method based on a replicate TRTR/RTRT design (or, more precisely, an adjusted variant intended to preserve the type I error rate, as shown by Labes and Schütz [10]) ...'

The authors did /not/ employ the adjusted variant [10] (also claimed in section 2.2, Page 4) but used a different approach. More details will follow.

Thanks for highlighting this topic. As we fully agree, we have performed again all our RSABE computations under EMA's [10] "Method A". More details follow.

In Methods, Section 2.2 of the article has been re-written accordingly.

Page 5, lines 40-43: 'Although much better than the unadjusted version, the RSABE adjusted EMA method (AdjEMA) surprisingly still seemed to have a slight type I error probability inflation, with a 0.054 value.'

Now, the type I error probability control is perfect. We fully agree with the forthcoming comments, although we would like you to consider our comments highlighted in grey.

First of all I would like to thank the authors to provide their R-code for inspection. It seems that in the R-code not 'Method A' (as suggested by the EMA in the Q&A-document) was implemented but intra-subject contrasts (as recommended by the FDA and Health Canada). If this is the case it would seriously limit its acceptability from a European regulatory perspective (i.e., 'They did not do what we want – so we don't have to care what they suggest...'). That would be a pity.

It should be noted that in package PowerTOST, function scABEL.ad() only 'Method A' is implemented (targeting at the EMA's approach [10]), which is based on ANOVA. Intra-subject contrasts can be approximated in the function power.scABEL() by modifiying the regulatory setup in the following way: original <- reg_const(regulator="EMA")
original$est_method
 # [1] "ANOVA"
modified <- original
modified$est_method <- "ISC"

Exploring an example: CVwR 30%, RTRT|TRTR, n 24: power.scABEL(alpha=0.05, CV=0.3, theta0=scABEL(CV=0.3, regulator="EMA")[["upper"]], regulator=original, design="2x2x4", n=24, nsims=1e6)
 # [1] 0.0804
Inflation of the TIE as reported [10] (Table II first row and Supplementary material 2, Table III a, first row).

power.scABEL(alpha=0.05, CV=0.3, theta0=scABEL(CV=0.3, regulator="EMA")[["upper"]], regulator=modified, design="2x2x4", n=24, nsims=1e6)
 # [1] 0.084143
Higher inflation of the TIE since intra-subject contrasts instead of ANOVA are employed. Seems that this value is close to the manuscript's Table 2 for 'Regulatory EMA' given with 0.085.

res <- scABEL.ad(alpha=0.05, CV=0.3, theta0=scABEL(CV=0.3, regulator="EMA")[["upper"]], regulator=original, design="2x2x4", n=24, nsims=1e6, print=FALSE) alpha.adj.ANOVA <- res$alpha.adj
cat(res$alpha.adj, res$TIE.adj, "\n")
 # 0.029331 0.050001
An adjusted α of 0.029331 maintains the TIE at 0.050001. See [10] (Table II first row and Supplementary material 2, Table III b, second row).
res <- scABEL.ad(alpha=0.05, CV=0.3, theta0=scABEL(CV=0.3, regulator="EMA")[["upper"]], regulator=modified, design="2x2x4", n=24, nsims=1e6, print=FALSE)
alpha.adj.ISC <- res$alpha.adj
cat(res$alpha.adj, res$TIE.adj, "\n")
 # 0.026937 0.050001
If intra-subject contrasts are used, more adjustment (α 0.026937) would be required to control the TIE!

power.scABEL(alpha=alpha.adj.ANOVA, CV=0.3, theta0=scABEL(CV=0.3, https://mc.manuscriptcentral.com/sim?PARAMS=xik_43rs7AeuybF5XR...a3CCpA8ZxMc TD6JMiQ7G1EPjN2VMxbrtpFLj2Kf2sqvgt1EpcpygTn9HfRqt2w Página 2 de 6 ScholarOne Manuscripts 15/4/17 11'30 regulator="EMA")[["upper"]], regulator=modified, design="2x2x4", n=24, nsims=1e6)
 # [1] 0.053694
If the adjusted α 0.029331 (from ANOVA) is used, inflation of the TIE is observed if the evaluation is done by ISCs. I assume that this is what the authors have done. I disagree with the authors that there is 'a slight type I error probability inflation, with a 0.054 value' (as reported in Table 2). This inflation is only due to a setup which is not the one proposed in [10] and in my opinion likely would not be acceptable for the EMA.

simRSABE.R line 143 contains this statement:
scABEL.ad(n = ssize, design = "2x2x4", print = FALSE)$alpha.adj
Note that:

1. The regulatory setting is not modified (as required for ISCs). Hence the adjustment is done for ANOVA (as already shown).
2. The CVwR is not specified. According to the documentation of scABEL.ad() in such a case the default CV = 0.3 is used. Consequently all simulations done by the authors for CV != 0.3 are flawed (i.e., too conservative).
I am considering to ask the maintainer of PowerTOST or the author of scABEL.ad() to remove this default and stop the execution if missing(CV) == TRUE. It might have been that the authors overlooked this detail and hence, were caught in a trap.

In fact, we were perfectly aware of the default value CV = 0.3. We had some a priori objections on the convenience of substituting this constant value (admittedly, representing the worst scenario with respect to TIE) with an estimation, subject to statistical error. Additional results seem to confirm these concerns. Assigning to CV a value estimated from the data seems to inflate the TIE. Please see the answer to the last comment concerning RSABE.

power.scABEL(alpha=alpha.adj.ISC, CV=0.3, theta0=scABEL(CV=0.3, regulator="EMA")[["upper"]], regulator=modified,
design="2x2x4", n=24, nsims=1e6)
 # [1] 0.050004
If the adjusted α 0.026937 (from ISC) is used, practically no more inflation of the TIE is observed.

I analogy for N 48 the TIE of 0.053 in the manuscript's Table 3 can be explained:
Adjusted α (ANOVA) 0.028057 (TIE 0.05000)
Adjusted α (ISC) 0.026489 (TIE 0.05000)
Adjusted α 0.028057 (from ANOVA) used in the evaluation by ISC: TIE 0.052402

Another important point to note is that the functions provided in PowerTOST simulate the respective statistics (log(GMR) and SE) via their associated distributions (normal and χ2) and not subjects' data. This gives (depending on the sample size) a boost in performance of 1,000 to 10,000. In the Supplementary material 1 [10] it was shown that the in the case of homoscedasticity the agreement with the 'gold standard' subject simulations is sufficiently good.
Unfortunately the statistical distributions of ISCs are not directly accessible. See the file Implementation_scaledABE_sims.pdf in [R installation]/library/PowerTOST/doc/
Therefore, a modified regulatory setup (as described above) is only approximate.
In order to assess a potential impact of a wrong assumption I performed subject simulations and compared results with the other methods:
CVwR 30%, empiric TIE (assessed at true GMR 1.25), 1 mio simulations each.
n alpha power.TOST Subj.sim. power.TOST Subj.sim. Manuscript authors'
(ANOVA) (ANOVA) (ISC) (ISC) Table # R-code
24 0.050000 0.08040 0.08029 0.084143 0.08357 2 0.085
24 0.029331 0.05000 0.04991 0.053694 0.05344 2 0.054 0.05397
24 0.026937 0.050001 0.04967
48 0.050000 0.08232 0.08208 0.084647 0.08467
48 0.028057 0.05000 0.04969 0.052402 0.05258 3 0.053 0.05241
48 0.026489 0.050000 0.050887
In my opinion the agreement of ISC (by subject simulations and via the modified regulatory setup in power.TOST) is sufficient. In 86% of 70 scenarios ([n 12, 18, 24, 30, 36, 48, 60] and CVwR [20, 25, 30, 40, 50, 60%]) the 95% confidence intervals (binomial test) of empiric TIEs overlapped. Runtimes on my machine for all scenario were 249 seconds for scABEL.ad() compared to 61 hours for subject simulations.

We fully agree. Now, in all simulations, either subject or distributional, we follow the ANOVA approach clarified in the EMA's 2015 Questions & Answers document.

Page 4, lines 14~21: '... the type I error probability has only one single maximum, at CVwR = 30%. This greatly facilitates adjusting the significance level at α = 0.05. To focus on an easy to use method for sponsors, and with chances to be included in the regulations, we consider the method already implemented in the function "scABEL.ad" in the R package PowerTOST [10]. However, as a consequence of adjusting the significance level, the EMA's scaled method (labeled AdjEMA in the table results) should lose some power.'
As shown above the authors used intra-subject contrast (which IMHO, has a low chance of regulatory acceptance) and adjust always as if the CVwR would be 30%. The loss in power (expressed in a higher sample size) is highest at CVwR 30% [10] but /decreases/ if the CVwR is lower or higher (outside the critical region of inflated TIEs no adjustment is performed). This is not the case with the authors' method. Example: Expected GMR 0.9, target power 80%, RTRT|TRTR-design; sample size (power).
CV (%) N EMA power for N with adj. alpha N for adj. alpha % incr. N 20 18 (0.8015) 0.8015 18 (0.8015) 0.0
25 28 (0.8116) 0.8069 28 (0.8069) 0.0
30 34 (0.8028) 0.7251 42 (0.8022) 23.5
35 34 (0.8118) 0.7728 38 (0.8100) 11.8
40 30 (0.8066) 0.7800 32 (0.8035) 6.7
45 30 (0.8112) 0.8112 30 (0.8112) 0.0
If the alpha would always be adjusted for CVwR 30% (and evaluated by ISCs), the loss in power would be substantial
CV (%) N ISC power for N with adj. alpha N for adj. alpha % incr. N
20 20 (0.8188) 0.7403 24 (0.8187) 20.0
25 30 (0.8224) 0.7435 36 (0.8183) 20.0
30 36 (0.8091) 0.7308 44 (0.8064) 22.2
35 36 (0.8181) 0.7284 42 (0.8015) 16.7
40 32 (0.8134) 0.6814 38 (0.8091) 18.8
45 30 (0.8173) 0.6558 36 (0.8234) 20.0
Whereas the method proposed by [10] looses power only in the critical region, the authors' approach looses power independent from the CVwR and would require ~20% higher sample size in order to compensate this loss. Taking this to the extreme: Beyond CVwR 50% ABEL behaves similar to TOST (TIE ≤0.05). The authors' approach would still 'compensate' for an inflated TIE which does not exist. I do not think that this is a desirable property.

We are not sure of the convenience of assigning to argument CV of scABEL.ad a value estimated from the data –or equivalently in simulations, assigning to it a (transformed to original scale) random chi-square value in each simulation replicate. This means substituting a fixed constant, 0.3 (which admittedly represents the most pessimistic scenario with respect to TIE and so the expected maximum loss in power), with a random value subject to statistical error. In our opinion, this deserves further study and possibly additional adjustments in order to fully preserve TIE, if possible.

We provide in the APPENDIX 1 (at the end of this document) the script "TIE_estimating_CV.R" along with some "Results" (pages 14 to 17), in which some preliminary simulations in the 2x2x4 design have been performed. These simulations are based on only 10,000 simulation replicates because they require adjusting the significance level in each simulation replicate in function of each random generated CV value, and so they are quite slow. As a consequence of the low simulation size, the point estimates of the TIE probabilities are very imprecise, but their associated confidence intervals suggest TIE inflation in the neighbourhood of a simulated CV of 0.3. On the other hand, TIE seems to be under control (but apparently too conservative, even at a simulated CV = 0.3 for both formulations, R and T) for a fixed significance level coming from scABEL.ad with CV = 0.3.

Similar inflation trends are observed in (even slower!) simulations from subjects' generated data when the significance level is adjusted in each simulation replicate, assigning to argument CV the estimated coefficient of variation (simulating under homoscedasticity conditions and using the pooled estimate). On the other hand, more intensive subject simulations (1,000,000 replicates) based on a fixed adjusted significance level at CV = 0.3 show less conservative results for all simulated CVs, with the resulting TIE probability very close to 0.05 for a simulated homoscedastic scenario at CV = 0.3.

So, perhaps it would better to restore the default 0.3 value of the CV argument of scABEL.ad.

B. Comments concerning the (modified) TSDs

Page 4, line 45: Change 'over 80%' to 'at least 80%'. See [17]. However, the flowchart in Figure 2 is correct.

Thank you for catching this imprecision. In fact, Section 2.3 has been re-written following your comments. So, finally, because the significance level of 0.0294 is not applicable anymore, we have removed this paragraph entirely.

Figure 1: The adjusted alpha in the boxes should be 0.0294 instead of 0.094

Thank you for catching this typo error. Now, the significance level has been updated according to your algorithm recommendation. In Figure 1, these adjusted significant values are shown.

In general there seems to be an inflation [sic] of capital letters denoting different Two-Stage methods. I recommend to follow the terminology as given in [5] (i.e., 'Type 1' and 'Type 2'). You made clear in the text that you modified Potvin et al. Method B and C by introducing a futility criterion (limiting the total sample with 150 subjects) and a minimum stage 2 sample size. However, in some other parts of the text (and tables) you call them simply 'Potvin B' or 'Potvin C'. This might confuse readers – as it did myself. Might be picky but the difference is important. The original 'Potvin B' for CV 60% and N1 36 reaches a total power of 79% (although with large total sample sizes), whereas the modified method gives just 31%. I agree that from the producer's perspective it might be desirable to stop a study early. However, from an ethical perspective even starting a study with such a low expected power might be questionable (see Fuglsang (2013, doi:10.1208/s12248-013-9540-0) and [5]).

We fully agree that this confuses the readers. We are now introducing the terminology you are proposing, i.e., type 1 and 2 as follow:

New paragraph (Methods Section 2.3. paragraph 2):

Among adaptive approaches to bioequivalence [15], we focused on those (almost partially) mentioned in regulations, considering two "Pocock-like" variants [16], as described by Potvin et al. and labelled A, B, C and D [17]. In particular, we studied ~~the Potvin B method as the base case, and Potvin C as a sensitive case~~ a Type 1 [5] Potvin B method consisting of using the same adjusted α in both stages regardless of whether a study stops in the first stage or proceeds to the second stage (Figure 1), and a Type 2 Potvin C method where an unadjusted α may be used in the first stage, dependent on interim power (Figure 2).

In general, throughout the whole article, we refer to "Potvin" algorithms as "modified Potvin" algorithms.

In addition, we have also added a sentence highlighting the convenience of not starting a BE study with such a low power. Therefore, in the 3rd paragraph of the Discussion section, we have added the following sentence:

However, from an ethical perspective, even starting a study with such a low expected power might be questionable [22].

By exploring the various published methods it is evident that for 'Type 2' TSDs the highest inflation of the TIE is observed at a combination of small stage 1 sample sizes and moderate CVs:
Potvin Method C [17]: α 0.0294, n1 12, CV 20% (TIE 0.0510; for GMR 0.95 power 84.7%)
Montague Method D [13]: α 0.0280, n1 12, CV 20% (TIR 0.0518; for GMR 0.90 power 81.9%)
Since in the published methods the adjusted α is chosen in such a way that the TIE should be controlled over the entire grid of n1/CV-combinations it is is evident that – since all methods aim to cover a wide range of n1/CV-combinations – in most cases the (global) adjusted α will be more conservative than required for the actual (local) n1/CV-combination, thus deteriorating power. I strongly suggest that the authors take another publication (Xu et al. 2016, doi:10.1002/pst.1721) into account which aimed at resolving this problem by recommending different alphas, futility regions, and maximum total sample sizes for low to moderate (10–30%) and high (30–55%) CVs.

We agree. We have reviewed/included the reference of *Xu et al.* 2016 (Section 2.3., paragraph 8). In addition, we followed your methodology about how to assess the adjusted significance level, which is applicable to both modified Potvin methodologies (type 1 and 2 methods). Please, see our extended answer to this comment further down in this document.

According to the provided R-code the authors estimated power (and sample size) by the exact method (Owen's Q). In most papers the shifted central t-distribution was used for speed reasons – which is an approximation (of the noncentral t-distribution). The noncentral t-distribution itself only approximates the exact method. In actual studies likely at least the noncentral t-distribution will be employed. Hence, the author's approach reflect what will be used in reality. This is a very good idea indeed but should be stated as such in the manuscript.

When we started this article, we were not aware of the R function power.2stage. At that moment, we started from the scratch programming both modified Potvin algorithms. To calculate the adjusted significance level in this review, we have used this function because it is very fast and accurate (using a *nct* approximation). However, for tables and figures, we have used our original R programs because we do adjust ANOVA models by including the "sequence-by-stage" interaction factor, following the EMA guidelines. Because our methodology is based on an "exact" method, we mention this fact in Section 2.5., second paragraph:

In addition, these TSD simulations were done using the "exact" method.

Page 4, lines 57/58 to Page 5 lines 4/5: It is a misconception that [5] suggested that the minimum total sample size N should be ≥ 1.5N1. In [5] Birkett and Day [ref. 48] were quoted, who recommended N2 ≥ 1.5N1. If I interpret lines 68/69 of the R-code auxSimTSD_Potvin.R, max(N, 1.5*n1) was correctly used – so this may be only a typo in the manuscript. Please clarify.

We are proposing to enrol in the second stage at least half of the patients used in the first stage, so that if, for example, N_1=24, then N_2 is of at least 12 subjects (thus, the final N is at least 36 subjects).

The comment of Birkett and Day about *"We show the danger of selecting n potentially too close to N_1 and so our recommendation would be that if N_0 = 25, then n should certainly near to 10"* is based on the assumption that a prior estimate S^2 of the true σ^2 is based on literature reviews, previous experience and so on, and gives an idea of the overall *N*, N_0. Based on this N_0 value, they propose choosing an initial sample size *n*, and N_0 is re-evaluated at the interim (N_1) look based on the estimation of σ^2, $\hat{\sigma}^2$. Then *N* is chosen as the *max(N_0, N_1)*. But we are reasoning a bit differently, as we are not assuming anything on the S^2 and just fixing the *n* as per the sponsor's consideration (though we finally are proposing that it not be too low, e.g., of at least 12 subjects per sequence in a 2x2 crossover trial). In our article, we are showing that the power is always at least 80% for *n* of at least 6 subjects per sequence, unless *CVw* is 60% or higher. Another difference is that they are assuming that σ^2 is the between-subject variability, but $\hat{\sigma}^2$ in parallel trials is usually a bit higher than in crossover trials (with within-subject variabilities).

In addition, if we are not wrong, we found that in the article: Schütz H. Two-stage designs in bioequivalence trials. European Journal of Clinical Pharmacology 2015; 71(3):271-281., Table 1 shows that other jurisdictions or organizations already follow this approach, *N2 ≥ N1/2*.

To make this clearer, we have detailed this approach a bit more.

Previous paragraph (Section 2.3.):

~~*However, according to the suggestions made by Schutz [5] and Karalis and Macheras [18,19], we considered two additional constraints:*~~
- A minimum of N ≥ 1.5N1 is required
- If N = N1 + N2 > 150, the trial fails and it is stopped at the first stage.

New paragraph (Section 2.3.):

We propose a modification to the original Potvin B and C algorithms, including two constraints consisting of using a minimum sample size in the second stage (like in other jurisdictions or organizations) [5], and a maximum overall number of 150 subjects enrolled [18,19] in ABE studies, as follows:
- A minimum of N ≥ 1.5N1 is required (or N2 ≥ 0.5N1)
- If N = N1 + N2 > 150, the trial fails and it is stopped at the first stage.

I agree that in many papers the expected total sample size E[N] is used to compare the performance characteristics of methods. However, by their very nature in TSD the distribution of total sample sizes is bimodal. In Xu's methods it might be even trimodal (i.e., studies where N1+N2 exceeds max.n do not stop in the 1st stage but are performed with max.n). Hence, E[N] is of limited value (as is the median). Even the mode would provide little (if any useful) information. The manuscript is the first work I am aware of which reports the standard deviation of E[N]. Given what I wrote before, the authors should reconsider whether it makes sense to report it. See also the attached code TSD_Examples.R which generates sample size distributions for various methods. Don't worry about the runtime: Couple of seconds.
See the 3rd example obtained by the code: Type 1, alpha 0.0294, GMR 0.95, CV 30%, N1 24, target power 80%, N2.min 36, N.max 150:
E[N] 47.0. This might give a completely false impression. Of the 100,000 simulated studies not a single one (!) had a total sample size in the range 26 to 58.
Similar the authors' example in Table 2, Potvin C (is this the original 'method C' but with exact power instead of the shifted t?). I got E[N] 39.8 with a SD of 16.9. Hence, E[N]±2SD stretches from 6–74. Note that since N1=24 this is not meaningful.

Yes, we fully agree. We should avoid using the *E[N]* because the distribution of our modified Potvin methods is bimodal (not trimodal, because studies exceeding N=150 are discarded, and the second stage is not reached). In fact, we've replicated your figure where this bimodality is shown by adding Figure 5 in the article [note that this figure is based on N2>=N1/2 (or N>=1.5N1) instead of N2>=1.5N1]. Instead, we have removed Table 3, where RSABE was evaluated based on E[N].

We also refer to this figure in the Discussion, as follows (5th paragraph):

In addition, the expected total sample size E[N] is usually used to compare the performance characteristics of different TSD methods. However, by their very nature in TSD, the distribution of total sample sizes N is bimodal, mainly due to the imposition of N ≥ 1.5N1. For example, using our modified Potvin B, with $\alpha_{adj.}$ = 0.03018396 at each stage, GMR = 0.95, CVw = 0.3, N1 = 24, and target power 80%, we obtain a E[N] of 40 subjects, but with 24 and 36 subjects having more likelihood of occurrence (Figure 5). As the average is skewed towards two sample values, we believe that the median of N is more useful to compare different TSD methods.

We also believe that the *Me[N]* suffers somehow from the same problem, but having a reference on a central trend is very useful to compare different TSD methodologies, and TSD versus RSABE too, so we are using the *Me[N]* instead of the *E[N]*, despite its limitation.

This figure has been run using our own algorithms for consistency reasons, though we've double checked that these results are quite similar to using the function power.2Stage.

A note on the Type I Error in TSDs. Kieser & Rauch [15] lament that [17] used Pocock's one-sided α 0.0294 instead of the 'correct' two-sided 0.0304. However, this argument is flawed. 0.0304 (as is 0.0294) is not some kind of a 'natural constant'. Since the published TSDs are based on different frameworks, the adjusted alpha is entirely empiric and has to be estimated in simulations. The fact that 0.0294 'worked' in Potvin B was no more than a lucky punch (and we see a slight inflation in Method C). It is unfortunate that Potvin et al. did not further consider Method D – which could control the TIE with an adjusted α of 0.0280 and would have avoided the current skepticism by European regulators towards 'Type 2' TSDs. See also http://bebac.at/lectures/Prague2016-2.pdf
Generally any futility criterion reduces the TIE. Hence, the methods by Karalis/Macheras are not problematic in this respect. On the other hand, a rule dictating a minimum stage 2 sample size is expected to /increase/ the TIE (studies which would have needed just a few more subjects are forced to larger sample size – effectively shrinking the CI and increasing the probability of passing BE). In other words, if a framework is modified (like the authors did) it is of utmost importance to find a suitable adjusted α! You must not assume that what 'worked' in one framework would work in another one as well.
Example (GMR 0.95, N1 12, CV 20%, target power 80%, adj. α 0.0294, exact method; TIE assessed at true GMR 1.25, 1 mio sim's):
Type 2 (Potvin C): 0.05122 (slight inflation of the TIE; significantly >0.05).
additionally Nmax 150: 0.05122 (same because with this CV Nmax has no impact).
additionally N2 ≥ 1.5N1: 0.05302 (larger TIE which is even above Potvin's 'negligible' inflation of 0.052).
Given that, the authors should assess the TIE for all combinations of n1/CV. I suspect that the adjusted α of 0.0294 will not control the TIE in all scenarios. If this is the case, a suitable one should be provided. I know, that this can be a cumbersome task. My algorithm:
1. Start with an arbitrary adjusted α. Maybe ~0.0265 is a good starting point.
2. Evaluate the TIE over the entire grid with a low number of sim's. Use "nct" instead of "exact" (since 40times faster). Depending on the speed of the machine 20,000–50,000 are enough.

3. Filter for the combinations of n1/CV where the TIE is at last 95% of the maximum TIE observed in the grid. Another approach is to filter for combinations of n1/CV where the TIE is significantly >0.05.
4. Only for these combinations assess the TIE with 1 mio sim's. Find the n1/CV with the highest TIE.
5. Set up a range of alphas close to the one which was used before (slightly lower to slightly higher). I generally use 10 values. Estimate the resulting TIEs (1 mio sim's each).
6. The TIE-surface is nonlinear. Try a linear and quadratic fit. Use the one with the lower AIC. Back- calculate the adjusted alpha which gives TIE = 0.05.
7. Evaluate the entire grid with this new adjusted α (1 mio sim's each). "nct" is good enough. If you want to play it safe, repeat for the highest TIEs with "exact" (generally the TIE will differ only in the 4-5th decimal figure).
8. If you do not find a TIE which is larger than 0.05 (plus a certain threshold), stop. Otherwise, degrease the adjusted α and start over with step 4.
It is important to use a narrow 'mesh size' in order not to miss the global maximum. I routinely use 2 (both for n1 and the CV).
Whereas in 'Type 1' TSDs the region of max. TIEs resembles a 'ridge', in 'Type 2' the region can be flat (studies passing in stage 1 are at least partly evaluated with α 0.05). See http://bebac.at/lectures/Prague2016-1.pdf (slides 14/15).
I strongly suggest to explore different alphas for 'Type 1' and 'Type 2' designs. 'Type 2' designs always need more adjustment than 'Type 1'. Maybe you find the R-package Power2Stage useful. Should be /much/ faster than your current R-code.

We want to thank you for such a detailed algorithm, which has served to obtain the adjusted alpha for our modified Potvin type 1 and 2 algorithms. We look for adjusted significance levels covering a wide range of n1/CVw because we believe that this can help regulators, since they can widely rely on the protection of patients against false positive results. Though sponsors can sometimes have an intuition about the true *CVw* value, since they might have the results of a previous bioavailability trial, we believe that it is important not to have any assumptions about the potential *CVw* true value. Therefore, we have found an adjusted significance level for a wide range of *CVw*.

The following paragraph has been included in Section 2.3., paragraph 8:

The adjusted significance level of α = 0.0294, used by Potvin et al. [13,16,17,18] at each stage, did not always control the overall type I error rate at a maximum 0.05 (e.g., when using our modified Potvin C algorithm with N_1 = 12 and considering a true unknown CV_W = 20%, the false positive rate would be inflated to 0.053). Like in Xu et al. [20], we did look for a significance level by strictly controlling the type I error rate below 0.05, which was useful for our specific modified Potvin B and C methodologies. Because the sponsor is unaware of the true CV_W value, we looked for a significance level which was applicable to a broad set of N_1 and CV_W, $\{N_1/CV_W\}$ (scenarios shown in Section 2.5.).

We used the method implemented in the function "power.2stage" (via non-central t-distribution) in the R package Power2Stage. The treatment effect was evaluated at the frontier 1.25, and assuming an expected GMR = 0.95 and a target power of 80%.

A short statement for assessing the adjusted significance level, α_{adj}:

(1) Define a grid with a set of $\{N_1/CV_W\}$
(2) Start with an arbitrary, e.g. α_{adj} = 0.0290
(3) Obtain the empirical probability of type I error, Pr{TIE}, over the grid (m = 30,000 simulation trials per scenario). Filter for the scenarios where Pr{TIE} is at least 95% of the max(Pr{TIE}) observed in the grid, let's say $\{N_1/CV_W\}_{TIE \geq P95\%}$

(4) For $\{N_1/CV_W\}_{TIE>P95\%}$, find the N_1/CV_W with max(Pr{TIE}) (m = 1,000,000)
(5) Set up a range of α_j close to the one used before, $\alpha_j \in \{\alpha_{adj} \pm \delta_j\}_{j=1...5}$ (e.g. by δ increments of 0.0001 units). By using the N_1/CV_W associated to max(Pr{TIE}), estimate the Pr{TIE} of all α_j (m = 1,000,000)
(6) Adjust linear $\alpha = g_{lin}(Pr\{TIE\})$ and quadratic $\alpha = g_{quad}(Pr\{TIE\})$ models, with and without the intercept. Choose the model with the lowest Akaike information criterion value (AIC)
(7) Use this model to predict a new α_{adj}, where $\alpha_{adj} = g(0.05)$
(8) Evaluate the entire grid of $\{N_1/CV_W\}$ with this new α_{adj} (m = 1,000,000)
(9) If Pr{TIE} < 0.05 for all $\{N_1/CV_W\}$, STOP and select this new α_{adj}; Otherwise, start again over with step (4)

We provide in APPENDIX 2 (at the end of this document) the script "Modified Potvin_Alpha.R" (pages 18 to 21).

Also, we have included a paragraph in the Discussion (2nd paragraph):

Statistical power is used to evaluate the performance of adaptive methodologies in ABE clinical trials. A power of at least 80% is desirable when considering N_1 subjects at the first stage, and assuming an expected but unknown within-subject coefficient of variation, CVw. In turn, this is always conditioned to not exceed the overall type I error rate of 0.05 for true bioinequivalent drugs. In our modified Potvin B and C methods, we found adjusted significance levels covering a wide range of N_1 and CVw combinations (i.e., $\alpha_{adj} = 0.03018396$ and $\alpha_{adj} = 0.02806472$ at each stage for Potvin B and C, respectively). This is useful to regulators, since they can widely rely on the protection of patients against false positive results. However, for a specific actual (local) N_1 and CVw combination, the power is slightly downgraded, although it is always above 80% in cases of true bioequivalence.

Only cosmetics:
I suggest to order columns of N in Table 1: Min 5% Median 95% Max

We have updated Table 1 according to your comments.

Page 5, line 5: Change 'MSO R' to 'Microsoft R Open'.

We have updated the previous wording in Section 2.5. (first paragraph): *"The results described in the next sections are based on simulations using ~~64 bits R and Microsoft R Open~~ 64 bits R and MSO R"*

Given my comments in (A) and (B) I suggest that you re-calculate what have lead to Figures 1 and 2.

Thank you. Done.

Reviewer: 2

Comments to the Author

1) Please amend the title to read, `European Scaled Average Bioequivalence.....'

Thank you for this comment, following your suggestion we propose this new title:

Two-Stage Designs versus European Scaled Average Designs in Bioequivalence Studies for Highly Variable Drugs: Which to Choose?

2) Page 4, Line 43: There are some known statistical issues with this approach from Potvin's methods. The authors should consider and determine whether these should impact the simulations and whether additional scenarios should be considered. See Patterson and Jones (2016), 2ed., Chapter 6.

We are addressing in this review some statistical issues regarding Potvin's methods. As suggested by Reviewer 1, we have re-evaluated the local significance level at each stage as per our modifications to the original Potvin algorithms. Therefore, Section 2.3 *"*Two-stage modified Potvin B and C designs*"* have been re-written. In addition, we also discuss a bit further the consequences of imposing some futility rules to the original Potvin methods (paragraphs 2 and 3 in the Discussion). Finally, we show the bimodality of the empirical distribution of the sample size *N*, along with the discussion/limitations of using the *E[N]* as the main indicator to compare the different Potvin methods (paragraph 5 in the Discussion).

3) Page 5, Line 5: BE trials in excess of this N=150 have been conducted. The authors - see Patterson, Zariffa, Howland et al. (2001) European Journal of Clinical Pharmacology, 57, 663-670.

Yes, you are right. In this article we consider that it is more ethical to have a decision rule(s) for stopping at an interim analysis if it involves too many subjects with a low power. Likely, these studies should not even be initiated.

We have made the following changes in the Discussion to include your comment.

Current version on paragraph 3:

Patterson et al. [23] explored the sample size N that provides 90% power (for true bioequivalent drugs) in cases of HVD. They showed that by using 2x2 crossover designs with conventional ABE limits of 0.8-1.25 and CVw of 60% or above, the required sample size exceeds 150 subjects (replicate designs require smaller sample size). Using adaptive designs, we avoid conducting studies with such a large sample size by imposing a futility criterion so that we can stop the trial at an interim look with only N1 subjects. According to Karalis and Macheras [19], we added a constraint to the original TSD methods, specifically by not recruiting more than 150 subjects overall. ~~In practice, this is a futility criterion, as sponsors should not conduct ABE studies using TSDs that require such a high number of subjects.~~ For example, in the case of a true bioequivalent drug with $0.95 \leq GMR \leq 1.05$*, and for highly variable drugs with an estimated within-subject coefficient of variation above 58% at the interim analysis, the final sample size needed for achieving a power of 80% at the second stage already exceeds 150 subjects. At first glance, this constraint represents some global loss of power, but this possibility of cancelling a study for futility may ultimately be considered a positive trait, since the sponsor is unaware of the true treatment effect value during the planning phase, and the overall sample size could unnecessarily soar above this threshold for a scenario of true bioinequivalence.*

And we have included the reference of Patterson SD, Zariffa N, Montague TH, Howland K.

4) Page 8, Line 52: The authors should remove the cost-benefit analysis statement. The cost of even very large BE studies (N>150) is negligible relative to the costs of a clinical development program.

Thank you. You are right, so we have removed this sentence.

APPENDIX 1. TIE_estimating_CV.R

```
library(PowerTOST)

standardBELim <- log(1.25)
prTIE.confid <- 0.95
# Critical two-tailed normal z value:
zPrTIE <- qnorm((1 - prTIE.confid) / 2, lower.tail = FALSE)

# ************************ rsabe.test ******************************
# From estimates (in log-scale) of formulation effect and R and T variance,
# perform two variants of the scaled bioequivalence tests (EMA) with adjusted
# significance level. Adjusting is performed by means of 'scABEL.ad'.
# In the first test, the argument 'CV' of 'scABEL.ad' corresponds to the
# CV estimates (taken from the input argument 'estims'); in the second test
# it corresponds to the constant (pesimistic with respect to TIE) 0.3 value.
#
# Arguments:
# estims       : a numeric vector of length 3, estimates of formulation effect,
#                R variance and T variance, respectively, all at logartihmic scale
# n            : sample sizes, with the same meaning as in functions like CI.BE or
#                scABEL.ad
# nominalAlpha : nominal significance level, defaults to 0.05
# adjAlphaCV0.3: adjusted significance level with CV = 0.3 in scABEL.ad (for
#                efficiency reasons, in simulations it is advisable to provide it
#                previously computed outside rsabe.test, otherwise it is computed
#                inside the function)
rsabe.test <- function(estims, n, nominalAlpha = 0.05, adjAlphaCV0.3)
{
  # Point estimate constraint to declare BE:
  if (abs(estims[1]) > standardBELim) {
    return(c(FALSE,FALSE))
  }
  # Point estimate of GMR, in original scale:
  peGMR <- exp(estims[1])
  # Point estimates of CV, in original scale:
  cv <- mse2CV(estims[-1])
  common.cv <- mse2CV(0.5 * sum(estims[-1]))
  # Adjust the significance level assuming CV = 0.3, if argument 'adjAlphaCV0.3' is
  # not provided (advisable providing it for efficiency reasons)
  if (missing(adjAlphaCV0.3)) {
    adjAlphaCV0.3 <- scABEL.ad(alpha = nominalAlpha, design = "2x2x4", n = seqSizes, print
= FALSE)$alpha.adj
    if (is.na(adjAlphaCV0.3)) {
      adjAlphaCV0.3 <- nominalAlpha
    }
  }
  # Adjusted significance level, according to CV = data or simulation estimate
  adj.alpha <- scABEL.ad(alpha = nominalAlpha, CV = cv, design = "2x2x4", n = n, print =
FALSE)$alpha.adj
  if (is.na(adj.alpha)) {
    adj.alpha <- nominalAlpha
  }
  # Confidence interval from adjusted significance level based on CV = estimate from data
  ci <- CI.BE(alpha = adj.alpha, pe = peGMR, CV = common.cv, n = n, design = "2x2x4")
```

```
  # Confidence interval from adjusted significance level based on CV = 0.3
  ciAdj0.3 <- CI.BE(alpha = adjAlphaCV0.3, pe = peGMR, CV = common.cv, n = n, design =
"2x2x4")
  beLims <- scABEL(cv[1])
  return(c((beLims[1] <= ci[1]) && (ci[2] <= beLims[2]),
           (beLims[1] <= ciAdj0.3[1]) && (ciAdj0.3[2] <= beLims[2])))
}

# Generate estimates (in log-scale) of formulation effect and R and T variance
# from their sampling distributions
generate.estims <- function(simSize = 10000,
                            formulEff = log(1.25), varsW.RT = c(0.3,0.3),
                            seqSizes = c(12,12))
{
  N <- sum(seqSizes)
  ddf <- N - 2
  c2 <- 0.25 * sum(1 / seqSizes)
  varW <- 0.5 * sum(varsW.RT)
  result <- rbind(
    rnorm(simSize, mean = formulEff, sd = sqrt(c2 * varW)),
    rchisq(simSize, df = ddf) * varsW.RT[1] / ddf,
    rchisq(simSize, df = ddf) * varsW.RT[2] / ddf
  )
  # rownames(result) <- c("fEff", "s2WR", "s2WT")
  return(as.data.frame(result))
}

# ************************ simul.rsabe ******************************
# Simulate to obtain the proportion of bioequivalence declarations
# for two variants of the scaled bioequivalence test (EMA) with adjusted
# significance level. Adjusting is performed by means of 'scABEL.ad'.
# In the first test variant, the significance level is adjusted "on the fly"
# from each one of the simulation generated CV estimates, assigning it to
# argument 'CV' of 'scABEL.ad'; in the second variant, it is always the
# constant (pesimistic with respect to TIE) 0.3 value.
#
# Arguments:
# simSize      : number of simulation replicates, defaults to 10000
# GMR          : simulated population geometric means ratio between T and R,
#                defaults to 1.25
# CV_W.RT      : a numeric vector o length 2, simulated within-subject coefficients
#                of variation of R and T at original scale, as proportions.
#                Defaults to c(0.3,0.3)
# seqSizes     : a numeric vector o length 2, sample sizein each sequence,
#                defaults to c(12,12)
# nominalAlpha : nominal significance level, defaults to 0.05
# adjAlphaCV0.3: adjusted significance level when CV = 0.3 in scABEL.ad (if
#                not provided, it is computed inside the function, only once)
simul.rsabe <- function(simSize = 10000,
                        GMR = 1.25, CV_W.RT = c(0.3,0.3),
                        seqSizes = c(12,12),
                        nominalAlpha = 0.05,
                        adjAlphaCV0.3)
{
  formulEff <- log(GMR)
```

```
  varsW.RT <- CV2mse(CV_W.RT)
  if (missing(adjAlphaCV0.3)) {
    adjAlphaCV0.3 <- scABEL.ad(alpha = nominalAlpha, design = "2x2x4", n = seqSizes, print
= FALSE)$alpha.adj
    if (is.na(adjAlphaCV0.3)) {
      adjAlphaCV0.3 <- nominalAlpha
    }
  }
  prTIE <- rowSums(
    vapply(
      generate.estims(simSize = simSize, formulEff = formulEff, varsW.RT = varsW.RT,
seqSizes = seqSizes),
      FUN = rsabe.test, FUN.VALUE = c(FALSE, FALSE),
      n = seqSizes, nominalAlpha = nominalAlpha, adjAlphaCV0.3 = adjAlphaCV0.3
    )
  ) / simSize
  prTIE <- rbind(prTIE,
                 matrix(prTIE, ncol = 2, nrow = 2, byrow = TRUE) +
                   outer(c(-zPrTIE, zPrTIE), sqrt(prTIE * (1 - prTIE) / simSize), "*"))
  colnames(prTIE) <- c("CV = estimated from data", "CV = 0.3")
  rowName <- paste0(round(prTIE.confid * 100, 0), "%", " conf.int", sep = "")
  rownames(prTIE) <- c("Pr{BE} estimate",
                       paste0("lower ", rowName, sep = ""),
                       paste0("upper ", rowName, sep = ""))
  return(prTIE)
}

# Significance level adjusted assuming constant default CV = 0.3
adj.alpha <- scABEL.ad(alpha = 0.05, CV = 0.3, design = "2x2x4", n = c(16,16), print =
FALSE)$alpha.adj
if (is.na(adj.alpha)) {
  adj.alpha <- 0.05
}

# For these GMRs and CVs, the probability of declaring BE should be (near) 0.05 or smaller.
# Simulations are very slow because the significance level is adjusted in each simulation
replicate

set.seed(17393)
simul.rsabe(10000, seqSizes = c(16,16), GMR = 1.25, CV_W.RT = c(0.2,0.2), adjAlphaCV0.3
= adj.alpha)

set.seed(17393)
simul.rsabe(10000, seqSizes = c(16,16), GMR = 1.25, CV_W.RT = c(0.29,0.29),
adjAlphaCV0.3 = adj.alpha)

set.seed(17393)
simul.rsabe(10000, seqSizes = c(16,16), adjAlphaCV0.3 = adj.alpha) # GMR = 1.25, CV_W.R
= CV_W.T = 0.3

set.seed(17393)
simul.rsabe(10000, seqSizes = c(16,16), GMR = scABEL(0.31), CV_W.RT = c(0.31,0.31),
adjAlphaCV0.3 = adj.alpha)

set.seed(17393)
```

```
simul.rsabe(10000, seqSizes = c(16,16), GMR = scABEL(0.4), CV_W.RT = c(0.4,0.4),
adjAlphaCV0.3 = adj.alpha)
############################################################
```

RESULTS:

```
> adj.alpha <- scABEL.ad(alpha = 0.05, CV = 0.3, design = "2x2x4", n = c(16,16), print =
FALSE)$alpha.adj
if (is.na(adj.alpha)) {
  adj.alpha <- 0.05
}

> set.seed(17393)
> simul.rsabe(10000, seqSizes = c(16,16), GMR = 1.25, CV_W.RT = c(0.2,0.2),
            adjAlphaCV0.3 = adj.alpha)
                CV = estimated from data   CV = 0.3
Pr{BE} estimate                0.0485000 0.02740000
lower 95% conf.int             0.0442896 0.02420044
upper 95% conf.int             0.0527104 0.03059956
>
> set.seed(17393)
> simul.rsabe(10000, seqSizes = c(16,16), GMR = 1.25, CV_W.RT = c(0.29,0.29),
            adjAlphaCV0.3 = adj.alpha)
                CV = estimated from data   CV = 0.3
Pr{BE} estimate               0.05570000 0.04060000
lower 95% conf.int             0.05120499 0.03673178
upper 95% conf.int             0.06019501 0.04446822
>
> set.seed(17393)
> simul.rsabe(10000, seqSizes = c(16,16), adjAlphaCV0.3 = adj.alpha)
            # GMR = 1.25, CV_W.R = CV_W.T = 0.3
                CV = estimated from data CV = 0.3
Pr{BE} estimate               0.06290000 0.046500
lower 95% conf.int             0.05814154 0.042373
upper 95% conf.int             0.06765846 0.050627
>
> set.seed(17393)
> simul.rsabe(10000, seqSizes = c(16,16), GMR = scABEL(0.31), CV_W.RT = c(0.31,0.31),
            adjAlphaCV0.3 = adj.alpha)
                CV = estimated from data   CV = 0.3
Pr{BE} estimate               0.05740000 0.04290000
lower 95% conf.int             0.05284102 0.03892849
upper 95% conf.int             0.06195898 0.04687151
>
> set.seed(17393)
> simul.rsabe(10000, seqSizes = c(16,16), GMR = scABEL(0.4), CV_W.RT = c(0.4,0.4),
            adjAlphaCV0.3 = adj.alpha)
                CV = estimated from data   CV = 0.3
Pr{BE} estimate               0.04960000 0.03350000
lower 95% conf.int             0.04534458 0.02997328
upper 95% conf.int             0.05385542 0.03702672
```

APPENDIX 2. Modified Potvin_Alpha.R (looks for the best alpha)

```
library(Power2Stage)
library(lattice)
library(plyr)

#FUNCTION POTVIN B (in each stage (row) of the object "d", n_sim simulations are
performed)

##########################################################
potvin = function(type, d, nmax, n_sim, targetpower, pmethod) {
 #min.n2 <- 0
 sapply(1:nrow(d), function(x) {
  return(power.2stage(method = type, alpha = c(d[x,"alpha1"],d[x,"alpha2"]),
                n1=d[x,"n1"], GMR=0.95, CV=d[x,"CV"], targetpower = targetpower, pmethod
= pmethod,
                Nmax = nmax, min.n2=d[x,"min.n2"], theta0=d[x,"t_effect"], #theta0: True
unknown GMR (t_effect); theta0=1.25 for T1E; theta0=0.95 for power
                npct = c(0.05, 0.5, 0.95), nsims = n_sim , setseed = 123))
 }
 )
}
##########################################################

a = function(alpha1, alpha2, CV, n1, potvin_type) {
  min.n2 <- sapply(n1, function(x) if((x/2) %% 2 != 0) (x/2)+1 else x/2)
  d_T1E <- expand.grid(t_effect=1.25, CV=CV, n1=n1, alpha1=alpha1, alpha2=alpha2)
  d_T1E <- cbind(d_T1E, min.n2 = rep(min.n2,each=length(d_T1E["t_effect"])*length(CV)))

  #EXECUTING SIMULATIONS FOR T1E AT t_effect=1.25 (n_sim = 30,000)
  #pmethod = c("nct", "exact")
  res_T1E <- potvin(type=potvin_type, d=d_T1E, nmax=150, n_sim=30000, targetpower=0.8,
pmethod = "nct")
  T1E <- res_T1E["pBE",]
  T1E <- t(matrix(as.numeric(T1E), nrow=length(CV), ncol=length(n1)))
  rownames(T1E) <- n1
  colnames(T1E) <- CV

  #T1E HIGHER OR EQUAL THAN QUANTILE 95% (n_sim = 1,000,000)
  T1E95 <- T1E[T1E >= quantile(T1E, 0.95)]
  index <- which(T1E >= quantile(T1E, 0.95), arr.ind=TRUE)
  n195 <- as.numeric(rownames(T1E)[index[,1]])
  min.n295 <- sapply(n195, function(x) if((x/2) %% 2 != 0) (x/2)+1 else x/2)
  d95 = data.frame(t_effect=1.25, CV=as.numeric(colnames(T1E)[index[,2]]), n1=n195,
alpha1=alpha1, alpha2=alpha2, min.n2=min.n295)
  res_T1E95 <-  potvin(type=potvin_type, d=d95, nmax=150, n_sim=1000000,
targetpower=0.8, pmethod = "nct")
  T1E_high <-
data.frame(cbind(d95[,2],d95[,3],d95[,4],d95[,5],as.numeric(res_T1E95["pBE",])))
  colnames(T1E_high) <- c("CV", "n1", "alpha1", "alpha2", "pBE")

  #MAX TIE (n_sim = 1,000,000)
  max_T1E <- T1E_high[T1E_high["pBE"]==max(T1E_high["pBE"]), ]
```

```
   #REDEFINING NEW ALPHA LEVELS (ARBITRARILY 5 VALUES BELOW AND 5
VALUES UPPER THE PREVIOUS ALPHA VALUE) AND PREPARING SIMULATIONS
   new_alpha <- seq(alpha1-0.0005, alpha1+0.0005, by = 0.0001)
   n_d <- data.frame(t_effect=1.25, CV=max_T1E["CV"], n1=max_T1E["n1"],
alpha1=new_alpha, alpha2=new_alpha,
                min.n2 <- if((max_T1E["n1"]/2) %% 2 != 0) (max_T1E["n1"]/2)+1 else
max_T1E["n1"]/2)
   colnames(n_d) <- c("t_effect","CV", "n1", "alpha1", "alpha2", "min.n2")

   #EXECUTING SIMULATIONS AT A t_effect=1.25 AND FOR A RANGE OF ALPHA1 =
ALPHA2 (n_sim = 1,000,000)
   res_new_t1e <- potvin(type=potvin_type, d=n_d, nmax=150, n_sim=1000000,
targetpower=0.8, pmethod = "nct")
   valpha1 <- sapply(1:ncol(res_new_t1e), function(x) res_new_t1e[["alpha",x]][1])
   valpha2 <- sapply(1:ncol(res_new_t1e), function(x) res_new_t1e[["alpha",x]][2])
   res_new_d_T1E <- data.frame(CV = unlist(res_new_t1e["CV",]),
                        n1 = unlist(res_new_t1e["n1",]),
                        alpha1 = valpha1,
                        alpha2 = valpha2,
                        min.n2 = unlist(res_new_t1e["min.n2",]),
                        T1E = unlist(res_new_t1e["pBE",]))

   d_power <- res_new_d_T1E
   d_power$t_effect <- 0.95 # To calculate the power
   res_power <- potvin(type=potvin_type, d=d_power, nmax=150, n_sim=1000000,
targetpower=0.8, pmethod = "nct")
   valpha1_power <- sapply(1:ncol(res_power), function(x) res_power[["alpha",x]][1])
   valpha2_power <- sapply(1:ncol(res_power), function(x) res_power[["alpha",x]][2])
   res_d_power <- data.frame(  CV = unlist(res_power["CV",]),
                        n1 = unlist(res_power["n1",]),
                        alpha1 = valpha1_power,
                        alpha2 = valpha2_power,
                        min.n2 = unlist(res_power["min.n2",]),
                        power = unlist(res_power["pBE",]))

   newd <- merge(res_d_power, res_new_d_T1E, by=c("n1", "CV", "alpha1", "alpha2",
"min.n2"), all.x = TRUE, sort = FALSE)

   #AIC models
   l.models1 <- lm(alpha1 ~ 0 + T1E, data = newd)
   l.models2 <- lm(alpha1 ~ T1E, data = newd)
   q.models1 <- lm(alpha1 ~ 0 + T1E + I(T1E^2), data = newd)
   q.models2 <- lm(alpha1 ~ T1E + I(T1E^2), data = newd)
   newd1 <- data.frame(AIC.l1=AIC(l.models1), AIC.q1=AIC(q.models1),
AIC.l2=AIC(l.models2), AIC.q2=AIC(q.models2),
                min.AIC=min(AIC(l.models1), AIC(q.models1), AIC(l.models2),
AIC(q.models2)))
   p <- data.frame(alpha1=NA, T1E=0.05)
   if (newd1$min.AIC == newd1$AIC.l1) {
      a_aic <- predict(l.models1, p)
   }
   else if (newd1$min.AIC == newd1$AIC.l2) {
      a_aic <- predict(l.models2, p)
   }
   else if (newd1$min.AIC == newd1$AIC.q1) {
```

```
    a_aic <- predict(q.models1, p)
   }
   else a_aic <- predict(q.models2, p)

 return(list(max_T1E=max_T1E, sensitivity=newd, new_alpha_aic=a_aic))
}
#############################################################

# Given the first iteration given with the "function a", the function result looks for the best alpha

result <- function(x) {
 success <- FALSE
 new_alpha1 = as.numeric(res[["new_alpha_aic"]])
 new_alpha2 = as.numeric(res[["new_alpha_aic"]])
 res <- a(new_alpha1, new_alpha2, CV, n1, potvin_type)
 while (!success) {
  if (res[["max_T1E"]]["pBE"] <= 0.05) {
   cat("Predicted alpha1=alpha2 with min(AIC) of lm(alpha ~ 0 + T1E) and lm(alpha ~ 0 +
T1E + I(T1E^2)","\n",
       "lm(alpha ~ T1E) and lm(alpha ~ T1E + I(T1E^2)","\n",
       "for the particular CV and n1 corresponding to max(T1E) from the cartesian product of",
"CV=", CV, "and", "n1=", n1,"\n",
       "Predicted alpha1=alpha2=", res[["max_T1E"]]$alpha1, "\n",
       "with max empirical T1E =",res[["max_T1E"]]$pBE[1],"\n",
       "ensuring that the predicted alpha protects all T1Es for cartesian product of n1 and CV
below 0.05 ")
   success <- TRUE
  } else {
   new_alpha1 = as.numeric(res[["new_alpha_aic"]])
   new_alpha2 = as.numeric(res[["new_alpha_aic"]])
   res <- a(new_alpha1, new_alpha2, CV, n1, potvin_type)
  }
 }
 return(res)
 break
}
#############################################################

# Parametrizaton

# Potvin B: alpha1=alpha2=0.03018396
# Potvin C: alpha1=alpha2=0.02806472
alpha1=0.03018396
alpha2=0.03018396
CV <- c(0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6)
n1 <- c(12, 18, 24, 30, 36, 48, 60)
potvin_type = "B"

#APPLYING POTVIN METHOD (FUNCTION a - Apply reviewer algorithm - first iteration
only)

res <- a(alpha1, alpha2, CV, n1, potvin_type)

# Given the first iteration given with the "function a" which returns data.frame "res", the
"function result" looks for the best alpha
```

```
#LOOKING FOR THE BEST ALPHA ("function result") WHICH CONTROLS T1E FOR
ALL COMBINATIONS OF CV AND n1

adj_alpha <- result(res)
```

Two-Stage Designs versus European Scaled Average Designs in Bioequivalence Studies for Highly Variable Drugs: Which to Choose?

Eduard Molins[a*†], Erik Cobo[a], Jordi Ocaña[b]

Abstract

The usual approach to determine bioequivalence for highly variable drugs is scaled average bioequivalence, which is based on expanding the limits as a function of the within-subject variability in the reference formulation. This requires separately estimating this variability, and thus using replicated or semi-replicated crossover designs. On the other hand, regulations also allow using common 2×2 crossover designs based on two-stage adaptive approaches with sample size re-estimation at an interim analysis. The choice between scaled or two-stage designs is crucial and must be fully described in the protocol. Using Monte Carlo simulations, we show that both methodologies achieve comparable statistical power, though the scaled method usually requires less sample size, but at the expense of each subject being exposed more times to the treatments. With an adequate initial sample size (not too low, e.g., 24 subjects), two-stage methods are a flexible and efficient option to consider: They have enough power (e.g., 80%) at the first stage for non-highly variable drugs and, if otherwise, they provide the opportunity to step up to a second stage that includes additional subjects.

[a]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya.
[b]Department of Genetics, Microbiology and Statistics. Universitat de Barcelona.
[*]Correspondence to: Eduard Molins, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya. Jordi Girona 1-3, 08034 Barcelona, Spain.
[†]E-mail: eduard.molins@astrazeneca.com

Include up to six keywords that describe your paper for indexing purposes: Average Bioequivalence (ABE), Reference Scaled Average Bioequivalence (RSABE), Two-Stage Designs (TSD), Group sequential designs (GSD), Highly Variable Drugs (HVD), Significance Level Adjustment

---

1. Introduction

Average bioequivalence (ABE) studies are conducted to demonstrate in vivo either that two products, say "test" *T* and "reference" *R*, are pharmaceutically equivalent (in the US) or that their rate and extent of absorption [1-3] are close enough to serve as alternative pharmaceutical products (in the EU). The most common measure of the rate of absorption is the bioavailability measure "maximum observed concentration" (*Cmax*), while the "area under the concentration curves" ($AUC_{0\text{-}t}$ and $AUC_{0\text{-}\infty}$) [4] are the most common bioavailability measures for the extent of absorption. To demonstrate ABE, regulatory guidelines recommend a single dose 2×2 crossover design, *RT/TR* that evaluates *T* and *R* on healthy volunteers. The most commonly used criterion to test (at a significance level of $\alpha = 0.05$) for ABE is the "interval inclusion rule", which is based on a 90% symmetric confidence interval for the formulation effect, say the mean difference between the bioavailabilities of formulations *T* and *R* at a log-transformed

scale. It is based on the Student's distribution, assuming data normality. In order to declare ABE, the back-transformed confidence interval for the geometric means ratio (*GMR*) should lie fully within the ABE limits of 0.80–1.25 (=1/0.80), corresponding to ±0.223 on the logarithmic scale [2,5].

Highly variable drugs (HVD) are characterized by high within-subject variability in the rate and/or extent of absorption of its active principle. This hinders researchers from declaring ABE when it really holds, unless unacceptably large sample sizes are used. Most regulations classify a drug as HVD if the within-subject coefficient of variation of the reference formulation *R* (CV_{WR}) is 30% or greater on the original scale. The percentage of HVD is not negligible. Davit et al. [6] collected data from all in vivo bioequivalence studies reviewed by the FDA's Office of Generic Drugs from 2003 to 2005, and they concluded that 31% of the studies (57/180) corresponded to highly variable drugs, many of them around CV_{WR} = 30%.

If HVD is suspected, the European Medicines Agency (EMA) allows linearly scaling the *Cmax* margins as a function of the *R* variability to a maximum plateau of 0.6984-1.4319, and it further allows application of the interval inclusion rule over these expanded limits [2]. Similarly, the FDA also allows researchers to re-scale the *AUC* limits [1,3]. These scaled approaches require the use of high order crossover designs like the replicated *TRTR/RTRT* or semi-replicated *TRR/RTR/RRT* designs [2,7,8]. However, these scaled methods, as defined by FDA and EMA regulations, do not adequately preserve the type I error rate in the neighborhood of CV_{WR} = 30% [9,10]. Thus, the proportion of non-ABE products erroneously declared as ABE is higher than its desired nominal value.

Regulators also allow using two-stage adaptive designs (TSD) with unblinded interim sample size re-estimation [2,5,11,12] based on the usual 2×2 crossover *RT/TR* design. Bioequivalence may be declared at the interim look with N_1 subjects; otherwise, the sample size can be increased on the basis of the estimated within-subject variability at the first stage, then ABE is tested again at a second stage with cumulated data $N = N_1 + N_2$. Two-stage designs preserve the type I error rate [13] by adjusting significance boundaries at each stage in various ways that are not fully specified in the regulations [14,15].

In turn, the planned sample size is crucial because it may lead to underpowered studies, as there is a high uncertainty about the assumed *GMR* and/or variability.

The main objective of this paper is to critically compare the EMA's original scaled method based on a replicate *TRTR/RTRT* design (or, more precisely, an adjusted variant intended to preserve the type I error rate, as shown by Labes and Schütz [10]) with two TSD methods based on the usual *RT/TR* crossover design.

Section 2 describes the compared methods and details the simulation methodology. Section 3 shows the results; and Section 4 discusses them in order to recommend the most appropriate approach.

2. Statistical methodology

2.1. 2010 Regulatory EMA reference scaled average bioequivalence approach (RSABE) (for Cmax only)

Replicate *TRTR/RTRT* designs allow separately estimating the CV_{WR} [9] and can easily be re-arranged for comparison with a 2×2 crossover design (needed for two-stage designs) once the first two periods are sliced (see Section 2.3).

We focus on the EMA regulation because the FDA's approach is based on scaled limits which are discontinuous at CV_{WR} = 30%. This discontinuity is associated with a sharp peak of type I probability around this *CV* value, which threatens its validity.

On the original scale, the null hypothesis of bioinequivalence is tested against an alternative of bioequivalence, as follows:

$$H_0: GMR \leq 0.80 \text{ or } GMR \geq 1.25$$
$$H_1: 0.80 < GMR < 1.25.$$

In the Reference Scaled Average Bioequivalence (RSABE) approach, the ABE limits are a function, say GMR_{EMA}, of the unknown population within-subject *R* coefficient of variation CV_{WR}, so the hypotheses being tested differ from the standard ones enunciated above:

$$H_0: GMR \leq 1/GMR_{EMA}(CV_{WR}) \text{ or } GMR \geq GMR_{EMA}(CV_{WR})$$
$$H_1: 1/GMR_{EMA}(CV_{WR}) < GMR < GMR_{EMA}(CV_{WR}).$$

If CV_{WR} < 30%, $GMR_{EMA}(CV_{WR}) = 1.25$; so the ABE limits are the usual 0.8–1.25. If CV_{WR} lies between 30% and 50%, the ABE limits grow as $GMR_{EMA}(CV_{WR}) = exp\{k_{EMA}\sqrt{log(CV_{WR}^2+1)}\}$, with $k_{EMA} = 0.76$. Otherwise, from CV_{WR}= 50%, $GMR_{EMA}(CV_{WR}) = 1.4319$; so the ABE limits stay constant at 0.6984 (= 1/1.4319).

A short statement of the EMA testing decision criterion is:

(1) Obtain the *GMR* estimate, $\widehat{GMR} = e^{\hat{\phi}}$, where $\hat{\phi}$ is the estimated formulation effect ϕ, the mean difference of test and reference products of the corresponding log *Cmax* scale;
(2) Point estimate constraint: If $\widehat{GMR}$ is outside the limits 0.8-1.25, do not declare bioequivalence and stop;
(3) Obtain the estimate of the within-subject coefficient of variation of the reference product, $\widehat{CV}_{WR} = \sqrt{e^{\hat{\sigma}_{WR}^2} - 1}$, where $\hat{\sigma}_{WR}^2$ is the estimated value of the reference residual standard deviation in the logarithmic scale;
(4) Obtain the 90% confidence interval for *GMR* around its estimate $\widehat{GMR}$, $CI_{\widehat{GMR}} = e^{[\hat{\emptyset}_L, \hat{\emptyset}_U]}$, where $\hat{\emptyset}_L$ and $\hat{\emptyset}_U$ are the estimated lower and upper limits of the confidence interval in the logarithmic scale, at a confidence level of $1 - 2\alpha$ for α = 0.05
(5) If $CI_{\widehat{GMR}}$ is fully included in the $GMR_{EMA}(\widehat{CV}_{WR})$ limits, declare ABE (reject H_0), otherwise do not declare ABE.

Note that the limits $GMR_{EMA}(\widehat{CV}_{WR})$ are random, not fixed constants like 0.8 or 1.25, since they depend on the random quantity $\widehat{CV}_{WR}$, which is not fixed in advance.

Muñoz J. et al. [9], among others, showed that the above decision criterion does not adequately control the type I error probability, or false positive rate (say, if bioequivalence is erroneously declared when in fact it does not hold) in the neighborhood of CV_{WR} = 30%.

2.2. Significance level adjustment on the Regulatory EMA scaled approach

As has been previously stated, the 2010 former EMA RSABE procedure does not control completely the type I error probability. To focus on an easy to use method for practitioners, and with chances to be included in the regulations, we considered the method already implemented in the function "scABEL.ad" in the R package PowerTOST [10]. As a consequence of adjusting the significance level, the EMA's scaled method (labeled AdjEMA in the table results) may lose some power. But this (small in general) loss of power is worth because it converts a potentially invalid procedure (with respect to the type I error probability) in a fully correct one.

As a function of the reference coefficient of variation, the type I error probability has only one single maximum at CV_{WR} = 30%. Consequently, though somewhat conservatively, we let the argument "CV" of scABEL.ad at its default value of 0.3. The alternative strategy of estimating the coefficient of variation from data and assigning this (random function of data, unknown in advance) value to the argument CV induces some type I error probability inflation.

In accordance with EMAs Questions & Answers guideline [11], section 10, the estimation of the required parameters was based on the ANOVA procedure labelled as "Method A" in this document, and not in the intra-subject contrasts, as are for example allowed in the FDA regulation for scaled average bioequivalence.

2.3. Two-stage modified Potvin B and C designs

We consider two adaptive two-stage designs (TSD) with one interim analysis (at the first stage) with N_1 subjects to either (1) establish equivalence early; or (2) stop for futility; or (3) recruit an additional group of N_2 subjects to repeat the bioequivalence assessment at a second stage with $N = N_1 + N_2$ subjects. Each stage is based on a 2×2 crossover balanced *RT/TR* design, and so the within-subject variability CV_W should be estimated by means of the pooled variability of *R* and *T*. Unlike the scaled approach, two-stage hypotheses always rely on the standard fixed limits 0.8–1.25.

Among adaptive approaches to bioequivalence [15], we focused on those (almost partially) mentioned in regulations, considering two "Pocock-like" variants [16], as described by Potvin et al. and labelled A, B, C and D [17]. In particular, we studied a Type 1 [5] Potvin B method consisting of using the same adjusted α in both stages regardless of whether a study stops in the first stage or proceeds to the second stage (Figure 1), and a Type 2 Potvin C method where an unadjusted α may be used in the first stage, dependent on interim power (Figure 2).

Both methods calculate N_2 as the minimum even number of additional subjects required for having a total sample size of N, which achieves a conditional power of at least 80% for declaring bioequivalence at the second stage. This is conditional on the estimated within-subject coefficient of variation $\widehat{CV}_W$ at the first stage for an assumed true *GMR* of 0.95.

Potvin A was discarded, as it did not adjust the significance boundaries; Potvin D was a more conservative variant of Potvin C, and therefore not recommended because it requires larger average sample sizes than Potvin C [13].

We propose a modification to the original Potvin B and C algorithms, including two constraints consisting of using a minimum sample size in the second stage (like in other jurisdictions or organizations) [5], and a maximum overall number of 150 subjects enrolled [18,19] in ABE studies, as follows:

- A minimum of $N \geq 1.5N_1$ is required (or $N_2 \geq 0.5N_1$)
- If $N = N_1 + N_2 > 150$, the trial fails and it is stopped at the first stage.

In any case, regardless of the method used, at least 12 evaluable subjects should be included in the first stage [1,11].

The adjusted significance level of $\alpha = 0.0294$ used by Potvin et al. [13,16,17,18] at each stage did not always control the overall type I error rate at a maximum 0.05 (e.g., when using our modified Potvin C algorithm with $N_1 = 12$ and considering a true unknown $CV_W = 20\%$, the false positive rate would be inflated to 0.053). Like in Xu *et al.* [20], we did look for a significance level by strictly controlling the type I error rate below 0.05, which was useful for our specific modified Potvin B and C methodologies. Because the sponsor is unaware of the true CV_W value, we looked for a significance level which was applicable to a broad set of N_1 and CV_W, $\{N_1/CV_W\}$ (scenarios shown in Section 2.5.).

We used the method implemented in the function "power.2stage" (via non-central *t*-distribution) in the R package Power2Stage. The treatment effect was evaluated at the frontier 1.25, and assuming an expected *GMR* = 0.95 and a target power of 80%.

A short statement for assessing the adjusted significance level, α_{adj}:

(1) Define a grid with a set of $\{N_1/CV_W\}$
(2) Start with an arbitrary, e.g. $\alpha_{adj} = 0.0290$
(3) Obtain the empirical probability of type I error, $Pr\{TIE\}$, over the grid (m = 30,000 simulation trials per scenario). Filter for the scenarios where $Pr\{TIE\}$ is at least 95% of the $max(Pr\{TIE\})$ observed in the grid, let's say $\{N_1/CV_W\}_{TIE \geq P95\%}$
(4) For $\{N_1/CV_W\}_{TIE \geq P95\%}$, find the N_1/CV_W with $max(Pr\{TIE\})$ (m = 1,000,000)
(5) Set up a range of α_j close to the one used before, $\alpha_j \in \{\alpha_{adj} \pm \delta_j\}_{j=1\ldots5}$ (e.g. by δ increments of 0.0001 units). By using the N_1/CV_W associated to $max(Pr\{TIE\})$, estimate the $Pr\{TIE\}$ of all α_j (m = 1,000,000)
(6) Adjust linear $\alpha = g_{lin}(Pr\{TIE\})$ and quadratic $\alpha = g_{quad}(Pr\{TIE\})$ models, with and without the intercept. Choose the model with the lowest Akaike information criterion value (AIC)
(7) Use this model to predict a new α_{adj}, where $\alpha_{adj} = g(0.05)$

(8) Evaluate the entire grid of $\{N_1/CV_W\}$ with this new α_{adj} (m = 1,000,000)
(9) If $Pr\{TIE\} < 0.05$ for all $\{N_1/CV_W\}$, STOP and select this new α_{adj}; Otherwise, start again over with step (4)

As the 2010 EMA guideline uses a Type 1 TSD method [2], we used the modified Potvin B as the main TSD approach and the modified Potvin C as a sensitive case.

2.5. Simulation methods

The results described in the next sections are based on simulations using 64 bits R and Microsoft R Open. The main outputs are: type I error rate, power and the number of trials stopping at the first stage for the TSD approach. For most scenarios, m = 100,000 datasets were generated, but m = 1,000,000 for those devoted to estimating the most crucial type I error probabilities, i.e., for simulated *GMR*s just on the bioequivalence limit.

In the simulations, we considered all combinations of 3 factors: sample size, true *GMR* and true within-subject variability under the homoscedasticity assumption that $CV_W = CV_{WR} = CV_{WT}$ (from now on, we use CV_W and CV_{WR} interchangeably, provided the assumed simulated homoscedasticity). The sample sizes were N_1 = 12, 18, 24, 30, 36, 48 and 60 subjects for RSABE methods and at the first stage for TSD methods, always considering a balanced design, i.e.: 6, 9, 12, 15, 18, 24 and 30 subjects per sequence. The simulated population *GMR* values were 0.95, 1.00, 1.12, 1.25 and 1.31; with the first three corresponding to scenarios under true bioequivalence (alternative hypothesis), and the last two corresponding to the true non-bioequivalence (null hypothesis). In fact, this statement is exactly true for the TSD approach, where the bioequivalence limits are the constants 0.80–1.25; see the next paragraph for clarification in the RSABE case. Finally, the simulated within-subjects coefficients of variation were 10%, 20%, 25%, 30%, 40%, 50% and 60%. A coefficient of variation of 30% or higher indicates an HVD. Section 3 reports only the results for a subset of the simulated values on sample size, true *GMR*, and true coefficient of variation. In addition, these TSD simulations were done using the "exact" method.

Provided that TSD and RSABE are based on different definitions of bioequivalence, comparing them is quite difficult. In order to have a reference case for comparison, we took the simulated true *GMR* values "on the frontier" of each approach (constant 1.25 in TSD or a function GMR_{EMA} in RSABE for varying simulated CV_{WR} values), which should provide similar proportions of bioequivalence declaration (near 0.05) if both approaches are adequately controlling the user's risk. For *GMR*s that are progressively inside or outside the corresponding bioequivalence regions, these probabilities should also be comparable. To define these concordant simulation scenarios, we reasoned at the logarithmic scale. The constant simulated *GMR* values in the TSD approach are 0.95, 1.00, 1.12, 1.25 and 1.31, and they correspond to formulation effects on the logarithmic scale of –0.0513, 0, 0.1133, 0.2231 and 0.2700, respectively. With respect to the (frontier) 0.2231 value, these formulation effects correspond to proportions λ = –0.230, 0, 0.508, 1 and 1.210, respectively. Then, λ = 1 refers to values on the frontier, $|\lambda| < 1$ to scenarios of true bioequivalence, and $|\lambda| > 1$ to scenarios of bioinequivalence. Therefore, the same λ value defines concordance in TSD and RSABE scenarios: the population *GMR*s in the original scale were taken as $exp\{\lambda\ 0.2231\}$ in the TSD

approaches, and for all simulated CV_{WR} values; while in the RSABE approach, they were taken as $exp\{\lambda\, 0.2231\}$ for $CV_{WR} < 30\%$, as $exp\left\{\lambda k_{EMA}\sqrt{log(CV_{WR}^2+1)}\right\}$ for CV_{WR} values between 30% and 50%, and as $exp\{\lambda\, 0.3590\}$ for a $CV_{WR} \geq 50\%$.

For simplicity, the simulated *GMR*s in the next sections will always be labeled as 0.95, 1.00, 1.12, 1.25 and 1.31; but it should be remembered that these values in the RSABE case correspond only to the simulated coefficients of variation below 30%.

Following the EMA Questions & Answers guideline [11], adjusted ANOVA models for analysis of the combined second stage data included the following terms: stage, sequence, interaction sequence*stage, subject nested in sequence*stage, period nested in stage, and formulation.

3. Simulation results

The adjusted significance level predicted for the modified Potvin B was assessed at α_{adj} = 0.03018396 at each stage; For the modified Potvin C, the adjusted significance level predicted was assessed at α_{adj} = 0.02806472 (Figures 1 and 2).

Both adaptive TSD modified Potvin B and C methods performed similarly in respect to the power achieved and the required median sample size *Me[N]* (Table 1). Because almost all simulated studies required stepping up to a second stage and resulted in large final sample sizes, it was not advisable to start with a too small sample size, like N_1 = 12, in scenarios with high variability ($CV_W \geq 30\%$).

On the other hand, when $N_1 \geq 24$, the global power (including both stages) was at least 80% when variabilities were raised up to 40%. Additionally, those sample sizes increased the likelihood of stopping for bioequivalence at the first stage. For the high value of CV_W = 60%, results were poor, with power always below 80%.

For the RSABE EMA method, a crucial variability value is at the threshold $CV_W = 30\%$, where there is a maximum type I error peak. Table 2 shows that for a true GMR of 1.25 the highest false positive rate is 0.085, confirming the already known risk control problems of the EMA scaled approach. On the other hand, the RSABE adjusted EMA method (AdjEMA) accurately respected the nominal 0.05 level. Both TSD approaches also respected the type I error at 0.05. In addition, for a sample size of N_1 = 24, all methods with a type I error close to the nominal 0.05 level provide satisfactory and similar powers on bioequivalent drugs (*GMR* = 0.95, 1.00, and 1.12). The apparently larger sample sizes required by TSD methods should be relativized: with half periods, they did not double mean size and reached a bioequivalence statement at the first stage in a notable proportion of times (approximately 41%, 47% and 24%).

Figure 3 shows a more comprehensive picture of the extended N_1 and CV_W values for a bioequivalent scenario fixed at *GMR* = 0.95. When N_1 = 12, TSD methods showed higher power than the RSABE adjusted EMA method for $CV_W > 20\%$, requiring relatively larger global sample sizes of *Me[N]* = 44 and around 70 for $CV_W = 30\%$ and 40%, respectively. For N_1 = 24 the RSABE adjusted EMA method showed a similar trend as both TSD methods; and for N_1 = 36, both methods showed power above 80%, for a true CV_W below 60%. For a true $CV_W \geq 60\%$, the power for both TSD

methods seriously suffered from the futility criterion of not allowing studies with more than 150 subjects, though for the RSABE adjusted EMA the power was still above 80%.

Figure 4 explores the power for different true levels of bioequivalence: *GMR* = 0.95, 1.00, and 1.12. It is remarkable that for a true value of *GMR* = 1.12, no methods reached 80% power for any HVD with $CV_W \geq 30\%$.

4. Discussion

Bioequivalence studies are the pivotal clinical studies submitted to regulatory agencies to support the marketing applications of new generic drug products. High levels of within-subject variability make it difficult to assess bioequivalence through standard procedures using reasonable sample sizes, thus delaying treatment. After many years of discussion, some agencies issued regulations describing those methods. In general, their approach is based on bioequivalence limits being scaled as a function of the reference formulation variability. This is the reference scaled average BE (RSABE) approach of the EMA regulation issued in 2010 [2]. Although also mentioned in the regulations, adaptive two-stage designs (TSD) are not used nearly as much as the widespread scaling methods, despite having some appealing characteristics. Deciding on the study's experimental design is crucial and must be done in advance (e.g., including it in the study protocol), generally without full knowledge of the within-subject variability. We compared two variants of well-known adaptive methods and an RSABE adjusted (type I error) EMA approach. Both methods showed similar statistical power, but the RSABE adjusted scaled method required less sample size, although at the expense of exposing subjects twice as long as TSD methods. For initial sample sizes of at least 24 subjects, TSDs are a good option to consider, as they have a power of around 80% at the first stage for non-highly variable drugs while at the same time they offer the opportunity for stepping up to the second stage (including additional subjects) for truly bioequivalent products.

Statistical power is used to evaluate the performance of adaptive methodologies in ABE clinical trials. A power of at least 80% is desirable when considering N_1 subjects at the first stage, and assuming an expected but unknown within-subject coefficient of variation, *CVw*. In turn, this is always conditioned to not exceed the overall type I error rate of 0.05 for true bioinequivalent drugs. In our modified Potvin B and C methods, we found adjusted significance levels covering a wide range of N_1 and *CVw* combinations (i.e. α_{adj} = 0.03018396 and α_{adj} = 0.02806472 at each stage for Potvin B and C, respectively). This is useful to regulators since they can widely rely on the protection of patients against false positive results. However, we understand that for a specific actual (local) N_1 and *CVw* combination, the power might be slightly downgraded, although it is always above 80% in case of true bioequivalence.

Patterson et al. [21] explored the sample size that provides 90% power (for true bioequivalent drugs) in case of HVD. They showed that by using 2x2 crossover designs with conventional ABE limits of 0.8-1.25 and *CVw* of 60% or above, the required sample size exceeds 150 subjects (though replicate designs require smaller sample size). Using adaptive designs, we avoid conducting studies with such a large sample size by imposing a futility criterion so that we can stop the trial at an interim look with only N_1 subjects. According to Karalis and Macheras [19], we added a constraint to the original TSD methods, specifically by not recruiting more than 150 subjects overall. For

example, in the case of a true bioequivalent drug with $0.95 \leq \text{GMR} \leq 1.05$, and for highly variable drugs with an estimated within-subject coefficient of variation above 58% at the interim analysis, the final sample size needed for achieving a power of 80% at the second stage already exceeds 150 subjects. At first glance this constraint represents some global loss of power, but this possibility of cancelling a study for futility may ultimately be considered a positive trait, since the sponsor is unaware of the true treatment effect value during the planning phase, and the overall sample size could unnecessarily soar above this threshold for a scenario of true bioinequivalence. However, from an ethical perspective even starting a study with such a low expected power might be questionable [22].

Kieser et al. [15] and Karalis and Macheras [19] pointed out a potential limitation of the original TSD methods stated by Potvin et al. [17] and Montague et al. [13], as although unblinded data are available after the first stage, the knowledge about the estimated *GMR* in the interim analysis is not used for sample size recalculation. We assumed a fixed true treatment effect of *GMR* = 0.95 after the first stage since Cui et al. [23] showed that a determination of the second stage sample size based on an interim estimate of the *GMR* can substantially inflate the probability of type I error in most practical situations.

In addition, the expected total sample size *E[N]* is usually used to compare the performance characteristics of different TSD methods. However, by their very nature in TSD, the distribution of total sample sizes N is bimodal, mainly due to the imposition of $N \geq 1.5N_1$. For example, using our modified Potvin B, with $\alpha_{adj} = 0.03018396$ at each stage, *GMR* = 0.95, *CVw* = 0.3, N1 = 24, and target power 80%, we obtain a *E[N]* of 40 subjects, but with 24 and 36 subjects having more likelihood of occurrence (Figure 5). As the average is skewed towards two sample values, we believe that the median of *N* is more useful to compare different TSD methods.

In general, regulators allow using adaptive methods, though they usually favor sample size re-estimation procedures that maintain the blinding of the treatment allocations throughout the trial, as shown by Golkowski et al. [24]. However, even though both TSD Potvin B and C methods studied in this article assume unblinded data at the interim analysis, the agencies do specifically also recommend using these two TSD methods [2], as they have demonstrated that they control the type I error rate in a strong way.

So, given that either the RSABE or TSD methods are suitable approaches for ABE studies, we have compared them through the behavior of the type I error rate and its power to facilitate the discussion about which to choose. In terms of power, both approaches perform similarly despite both adaptive methods requiring a higher mean sample size to reach the same power, especially for clearly variable drugs. Nevertheless, they demonstrate suitable power at the first stage in some cases. However, as RSABE relies on replicate designs, double exposure of subjects is needed. The crucial point to consider is the assessment made by sponsors regarding the relative importance of the number of required subjects (an argument favoring the scaled approach) and the exposure of these subjects (which tips the balance in favor of the TSD approach).

The applicability of the TSD approaches is essentially the same as the classical approach, in that they have the same *RT/TR* design and fixed standard limits [25]. The

RSABE approaches (with type I error adjustment) are appropriate for drugs with low to moderate variability, because dose-to-dose variability within a patient is comparable to the width of the criteria. However, with HVD, dose-to-dose variability within a patient is greater than the width of the standard criteria, and it is usually characterized by flat dose response curves and wide safety margins. Therefore, broadening the acceptance limits in the RSABE approach is at the very least controversial, since clinically sound criteria should be used to clearly prove if a greater difference in *Cmax* (and also in *AUC* for the FDA) is irrelevant.

In conclusion, the RSABE approach is well powered and usually requires enrolling fewer patients than adaptive TSD methods, even though scaling the ABE limits ultimately depends on additional clinical judgment. For HVD in general, samples of 36 subjects provided well-powered studies using RSABE methods. As there is a considerable chance of declaring ABE at the first stage in adaptive approaches, sponsors should consider them because they imply less subject exposure and less treatment duration.

Acknowledgments

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

We would like to thank the reviewers who identified areas of the manuscript that needed corrections or modifications.

References

1. FDA. *Guidance for Industry: Bioavailability and Bioequivalence Studies submitted in NDAs or INDs - General considerations*. U.S. Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER): Rockville, MD, 2014. Available from: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm389370.pdf [Accessed on 18 October 2016].
2. EMA. *Guideline on the investigation of bioequivalence*. *CPMP/EWP/QWP/1401/98 Rev.1/Corr.* European Medicines Agency: London, 2010. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf [Accessed on 18 October 2016].
3. Tothfalusi L, Endrenyi L, Garcia Arieta A. Evaluation of bioequivalence for highly variable drugs with scaled average bioequivalence. *Clinical Pharmacokinetics* 2009; **48**(11):725-743. DOI:10.2165/11318040-000000000-00000.
4. FDA. *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence*. U.S. Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER): Rockville, MD, 2001. Available from: http://www.fda.gov/downloads/Drugs/.../Guidances/ucm070244.pdf [Accessed on 18 October 2016].
5. Schütz H. Two-stage designs in bioequivalence trials. *European Journal of Clinical Pharmacology* 2015; **71**(3):271-281. DOI:10.1007/s00228-015-1806-2.
6. Davit BM, Conner DP, Fabian-Fritsch B, Haidar SH, Jiang X, Patel DT, Seo PR, Suh K, Thompson CL, Yu LX. Highly variable drugs: observations from bioequivalence data submitted to the FDA for new generic drug applications. *The AAPS Journal* 2008; **10**(1):148-156. DOI:10.1208/s12248-008-9015-x.
7. FDA. *Guidance for Industry: Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA*. U.S. Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER): Rockville, MD, 2013. Available from: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm377465.pdf [Accessed on 18 October 2016].
8. FDA. *Draft Guidance on Progesterone*. U.S. Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER): Rockville, MD, 2011. Available from: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm209294.pdf [Accessed on 18 October 2016].
9. Muñoz J, Alcaide D, Ocaña J. Consumer's risk in the EMA and FDA regulatory approaches for bioequivalence in highly variable drugs. *Statistics in Medicine* 2016; **35**(12):1933-1943. DOI:10.1002/sim.6834.
10. Labes D, Schütz H. Inflation of Type I error in the evaluation of scaled average bioequivalence, and a method for its control. *Pharmaceutical Research* 2016; **33**(11):1-10.
11. EMA. *Questions & Answers: Positions on specific questions addressed to the pharmacokineticsworking party*. European Medicines Agency: London, 2015. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002963.pdf [Accessed on 18 October 2016].
12. Bandyopadhyay N, Dragalin V. Implementation of an adaptive group sequential design in a bioequivalence study. *Pharmaceutical Statistics* 2007; **6**(2):115-122. DOI:10.1002/pst.252.

13. Montague TH, Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ. Additional results for "Sequential design approaches for bioequivalence studies with crossover designs" *Pharmaceutical Statistics* 2012; **11**(1):8-13. DOI:10.1002/pst.483.
14. Davit B, Braddy AC, Conner DP, Yu LX. International guidelines for bioequivalence of systemically available orally administered generic drug products: a survey of similarities and differences. *The AAPS Journal* 2013; **15**(4):974-990. DOI:10.1208/s12248-013-9499-x.
15. Kieser M, Rauch G. Two-stage designs for cross-over bioequivalence trials. *Statistics in Medicine* 2015; **34**(16):2403-2416. DOI:10.1002/sim.6487.
16. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**(2):191-199. DOI:10.1093/biomet/64.2.191.
17. Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ, Smith RA. Sequential design approaches for bioequivalence studies with crossover designs. *Pharmaceutical Statistics* 2008; **7**(4):245-262. DOI:10.1002/pst.294.
18. Karalis V, Macheras P. On the statistical model of the two-stage designs in bioequivalence assessment. *Journal of Pharmacy and Pharmacology* 2014; **66**(1):48-52. DOI:10.1111/jphp.12164.
19. Karalis V, Macheras P. An insight into the properties of a two-stage design in bioequivalence studies. *Pharmaceutical Research* 2013; **30**(7):1824-1835. DOI:10.1007/s11095-013-1026-3.
20. Xu J, Audet C, DiLiberti CE, Hauck WW, Montague TH, Parr AF, Potvin D, Schuirmannh DJ. Optimal adaptive sequential designs for crossover bioequivalence studies. *Pharmaceutical Statistics* 2016; **15**(1):15-27. DOI: 10.1002/pst.1721.
21. Patterson SD, Zariffa N, Montague TH, Howland K. Non-traditional study designs to demonstrate average bioequivalence for highly variable drug products. *Eur J Clin Pharmacol.* 2001; **57**(9):663-70. DOI:10.1007/s002280100371.
22. Fuglsang A. Futility rules in bioequivalence trials with sequential designs. *The AAPS Journal* 2014; **16**(1):79-82. DOI: 10.1208/s12248-013-9540-0.
23. Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**(3):853-857. DOI:10.1111/j.0006-341X.1999.00853.x.
24. Golkowski D, Friede T, Kieser M. Blinded sample size re-estimation in crossover bioequivalence trials. *Pharmaceutical Statistics* 2014; **13**(3):157-162. DOI:10.1002/pst.1617.
25. EGA. *Revised EMA Bioequivalence Guideline: Questions and Answers. Summary of the discussions held at the 3rd symposium on bioequivalence.* European Generic Medicines Association: London, 2010. Available from: http://www.medicinesforeurope.com/wp-content/uploads/2016/03/EGA_BEQ_QA_WEB_QA_1_32.pdf [Accessed on 18 October 2016].

Table 1. Two-stage design (TSD) modified Potvin B and C: Bioequivalence, sample size, and percentage of studies stepping up to second stage for true *GMR* = 0.95, and under different fixed N_1 and a true *CVw*

		Modified Potvin B								Modified Potvin C							
Fixed a priori		ABE		Step to St2	*N*					ABE		Step to St2	*N*				
N_1	True *CVw*	% St1	% St1+St2	%	Min	5%	Me	95%	Max	% St1	% St1+St2	%	Min	5%	Me	95%	Max
12	20	41.92	85.00	55.69	12	12	18	40	104	41.56	84.76	54.44	12	12	18	40	106
12	30	7.03	78.61	92.71	12	12	44	84	150	6.40	78.34	93.05	12	12	44	84	150
12	40	1.03	71.65	95.68	12	22	70	128	150	0.90	70.96	95.28	12	20	72	130	150
12	60	0.05	29.43	51.00	12	12	44	142	150	0.05	27.76	49.06	12	12	12	142	150
24	20	83.76	90.16	8.20	24	24	24	36	62	87.89	91.19	4.22	24	24	24	24	64
24	30	41.86	83.86	57.47	24	24	36	70	138	40.47	83.38	57.69	24	24	38	72	140
24	40	10.12	79.79	89.45	24	24	76	118	150	8.93	79.44	90.49	24	24	78	120	150
24	60	0.19	31.19	46.47	24	24	24	146	150	0.15	28.83	43.59	24	24	24	146	150
36	20	95.68	95.75	0.07	36	36	36	36	54	97.51	97.51	0.01	36	36	36	36	54
36	30	68.13	87.23	28.33	36	36	36	60	120	69.94	85.77	22.95	36	36	36	62	124
36	40	34.32	82.42	65.54	36	36	68	110	150	32.40	82.14	67.16	36	36	72	112	150
36	60	1.53	31.28	42.66	36	36	36	146	150	1.20	28.35	39.37	36	36	36	146	150

ABE, average bioequivalence; TSD, two-stage design; *GMR*, geometric mean ratio; N_1, initial and fixed sample size (Stage 1); *CVw*, within-subject coefficient of variation; %St1, proportion of simulations declaring bioequivalence at Stage 1; %St1+St2, cumulative proportion of simulations declaring ABE at Stage 2, Step up to St2, proportion of simulations requiring stepping up from Stage1 to Stage 2; Min, min of *N*; 5%, percentile 5 of *N*; Me, median of N; 95%, percentile 95 of *N;* Max, max of *N*

Table 2. Probability of bioequivalence acceptance according to the regulatory reference scaled ABE (RSABE) EMA and an adjusted EMA method compared to two-stage designs (TSD) modified Potvin B and C (true *CVw* = 30%)

		Probability ABE acceptance			Type I error	
		True *GMR*				
	Method	**0.95**	**1.00**	**1.12**	**1.25**	**1.31**
RSABE method	Regulatory EMA (N_1 = 24)	0.896	0.963	0.631	0.085	0.021
	AdjEMA (N_1 = 24)	0.864	0.948	0.559	0.050	0.009
TSD method	Modified Potvin B (N_1 = 24 at Stage 1)	0.419	0.484	0.242	0.029	0.008
	Modified Potvin B (Stage 1 + Stage 2 with 36≤*N*≤150)	0.839	0.926	0.527	0.050	0.012
	Modified Potvin C (N_1 = 24 at Stage 1)	0.405	0.468	0.236	0.030	0.009
	Modified Potvin C (Stage 1 + Stage 2 with 36≤*N*≤150)	0.834	0.922	0.519	0.048	0.012

ABE, average bioequivalence; RSABE, reference scaled average bioequivalence; TSD, two-stage design; *GMR*, geometric mean ratio; *CVw*, within-subject coefficient of variation; N_1, initial and fixed sample size fixed at 24 subjects (Stage 1 with modified Potvin B and C); Regulatory EMA, regulatory European Medicines Agency approach; AdjEMA, adjusted EMA type I error

Legends

Figure 1. Type 1 TSD Modified Potvin B algorithm

Adapted from the figure depicted in detail by Montague, Potvin et al. [13], with the restriction of not including more than 150 subjects [18] and $N \geq 1.5N_1$;
ABE, average bioequivalence; N_1, Initial fixed sample size; N_2, the additional number of subjects recruited at Stage 2; *GMR*, assumed geometric mean ratio; $\widehat{CV}_W$, estimated within-subject coefficient of variation

Figure 2. Type 2 TSD Modified Potvin C algorithm

Adapted from the figure depicted in detail by Montague, Potvin et al. [13], with the restriction of not including more than 150 subjects [18] and $N \geq 1.5N_1$;
ABE, average bioequivalence; N_1, Initial fixed sample size; N_2, the additional number of subjects recruited at Stage 2; *GMR*, assumed geometric mean ratio; $\widehat{CV}_W$, estimated within-subject coefficient of variation

Figure 3. Bioequivalence acceptance of the adjusted reference scaled ABE (RSABE) EMA method and two-stage designs (TSD) modified Potvin B and C at stages 1 and 2, for a true *GMR* of 0.95, and a progressive increase of the within-subject variability

ABE, average bioequivalence; RSABE, reference scaled average bioequivalence; TSD, two-stage design; *GMR*, geometric mean ratio; HVD, highly variable drugs; N_1, initial and fixed sample size used for the modified EMA method and both TSD methods at Stage1; *CVw*, within-subject coefficient of variation; *Me[N]*, TSD media total sample size at Stage 2 (beside the squares in the figure); AdjEMA, type I error adjusted EMA method

Figure 4. Bioequivalence acceptance of the adjusted reference scaled ABE (RSABE) EMA method and two-stage designs (TSD) modified Potvin B for different levels of true bioequivalence and a progressive increase in the within-subject variability

ABE, average bioequivalence, RSABE, reference scaled average bioequivalence; TSD, two-stage design; HVD, highly variable drugs; N_1, initial and fixed sample size (EMA method); *GMR*, geometric mean ratio; *CVw*, within-subject coefficient of variation; *Me[N]*, TSD median total sample size (beside the squares in the figure); AdjEMA, type I error adjusted EMA

Figure 5. Type 1 TSD modified Potvin B distribution of *N* (Stag1 + Stage 2) *GMR*=0.95*; CVw*=30%; N_1=24; α_{adj} =0.03018396; *P*=0.8; *m*=1,000,000 simulations

GMR, true geometric mean ratio; *CVw*, true within-subject coefficient of variation; N_1, Initial fixed sample size; N_2, the additional number of subjects enrolled at stage 2; $N=N_1+N_2$, total sample size (stage 1 + stage 2); α_{adj}, significance level used in each stage; *P*. target power; *m*, number of simulations

Suplementary Material Not For Review is included:

1. TIE_estimating_CV.R; and 2. *TIE_estimating_CV_results.pdf*
3. *Modified Potvin_Alpha.R*; and 4. *potvin.R*

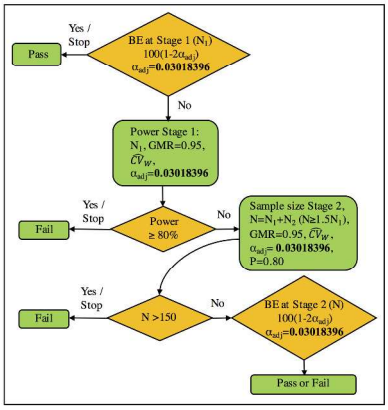

Figure 1. Type 1 TSD Modified Potvin B algorithm

312x220mm (300 x 300 DPI)

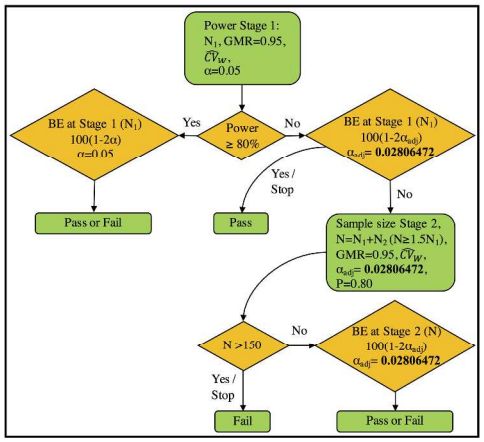

Figure 2. Type 2 TSD Modified Potvin C algorithm

338x190mm (300 x 300 DPI)

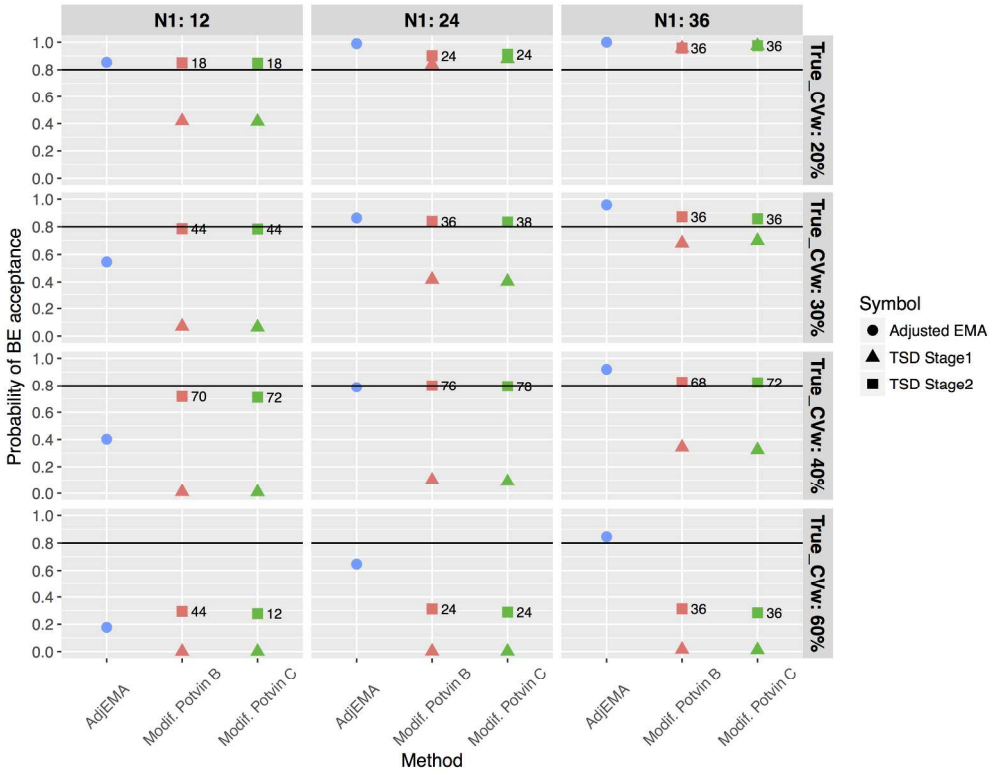

Figure 3. Bioequivalence acceptance of the adjusted reference scaled ABE (RSABE) EMA method and two-stage designs (TSD) modified Potvin B and C at stages 1 and 2, for a true GMR of 0.95, and a progressive increase of the within-subject variability

228x177mm (300 x 300 DPI)

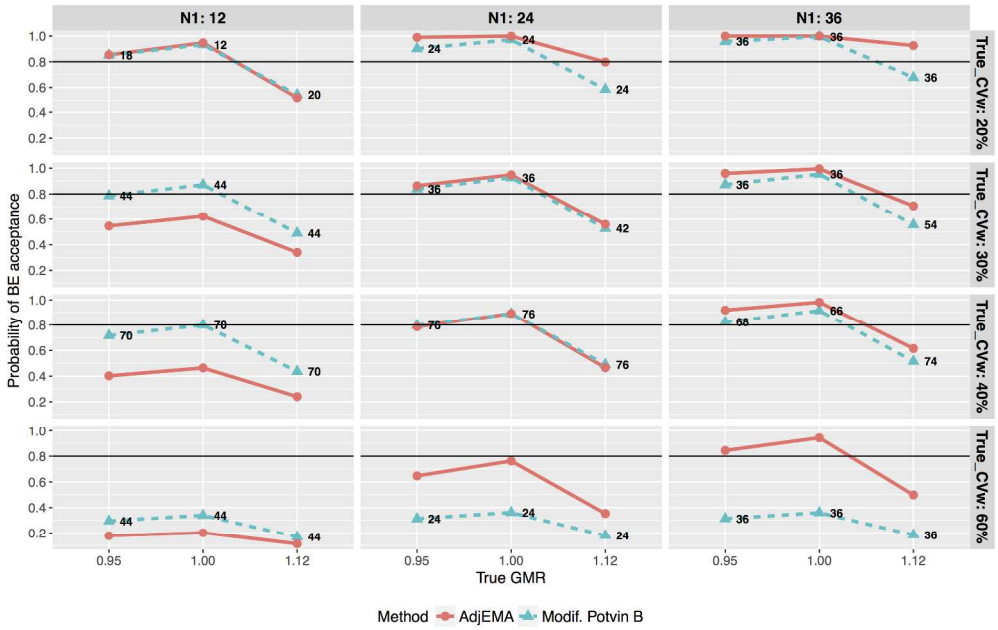

Figure 4. Bioequivalence acceptance of the adjusted reference scaled ABE (RSABE) EMA method and two-stage designs (TSD) modified Potvin B for different levels of true bioequivalence and a progressive increase in the within-subject variability

276x177mm (300 x 300 DPI)

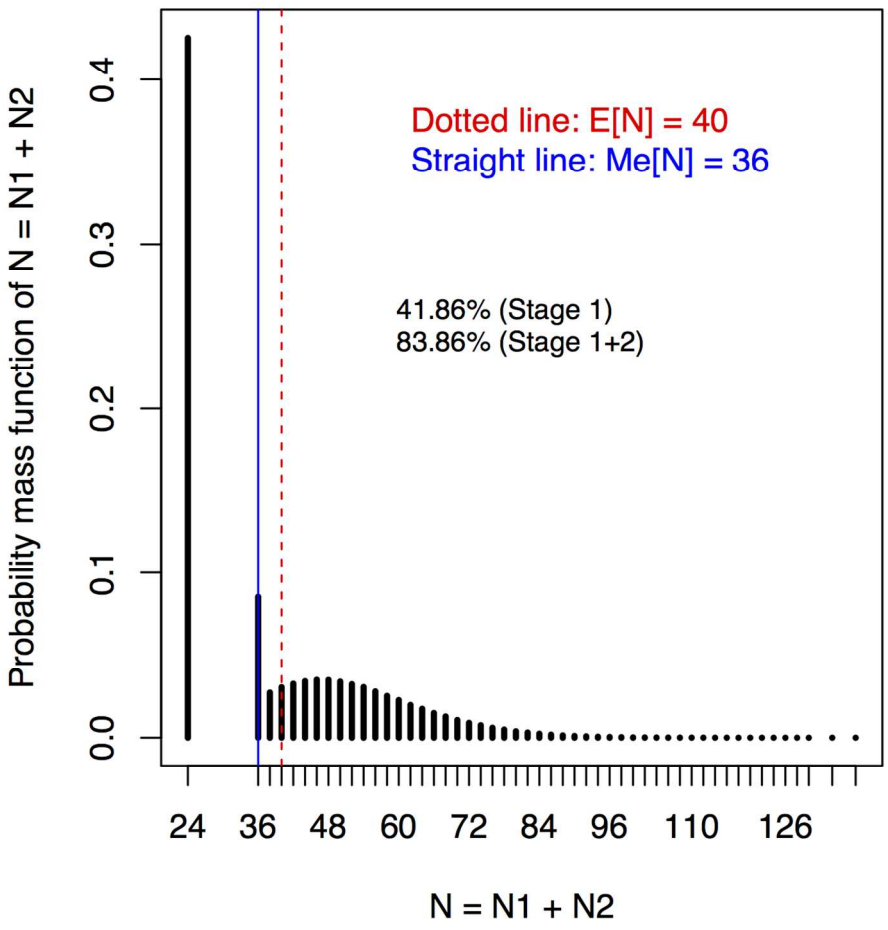

Figure 5. Type 1 TSD modified Potvin B distribution of N (Stag1 + Stage 2)
GMR=0.95; CVw=30%; N1=24; αadj =0.03018396; P=0.8; m=1,000,000 simulations

127x126mm (300 x 300 DPI)