Proposing some innovative study design features to regulatory agencies (EMA and FDA) in bioequivalence trials

Reference Scaled Average Bioequivalence, and Two-Stage Adaptive Designs

Eduard Molins Lleonart

Thesis director

Dr. Jordi Ocaña Rebull

Thesis tutor

Dr. Erik Cobo Valeri

Thesis submitted to obtain the title of Doctor by the
Universitat Politècnica de Catalunya.
Department of Statistics and Operations Research
Barcelona, March 6, 2021

Proposing some innovative study design features to regulatory agencies (EMA and FDA) in bioequivalence trials

Reference Scaled Average Bioequivalence, and Two-Stage Adaptive Designs

PhD in Biostatistics and Operations Research

Statistics and Operations Research Department, Universitat Politècnica de Catalunya

Doctoral thesis by

Eduard Molins Lleonart - Department of Statistics and Operations Research. Universitat Politècnica de Catalunya

Thesis director

Dr. Jordi Ocaña Rebull - Department of Genetics, Microbiology and Statistics - Statistics Section. Universitat de Barcelona

Thesis tutor

Dr. Erik Cobo Valeri- Department of Statistics and Operations Research. Universitat Politècnica de Catalunya

Barcelona, March 2021

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

FUNDING INFORMATION

This research was supported by: Ministry of Economy and Competitiveness (Spain) under grant MTM2015-64465-C2-1-R (MINECO/FEDER, UE); Generalitat de Catalunya under grant 2014 SGR 464; Ministry of Science and Innovation (Spain) under grant PID2019-104830RB-I00 from, DOI (AEI): 10.13039/501100011033.

AGRAÏMENTS

A finals del 2014, després de gairebé quinze anys dedicant-me a la recerca mèdica, la companyia biofarmacèutica AstraZeneca va adquirir els productes respiratoris d'Almirall i part de la seva plantilla, de la qual jo formava part. Començava una nova etapa de futur incert, i després de pensar-ho molt, vaig decidir començar el doctorat en estadística per complementar la meva formació acadèmica i professional.

Sempre estaré molt agraït al Dr. **Jordi Ocaña** perquè va acceptar ser el meu director de tesi. Als anys noranta va ser professor meu a la diplomatura d'estadística, i a principis del 2015, ja en el marc del doctorat, vaig tenir novament l'oportunitat de gaudir-lo com a professor d'una assignatura de bioequivalència. I des de llavors i durant els següents anys, vam passar moltes hores reunits discutint llargament sobre els dissenys escalats i adaptatius. En tinc molt bon records ja que vaig aprendre i fruir a parts iguals.

També vull agrair especialment aquesta tesi al meu tutor, al Dr. **Erik Cobo**, qui també va ser professor meu en assajos clínics durant la llicenciatura d'estadística. La seva experiència i dots comunicatives han estat claus en la discussió dels articles. Agraeixo les seves reflexions que sempre han estat pertinents i punyents.

Els agreixo el propòsit de continuar fent recerca tots tres un cop s'acabi aquesta tesi.

Una part clau d'aquesta tesi es deu a la qualitat de les revisions dels articles sotmesos en revistes internacionals. Després de la publicació del primer article a *Statistics in Medicine*, el Sr. **Helmut Schütz** ens va enviar un missatge de felicitació tot identificant-se com un dels revisors. Ell havia estat un dels autors referents més citats al nostre article, així que després de redactar el segon, el vam contactar per oferir-li autoria, cosa que va acceptar gratament. Per recomanació d'ell vam tenir la sort d'incorporar també al Sr. **Detlew Labes**. Als dos els vull agrair sincerament les hores dedicades a la millora de la metodologia i al codi de programació, essencials per a la publicació de l'article a *Biometrical Journal*.

Gràcies als companys del Departament d'Estadística i Investigació Operativa i al *GRBIO* per incloure'm al seu grup de recerca. Espero poder seguir formant part del vostre grup en el futur!

Per últim, gràcies a la meva família. Vull agrair a la **Julieta** que durant tots aquests anys sempre m'ha animat a continuar. Agraeixo el seu suport i opinions des de la doble perspectiva, d'economia i salut púbica. Gràcies per estar al meu costat! I també a la nostra estimada filla **Emma,** que va néixer enmig d'aquesta tesi, i que ara, en temps de pandèmia, estem gaudint dia rere dia! Gràcies als meus **pares, germà** i **germana** que sempre han estat al meu costat. Gràcies a tots!

ACKNOWLEDGMENTS

By the end of 2014, after nearly fifteen years working in medical research, biopharmaceutical company AstraZeneca acquired Almirall's respiratory pipeline as well as part of the staff of which I was a member. I began an uncertain new stage and so, after giving it a lot of thought, I decided to start a doctoral research in Statistics which would complement my academic and professional career.

I much appreciate that Dr. **Jordi Ocaña** agreed to be my thesis director. In the 1990s he was my teacher during the *BSc* in Statistics, and at the beginning of the year 2015, as part of this thesis, I enjoyed his lectures in a bioequivalence course. Since then and for the following years, we spent many hours in long meetings where we discussed over the scalded and adaptive designs. I have very good memories as I learned and enjoyed from Dr. **Jordi Ocaña** in equal parts.

I also want to especially thank my tutor, Dr. **Erik Cobo**, who was also my teacher in Clinical Trials during the Bachelor's degree in Statistics. His experience and communicative skills have been key in the discussion of articles. I appreciate his thoughts that have always been relevant and sharp.

I appreciate the proposal to continue to research all three once this thesis is finished.

A key part of this thesis is the quality of the reviews of the articles submitted in international journals. After the publication of the first article in *Statistics in Medicine,* Mr. **Helmut Schütz** sent us a message of congratulations while identifying himself as one of the reviewers. He had been one of the most cited authors in our article, so after writing the second one, we contacted him to offer a co-authorship, something he willingly accepted. On his recommendation, we were lucky to also include Mr. **Detlew Labes** as a co-author. I would like to sincerely thank them for the hours dedicated to improving methodology and programming, essential for the publication of the article in *Biometrical Journal*.

Thanks to colleagues from the *Department of Statistics and Operational Research* and *GRBIO* for including me in the research group. I look forward to continuing to be part of your group in the future!

Finally, I would like to thank my family. I want to thank **Julieta** who always encouraged me to continue during these years. I appreciate her support and opinions from a dual perspective, Economics and Public Health. Thank you for staying with me! And also to our beloved daughter **Emma**, who was born during this thesis preparation, and in times of pandemic, we are seeing her growing day after day! Thank you to my **parents, brother and sister** who have always been by my side. Thank you all!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

RESUM

En aplicacions per a medicaments genèrics el concepte de bioequivalència és fonamental. Dos productes, un 'test' i un de 'referència', amb el mateix principi actiu, es consideren bioequivalents si la seva biodisponibilitat (quantitat 'C_{max}' i velocitat 'Tmax' d'una substància activa que s'absorbeix d'un fàrmac i està disponible en el seu lloc d'acció) després de l'administració d'ambdós productes produeix un efecte terapèutic similar. Per això, l'interval de confiança del 90% per a la ràtio de les mitjanes (mitjanes geomètriques poblacionals) dels productes test i referència de les mesures farmacocinètiques han d'estar dins dels límits de bioequivalència 80%-125%. Es recomana utilitzar dissenys aleatoritzats encreuats 2x2, és a dir, de dos períodes i dues seqüències (en anglès 2x2 crossover designs).

El nombre de subjectes que s'inclouen es basa en un càlcul adequat de la grandària mostral, tot i que aquest nombre sol ser petit però mai inferior a 12 subjectes.

Però en cas de productes/fàrmacs d'alta variabilitat cal incloure molts més subjectes per aconseguir una potència estadística adequada, de manera que la bioequivalència es determina amb pocs subjectes però a través de l'escalat dels límits de bioequivalència (*RSABE*, 'Reference Scaled Average Bioequivalence'), expandits en funció de la variabilitat intra-subjecte en el grup de referència. En aquest cas, amb dissenys 2x2 no és possible estimar per separat la variabilitat dels productes test i referència i cal fer servir dissenys més complexos com ara dissenys encreuats replicats o semi-replicats.

Les agències reguladores també permeten utilitzar dissenys encreuats 2x2 adaptatius de dues etapes amb re-estimació de la grandària mostral en la primera (anàlisi provisional, en anglès interim analysis). Llavors, si no podem declarar bioequivalència a la primera etapa amb una grandària mostral inicial petita, podem incrementar la mostra en funció de la variabilitat intra-subjecte estimada i afegir nous subjectes en la segona, o parar l'estudi per futilitat si la probabilitat de declarar bioequivalència és finalment petita. Aquesta estratègia ha d'estar definida en el protocol, i prèviament acordada amb les agències reguladores amb especial èmfasi en el control de l'error de tipus I.

Mitjançant simulacions de Monte Carlo, mostrem que les metodologies basades en *RSABE* i dissenys adaptatius bietàpics proporcionen una potència estadística similar, tot i que els mètodes escalats normalment requereixen menys grandària mostral tot i que cal exposar més vegades els subjectes als tractaments. Amb una grandària mostral inicial adequada (no molt petita, per exemple 24 subjectes), els dissenys bietàpics són una opció molt flexible i eficient a considerar: proporcionen una potència raonable (per exemple del 80%) a la primera etapa per fàrmacs que no són altament variables, i en cas contrari, proporcionen l'oportunitat de saltar a una segona etapa que inclou subjectes addicionals.

Basant-nos en aquests dissenys adaptatius bietàpics, presentem un mètode iteratiu per ajustar el nivell de significació a cada etapa que preserva l'error de tipus I global per a un conjunt d'escenaris que molt probablement inclouen el vertader valor desconegut de la variabilitat intra-subjecte, i que proporciona una potència estadística d'almenys el 80%. Aquests dissenys funcionen particularment bé per coeficients de variació per sota de 0.3 pel balanç que proporcionen entre la potència estadística i el percentatge d'estudis que salten a la segona etapa. Presentem un paquet d'*R* que ens permet ajustar els nivells de significació a cada etapa i que controla l'error de tipus I global.

RESUMEN

En aplicaciones para medicamentos genéricos el concepto de bioequivalencia es fundamental. Dos productos, uno 'test' y uno de 'referencia', con el mismo principio activo, se consideran bioequivalentes si su biodisponibilidad (cantidad 'C_{max}' y velocidad 'Tmax' de una sustancia activa que se absorbe de un medicamento y está disponible en su lugar de acción) después de la administración de ambos productos produce un efecto terapéutico similar. Para ello, el intervalo de confianza del 90% para el ratio de las medias (medias geométricas poblacionales) de los productos test y referencia de las medidas farmacocinéticas tienen que estar dentro de los límites de bioequivalencia 80%-125%. Se recomienda utilizar diseños aleatorizados cruzados 2x2, de dos períodos y dos secuencias (en inglés 2x2 crossover designs).

El número de sujetos que se incluyen se basa en un cálculo adecuado del tamaño de muestra, aunque este número suele ser pequeño pero nunca inferior a 12 sujetos.

Pero en el caso de productos/medicamentos de alta variabilidad es necesario incluir muchos más sujetos para conseguir una potencia estadística adecuada, de forma que la bioequivalencia se determina con pocos sujetos pero a través del escalado de los límites de bioequivalencia (*RSABE*, 'Reference Scaled Average Bioequivalence'), expandidos en función de la variabilidad intra-sujeto en el grupo de referencia. En este caso, con diseños 2x2 no es posible estimar por separado la variabilidad de los productos test y referencia y se requieren diseños más complejos como diseños cruzados replicados o semi-replicados.

Las agencias reguladoras también permiten usar diseños cruzados 2x2 adaptativos de dos etapas con re-estimación del tamaño del a muestra en la primera (análisis provisional, en inglés interim analysis). Entonces, si no podemos declarar bioequivalencia en la primera etapa con un tamaño de muestra inicial pequeño, podemos incrementar la muestra en función de la variabilidad intra-sujeto estimada y añadir nuevos sujetos en la segunda, o parar el estudio por futilidad si la probabilidad de declarar bioequivalencia es finalmente pequeña. Esta estrategia se define en el protocolo, previo acuerdo con las agencias reguladoras con especial énfasis en el control del error de tipo I.

Mediante simulaciones de Monte Carlo, mostramos que las metodologías basadas en *RSABE* y diseños adaptativos bietápicos proporcionan una potencia estadística similar, aunque los métodos escalados habitualmente requieren menos tamaño de muestra aún siendo necesario exponer más veces al sujeto a los tratamientos. Con un tamaño de muestra inicial adecuado (no muy pequeño, por ejemplo 24 sujetos), los diseños bietápicos son una opción muy flexible y eficiente a considerar: proporcionan una potencia razonable (por ejemplo del 80%) en la primera etapa para medicamentos que no son altamente variables, y en caso contrario, proporcionan la oportunidad de saltar a una segunda etapa e incluir sujetos adicionales.

Basándonos en éstos diseños adaptativos bietápicos, presentamos un método iterativo para ajustar el nivel de significación en cada etapa que preserva el error de tipo I global para un conjunto de escenarios que muy probablemente incluyen el verdadero valor desconocido de la variabilidad intra-sujeto, y que proporciona una potencia estadística de al menos el 80%. Estos diseños funcionan particularmente bien para coeficientes de variación por debajo de 0.3 dado el balance que proporcionan entre la potencia estadística y el porcentaje de estudios que saltan a la segunda etapa. Presentamos un paquete de *R* que nos permite ajustar los niveles de significación en cada etapa y que controla el error de tipo I global.

ABSTRACT

In applications for generic medicinal products the concept of bioequivalence is fundamental. Two medicinal products, i.e. a test and a reference drugs containing the same active substance are considered bioequivalent if their bioavailability (rate and extent of absorption of an active substance that is absorbed from a drug product and becomes available at the site of action) after the administration of both products produce a similar therapeutic effect. The assessment of bioequivalence is based upon 90% confidence intervals for the ratio of the population geometric means (test/reference) for the parameters under consideration which should be contained within the limits 80%-125%. It is recommended using randomized, two-period, two-sequence, single dose crossover designs (2x2 crossover designs).

The number of subjects to be included should be based on an appropriate sample size calculation, though the number of evaluable subjects should not be less than 12.

Sometimes, there are drugs whose rate and extent of absorption is highly variable dose to dose within the same subject. The main problem with highly variable drugs is that to declare bioequivalence it requires a study with an unacceptably larger sample size. In this case, the usual approach to determine bioequivalence is 'Reference Scaled Average Bioequivalence' (*RSABE*), which is based on expanding the limits as a function of the within-subject variability in the reference formulation. But, using 2x2 crossover designs, it is not possible to estimate separately the test and reference variabilities, and thus it requires using more complex designs like replicated or semi-replicated crossover designs.

On the other hand, regulations also allow using common 2×2 crossover designs based on two-stage adaptive designs (*TSD*) with sample size re-estimation at an interim analysis. At an interim look (stage 1), if average bioequivalence is not declared with an initial sample size, they allow to increase it based on the intra-subject estimated variability and to enroll additional subjects at a stage 2, or to stop for futility in case of poor likelihood of bioequivalence. This is crucial because both parameters must clearly be pre-specified in protocols, and the strategy agreed with regulatory agencies in advance with emphasis on controlling the overall type I error.

Using Monte Carlo simulations, we show that *RSABE* and *TSD* 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), *TSD*s are a flexible and efficient option to consider: They have enough power (e.g., 80%) at the stage 1 for non-highly variable drugs and, if otherwise, they provide the opportunity to step up to a stage 2 that includes additional subjects.

Based on *TSD*s, we also present an iterative method to adjust the significance levels at each stage which preserves the overall type I error for a wide set of scenarios which should include the true unknown variability value, and which provides a power of at least 80%. *TSD*s work particularly well for coefficients of variation below 0.3 which are especially useful due to the balance between the power and the percentage of studies proceeding to stage 2. We present an *R* package to adjust the significance levels at each stage in order to control the overall type I error.

1. INTRODUCTION

1.1. Development of a novel drug

A novel/innovator drug is one that contains an active ingredient that has not yet been approved. Any novel drug goes through different development phases where thousands of patients are included in clinical trials performing a full collection of safety and efficacy information. Property rights (patents) are granted during the development of a drug which expire 20 years from the date on which the application for the patent was applied (1).

A development period may last more than 10 years, and when the commercialization of a new drug is authorized, negotiations to stablish pricing and reimbursement begins at country level, between sponsors and local governments. From this point and up to the patent expiration, sponsors have all rights and selling exclusivities.

1.2. Generic drugs

When the patent on a novel drug nears expiration, drug companies (sponsors) that want to manufacture a copy drug can apply to the regulatory agencies to sell a generic version of the drug.

The development cost of a generic drug is much lower than a novel drug because it is not necessary to pass through all the development phases, and usually are based on the evaluation of the bioequivalence, assessed only in some tens of healthy volunteers. When generic drugs enter the market, the supply extends and drug prices fall making easier access to consumers. Currently, generic drugs represent 65% of all the US medical prescriptions (2).

1.3. Regulatory evaluation of the bioequivalence of generic drugs

The details of the evaluation process for bioequivalence can slightly differ between main jurisdictions such as the EU (EMA), US (FDA), China (NMPA) and Japan (PMDA) but the general approach is the same.

A generic drug is essentially a duplicate of an approved novel drug. There may be differences in the way a generic and innovator drug look (e.g., size, shape, color), but

they are expected to share the same active ingredients, strength, safety, effectiveness, and quality characteristics (3).

The approval pathways for generic drugs are shrank to encourage a quicker time to market. They allow applicants to use existing knowledge instead of performing a full collection of safety and efficacy studies, as would be required for an innovator drug.

When a sponsor submits a generic drug for marketing approval, they submit an Abbreviated New Drug Application (ANDA) instead of a full NDA. In an ANDA, the applicant is claiming that their drug is a duplicate of an already-approved drug.

Within the ANDA, the innovator drug is specified. Innovator drug is commonly referred to as the "Reference Listed Drug, *RLD*" (*RLDs* are listed in FDA's electronic Orange Book). Because the Agency has already approved the *RLD* to be safe and effective, the goal of an ANDA is to demonstrate "sameness" with the *RLD*.

Sameness is demonstrated via a bioequivalence assessment where differences in systemic drug exposures in test and *RLD* are considered not clinically important. According to regulatory section *505(j)*, the assessment of bioequivalence is based upon 90% confidence intervals for the ratio of the population geometric means (test/*RLD*) for the rate and extent of absorption parameters (see section 1.4).

In turn, drugs approved under an ANDA must be therapeutically equivalent to the *RLD*. This means that if any ANDA-approved drug is exchanged with the *RLD*, patients should experience the same clinical effect and safety profile. The active ingredient, dosage form, route of administration, and strength must all be the same, and the product labeling is usually the same (Table 1) (4).

In addition, the sameness requirement also extends to the inactive ingredients in a generic drug product. An inactive ingredient is any component of a drug product other than the active ingredient, i.e., preservatives, buffers, and antioxidants.

Because a generic drug is intended to act as a duplicate of the *RLD*, no new safety or efficacy studies are performed and only small confirmatory studies are allowed to support the ANDA. The precise scope and type of information necessary for approval will vary and may be the subject of discussion between the applicant and regulatory agency during the drug development process. If additional preclinical or clinical data

are needed to support safety or efficacy, then the ANDA route is not appropriate and a *505(b)(2)* NDA should be pursued.

Table 1. Requirements for NDA vs. ANDA

Brand name drug requirements (NDA)	**Generic drug requirements (ANDA)**
Chemistry (physical and chemical characteristics)	Chemistry (physical and chemical characteristics)
Manufacturing (residues and addition impurities)	Manufacturing (residues and addition impurities)
Controls (inspections)	Controls (inspections)
Labeling (prescribing information)	Labeling (prescribing information)
Dissolution testing	Dissolution testing
Animal studies	Bioequivalence
Clinical Studies	
Bioavailability	

NDA: New Drug Application
ANDA: Abbreviated New Drug Application

1.4. Average bioequivalence

Average bioequivalence (ABE) studies typically involve testing two products, test (potential generic drug), *T*, and reference (already marketed novel drug), *R*, against each other. Usually, these studies are based on the usual 2×2 crossover *RT/TR* design involving just some tens of healthy volunteers (Table 2) (5).

Table 2. 2x2 Crossover design

	Period	
Sequence	$j = 1$	$j = 2$
$k = 1$	Y_{111} Y_{211} ..... Y_{n11} } R	Y_{121} Y_{221} ..... Y_{n21} } T
$k = 2$	Y_{112} Y_{212} ..... Y_{n12} } T	Y_{122} Y_{222} ..... Y_{n22} } R

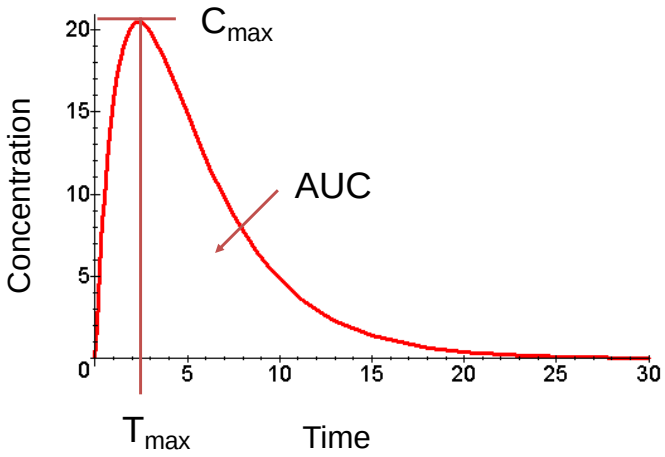
Note: Subjects are randomly assigned to sequence 1 or 2

Where Y_{ijk} is the *log* of the bioavailability measure (subject *i*, period *j*, sequence *k*)

Bioavailability is usually determined by some pharmacokinetic measurements that can be estimated from the blood or plasma concentration-time curve obtained following drug administration (6). Primary pharmacokinetic metrics are C_{max}, maximum observed

plasma concentration (rate), and the area under the concentration time curve (extent), $AUC_{0\text{-}t}$ and $AUC_{0\text{-}\infty}$, Figure 1 (7,8).

Figure 1. Concentration (µg/mL) of drug as a function of time (min.)



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

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

$$H_1: -\delta < \phi < +\delta, \text{ i.e. } |\phi| < \delta,$$

where ϕ is the difference between population bioavailability means $\mu_T - \mu_R$ (in log scale) of *T* and *R* (treatment effect), and usually $\delta = log(1.25) = 0.223$, or equivalently, the back exponentially transformed geometric mean ratio, $GMR = e^{\phi}$ should lie fully within 0.80 to 1.25 (=1/0.80). The basis for the 0.8-1.25 range is arbitrary. The FDA (and other regulatory bodies) 'decided' by consensus that differences in systemic drug exposure up to 20% are not clinically important.

Schuirmann *et al.* (9) proposed conducting 'Two One Sided Tests' (TOST) at significance level, α.

$$H_{01}: \phi \leq -\delta \text{ vs. } H_{11}: \phi > -\delta, \text{ and,}$$

$$H_{02}: \phi \geq +\delta \text{ vs. } H_{12}: \phi < +\delta.$$

The estimation of the treatment effect ϕ is based on the difference contrast $d_{ik} = \frac{1}{2}(Y_{i2K} - Y_{i1K})$, accounting for the within-subject bioavailability measure between

period one and two. In absence of carryover, an unbiased estimator of ϕ is $\overline{D}$, with $\overline{D} = \bar{d}_1 - \bar{d}_2$, being $\bar{d}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} d_{ik}$, *k*=1,2.

So, provided that $T = \frac{\overline{D} - \phi}{SE_{\overline{D}}} \sim t(N-2)$ the TOST procedure may be implemented as two one-sided *t* tests for H_{01} and H_{02}. To declare ABE both null hypotheses must be rejected at a significance level α. This ensures a test of level α for H_0.

$SE_{\overline{D}} = \hat{\sigma}_d\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is the standard error estimate of $\overline{D}$, where $\hat{\sigma}_d$ is the estimator of the standard deviation of d_{ik}, and n_1 and n_2 are the number of patients included in each sequence.

Analogously, an alternative way of assessing the equivalence test problem is based on the "interval inclusion rule". To declare bioequivalence (i.e., to reject the null hypothesis of bioinequivalence) at a significance level α = 0.05, based on a normal *ln*-linear model, the two-sided 1 – 2α = 0.9 symmetric confidence intervals for $\mu_T - \mu_R$, ϕ, should lie fully within the constant bioequivalence limits of ±0.223, or equivalently, the back exponentially transformed confidence interval for the geometric mean ratio, $GMR = e^{\phi}$ should lie fully within 0.80 to 1.25 (=1/0.80) (7,10).

1.5. ANOVA model for 2x2 crossover design

The pharmacokinetic parameters under consideration are usually modeled using ANOVA. The data should be transformed prior to analysis using a logarithmic transformation. A confidence interval for the difference between formulations on the log-transformed scale is obtained from the ANOVA model. This confidence interval is then back-transformed to obtain the desired confidence interval for the ratio on the original scale (11).

Bioavailability measures are usually modeled through the following linear statistical model:

$$Y_{ijk} = \mu + S_{i(k)} + P_j + F_{(j,k)} + C_{(j-1,k)} + e_{ijk,}$$

where:

- Y_{ijk} is the *log* of the bioavailability measure (subject *i*, period *j*, sequence *k*).
- μ overall mean.
- $S_{i(k)}$ random effect of subject *i* within the sequence *k=1,2*. Accounts for the inter-subject variability, $S_{i(k)} \sim N(0, \sigma_B^2)$.
- P_j is the period fixed effect, *j=1,2.*
- $F_{(j,k)}$ fixed effect of treatment at period *j* and sequence *k.*
- $C_{(j-1,k)}$ residual fixed effect (carryover) of period *j-1* within sequence *k*: C(1,1) = C_R, C(1,2) = C_T.
- e_{ijk} residual, accounts for the intra-subject variability, and $e_{ijk} \sim N(0, \sigma_W^2)$.
- $S_{i(k)}$ and e_{ijk} are mutually independent.

In this model 'subject' is specified as a random effect and so there are two variance terms (within and between) designated as σ_W^2 and σ_B^2.

1.6. Highly variable drugs and scaled methods

Sometimes, there are drugs whose rate and extent of absorption is highly variable dose to dose within the same patient *(HVD)*. Within-subject variability refers to variability in a response (e.g., plasma drug concentration) within the same subject, when the subject is administered two doses of the same drug on two different occasions. Most regulations classify a drug as *HVD* if the within-subject coefficient of variation of the reference formulation *R* is 30% or greater on the original scale.

The main problem with *HVD* is that to declare bioequivalence it requires a study with an unacceptably larger sample size.

In 2003-2005, the Open Government Data, *OGD*, reviewed 1,010 acceptable bioequivalence studies of 180 different drugs, of which 31% (57/180) were highly variable (12).

1.6.1. EMA approach on average bioequivalence in highly variable drugs

If *HVD* is suspected, the EMA allows linearly scaling the C_{max} margins as a function of the *R* variability (σ_{WR} or CV_{WR}). To declare bioequivalence, we need to estimate $\hat{\sigma}_{WR}$ using higher order crossover designs, e.g. replicate *TRTR/RTRT* or semi-replicate

TRR/RTR/RRT. Three models for analyzing data are considered, EMA's *Methods A* and *B* (11) and FDA's *Method C* (13):

EMA - *Method A* uses the same analysis method for replicate designs as is used for 2×2 crossover trials, and considers that subject is a fixed effect and each subject is treated as being selected in some way rather than being sampled from a random distribute. For this model there is only one variance term estimated, σ_W^2, the within subject variability.

EMA - *Method B* considers the same model as specified above but where subject is specified as a random effect and so there are two variance terms (within and between) estimated σ_W^2 and σ_B^2. Both models give the same results if all subjects included in the analysis provide data for all treatment periods.

FDA - *Method C* allows a different subject effect for each formulation (i.e. a subject by formulation interaction), and therefore has 5 variance terms (within subject for reference, within subject for test, between subject for test, between subject for reference, covariance for between subject test and reference – the last three are combined to give the subject x formulation interaction variance component). This model will provide the same point estimate as *Methods A and B* if all subjects provide data for all treatment periods. However, it will generally give wider confidence intervals than those produced by *Methods A and B*.

In the EMA 2010 regulation (7) the Reference Scaled Average bioequivalence limits (*RSABE*) for C_{max} are specified as follows (Figure 2):

- First, constant limits, the usual $\phi(\sigma_{WR})$ = ±0.223 for $\sigma_{WR} < 0.2935$ (corresponding to C_{WR} below 30%);
- Next, scaled limits, $\phi(\sigma_{WR}) = \pm k_{EMA}\sigma_{WR}$, for $0.2935 \leq \sigma_{WR} < 0.4724$ (from C_{WR} of 30% to C_{WR} of 50%);
- Finally, constant limits $\phi(\sigma_{WR})$ = ±0.3590 from $\sigma_{WR} \geq 0.4724$ (from C_{WR} of 50%),

where k_{EMA} = 0.760 = log(1.25)/0.2935.

To declare bioequivalence:

- Estimate the parameter $\hat{\sigma}_{WR}$.
- Point estimate constraint: $\hat{\phi}$ must be within the limits ±0.223.
- At a significance level α = 0.05, based on a normal *ln*-linear model, the two-sided $1 - 2\alpha$ = 0.9 symmetric confidence intervals for $\hat{\phi}$ should lie fully within the constant bioequivalence scaled limits: $\pm\phi(\hat{\sigma}_{WR})$.

To estimate $\hat{\sigma}_{WR}$ higher order crossover designs are needed, e.g. *TRTR/RTRT or TRR/RTR/RRT*. The replicate design has the advantage of using fewer subjects although each subject should receive more treatments than in the two-treatment, crossover design (3,7,14-17).

The table below gives examples of how different levels of variability lead to different acceptance limits in the original scale, where $CV_{WR} = \sqrt{e^{\sigma^2_{WR}} - 1}$ (Table 3).

Table 3. EMA scaled limits in the original scale

Within-subject CV_{WR} (%)	Lower Limit	Upper Limit
30	80.00	125.00
35	77.23	129.48
40	74.62	134.02
45	72.15	138.59
>=50	69.84	143.19

1.6.2. FDA approach on average bioequivalence in highly variable drugs

The FDA also allows researchers to re-scale the C_{max} and *AUC* limits in case of *HVD* (8). As in the case of the EMA, FDA bioequivalence limits are of ±0.223 for $\sigma_{WR} < 0.2935$, and scaled limits are applied if $\sigma_{WR} \geq 0.2935$, thus, the scaled approach does not have an upper bound limit. The FDA scaling constant is k_{FDA}= log(1.25)/0.25 = 0.892 and thus the scaled limits are discontinuous at CV_{WR} = 30% at original scale (i.e., at σ_{WR} = 0.2935 in logarithmic scale). On the other hand, it is worth pointing that this interpretation of the scaled FDA limits is not universally accepted (18), an alternative definition puts the starting point of scaling at σ_{WR} = 0.25 which avoids any discontinuity (Figure 2).

The decision procedure may be described as follows: The null hypotheses:

H_0: $|\phi| \geq 0.223$ if $\sigma_{WR} < 0.2935$ (or 0.25 according to the alternative interpretation)

and $|\phi| \geq k_{FDA}\sigma_{WR}$ if $\sigma_{WR} \geq 0.2935$ (or 0.25)

is reformulated in its scaling region using a new parameter τ (Howe's method) (19,20):

$$\tau = \phi^2 - k_{FDA}^2\sigma_{WR}^2 \geq 0,$$

provided that $-k_{FDA}^2\sigma_{WR} < \phi < +k_{FDA}^2\sigma_{WR}$,

where $\tau = \phi^2 - k_{FDA}^2\sigma_{WR}^2 = \sum c_j\theta_j$ is a linear combination with c_1=1, $\theta_1 = \phi^2$, $c_2 = k_{FDA}^2$, $\theta_2 = \sigma_{WR}^2$.

Then, considering the 95% confidence interval (CI) for the parameter τ $(-\infty, \tau_U)$, we declare bioequivalence if $\tau_U < 0$.

To obtain τ_U:

- Estimate the parameter $\hat{\sigma}_{WR}$.
- Obtain two point estimates: $E_1 = c_1\hat{\theta}_1 = \hat{\phi}^2$ and $E_2 = c_2\hat{\theta}_2 = -k_{FDA}^2 * \hat{\sigma}_{WR}^2$.
- Calculate upper limits of the $1-\alpha$ one-sided 95% CI for E_1 and E_2 as follow:

$$U_1 = (|\hat{\phi}| + t_{1-\alpha,N-s}se_{\hat{\phi}})^2$$

$$U_2 = \frac{k_{FDA}^2\sigma_{WR}^2(N-s)}{\chi_{1-\alpha,N-s}^2},$$

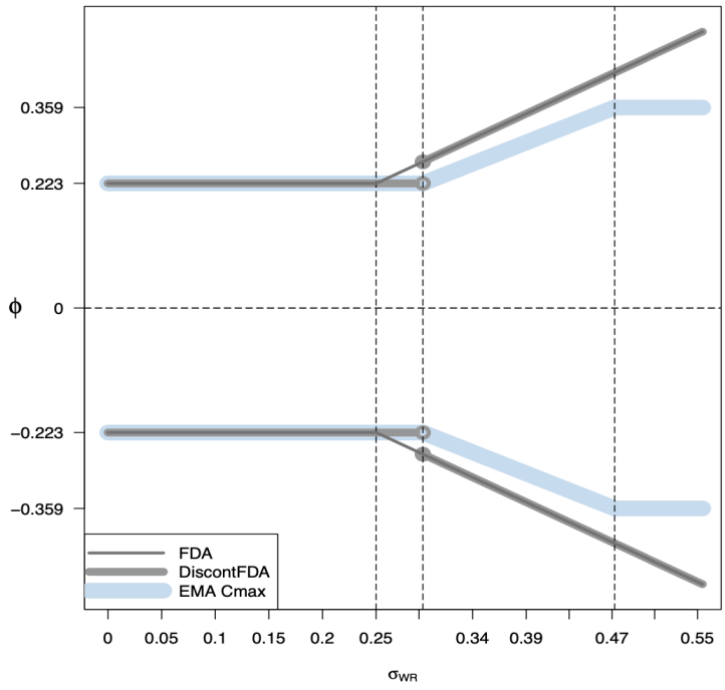
  where *N* is the total sample size, and *s* is the number of sequences.
- Calculate $D_j = (U_j - E_j)^2$.
- Then, $U_\tau = \sum E_j + \sqrt{D_j}$ is the upper limit of an approximate one-sided $1-\alpha$ confidence interval for τ.

Based on the confidence interval inclusion rule, the upper limit of the confidence interval (-∞,U_τ], U_τ, for the parameter τ, at a one-sided confidence level $1-\alpha$, should lie entirely < 0 to declare bioequivalence.

Irrespectively of the scaled equivalence limits being considered (either discontinuous or continuous), in practice, Howe's method is applied only if the estimation of σ_{WR} from data is greater or equal to 0.2935, otherwise the standard limits ±0.223 and the standard bioequivalence criterion are applied.

Again, this scaled approach requires the use of high order crossover designs like the replicated *TRTR/RTRT* or semi-replicated *TRR/RTR/RRT* design (3,7,14).

Figure 2. Bioequivalence limits according to the EMA and the US FDA regulations, scaled in function of the within-subject variability of the reference *R* formulation



Source: Ocaña J., Muñoz, J. (2019) (18)

1.7. Two-stage adaptive designs

Regulators also allow using two-stage adaptive designs (*TSD*) with unblinded interim sample size re-estimation based on the usual 2×2 crossover *RT/TR* design with bioequivalence limits 0.80-1.25 (in the original scale), whose application is becoming increasingly popular. This design is also useful for *HVD* (7,10,11,21-25).

The study starts with a few number of healthy volunteers (e.g. 12 subjects), and at an interim look, if it is not possible to declare bioequivalence, but the results are promising, they allow to add some new subjects on a stage 2 based on the variability observed in the stage 1 (interim), so finally increasing the likelihood of declaring bioequivalence at a stage 2 with cumulated data.

*TSD*s provide investigators with an attractive solution to address some of the uncertainty that exists when the trial is originally designed (26), allowing stopping the

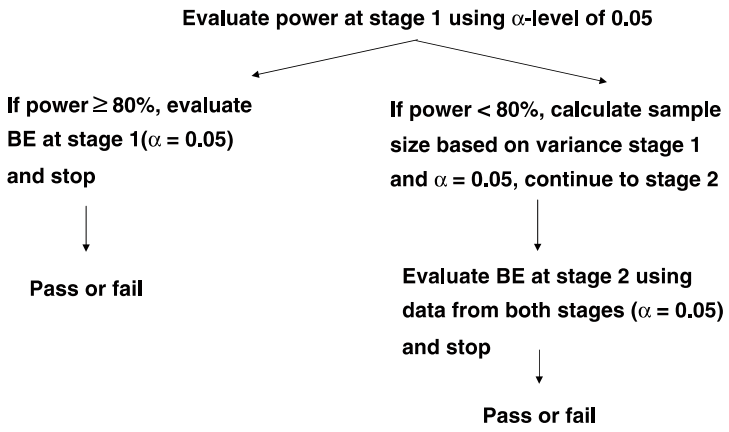
study at stage 1 with an small initial sample size, avoiding to unnecessarily soar the sample size at a stage 2 above what is reasonable to attain a desired power. And they are especially useful in case of drugs with little evidence about the true within-subject variability, and for *HVD*s.

1.7.1. Sample size re-estimation

*TSD*s allow to add some new subjects on a stage 2 based on the variability observed in the stage 1 (interim), so finally increasing the likelihood of declaring bioequivalence at a stage 2 with cumulated data.

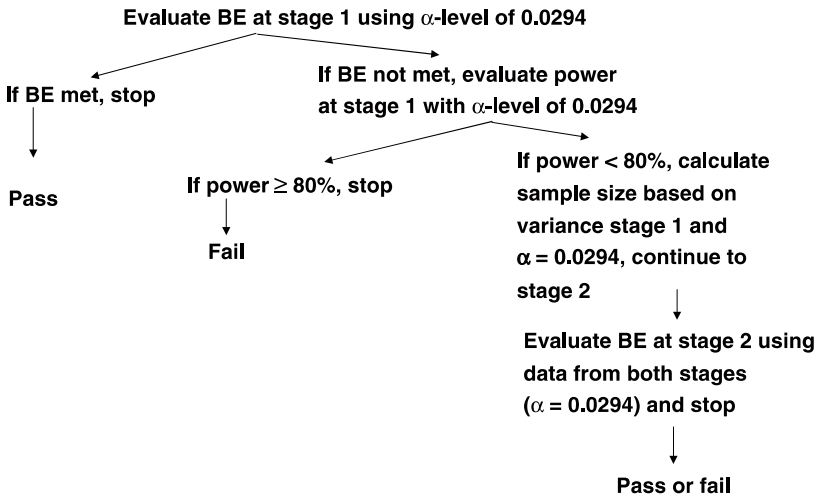
The following four approaches have been proposed (27):

Potvin A. Evaluate the power at stage 1 using the variance estimate from stage 1 with a α level of 0.05. If the power is ≥ 80% at stage 1, evaluate bioequivalence at stage 1, using a α level of 0.05 and stop whether bioequivalence is concluded or not. If the power is less than 80%, calculate the sample size based on the variance estimated at stage 1 using a level of 0.05 and continue to stage 2. Evaluate bioequivalence at stage 2 using data from both stages and a α level of 0.05. Stop here whether bioequivalence is met or not and regardless of what power was achieved. Note that bioequivalence is evaluated only once in any event. This is sometimes referred to as an internal pilot study design.

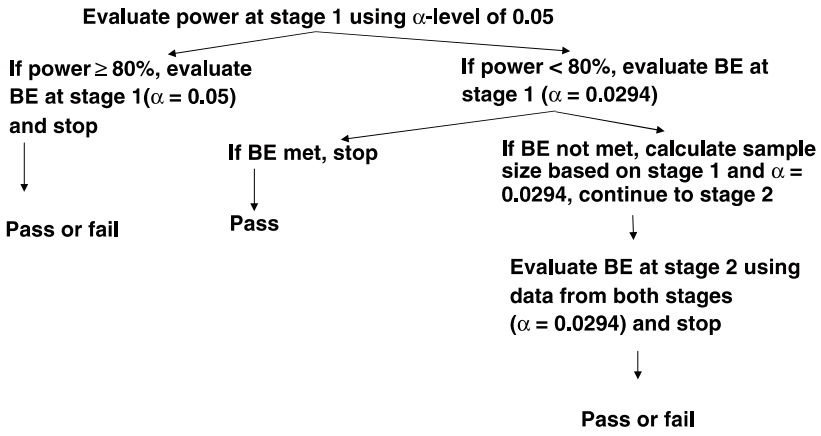


Potvin B. Evaluate bioequivalence at stage 1 using a level of 0.0294, regardless of the power achieved. If the bioequivalence criterion is met, stop. If the bioequivalence criterion is not met, calculate the sample size based on the variance estimated at stage 1 at a α level of 0.0294. If stage 1 already has at least 80% power, then stop. If not,

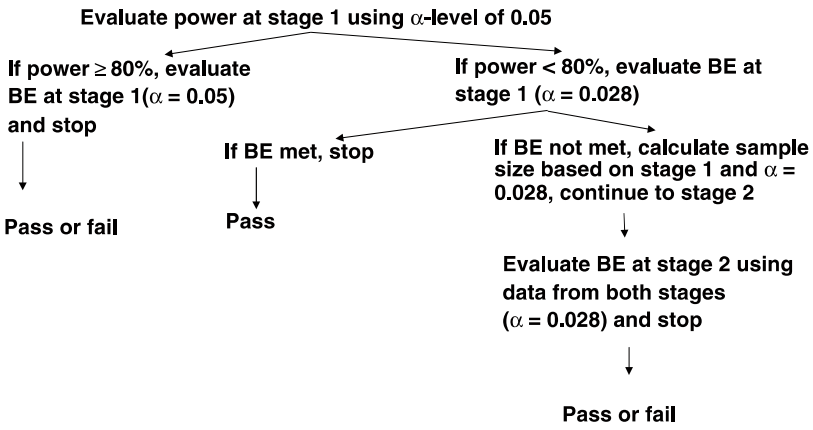
continue to stage 2. Evaluate bioequivalence at stage 2 using data from both stages at a α level of 0.0294. Stop here whether bioequivalence is met or not and regardless of the power achieved.



Potvin C. Evaluate the power at stage 1 using the variance estimate from stage 1 and an α level of 0.05. If the power is greater than or equal to 80%, evaluate bioequivalence at stage 1 using a α level of 0.05 and stop whether bioequivalence is met or not. If the power is less than 80%, evaluate bioequivalence using a α of 0.0294. If the bioequivalence criterion is met, stop. If the bioequivalence criterion is not met, calculate the sample size based on the variance estimated at stage 1 and a α level of 0.0294 and continue to stage 2. Evaluate bioequivalence at stage 2 using data from both stages at a α level of 0.0294. Stop here whether bioequivalence is met or not and regardless of the power achieved.



Potvin D. Evaluate the power at stage 1 using the variance estimate from stage 1 and at a α level of 0.05. If the power is greater than or equal to 80%, evaluate bioequivalence at stage 1 using a α level of 0.05 and stop whether bioequivalence is met or not. If the power is less than 80%, evaluate bioequivalence using at a α of 0.028. If the bioequivalence criterion is met, stop. If the bioequivalence criterion is not met, calculate the sample size based on the variance estimated at stage 1 and at a α level of 0.028 and continue to stage 2. Evaluate bioequivalence at stage 2 using data from both stages at a α level of 0.028. Stop here whether bioequivalence is met or not and regardless of the power achieved.



1.8. Type I error control

The process of declaring bioequivalence (or not), as in any other probabilistic study, is subject to errors. It is important to control the type I error or false positive rate or consumer risk, i.e. the probability of declaring bioequivalence when it is not the case.

The scaled methods defined by EMA and FDA regulations do not preserve adequately the type I error rate (e.g. at a maximum significance level α = 0.05) under some variability conditions, like in the neighborhood of CV_{WR} = 30% (Figure 2) (28,29).

Similarly, *TSD*s must preserve the type I error (*T1E*) rate at an overall significance level, e.g. α = 0.05. Pocock and Potvin *et al.* proposed using a significance level of 0.0294 at both stages (27,30). But this constant does not always control the *T1E* rate at a maximum 5%. In fact, the *T1E* depends on the treatment effect, variability, target power, or sample size (31,32). Though there are various ways of preserving these significance levels, the methodologies are not fully specified in the regulations (33,34).

1.9. Justification of the investigation

Highly variable drugs (*HVD*) are characterized by a high within-subject variability in the rate and/or extent of absorption of its active principle. This hinders researchers from declaring bioequivalence when it really holds, unless unacceptably large sample sizes are used.

If *HVD* is suspected, regulatory agencies (EMA and FDA) allow linearly scaling the pharmacokinetic metrics margins as a function of the reference variability product, and it further allows application of the interval inclusion rule over the expanded limits.

We compare a variant of the EMA Reference Scaled Bioequivalence *(RSABE)* method based on a replicate *TRTR/RTRT* design which preserves the type I error rate (especially for *HVD* with intra-subject coefficient of variation in the neighborhood of CV_{WR} = 30%) with two Two-Stage Adaptive Designs *(TSDs)* methods based on the usual 2x2 *RT/TR* crossover design which also preserve the type I error rate, through the power achieved, sample size required, blood sample extraction, and exposure time.

These results were discussed and published in Statistics in Medicine, Dec 2017. DOI: 10.1002/sim.7452

Regulators allow using *TSD* with unblinded interim sample size re-estimation. In this case, bioequivalence may be declared at the interim look with a reduced number of subjects; otherwise, the sample size can be increased on the basis of the estimated within-subject variability at the stage 1, then bioequivalence is tested again at a stage 2 with cumulated data. At the interim look we also have the possibility of cancelling a study for futility. *TSD* preserve the type I error rate by adjusting significance boundaries at each stage in various ways that are not fully specified in the regulations.

We present an iterative method to adjust the significance levels at each stage, α_1 and α_2, which preserves the overall type I error (usually at 5%) for a wide set of CV_{WR} scenarios. Also, we propose an extended feature by allowing α_1 being different than α_2.

These results were discussed and published in published in Biometrical Journal, Oct 2020. DOI: 10.1002/bimj.201900388.

This method was implemented in an *R* package called 'betsd' which includes the function 't1e.tsd'. We present this package which is useful to preserve the overall type I error rate in a strong sense and which provides the simulated significance levels to use at stages 1 and 2. It also includes an accurate description of all the arguments of the function 't1e.tsd' and allows calculating the probability to jump to a stage 2, the sample size at the stage 2, and the power at stages 1 and 2.

The function is described in the Biometrical Journal supporting information: bimj2181-sup-0001-SuppMat.docx. Also, source code to reproduce the results is available as Supporting Information on the journal's web page bimj2181-sup-0002-SuppMat.zip. The package is hosted on *GitHub* *https://github.com/eduard-molins/betsd*.

We also present a future line of research based on population and individual bioequivalence, and biosimilars.

2. HYPOTHESIS AND OBJECTIVES

2.1. Hypothesis

The scaled methods to test bioequivalence defined by FDA and EMA regulations do not adequately preserve the type I error rate (false positive or consumer risk) below the significance level in the neighborhood of CV_{WR} = 30%. A significance-level adjustment procedure may lose some power but it should convert a potential invalid procedure in a fully correct one.

Two-Stage designs (*TSD*s) based on adjusted significance level of $\alpha_1 = \alpha_2 = 0.0294$ at each stage (27) did not always control the overall type I error rate at a maximum α = 0.05. Since the type I error depends on the study framework, i.e., on the design, treatment effect, variability, target power, or sample size, the adjusted significance levels at each stage are entirely empiric and must be estimated in simulations.

2.2. Objectives

To discuss and improve some regulatory features (EMA and FDA) when assessing average bioequivalence (ABE).

Specific objectives:

1. To protect the type I error rate

    1.1. To present 2 'modified' *TSD* based on the usual 2x2 RT/TR crossover design which also preserve the type I error rate.

    1.2. To compare the properties between an 'adjusted' EMA *RSABE* method and modified *TSD*s using simulations by means of the power achieved, sample size required, blood sample extraction, and exposure time.

2. To extend the methodology to preserve the overall type I error for *TSD* (e.g. at α = 0.05) by means of a new iterative method (based on simulations) which covers a wide set of initial sample size, and intra-subject variability scenarios.

3. To present a new *R* package for *TSD* called ‘betsd’ along with the function ‘t1e.tsd’ to help on calculating the significance levels of each stage, the probability to jump to a stage 2, the sample size at the stage 2, and the power at stages 1 and 2.

3. TWO-STAGE DESIGNS VERSUS EUROPEAN SCALED AVERAGE DESIGNS IN BIOEQUIVALENCE STUDIES FOR HIGHLY VARIABLE DRUGS

This chapter is based on the following published research paper:

- **Title:** Two-Stage Designs versus European Scaled Average Designs in bioequivalence Studies for Highly Variable Drugs: Which to Choose? (21)
- **Published in:** Statistics in Medicine, Dec 2017
- **DOI**: 10.1002/sim.7452
- **PubMed ID:** 28853164
- **Authors:** Eduard Molins, Erik Cobo, Jordi Ocaña

3.1. 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 (7,8,15) 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" (C_{max}), while the "area under the concentration curves" (AUC_{0-t} and $AUC_{0-\infty}$) (13) are the most common bioavailability measures for the extent of absorption. As explained in more detail in the introductory section, to demonstrate bioequivalence, 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 α = 0.05) for bioequivalence 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 bioequivalence, the back-transformed confidence interval for the geometric means ratio (*GMR*) should lie fully within the bioequivalence limits of 0.80-1.25 (=1/0.80), corresponding to ±0.223 on the logarithmic scale (7,10).

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 bioequivalence 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.* (12) 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 *HVD*s, many of them around CV_{WR} = 30%.

If high variability is suspected, the European Medicines Agency (EMA) allows linearly scaling the C_{max} 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 (7). Similarly, the FDA also allows researchers to re-scale the *AUC* limits (8,15). These scaled approaches require the use of high order crossover designs like the replicated *TRTR/RTRT* or semi-replicated *TRR/RTR/RRT* designs (3,7,14). 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%. (28,29). Thus, the proportion of non-ABE products erroneously declared as bioequivalence is higher than its desired nominal value.

Regulators also allow using two-stage adaptive designs (*TSD*s) with unblinded interim sample size re-estimation (7,10,11,25) based on the usual 2×2 crossover *RT/TR* design. ABE 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 stage 1, then bioequivalence is tested again at a stage 2 with cumulated data $N = N_1 + N_2$. Two-stage designs preserve the type I error rate (31) by adjusting significance boundaries at each stage in various ways that are not fully specified in the regulations (33,34).

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.

3.2. Study objectives

The main objective was 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 (29)) with two *TSD* methods based on the usual *RT/TR* crossover design.

3.3. Statistical methodology

3.3.1. 2010 Regulatory EMA Reference Scaled approach (for C_{max} only)

Replicate *TRTR/RTRT* designs allow separately estimating the CV_{WR} (28,29) and can easily be re-arranged for comparison with a 2×2 crossover design (needed for *TSD*s) once the first two periods are sliced.

We focus on the EMA regulation because the interpretation of FDA scaled limits are controversial. Some authors (18,35-37) consider a high variability threshold of CV_{WR} = 25% though the scaling criterion starts from an observed $CV_{WR} \geq 30\%$. As a consequence, the scaling region starts at a lower value of 25%, and the limits are always continuous.

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

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

$$H_1\text{: } 0.80 < GMR < 1.25.$$

In the Reference Scaled Average bioequivalence (*RSABE*) approach, the bioequivalence 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:

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

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

If CV_{WR} < 30%, $GMR_{EMA}(CV_{WR}) = 1.25$; so the bioequivalence limits are the usual 0.8-1.25. If CV_{WR} lies between 30% and 50%, the bioequivalence 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 bioequivalence 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 C_{max} 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, $\widehat{CV}_{WR} = \sqrt{e^{\hat{\sigma}^2_{WR}} - 1}$, where $\hat{\sigma}^2_{WR}$ 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 $\widehat{GMR}$, $CI_{\widehat{GMR}} = e^{[\widehat{\emptyset}_L, \widehat{\emptyset}_U]}$, where $\widehat{\emptyset}_L$ and $\widehat{\emptyset}_U$ 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 bioequivalence (reject H_0), otherwise do not declare bioequivalence.

Note that the limits $GMR_{EMA}\left(\widehat{CV}_{WR}\right)$ 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 *et al*. (28) among others (29), 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%.

3.3.2. 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 (29). 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 (see 1.6.1), and not in the intra-subject contrasts, as are for example allowed in the FDA regulation for scaled average bioequivalence.

3.3.3. Two-stage modified Potvin B and C designs

We consider two *TSD*s with one interim analysis (at the stage 1) 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 stage 2 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 (34), we focused on those (almost partially) mentioned in regulations, considering two "Pocock-like" variants (30), as described by Potvin *et al.* and labelled A, B, C and D (27). In particular, we studied a type 1 Potvin B method (10) consisting of using the same adjusted α in both stages regardless of whether a study stops in the stage 1 or proceeds to the stage 2 (Figure 3), and a Type 2 Potvin C method where an unadjusted α may be used in the stage 1, dependent on interim power (Figure 4).

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 stage 2. This is conditional on the estimated within-subject coefficient of variation $\widehat{CV}_W$ at the stage 1 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 (31).

We propose a modification to the original Potvin B and C algorithms, including two constraints consisting of using a minimum sample size in the stage 2 (like in other jurisdictions or organizations) (10), and a maximum overall number of 150 subjects enrolled (32,38) in bioequivalence studies, as follows:

- A minimum of $N \geq 1.5N_1$ is required (or $N_2 \geq 0.5N_1$)

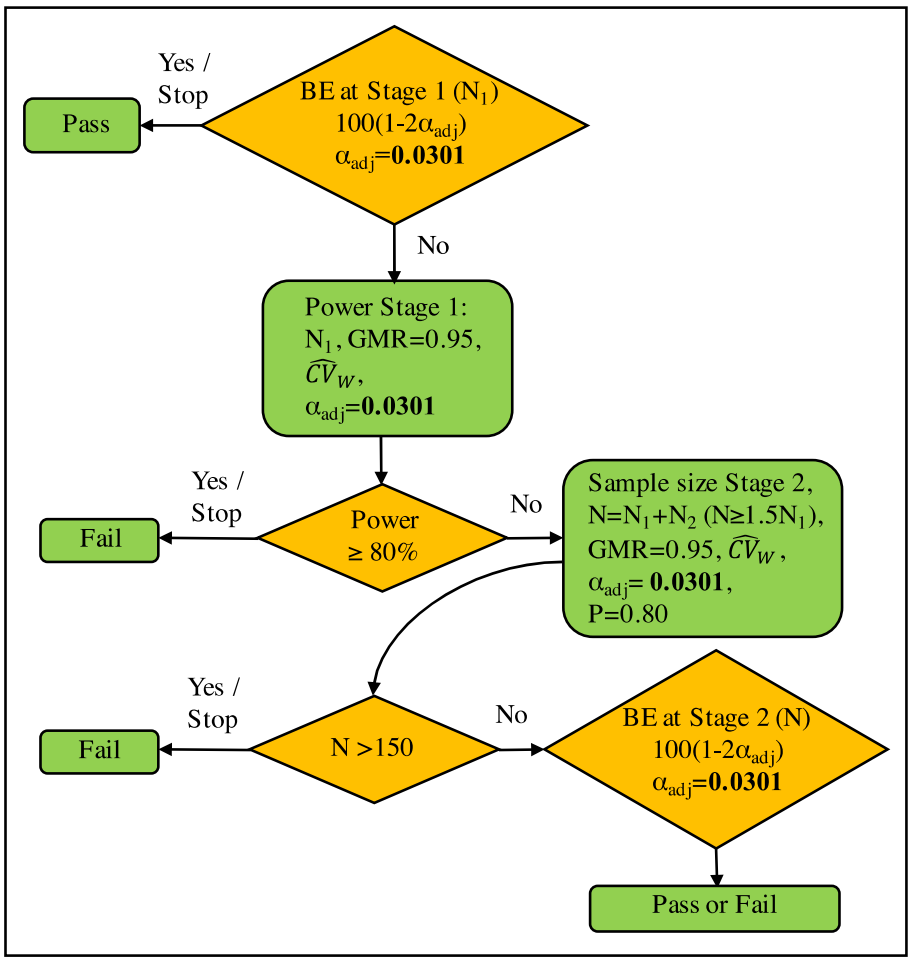
- If $N = N_1 + N_2 > 150$, the trial fails and it is stopped at the stage 1.

In any case, regardless of the method used, at least 12 evaluable subjects should be included in the stage 1 (8,11).

The adjusted significance level of α = 0.0294 used by Potvin *et al*. (27,30-32) 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*. (39) 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 3.3.4).

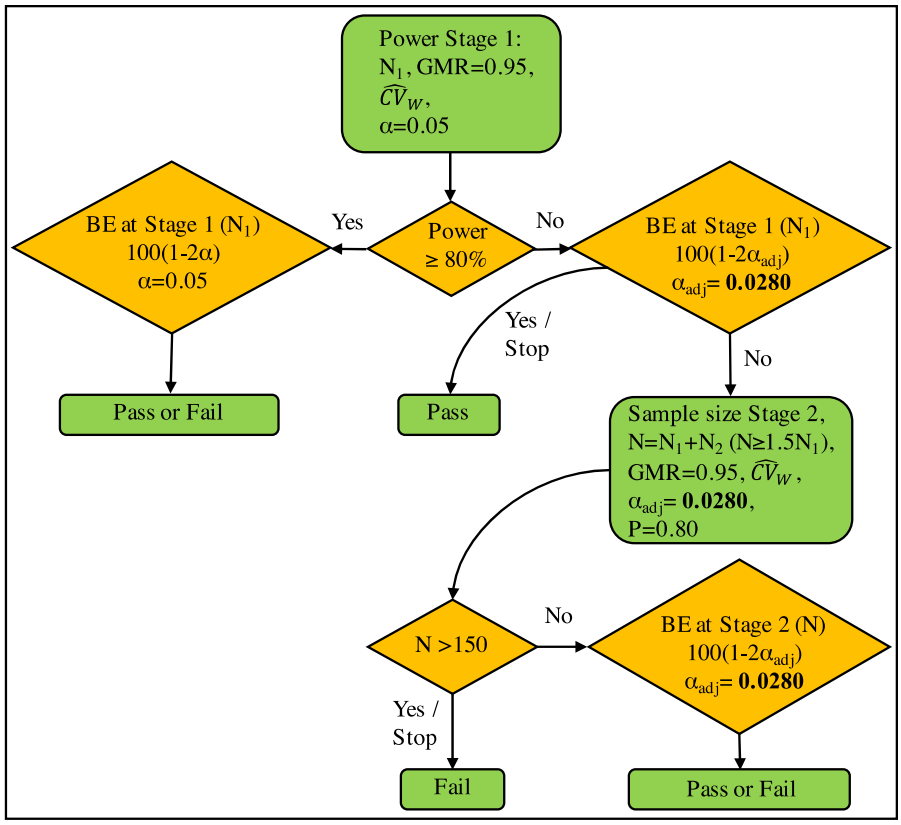
We used the method implemented in the function 'power.tsd' (via non-central *t*-distribution) in the *R* package 'Power2Stage' (40). The treatment effect was evaluated at the frontier 1.25, and assuming an expected *GMR* = 0.95 and a target power of 80%.

Figure 3. Type 1 Two-Stage Design - Modified Potvin B algorithm



Adapted from the figure depicted in detail by Montague *et al*. (31), with the restriction of Karalis and Macheras (32) of not including more than 150 subjects 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 4. Type 2 Two-Stage Design - Modified Potvin C algorithm



Adapted from the figure depicted in detail by Montague *et al.* (31), with the restriction of Karalis and Macheras (32) of not including more than 150 subjects 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

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*{*T1E*}, over the grid (m = 30,000 simulation trials per scenario). Filter for the scenarios where *Pr*{*T1E*} is at least 95% of the *max*(*Pr*{*T1E*}) observed in the grid, let's say $\{N_1/CV_W\}_{T1E \geq P95\%}$.
4) For $\{N_1/CV_W\}_{T1E \geq P95\%}$, find the N_1/CV_W with *max*(*Pr*{*T1E*}) (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*{*T1E*}), estimate the *Pr*{*T1E*} of all α_j (m = 1,000,000).

6) Adjust linear $\alpha = g_{lin}(Pr\{T1E\})$ and quadratic $\alpha = g_{quad}(Pr\{T1E\})$ 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 $Pr\{T1E\} < 0.05$ 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 (7), we used the modified Potvin B as the main *TSD* approach and the modified Potvin C as a sensitive case.

3.3.4. 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 stage 1 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 stage 1 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*{λ 0.2231} in the *TSD* approaches, and for all simulated CV_{WR} values; while in the *RSABE* approach, they were taken as *exp*{λ 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*{λ 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 stage 2 data included the following terms: stage, sequence, interaction sequence*stage, subject nested in sequence*stage, period nested in stage, and formulation.

3.4. Simulation results

The adjusted significance level predicted for the modified Potvin B was assessed at α_{adj} = 0.0301 at each stage; For the modified Potvin C, the adjusted significance level predicted was assessed at α_{adj} = 0.0280 (Figures 3 and 4).

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 4). Because almost all simulated studies required stepping up to a stage 2 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 ≥ 30%).

On the other hand, when N_1 ≥ 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 stage 1. 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 5 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 stage 1 in a notable proportion of times (approximately 41%, 47% and 24%).

Table 4. Two-stage design modified Potvin B and C: bioequivalence, sample size, and percentage of studies stepping up to stage 2 for true GMR = 0.95, and under different fixed N_1 and a true CV_W

		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; *GMR*, geometric mean ratio; N_1, initial and fixed sample size (Stage 1); CV_w, within-subject coefficient of variation; %St1, proportion of simulations declaring bioequivalence at Stage 1; %St1+St2, cumulative proportion of simulations declaring bioequivalence 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 5. Probability of bioequivalence acceptance according to the regulatory reference scaled bioequivalence ABE EMA and an adjusted EMA method compared to two-stage designs modified Potvin B and C (true CV_W = 30%)

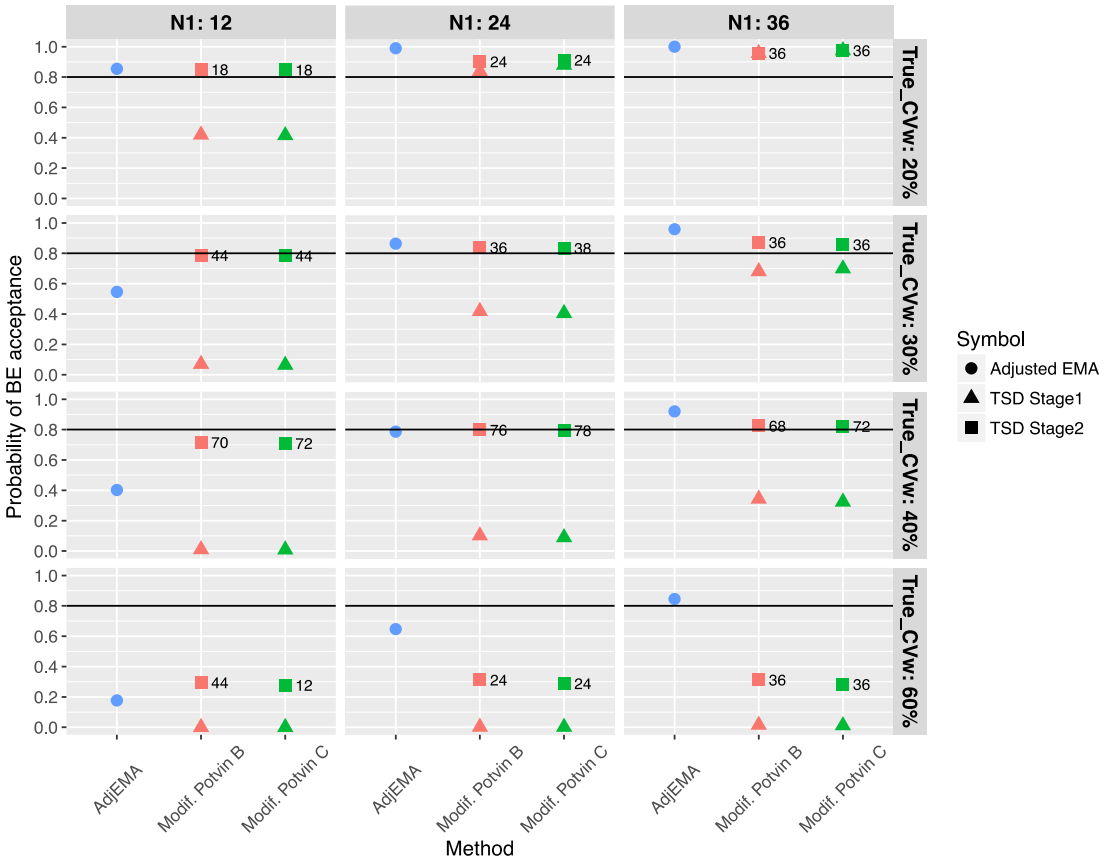
		Probability ABE acceptance			Type I error	
		True *GMR*				
	Method	**0.95**	**1.00**	**1.12**	**1.25**	**1.31**
RSABE method	Regulatory EMA (N_1 = 24)	0.896	0.963	0.631	0.085	0.021
	AdjEMA (N_1 = 24)	0.864	0.948	0.559	0.050	0.009
TSD method	Modified Potvin B (N_1 = 24 at Stage 1)	0.419	0.484	0.242	0.029	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; CV_w, within-subject coefficient of variation; N1, 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

Figure 5 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%.

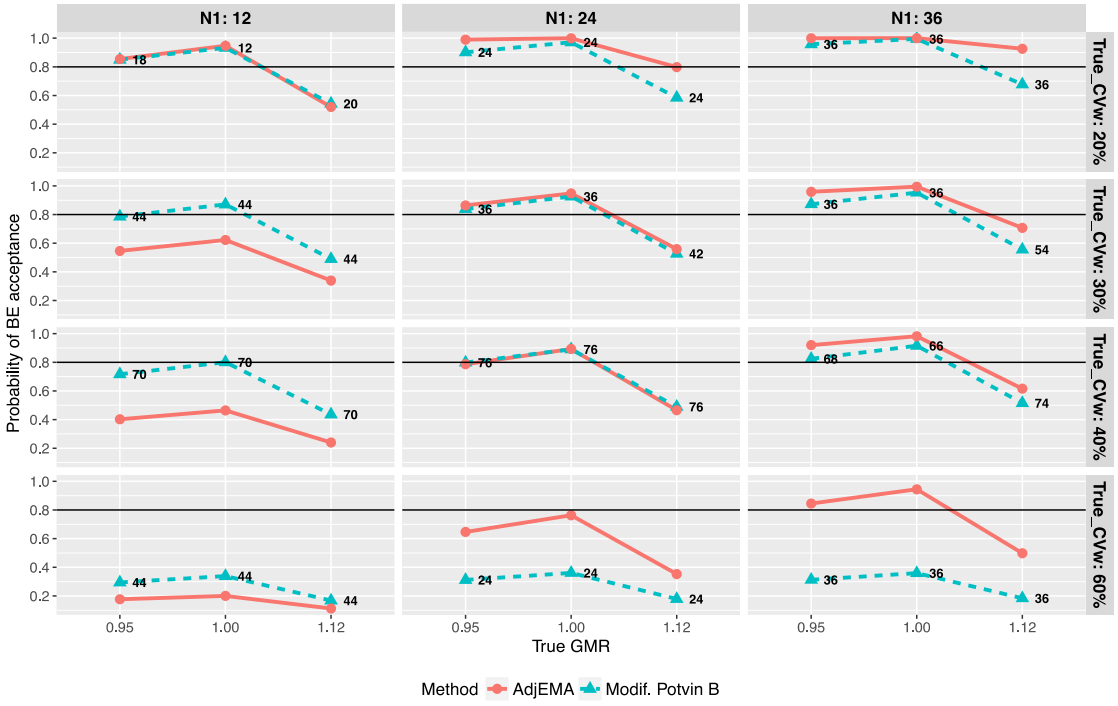
Figure 6 explores the power for different true levels of bioequivalence: *GMR* = 0.95, 1.00, and 1.12. It is remarkable that for a true value of *GMR* = 1.12, no methods reached 80% power for any *HVD* with $CV_W \geq 30\%$.

Figure 5. Bioequivalence acceptance of the adjusted reference scaled ABE EMA method and two-stage designs 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; *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 6. Bioequivalence acceptance of the adjusted reference scaled ABE EMA method and two-stage designs modified Potvin B for different levels of true bioequivalence and a progressive increase in the within-subject variability



ABE, average bioequivalence, *HVD*, highly variable drugs; N1, 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

3.5. 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 bioequivalence (*RSABE*) approach of the EMA regulation issued in 2010 (7). Although also mentioned in the regulations, adaptive *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, *TSD*s are a good option to consider, as they have a power of around 80% at the stage 1 for non-highly variable drugs while at the same time they offer the opportunity for stepping up to the stage 2 (including additional subjects) for truly bioequivalent products.

Statistical power is used to evaluate the performance of adaptive methodologies in bioequivalence clinical trials. A power of at least 80% is desirable when considering N_1 subjects at the stage 1, and assuming an expected but unknown within-subject coefficient of variation, CV_W. 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 CV_W combinations (i.e. α_{adj} = 0.0301 and α_{adj} = 0.0280 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 CV_W combination, the power might be slightly downgraded, although it is always above 80% in case of true bioequivalence.

Patterson *et al.* (41) 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 bioequivalence limits of 0.8-1.25 and CV_W 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 (38), 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 $0.95 \leq GMR \leq 1.05$, 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 stage 2 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 (42), 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.

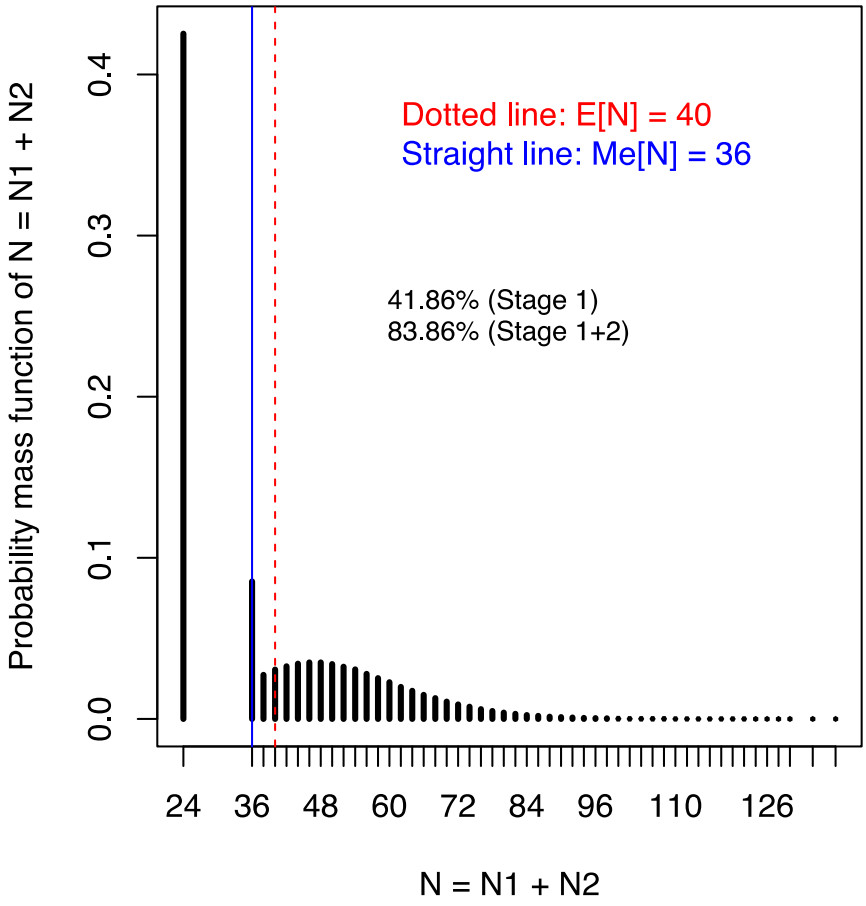
Kieser and Rauch (34) and Karalis and Macheras (38) pointed out a potential limitation of the original *TSD* methods stated by Potvin *et al*. (27) and Montague *et al*. (31) as although unblinded data are available after the stage 1, the knowledge about the estimated *GMR* in the interim analysis is not used for sample size recalculation. We assumed a fixed true treatment effect of *GMR* = 0.95 after the stage 1 since Cui *et al*. (43) showed that a determination of the stage 2 sample size based on an interim estimate of the *GMR* can substantially inflate the probability of type I error in most practical situations. The use of observed treatment effects at first stage is at least controversial (44), in particular when N_1 is low and CV_W is high. The use of a fixed *GMR* = 0.95 is a balance between a value which is not much optimistic and provides a certain level of stability for sample size re-estimation.

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 α_{adj} = 0.0301 at each stage, *GMR* = 0.95, CV_W = 0.3, N_1 = 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 7). 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*. (45). 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 (7), as they have demonstrated that they control the type I error rate in a strong way.

Figure 7. Type 1 Two-Stage Designs modified Potvin B distribution of N (Stage1 + Stage 2)

GMR=0.95; CV_w=30%; N1=24; alpha_adj =0.03018396; P=0.8; m=1,000,000 simulations



GMR, true geometric mean ratio; CV_w, 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

So, given that either the *RSABE* or *TSD* methods are suitable approaches for bioequivalence 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 stage 1 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 (46). 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 C_{max} (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 bioequivalence 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 bioequivalence at the stage 1 in adaptive approaches, sponsors should consider them because they imply less subject exposure and less treatment duration.

4. AN ITERATIVE METHOD TO PROTECT THE TYPE I ERROR RATE IN BIOEQUIVALENCE STUDIES UNDER TWO-STAGE ADAPTIVE DESIGNS

This chapter is based on the following published research:

- **Title:** An iterative method to protect the type I error rate in bioequivalence studies under two-stage adaptive 2x2 crossover designs (22)
- **Published in:** Biometrical Journal, Oct 2020
- **DOI**: 10.1002/bimj.201900388
- **PubMed ID:** 33000873
- **Authors:** Eduard Molins, Detlew Labes, Helmut Schütz, Erik Cobo, Jordi Ocaña

4.1. Introduction

Bioequivalence studies typically involve testing two products, test, *T,* and reference, *R,* against each other in a two-period, two-sequence 2×2 crossover *RT/TR* trial. Primary pharmacokinetic metrics are C_{max} (maximum observed plasma concentration) and the area under the concentration time curve, $AUC_{0\text{-}t}$ and $AUC_{0\text{-}\infty}$ (7,8).

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

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

$$H_1\colon -\delta < \phi < +\delta.$$

Based on the "interval inclusion rule", to declare bioequivalence (i.e., to reject the null hypothesis of bioinequivalence) at a significance level α = 0.05, based on a normal *ln*-linear model, the two-sided 1 – 2α = 0.9 symmetric confidence intervals for $\mu_T - \mu_R$, ϕ, should lie fully within the constant bioequivalence limits of ±0.223, or equivalently, the back exponentially transformed confidence interval for the geometric mean ratio, $GMR = e^{\phi}$ should lie fully within 0.80 to 1.25 (=1/0.80) (7,10).

Regulatory agencies usually accept conducting studies based on *RT/TR* two-stage adaptive 2×2 crossover designs (*TSD*) (7,10,11,25,47,48), whose application is

becoming increasingly popular (23,24). *TSD*s allow declaring bioequivalence at an interim look (or stage 1) with a small number of N_1 subjects; and if bioequivalence is not met due to insufficient power, the sample size can be increased in a stage 2 based on the estimation of the within-subject variability, calculated by means of the pooled coefficient of variation of *R* and *T*, considering $CV_W = \sqrt{e^{\sigma^2_W} - 1}$, where σ^2_W is the estimated value of the residual variance obtained from an ANOVA model on *ln*-transformed data. Then bioequivalence is tested again at stage 2 with cumulated $N = N_1 + N_2$ sample size.

Also, *TSD*s provide investigators with an attractive solution to address some of the uncertainty that exists when the trial is originally designed (26), allowing stopping the study at stage 1 with an small N_1, avoiding to unnecessarily soar *N* above what is reasonable to attain a desired power, e.g., 80%. And they are especially useful in case of drugs with little evidence about the true within-subject variability, and for highly variable drugs (*HVD*), i.e. with a $CV_W \geq 0.3$ (21,49). This discussion is important because the precise model for analysis must be pre-specified in the protocol including the sources of variation that reasonably influence primary metrics (27,47). However, little guidance exists yet on how investigators should proceed when designing and planning an adaptive clinical trial (50).

The critical point about using *TSD*s is the difficulty to preserve the type I error rate (*T1E*) (7,34,51,52). Significance level boundaries can be adjusted in various ways that are not fully specified in the regulations (7). Using an a priori fixed sample size split at equal sequential groups (30), decision to stop the trial or continue was based on repeated significance tests of the accumulated data after each group was evaluated. Based on Pocock's method but using sample size re-assessment, i.e. *TSD*s, Potvin *et al*. (27) and Montague *et al*. (31) proposed two methods to control the overall *T1E* rate: Their "type 1" method consists on using the same adjusted significance level at stages 1 and 2, i.e. $\alpha_{adj} = \alpha_1 = \alpha_2$; and "type 2" method consists on using an unadjusted $\alpha = 0.05$ in the stage 1, if the interim power is of at least of 80% at stage 1, or else an adjusted α_1 and α_2 at stages 1 and 2, respectively.

Using simulations, Xu *et al*. (39) implemented two methods (called E and F) to find optimal solutions of α_1, α_2, N_1 (and a futility parameter) by means of average cost

functions of *GMR* and CV_W combination values. They presented optimal solutions for CV_W ranging from 10-30%, and for 30-55%. Maurer, Jones, and Chen (52) used a principled approach using a standard inverse-normal *p*-value combination test, in conjunction with standard group sequential techniques (called it maximum combination test) to guarantee the control of *T1E* rate.

4.2. Study objectives

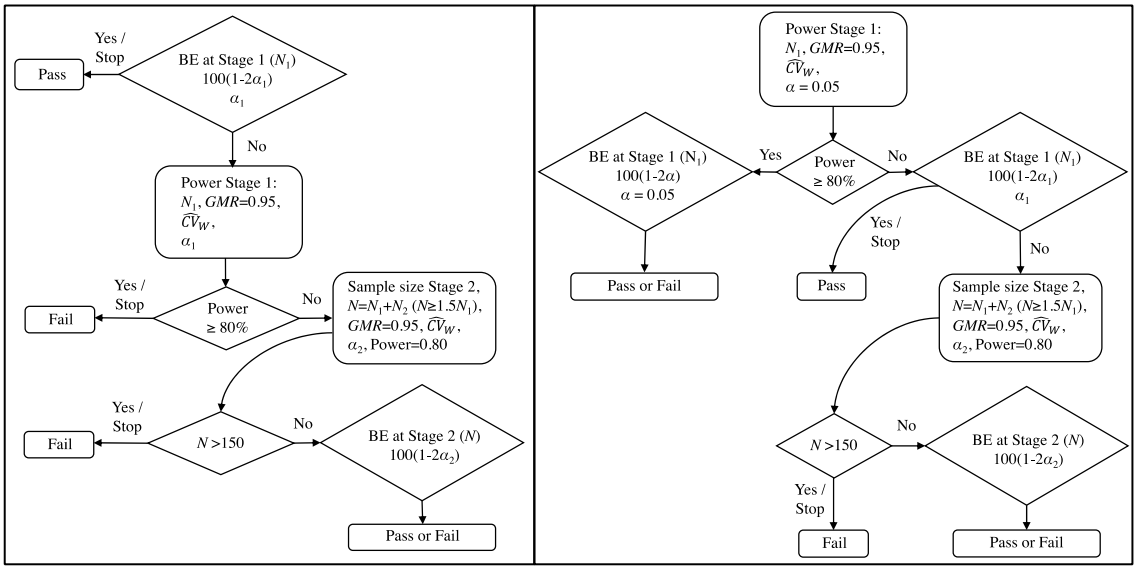
We present an iterative method, which is based on simulations, to adjust the significance levels at each stage, α_1 and α_2, which preserves the overall *T1E* (usually at 5%) for a wide set of scenarios which should include the true unknown variability value. Additionally, we propose an extended feature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

In the next section we present the methodology to obtain the adjusted significance levels using simulated samples; Then we present the simulation results where we provide comparisons of our method with the most recent articles released by Xu *et al*. (39), and Maurer, Jones, and Chen (52), and we finalize with a discussion.

4.3. Methodology to obtain the adjusted significance levels

Figure 8 shows two algorithms to test bioequivalence using *TSD* by means of the type 1 and 2 methodologies. They include two constraints, first on the minimum sample size at stage 2 of at least $N_2 \geq 0.5N_1$, and secondly, as previously discussed in Molins *et al*. (21), Xu *et al*. (39) and Karalis and Macheras (32), with a futility criterion to stop the study at stage 1 based on a total study size upper limit, in our case of 150 subjects maximum. In contrast to the algorithms proposed by Potvin *et al*. (27) we allow α_1 and α_2 being different.

Figure 8. Testing ABE using two-stage designs by means of type 1 (on the left) and type 2 (on the right) methodologies, with significance levels α_1 and α_2 at each stage



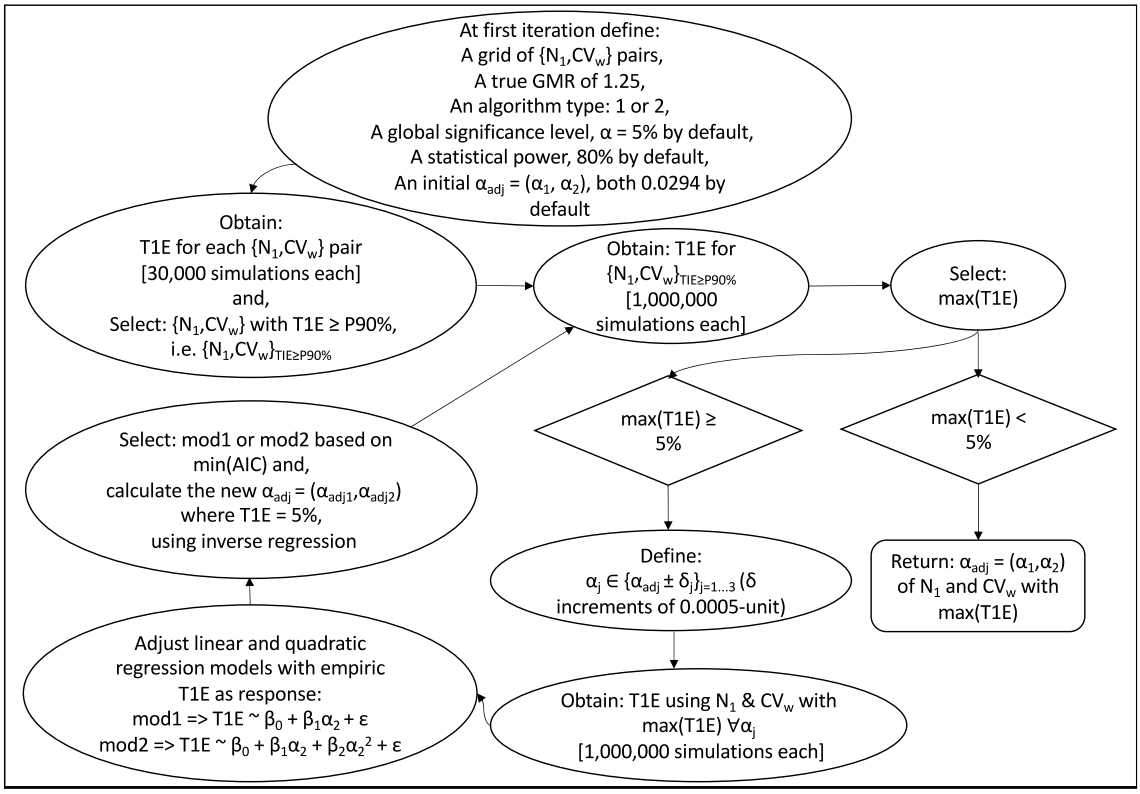
Adapted from the figure depicted in detail by Montague *et al.* (31) with the restriction of Karalis and Macheras (32) of not including more than *Nmax* subjects (150 by default), and min(N_2) ($N_2 \geq 0.5N_1$ by default); α_1 and α_2, adjusted significance levels at stages 1 and 2 (α_1 may be different than α_2); ABE, average bioequivalence; N_1, initial fixed sample size; N_2, additional number of subjects recruited at stage 2; *GMR*, geometric mean ratio; $\widehat{CV}_W$, simulation-based estimated within-subject coefficient of variation at stage 1

Figure 9 shows the iterative method used to find an optimal significance level adjustment at stages 1 and 2, $\alpha_{adj} = (\alpha_1, \alpha_2)$, granting a global significance level below α (usually α = 5%).

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

1) Arbitrary starting initial significance levels at each stage, e.g., (α_1, α_2) = (0.0294, 0.0294) at stages 1 and 2, respectively (based on Potvin *et al*. (27) constant).
2) An initial fixed sample size N_1. A minimum of 12 subjects are required (8,46).
3) A meaningful set of CV_W values trying to cover the true unknown variability value, a scalar or vector (larger set in case of higher uncertainty), e.g. CV_W = 0.2.
4) An expected *GMR* for power calculation, e.g. 0.95.
5) A true *GMR* for type I error assessment, let's say θ_0, fixed at 1.25.
6) Type 1 or type 2 methodology (as shown in Figure 8).
7) A global significance level, e.g. α = 0.05.

Figure 9. Iterative method to obtain adjusted α_1 and α_2 at each stage to grant a global T1E below α



α, desired global significance level (be default, 5%); $\alpha_{adj} = (\alpha_1, \alpha_2)$ adjusted significance levels at each stage; N_1, sample size at stage 1; CV_W, within-subject coefficient of variation; *T1E*, Type I error rate, assessed by means of *R* function 'power.tsd' of Labes and Schütz (29), and following Figure 8; P90%, percentile 90% of *T1E*; *AIC*, Akaike information criterion

By means of a 'current' arbitrary significance level α_{adj} = (0.0294, 0.0294) at stages 1 and 2, respectively, Figure 9 shows the algorithm which starts with a warm up period assessing the empiric *T1E* with 30,000 simulations for each test point at a grid of pre-defined {N_1, CV_W} combinations (corresponding to N_1 x CV_W cartesian product), and selecting those pairs exceeding the percentile 90%. For more accuracy, simulations are repeated for this subgroup 1,000,000 times each. The N_1 and CV_W pair with the maximum empiric *T1E* is selected. Six new significance levels of adjustment are then defined at ±0.0005 distance from the 'current' significance level, and the empiric *T1E* rate is assessed for each one (with 1,000,000 simulations each time), using the previous maximum N_1 and CV_W pair. This is the base to find the new adjusted significance level, $\alpha_{adj} = (\alpha_1, \alpha_2)$. To do so, regression models are adjusted with 'empiric *T1E*' as response and 'significant level' as covariate (linear and quadratic) as shown in

the Figure 9. The model with the minimum Akaike Information Criterion (*AIC*) is selected, and the 'adjusted' $\alpha_{adj} = (\alpha_1, \alpha_2)$ is established by isolating α_2 using the estimated parameters at a fixed *T1E* equals to 0.05. In summary, we output the solution if the 'current' significance level protects the *T1E* below α, otherwise, the algorithm starts again from the beginning with the assignment of 'current' = 'adjusted' significance level.

To obtain the adjusted significance levels to preserve the overall *T1E* below α, simulations were performed with a true effect ratio θ_0 of 1.25 (i.e., considering the null hypothesis of bioinequivalence true), where the treatment effect is just on the bioequivalence frontier so that the likelihood of leading to a false positive result is highest. Under θ_0 equals to 1.25, we used an expected *GMR* at 0.95 to test for ABE following both *TDS* algorithms shown in Figure 8. Once the adjusted significance levels were obtained and fixed, we conducted new simulations with θ_0 and *GMR* at 0.95 to predict the power at stage 1 and overall (stage 1 plus stage 2), the percentage of studies switching to stage 2, and the percentiles 5, 50, and 95 of $N = N_1 + N_2$ subjects.

Parameters N_1 and CV_W can be scalars or vectors. If they are vectors, e.g., $N_1 = (12, 24)$ and $CV_W = (0.1, 0.15, 0.2, 0.25)$, then the (N_1, CV_W) combination of all $\{N_1, CV_W\}$ combinations with maximum *T1E* is selected for α_{adj} adjustment (see Figure 9). Additionally, we propose an extended feature by allowing α_1 being different than α_2. When this occurs, α_1 is considered fixed, and the adjustment is only based on α_2. Since the true CV_W is unknown at the time that the simulations are conducted (before the study starts), and to avoid imprecise specifications for simulations based on tight ranges of CV_W, (or a vague idea about the true/unknown CV_W) our methodology controls the *T1E* considering CV_W below and upper 0.05 from the values specified/considered.

By means of the function 'power.tsd' included in the *R* package 'Power2Stage', developed by Labes and Schütz (29), and hosted on CRAN (40), we developed an open *R* package called 'betsd', and hosted on *GitHub* https://github.com/eduard-molins/betsd to allow traceability of all versions. This package includes an accurate description of all functionalities of the 't1e.tsd' function which serves to calculate the adjusted significance levels at stages 1 and 2. This function implements both

methodologies shown in Figure 8, including the modifications proposed in Molins *et al*. (21). Also, source code to reproduce the results is available as Supporting Information on the journal's web page bimj2181-sup-0002-SuppMat.zip. In order to allow reproducibility of simulations, we used seed number 1234567.

In turn, this package follows the EMA Questions & Answers document (11), so that, in stage 1, the terms used in the ANOVA model are sequence, subject within sequence, period and formulation. Fixed effects, rather than random effects, are used for all terms. In stage 2, the adjusted ANOVA model includes sequence, stage, sequence × stage, subject within sequence × stage, period within stage, and formulation. Note that models do not include carryover effects or treatment-by-period interactions.

4.4. Simulation results

Using simulated samples, we found the adjusted significance levels when α_1 equals α_2, with *T1E* rates always strictly below 5%. We assumed some credible scenarios for CV_W and N_1. Table 6 shows the results for 16 scenarios corresponding to a pre-planned fixed initial sample size N_1 of 12 and 24, and a priori true intra-subject CV_W in the following ranges: from 0.10 to 0.19 (a vector of discrete values analyzed at intervals of 0.01-units, i.e. 0.10, 0.11, 0.12, ...0.19), from 0.20 to 0.29, from 0.30 to 0.39, and from 0.40 to 0.49. We found the adjusted significance levels, *T1E*, % power at stage 1, % of studies jumping to stage 2, % overall power, and percentiles 5, 50 and 95 of *N.* 10E6 simulations were conducted per scenario.

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

Table 6. Adjusted α_1 and α_2 in both stages preserving the overall *T1E* below 5%

N_1	CV_W LB – UB	Adjusted $\alpha_1 = \alpha_2$	*T1E*	% power Stg. 1	% to Stg. 2	% overall Power	*P*: 5, 50, 95
Type 1 methodology							
12	0.10-0.19	0.0299	0.046063	47.93	49.08	85.96	12, 12, 36
12	0.20-0.29	0.0307	0.049771	15.49	83.91	80.85	12, 34, 64
12	0.30-0.39	0.0303	0.044972	7.02	92.74	78.63	12, 44, 84
12	0.40-0.49	0.0377	0.044389	1.60	96.30	73.56	24, 66, 124
24	0.10-0.19	0.0381	0.039430	89.59	2.85	91.95	24, 24, 24
24	0.20-0.29	0.0306	0.048095	50.95	47.67	84.87	24, 24, 60
24	0.30-0.39	0.0302	0.049831	29.86	69.90	82.63	24, 50, 84
24	0.40-0.49	0.0306	0.045264	10.55	89.01	79.98	24, 76, 118
Type 2 methodology							
12	0.10-0.19	0.0280	0.049858	55.12	39.34	86.54	12, 12, 34
12	0.20-0.29	0.0280	0.049787	35.58	61.06	84.10	12, 22, 44
12	0.30-0.39	0.0295	0.044164	6.88	92.57	78.61	12, 44, 84
12	0.40-0.49	0.0377	0.044501	1.61	96.27	73.68	25, 66, 124
24	0.10-0.19	0.0314	0.049608	96.08	0.23	96.28	24, 24, 24
24	0.20-0.29	0.0301	0.049985	46.96	50.66	83.94	24, 36, 66
24	0.30-0.39	0.0303	0.049815	26.47	72.98	82.13	24, 54, 88
24	0.40-0.49	0.0306	0.044950	10.56	88.95	79.99	24, 76, 118

Notes: Burn-in α_1 and α_2 values were initially set at 0.0294; N_1, Initial fixed sample size; CV_W LB-UB, lower and upper bound (-/+ 0.05) range of the within-subject coefficient of variation, analyzed at increments of 0.01-units; Adjusted $\alpha_1 = \alpha_2$, same adjusted significance levels at stages 1 and 2; *T1E*, empirical type I error; % power Stg. 1, power at stage 1; % to Stg. 2, percentage of studies which switch to stage 2; % overall power, overall power; *P:* 5, 50, 95, percentiles 5, 50, and 95 of $N = N_1 + N_2$

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

In the Table 7 we found the adjusted α_2, *T1E*, % power at stage 1, % of studies jumping to stage 2, % overall power, and percentiles 5, 50 and 95 of *N*, for 4 scenarios with initial sample sizes N_1 of 12 and 24, a priori assumption on the true intra-subject CV_W ranging from 0.20 to 0.29 (at intervals of 0.01-units) and given a fixed a priori α_1. 10E6 simulations were conducted per scenario. Results of *T1E* rates were always below 5%. We considered both possibilities, to be more permissive at stage 1 with $\alpha_1 \leq \alpha_2$, or at stage 2 with $\alpha_1 \geq \alpha_2$. We can compare these results to the ones obtained in the Table 6 where $\alpha_1 = \alpha_2$.

For N_1 equals to 12, and a fixed $\alpha_1 = 0.0294 < \alpha_2$, the significance level at stage 2 was adjusted at 0.0310. These results contrast with the ones obtained in Table 6 with $\alpha_1 = \alpha_2 = 0.0307$, being the test less permissive at stage 1 and more permissive at the stage 2. Additionally, a power of 81% was reached, with a likelihood of 81% of stepping up to stage 2, and with a percentile 95 of *N* equals to 58 subjects. Similarly, for N_1 equals to 12, and when $\alpha_1 = 0.0320 < \alpha_2$, α_2 was adjusted at 0.0279. This test is more permissive at stage 1 and less permissive at stage 2.

Table 7. Type 1 method to adjust α_2 for a fixed α_1 preserving the overall *T1E* below 5%

N_1	CV_W LB – UB	Adjusted α_2	*T1E*	% power Stg. 1	% to Stg. 2	% overall Power	*P*: 5, 50, 95
$\alpha_1 = 0.0294 < \alpha_2$							
12	0.20-0.29	0.0310	0.049891	17.92	81.31	81.45	12, 32, 58
24	0.20-0.29	0.0318	0.048936	45.60	53.35	84.28	24, 36, 64
$\alpha_1 = 0.0320 > \alpha_2$							
12	0.20-0.29	0.0279	0.049767	27.96	71.02	83.25	12, 26, 52
24	0.20-0.29	0.0285	0.048875	47.88	51.30	84.54	24, 36, 66

Notes: Burn-in α_2 value was set at 0.0300 for α_1 = 0.0294, and at 0.0294 for α_1 = 0.0320; N_1, Initial fixed sample size; CV_W LB-UB, lower and upper bound (-/+ 0.05) range of the within-subject coefficient of variation, analyzed at increments of 0.01-units; Adjusted α_2, adjusted significance level at stage 2; *T1E*, empirical type I error; % power Stg. 1, power at stage 1; % to Stg. 2, percentage of studies which switch to stage 2; % overall power, overall power; *P:* 5, 50, 95, percentiles 5, 50, and 95 of $N = N_1 + N_2$

Given adjusted significance levels, Table 8 shows the empiric *T1E* rate and power for CV_W at 0.05 above and below the upper and lower CV_W bounds using the type 1 and 2 methodologies. Type 1 error and % overall power were calculated by means of 10E6 and 10E5 simulations per scenario, respectively. We can see that *T1E* never exceed the 5% global significance level and the power was around 80% or higher, except for CV_W values of 0.54 affected by the constraint of max($N = N_1 + N_2$) = 150.

Table 8. Empiric type 1 error and power for CV_W at 0.05 below and above LB and UB

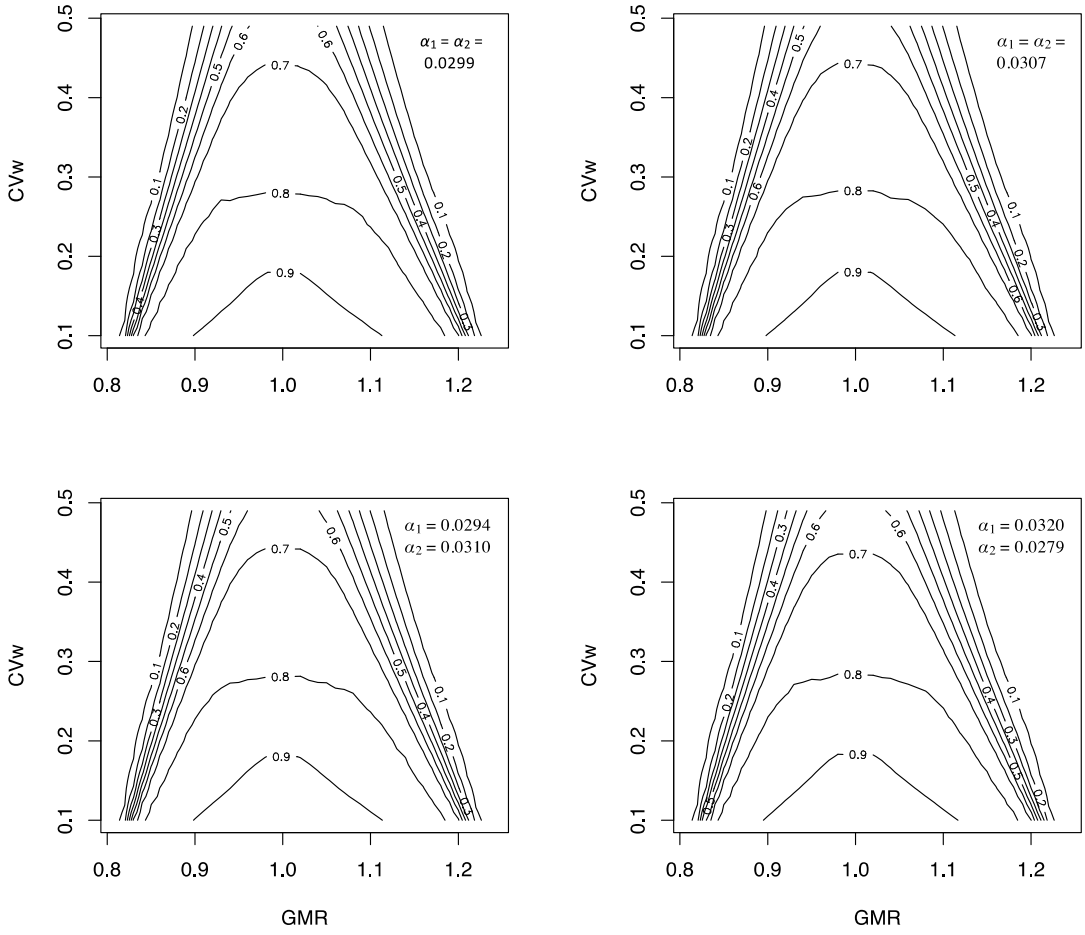
			Type 1 error		% overall power	
N_1	CV_W LB – UB	Adjusted $\alpha_1 = \alpha_2$	CV_W LB – 0.05	CV_W UB + 0.05	CV_W LB – 0.05	CV_W UB + 0.05
Type 1 methodology						
12	0.10-0.19	0.0299	0.0299	0.0498	99.99	82.07
12	0.20-0.29	0.0307	0.0379	0.0411	90.09	76.61
12	0.30-0.39	0.0303	0.0498	0.0314	81.63	65.86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

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

Based on our method, protocols for bioequivalence must include an initial N_1, a method type (1 or 2) with constraint max($N = N_1 + N_2$) = 150 (if $N > 150$, bioequivalence fails), and $N_2 \geq N_1/2$, a target power, and the significance levels to use, obtained by means of the function 't1e.tsd'. Figure 10 shows power contour plots, considering N_1 set to 12, the type 1 method, and a target power of 0.8. True unknown CV_W values range from 0.10 to 0.49 (y-axis), and true unknown GMRs between 0.80-1.25 (x-axis, extremes not included), both at increments of 0.05. Significance levels were taken from Table 6 and 7: $\alpha_1 = \alpha_2 = 0.0299$; $\alpha_1 = \alpha_2 = 0.0307$; $\alpha_1 = 0.0294$ $\alpha_2 = 0.0310$; $\alpha_1 = 0.0320$ $\alpha_2 = 0.0279$. We tested 1,760 scenarios per graph (40 CV_W x 44

GMRs) using the function 'power.tsd' with 10E5 simulations for scenario. We can see in all graphs that the constraint of a maximum of 150 subjects provokes a power decrease of at least 70% for CV_W values above 40%.

Figure 10. Power assessment based on true GMR and CV_W with N_1 = 12 and type 1 methodology



Note: All combinations of *GMR* between 0.80 and 1.25 (extremes not included), and CV_W between 0.10 and 0.49, both defined as vectors of discrete values at intervals of 0.01-units, resulted on 1,760 scenarios which were simulated 10E5 times each

Xu *et al*. (39) obtained α_1, α_2, N_1, and a futility criterion (*f*) by means of average cost functions for *GMR* and CV_W combination values at increments of 5%. They varied (and fixed) the two significance levels α_1 and α_2, N_1, and a futility criterion (*f*), and checked whether the power was of at least of 80% (at a true *GMR* of 0.95) and the type I error rate (at a true *GMR* of 0.8 of bioinequivalence) controlled for each *GMR* and CV_W combination value. They obtained optimal designs based on the lowest cost among

valid combinations of α_1, α_2, N_1, and f. We obtained type 1 and 2 α_1 and α_2 using the function 't1e.tsd' based on the N_1 and CV_W obtained by Xu *et al.* (Table 9). Due to design similarities, type 1 method (modified Potvin B) can be compared with Xu *et al.* Method E, and type 2 (modified Potvin C) to compare with Method F.

Table 9. Xu et al. Optimal two-stage designs of methods E and F and our methodology (type 1 and 2 methods)

	CV_W range: 0.10-0.30 N_1 = 18	**CV_W range: 0.30-0.55 N_1 = 48**
Method E (Xu et al.)	α_1: 0.0249 α_2: 0.0363 f: 93.74 – 106.67	α_1: 0.0254 α_2: 0.0357 f: 93.05 – 107.47
Method F (Xu et al.)	α_1: 0.0248 α_2: 0.0364 f: 94.92 – 105.35	α_1: 0.0259 α_2: 0.0349 f: 93.50 – 106.95
Type 1 method	$\alpha_1 = \alpha_2$ = 0.0303	$\alpha_1 = \alpha_2$ = 0.0305
Type 2 method	$\alpha_1 = \alpha_2$ = 0.0331	$\alpha_1 = \alpha_2$ = 0.0331

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

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

Table 10. Percentiles of *N* (5th, 50th, 95th) and % of Studies in Stage 2

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

CV_W LB-UB, lower and upper bound values of the within-subject coefficient of variation
Type 1 method: Modified Potvin B method with max(N = N1+ N2) = 150, and N2 ≥ N1/2
Type 2 method: Modified Potvin B method with max(N = N1+ N2) = 150, and N2 ≥ N1/2
Type 1 method is compared with Method E and type 2 with method F
Target power = 0.80 and planned and true *GMR* = 0.95

Maurer, Jones, and Chen (52) used a standard inverse-normal *p*-value combination test, in conjunction with standard group sequential techniques (called it maximum combination test), to guarantee the control of type I error rate at any given significance level. The sample size N_2 at the stage 2 was based on comparing a 'target conditional power', with the power achieved at stage 1, versus a 'conditional power', with the conditional errors for maximum combination test (using the CV_W estimation

at interim), a formulation effect of 0.95, and N_2. Starting on an initial N_2 set to 4, the 'conditional power' was assessed at increments of 2 subjects until it exceeded the 'target conditional power'.

Table 11 shows the power and sample size of different methods for *HVD*. Results from Potvin *et al*. (27) ($\alpha_1 = \alpha_2 = 0.0294$) and Maurer, Jones, and Chen (52) ($\alpha_1 = \alpha_2 = 0.0263$ for maximum combination test with (*w, w**): (0.5, 0.25)) were taken from Maurer, Jones, and Chen (52) manuscript. Type 1 significance levels were obtained using the function 't1e.tsd', considering N_1 = (12, 24, 36), CV_W between 0.4 and 0.8 at increments of 0.01; and constraints $N \leq 4000$ and $N_2 \geq 0.5\ N_1$. The result was $\alpha_1 = \alpha_2 = 0.0302$. Then, we used the 'power.tsd' function with 10E6 simulations per N_1 and CV_W pair with target power 80% and planned and true *GMR* 0.95 to calculate the power achieved and mean *N*. Results show a power and sample size which are comparable across methods.

Table 11. Power and mean sample size with constraint N ≤ 4000 for HVD

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

Type 1 method: Modified Potvin B method with $max(N = N_1 + N_2) = 4000$, and $N_2 \geq N_1/2$
MCT: Maximum Combination Test; *HVD*: Highly variable drugs
Target power = 0.80 and planned and true *GMR* = 0.95

4.5. Discussion

Average bioequivalence (ABE) studies using *TSD* offer several advantages over conventional crossover trials. They provide an attractive solution to address some of the uncertainty that exists on the true variability value when the trial is originally designed, although they are typically more complex and exhaustive and require more efforts and time for planning and implementing (50). *TSD*s should be standardized and agreed between the pharmaceutical industries and the agencies, in particular, about the specific pathways to control the type I error *(T1E)* rate, usually at 5%. We adapted two methodology types proposed initially by Potvin *et al*. (27) to adjust the significance levels at each stage which controls the *T1E*. Adjusted significance levels were higher

than 0.0300 in most cases with a power of at least 80%. We also adapt and compare our approach with Xu *et al*. (39) and Maurer, Jones, and Chen (52) to conclude that operating characteristics are comparable.

Our approach is implemented using our own function. In summary, given a grid of { N_1, CV_W} and an initial warm-up α_1 and α_2 values, we found adjusted α_1 and α_2 and the (N_1, CV_W) pair with maximum empiric *T1E* (Tables 6 and 7). In the grid, we should cover an important range of CV_W values to ensure that the true/population CV_W is included. In Molins *et al*. (21) we assessed a particular case assuming that the degree of uncertainty was encompassed by evaluating CV_W at 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, and 0.6. We have now improved this feature sweeping CV_W range values at intervals of 0.01-units. In addition, we have considered the case of an applicant/sponsor who assumes CV_W values which unfortunately do not contain the true unknown CV_Wvalue. So, the *T1E* is controlled, by default, at an overall significance level considering the CV_W assumed ± 0.05. We admit that though it is sometimes necessary to cover such a range of CV_W values, there is always a risk of losing some power.

We provide a methodology which usually adjusts significance levels above 0.0294 and strictly controls the *T1E*. The significance levels of 0.0294 at both stages (27,30) are not some kind of a 'natural constant', because they depend on the design, treatment effect, variability, target power, or sample size, and so they are entirely empiric and must be estimated in simulations. In addition, they did not always control the overall *T1E* rate at a maximum 5% (31,32). For example, the original 'Potvin D' method only grants the maintenance of the *T1E* rate below 5.2%. And, by using the modified 'Potvin C' method, with *GMR* = 0.95, $\alpha_1 = \alpha_2 = 0.0294$, $N_1 = 12$, and a true $CV_W = 0.2$, the *T1E* is assessed at 5.3% (5.5% in case of *GMR* = 0.90).

Other methods to adjust significance levels are discussed by some authors and regulatory instances (34,39,47,48,51-53). In order to see how some operating characteristics compare to each other, we followed the frameworks (N_1 and CV_W range) used by Xu *et al*. (39) and Maurer, Jones, and Chen (52) to calculate the significance levels. We saw comparable results on the overall sample size, the percentage of studies jumping to stage 2 or the overall power (Tables 10 and 11). We

highlight that our method is very flexible because it is customizable in many different ways.

We also allow α_1 and α_2 being different from each other. O'Brien and Fleming (54) proposed a group sequential procedure with boundary values that decreased over the stages to make early stopping less likely. Xu *et al*. (39) also found significance levels where $\alpha_1 < \alpha_2$ with α_1 at stage 1 close to 0.025. Adaptive strategies are persuasive because they allow stopping the trial at stage 1 and declare bioequivalence with a low number of N_1 subjects. However, it will be difficult to declare bioequivalence at stage 1 if α_1 is very conservative. Though $\alpha_1 < \alpha_2$ seems the most natural way of proceeding, an applicant may be interested in being more permissive at stage 1, e.g. Lan and DeMets α-spending function. We allowed both $\alpha_1 < \alpha_2$: 0.0294, 0.0310, and $\alpha_1 > \alpha_2$: 0.0320, 0.0279 (Table 7 and Figure 10).

Maurer, Jones, and Chen (52) provided an attractive principled solution based on a maximum combination test to control the *T1E* inflation. While simulation-based approaches are criticized because require the investigation of many scenarios (in our case, CV_W range values should be large enough) to ensure the control of this error, this principled method also relies on specifying two weights *w* and *w** which need to be pre-defined a priori, and an initial guess on the CV_W. Additionally, there is no a simple formula of obtaining the power which is desirable to compare the different settings (N_2, weights, futility criteria). In analogy to the Potvin *et al*. (27) methods, it is needed to undertake simulations to gain those values.

Some other differences between methodologies lie on the specifics of futility rules to stop the trial at stage 1. Xu *et al*. (39) and Maurer, Jones, and Chen (52) specified futility rules based on 90% CI of the formulation effect completely outside of some margins. Also, Xu *et al*. (39) and Karalis and Macheras (32), included a futility criterion to stop the study at stage 1 based on a total study size upper limit. We included an upper limit for *N* of 150 subjects.

We consider *HVD* a special case under investigation (28,49,55-58). We compared *EMA* Reference Scaled Average bioequivalence (*RSABE*) based on replicate *TRTR/RTRT* designs and *TSD* methods (21). In terms of power, we saw that both approaches

perform similarly despite adaptive methods usually requires a higher mean sample size to reach the same power, especially for clearly *HVD*. Nevertheless, we demonstrated suitable power at the stage 1 in some cases. But for true CV_W values above 0.29, the power at stage 1 is low and the proportion of studies switching to stage 2 high. In addition, assertion of bioequivalence becomes difficult for CV_W greater than 0.5 (data shown in Tables 10 and 11), as bioequivalence seldom can be declared at stage 2. It is arguable launching a drug into the market with such a within-subject variability, or even starting a study with such a low expected power (42).

We calculated CV_W by means of the coefficient of variation under homoscedasticity assumption $CV_{WR} = CV_{WT} = CV_W$. Kang *et al*. (59) showed that bioequivalence testing with heterogeneous residual variances gives similar performance for CV_W lower than 0.4. In fact, power curves (Figure 10) show that the constraint that we are using of a maximum of 150 subjects provokes a power decrease for CV_W values above 0.4.

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

5. FUNCTION 'T1E.TSD' TO PRESERVE THE TYPE I ERROR RATE USING TWO-STAGE DESIGNS

Based on simulations, we created the function 't1e.tsd' included within the package 'betsd' to preserve the overall type I error. The function provides with the adjusted significance levels to be used in each stage, α_1 and α_2, the probability to jump to a stage 2, the sample size at the stage 2, and the power at stages 1 and 2.

5.1. Introduction

The use of two-stage adaptive 2×2 crossover designs (*TSD*) in bioequivalence studies seems to be a beneficial alternative to the 2×2 crossover design. In accordance with the EMA guideline (7), the number of participants can be expanded if average bioequivalence (ABE) has not been demonstrated in the first group of subjects. The results for the initial and the second group are combined for the final assessment. They are especially useful in case of drugs with little evidence about the true within-subject variability, and for highly variable drugs (*HVD*), i.e. with a within-subject coefficient of variation, $CV_W \geq 0.3$ (21,49).

The critical point about using *TSD*s is the difficulty to preserve the type I error rate (*T1E*) (7,34,51,52). Significance level boundaries α_1 and α_2 at each stage can be adjusted in various ways that are not fully specified in the regulations (7,40).

- (α_1, α_2) = (0.0294, 0.0294): Potvin *et al.* 'Method B' and *GMR*=0.95 (27).
- (α_1, α_2) = (0.0294, 0.0294): Potvin *et al.* 'Method C', alpha0=0.05, and *GMR*=0.95.
- 'Method D' = 'Method C' but with (α_1, α_2) = (0.028, 0.028), alpha0=0.05, and *GMR*=0.9.
- (α_1, α_2) = (0.0269, 0.0269): Fulgsang 'Method C/D' (method="C", *GMR*=0.9, targetpower=0.9) (42,51,60).
- (α_1, α_2) = (0.0274, 0.0274): Fuglsang 'Method C/D' (method="C", targetpower=0.9) (42,51,60).
- (α_1, α_2) = (0.0280, 0.0280): Montague *et al.* 'Method D' (method="C", *GMR*=0.9) (31).

- (α_1, α_2) = (0.0284, 0.0284): Fulgsang 'Method B' (*GMR*=0.9, targetpower=0.9) (42,51,60).
- - (α_1, α_2) = (0.0304, 0.0304): Kieser & Rauch (34).

But, α_1 and α_2 are not some kind of a 'natural constant', because they depend on the design method (27,30,31), treatment effect (*GMR*), variability (CV_w), target power, or sample size (N_1). Yet, in some circumstances, these method do not grant in strong sense the maintenance of the T1E rate below 5% (31,32).

In turn, like in Xu *et al*. (39) we assumed that the adjusted significance levels at both stages may be different, $\alpha_1 \neq \alpha_2$. In the Chapter 4 we present some scenarios with adjustments of (α_1, α_2) = (0.0294, 0.0310) or (α_1, α_2) = (0.0320, 0.0279) at stages 1 and 2, respectively.

5.2. Study objectives

We present and open *R* package called 'betsd', based on an iterative simulation method, which preserves in a strong sense the overall *T1E*. It includes an accurate description of all properties of the function 't1e.tsd' which serves to calculate the adjusted significance levels at stages 1 and 2. The function allows testing as many (N_1, CV_W) scenarios (combination of pairs) as considered. It provides with the adjusted significance levels to be used in each stage, α_1 and α_2, the probability to jump to a stage 2, the sample size at the stage 2, and the power at stages 1 and 2. It is flexible and intuitive because the applicant can adapt it to any real situation even with little knowledge on the multiplicity issue.

This package is hosted on *GitHub https://github.com/eduard-molins/betsd*. Also, source code to reproduce the results is available as Supporting Information on the journal's web page bimj2181-sup-0002-SuppMat.zip.

5.3. Function t1e.tsd usage

5.3.1. Description

The function 't1e.tsd' calculates, by iterative search, the adjusted significance levels to be used in each stage of *TSD*s, to ensure an overall *T1E* below a specified significance level.

This function calculates the empiric *T1E* and power of stage 2 according to a 'modified' Potvin *et al*. methodology (Figure 8) (22,27). But instead of simulating individual subject data, the statistics point estimate at stage 1, mean square error at stage 1 (or intra-subject residual variance calculated from CV_W), and point estimate at stage 2, and sum of square at stage 2 are simulated via their associated distributions (normal and χ2 distributions).

The function 't1e.tsd' calls the functions 'power.tsd' and 'sampleN2.TOST' both included in the package 'Power2Stage' (40) hosted on CRAN.

Using simulations, the function 'power.tsd' allows calculating the power, type I error, the probability to jump to a stage 2, the sample size at the stage 2, and the power at stages 1 and 2. The sample size re-estimation is performed during the interim analysis using the function 'sampleN2.TOST', given the method/design used according to Figure 8, based on the initial sample size, the estimated within subject variability observed in the interim look, the significance level, and the target power. The knowledge about the estimated treatment effect in the interim analysis is not used for sample size re-estimation/recalculation. We assumed a fixed true treatment effect of *GMR* = 0.95 after the stage 1 since Cui *et al.* (43) showed that a determination of the stage 2 sample size based on an interim estimate of the *GMR* can substantially inflate the probability of type I error in most practical situations.

Note that both functions 'power.tsd' and 'sampleN2.TOST' require the specification of the significance level argument. We should ensure that the *T1E* never exceeds the significance level. However, the *T1E* depends on the study framework, i.e., on the design, treatment effect, variability, target power, or sample size, and so they are entirely empiric and must be estimated in simulations. But, there is not a 'natural constant' that serves in all situations, so, given a framework, the function 't1e.tsd' allows calculating the significance levels to be used at each stage.

5.3.2. Usage

Our function 't1e.tsd' is found in the Biometrical Journal supporting information: bimj2181-sup-0001-SuppMat.docx (see Appendix 1).

```
t1e.tsd(N1, CV, GMR = 0.95, Nmax = 150, min.N2 = N1/2, type = 1,
        alpha = 0.05, alpha1, alpha2, targetpower = 0.8,
        setseed = TRUE, theta1, theta2,
        details = TRUE, print = TRUE, ...)
```

5.3.3. Arguments

N1	Sample size of stage 1.
CV	Within subject coefficient of variation (use *e.g.*, 0.3 for 30%).
GMR	Expected geometric mean ratio to be used in decision scheme (power calculations in stage 1 and sample size estimation for stage 2). By default 0.95.
Nmax	Overall maximum number of subjects (sum of sample sizes in both stages, *i.e.*, N = N1 + N2). By default 150, see Chapter 4.
min.N2	Minimum number of subjects at stage 2. By default N1/2, see Chapter 4. Set min.N2 = 0 to cancel any limitation on the sample size at the stage 2, N2.
type	Type 1 or 2 methodology. By default 1.
alpha	Target overall significance level (both stages). By default 0.05.
alpha1	Initial significance level at stage 1. By default 0.0294.
alpha2	Initial significance level at stage 2. By default 0.0294.
targetpower	Power threshold in the power monitoring steps and power to achieve in the sample size estimation step. By default 0.8.
setseed	Simulations are dependent on the starting point of the (pseudo) random number generator. To avoid differences in power for different runs a setseed(1234567) is issued if set to TRUE, the default. Set this argument to FALSE for a random seed.

theta1	Lower limit of the bioequivalence range. By default 0.8.
theta2	Upper limit of the bioequivalence range. By default 1.25.
details	If set to TRUE (default) shows intermediate results in the console. Set this argument to FALSE to suppress intermediate results.
print	If set to TRUE (default) shows final results in the console. Set this argument to FALSE to return a list of final results.
...	Optional additional arguments. See package Power2Stage, function power.tsd.

5.3.4. Details

The type 1 method (Figure 8) uses the same adjusted alpha1 and alpha2 in stages 1 and 2. The type 2 method uses an unadjusted alpha if interim power is at least the target power, or adjusted alpha1 and alpha2 in stages 1 and 2 otherwise.

By default, the maximum sample size *Nmax* in both stages, max($N = N_1 + N_2$), was restricted to 150 healthy volunteers, and we consider that the minimum number of healthy volunteers to be enrolled in the stage 2, N_2, was $N_1/2$. After computing the required sample size in the second step, say N_2 (to ensure the required power), the number of additional subjects is computed as $N_2 = \max(N_1/2, N_2)$, but if $N_1 + N_2 >$ *Nmax*, the study is terminated due to futility. These criteria and default values are based on Molins *et al.* (21) and Molins *et al*. (22).

5.4. The iterative method

The iterative method is described in Figure 9.

We start with a 'current_alpha' (arbitrary) initial α_1 and α_2. By default, we set these values to Potvin's constant, i.e. $\alpha_1 = \alpha_2 = 0.0294$. We evaluate some significance levels to find the final empiric significance level at each stage which controls the overall *T1E* below 5%:

```
new_alpha1 <- seq(current_alpha[1, 1] - 0.0005, current_alpha[1, 1] +
0.0005, length.out = 7)
new_alpha2 <- seq(current_alpha[1, 2] - 0.0005, current_alpha[1, 2] +
0.0005, length.out = 7)
```

```
> new_alpha1
[1] 0.02890000 0.02906667 0.02923333 0.02940000 0.02956667 0.02973333
0.02990000
> new_alpha2
[1] 0.02890000 0.02906667 0.02923333 0.02940000 0.02956667 0.02973333
0.02990000
```

For simplicity, we fixed α_1 to its initial value, and the adjustment is only done for α_2. For example, if 'current_alpha' is set to 0.027 for α_1, and 0.0300 for α_2, the algorithm will again evaluate scenarios with $\alpha_1 = 0.070$ and α_2 as follow:

```
new_alpha2 <- seq(current_alpha[1, 2] - 0.0005, current_alpha[1, 2] +
0.0005, length.out = 7)
```

```
> new_alpha2
[1] 0.02950000 0.02966667 0.02983333 0.03000000 0.03016667 0.03033333
0.03050000
```

And, the pairs evaluated are: (0.027, 0.02950000), (0.027, 0.02966667), … , (0.027, 0.03050000).

T1E are simulated/obtained for each of these significance levels by means of the function 'power.tsd' (from repository CRAN). Then, we fit linear and quadratic

regression models with *T1E* as independent variable, which is adjusted by (α_1, α_2) pairs, and we choose the 'best' model based on the minimum Akaike Information Criterion (*AIC*). To obtain the 'adj_alpha', we isolate (α_1, α_2) for *T1E* = 5%. For example, if *e.g.* if $\alpha_1 = \alpha_2$ and the min(*AIC*) is obtained from the linear regression model, then 'adj_alpha' is obtained as:

$$\alpha_{adj1} = \alpha_{adj2} = \frac{(0.05 - \widehat{\beta}_0)}{\widehat{\beta}_1}.$$

Details can be followed in the *R*-code, but 4 scenarios will determine if we already reached the desired empiric significance level, or if the algorithm should start again with a new 'current_alpha' = 'adj_alpha'.

SCENARIO 1/4: *T1E* < 0.05 and *diff* <= 2E-04

SCENARIO 2/4: *T1E* > 0.05 and *diff* <= 2E-04

SCENARIO 3/4: *T1E* < 0.05 and *diff* > 2E-04

SCENARIO 4/4: *T1E* > 0.05 and *diff* > 2E-04

Where *diff* is the difference between the 'current_alpha' and the 'adj_alpha'.

5.5. Computing time

Please, see our package 'betsd' on https://github.com/eduard-molins/betsd, in folder 'man':

For example:

t1e.tsd(N1 = 24, CV = c(0.3, 0.4, 0.5, 0.6), GMR = 0.95, type = 1).

After 3 iterations:

We obtained adjusted alpha levels of 0.0307 at both stages and a maximum empirical type I error of 0.04998.

Note: runtime ~15 minutes on a Xeon E3-1245 Quadcore 3.4 GHz,

runtime ~6 minutes on a MacBook 2.6 GHz Intel Core i5.

For example:

t1e.tsd(N1 = 12, CV = 0.2, GMR = 0.9, alpha1 = 0.0280, alpha2 = 0.0280, type = 2).

After 2 iterations:

We obtained adjusted alpha levels of 0.0268 at both stages and a maximum empirical type I error of 0.049591.

Note: runtime ~7 minutes on a Xeon E3-1245 Quadcore 3.4 GHz,

runtime ~5 minutes on a MacBook 2.6 GHz Intel Core i5.

5.6. Case study

As an example, let's imagine someone interested in obtaining the adjusted significance levels, considering the following assumptions on N_1 and CV_W:

N_1: 12, 18, 24.

CV_W: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6.

> t1e.tsd(N1 = c(12, 18, 24), CV = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6), GMR = 0.95, targetpower = 0.8, type = 1, alpha = 0.05).

In this case, the cartesian products of N_1 and CV_W will be evaluated, i.e. the following 18 scenarios:

(12, 0.1), (12, 0.2), (12, 0.3), (12, 0.4), (12, 0.5), (12, 0.6), (18, 0.1), (18, 0.2), (18, 0.3), (18, 0.4), (18, 0.5), (18, 0.6), (24, 0.1), (24, 0.2), (24, 0.3), (24, 0.4), (24, 0.5), (24, 0.6).

Using the function 't1e.tsd' and the debugger we obtain:

t1e.tsd(N1 = c(12, 18, 24), CV = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6), GMR = 0.95, targetpower = 0.8, type = 1, alpha = 0.05).

Browse[2]>

	N1	CV	GMR	alpha1	alpha2	min.N2
1	12	0.1	0.95	0.0294	0.0294	6
2	18	0.1	0.95	0.0294	0.0294	10
3	24	0.1	0.95	0.0294	0.0294	12
4	12	0.2	0.95	0.0294	0.0294	6
5	18	0.2	0.95	0.0294	0.0294	10
6	24	0.2	0.95	0.0294	0.0294	12
7	12	0.3	0.95	0.0294	0.0294	6
8	18	0.3	0.95	0.0294	0.0294	10

9	24	0.3	0.95	0.0294	0.0294	12
10	12	0.4	0.95	0.0294	0.0294	6
11	18	0.4	0.95	0.0294	0.0294	10
12	24	0.4	0.95	0.0294	0.0294	12
13	12	0.5	0.95	0.0294	0.0294	6
14	18	0.5	0.95	0.0294	0.0294	10
15	24	0.5	0.95	0.0294	0.0294	12
16	12	0.6	0.95	0.0294	0.0294	6
17	18	0.6	0.95	0.0294	0.0294	10
18	24	0.6	0.95	0.0294	0.0294	12

These are the *T1E* assessed with 30,000 simulations for each N_1 and CV_W pair using the function 'power.tsd' nested within the function 't1e.tsd':

Browse[2]> T1E

	0.1	0.2	0.3	0.4	0.5	0.6
12	0.0309	0.04846667	0.04600000	0.03546667	0.02546667	0.01713333
18	0.0305	0.04103333	0.04953333	0.03770000	0.02863333	0.01640000
24	0.0299	0.03406667	0.04946667	0.04430000	0.03380000	0.01686667

These are the N_1, CV_W combinations where *T1E* >= *P90%* (percentile 90%):

Browse[2]> d90

	N1	min.N2	CV	GMR	alpha1	alpha2
1	18	10	0.3	0.95	0.0294	0.0294
2	24	12	0.3	0.95	0.0294	0.0294

Now, for accuracy purposes these two scenarios are simulated 10E6 times each;

Browse[2]> T1E_high

	N1	CV	GMR	min.N2	alpha1	alpha2	pbioequivalence
1	18	0.3	0.95	10	0.0294	0.0294	**0.048160**
2	24	0.3	0.95	12	0.0294	0.0294	**0.047816**

Where the max(T1E) is:

Browse[2]> max_T1E

	N1	CV	GMR	min. N2	alpha1	alpha2	pBE	pbioequivalence_s1	pct_s2	Nperc
1	18	0.3	0.95	10	0.0294	0.0294	**0.04816**	0.02445	96.7072	28, 46, 76

With N_1 = 18 and CV_W = 0.3, with significance levels of 0.0294, 3 new significance levels below and above are tested as follow:

Browse[2]> N_d

	N1	CV	GMR	min.N2	alpha1	alpha2
1	18	0.3	0.95	10	0.02890000	0.02890000
2	18	0.3	0.95	10	0.02906667	0.02906667
3	18	0.3	0.95	10	0.02923333	0.02923333
4	18	0.3	0.95	10	0.02940000	0.02940000
5	18	0.3	0.95	10	0.02956667	0.02956667
6	18	0.3	0.95	10	0.02973333	0.02973333
7	18	0.3	0.95	10	0.02990000	0.02990000

Type I errors are evaluated for each scenario with 10E6 simulations:

Browse[2]> res_new_d_T1E

	CV	N1	GMR	min.N2	alpha1	alpha2	T1E
1	0.3	18	0.95	10	0.02890000	0.02890000	0.047318
2	0.3	18	0.95	10	0.02906667	0.02906667	0.047676
3	0.3	18	0.95	10	0.02923333	0.02923333	0.047847
4	0.3	18	0.95	10	0.02940000	0.02940000	0.048160
5	0.3	18	0.95	10	0.02956667	0.02956667	0.048481
6	0.3	18	0.95	10	0.02973333	0.02973333	0.049002
7	0.3	18	0.95	10	0.02990000	0.02990000	0.049108

And, after the regression model, we obtain the following adjusted significance levels, alpha1 and alpha2, at stages 1 and 2, respectively:

Browse[2]> alpha_adj

	alpha1	alpha2
[1,]	0.03035564	0.03035564

We start again with the 'current_alpha' at 0.03035564:

Browse[2]> d90

	N1	min.N2	CV	GMR	alpha1	alpha2
1	18	10	0.3	0.95	0.03035564	0.03035564
2	24	12	0.3	0.95	0.03035564	0.03035564

Final results are:

Run-in with alphas 0.0294, 0.0294
- max.TIE 0.048160 at N1 = 18 and CV = 0.3
Iteration 1 with alphas 0.03035, 0.03035
- max.TIE 0.049798 at N1 = 18 and CV = 0.3

Method type = 1 (B)
setseed = 1234567
bioequivalence acceptance range (theta1, theta2) = 0.8 ... 1.25
Power calculation method = nct
N1 = 12 18 24
GMR = 0.95
CV = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6
Nmax = 150
min.N2 = 6 10 12
Adjusted alpha at stage 1 = 0.0303 and alpha at stage 2 = 0.0303
Maximum empirical type I error = 0.049641 at N1 = 18 and CV = 0.3
Power: Overall probability of bioequivalence = 0.81975
Power: Probability of bioequivalence at stage 1 = 0.22431
Studies in stage 2 = 77.246%
5% 50% 95% percentiles of N = 18, 44, 76
max.iter = FALSE

(Note: Runtime was 5' 16'' on my computer, a MacBook 2.6 GHz Intel Core i5)

5.7. Discussion

The significance level proposed by Pocock's and Potvin *et al.* approaches (27,30), i.e. based on a significance level of 0.0294, is not some kind of a 'natural constant', because *TSD*s can be based on different frameworks, and the adjusted significance levels are entirely empiric and must be estimated in simulations. The fact that 0.0294 'worked' in Potvin B was fortunate (and we saw a slight inflation in Method C). That's why Potvin *et al*. (27) wrote they did not seek to find the best possible *TSD* solution, but rather to find good significance levels that could be used by sponsors without further validation. When the framework is modified (our case) it is of utmost importance to find suitable adjusted significance levels. Significance level boundaries can be adjusted in various ways that are not fully specified in the regulations. And we propose a methodology which controls the overall *T1E* in a strong sense below a significance level, e.g. 5%. We find significance levels which usually are above 0.0294, providing much power, and always ensuring that the *T1E* is controlled for any parameter assumption/framework. Also, the algorithm also allows α_1 being different than α_2.

*TSD*s provide investigators with an attractive solution to address some of the uncertainty that exists when the trial is originally designed, allowing stopping the study at stage 1 with a small N_1, avoiding to unnecessarily soar *N* above what is reasonable to attain a desired power, *e.g.*, 80%. And they are especially useful in case of drugs with little evidence about the true within-subject variability, and for highly variable drugs (*HVD*), i.e. with a $CV_W \geq 30\%$.

6. GENERAL 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 difficult to assess bioequivalence through standard procedures using reasonable sample sizes, thus delaying treatment. After many years of discussion, some regulatory 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, called reference scaled average bioequivalence (*RSABE*) (7).

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. In this case, 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.

In general, average bioequivalence (ABE) studies using *TSD*s offer several advantages over conventional crossover trials. They provide an attractive solution to address some of the uncertainty that exists on the true within subject CV_W value when the trial is originally designed, although they are typically more complex and exhaustive and require more efforts and time for planning and implementing.

A described limitation of *TSD* simulation methods is that unless all possible CV_W scenarios for the intended design and analysis are investigated, it is impossible to be sure that the type I error rate is controlled. In our manuscript, we dealt this issue being a bit conservative, considering a broad bunch of CV_W values (where the true unknown CV_W is quite likely be included), from a lower to an upper bound, where the type I error was assessed at CV_W increment values of 1%, e.g. from CV_W between 10% to 20%, we evaluated 10%, 11%, 12%, …, 20%. This method is simple to apply and useful from both the sponsors as well as from the regulatory bodies.

We showed that our results are quite comparable with the latest proposed methodologies by Xu *et al*. (Table 10) (39) and Maurer, Jones, and Chen (Table 11)

(52), in particular on the main features like the overall sample size, percentage of studies jumping to stage 2, and power.

Xu *et al*. (39) implemented two innovative Methods to calculate bioequivalence using *TSD* using simulations and optimizations to look for optimal solutions. They studied *TSD* Methods, E and F, obtaining cost functions comparing the sample size achieved using *TSD*s versus the sample size resulting from conventional single stage 2x2 crossover designs. They created an average cost function for each *GMR* and CV_W combination value. For *TSD*s, *GMR*s ranged 70-100% at increments of 5% and CV_W were split and evaluated into two design spaces, one ranged 10-30%, and the other 30-55%, both at increments of 5%. For conventional, single stage designs, they used the same CV_W values but *GRM*s were fixed at 95%. Using simulations, they were varying (and fixing) the two significance levels α_1 and α_2, the stage 1 sample size (N_1), and a futility criterion (*f*), and checking whether for all *GMR* and CV_W combination values the power was of at least of 80% (at a true *GMR* of 0.95) and the type I error rate (at a true *GMR* of 0.8 of bio-inequivalence) of 5% maximum. They resolved an optimization problem, obtaining the optimal design based on the lowest cost among valid combinations of α_1, α_2, N_1, and *f*. Note that maximum number of subjects allowed by Xu et. al was not optimized but rather fixed based on practical considerations (42 for the CV_W range 10–30% and 180 for the CV_W range 30-55%). In our manuscript this ceiling sample size was set at 150 subjects for any CV_W.

Also, we admit that the well-founded method of Maurer, Jones, and Chen (52) gives a desirable solution. They used a standard inverse-normal *p*-value combination test, in conjunction with standard group sequential techniques (called it maximum combination test), to guarantee the control of type I error rate at any given significance level.

But Maurer, Jones, and Chen (52) methodology has also some limitations in practice. The maximum combination test is based on a weighted average of a transformation of the *z*-values from stages 1 and 2 and use this as the final test statistic at the end of the trial (using N_2 data), where *w* is the weight for stage 1 and *1-w* the weight for stage 2. In fact, they propose using two sets of weights, *w* and *w**, where *w** is below *w* for $N_2 > N_1$ (i.e. during the sample size re-estimation). In the case study, they are proposing

using *w*=0.5 and *w**=0.25. But, the optimal choice of the weights for the combination tests depend on the expected magnitude of a sample size increase of the stage 2 compared to the stage 1, which in practice is unknown because *w* and *w** must be pre-specified before the trial starts (in the protocol). They show that small differences on *w* and *w** assumption lead to a (non-negligible) different nominal adjusted significance levels.

In addition, Maurer, Jones, and Chen (52) method also depends on the CV_W assumption. An initial CV_W guess is used to obtain an overall sample size *n* (with target power usually 0.8) from where *N1* = *N*/2, so the *z1*-value (quantile) assessed at stage 1 depends on *N1* and therefore on the initial CV_W considered.

In comparison with Potvin *et al*. (27), Maurer, Jones, and Chen (52) usually provides a lower sample size N_2 at the stage 2 to reach a global desired power. The assessment is based on comparing a 'target conditional power', which uses power achieved at stage 1, versus a 'conditional power', which uses the conditional errors for maximum combination test (using the CV_W estimated at interim), a formulation effect of 0.95, and N_2. Starting on an initial N_2 set to 4, the 'conditional power' is assessed at increments of 2 subjects until it exceeds the 'target conditional power'. In any case, there is no a simple formula of obtaining the power which is desirable to compare the different settings (N_2, weights, futility criteria). In analogy to the Potvin *et al.* methods, it is needed to undertake simulations to gain those values.

Both, Potvin *et al*. (27) and us calculate N_2 based on a non-conditional pre-defined target power (usually 0.8), using CV_W estimation at interim, and a formulation effect of 0.95. Then, the power finally achieved is calculated using pooled data from both stages.

Like in Xu *et al*. (39) we assumed that the adjusted significance levels at both stages may be different, $\alpha_1 \neq \alpha_2$. We discussed that this is particularly interesting for someone interested in being more permissive at stage 1 than at stage 2, or vice versa. In Table 7, we present some scenarios with adjustments of (α_1, α_2) = (0.0294, 0.0310) or (α_1, α_2) = (0.0320, 0.0279) at stages 1 and 2, respectively.

Also, should be considered that regardless of the methodology used, principled or simulations, the overall power depends on a non-biased estimated CV_W at stage 1. So, there is no guarantee that the power is finally achieved in any case.

7. CONCLUSIONS

We compared two variants of two-stage adaptive 2×2 crossover designs (*TSD*) 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, *TSD*s are a good option to consider, as they have a power of around 80% at the stage 1 for non-highly variable drugs while at the same time they offer the opportunity for stepping up to the stage 2 (including additional subjects) for truly bioequivalent products.

*TSD*s can be applied to bioequivalence studies more widely. We provide a function, which is quite flexible and easy to use for any applicant, to adjust the significance levels at each stage which strictly grant the control of the type I error for different assumptions on the *GMR*, N_1, and CV_W. In addition, we provide operating characteristics like the probability to jump to a stage 2, the sample size at the stage 2, and the power at stages 1 and 2.

In turn, *TSD*s provide investigators with an attractive solution to address some of the uncertainty that exists when the trial is originally designed, allowing stopping the study at stage 1 with a small N_1, avoiding to unnecessarily soar *N* above what is reasonable to attain a desired power, *e.g.*, 80%. And they are especially useful in case of drugs with little evidence about the true within-subject variability.

*TSD*s are supposed to be used when there is a lack of knowledge regarding some aspects of the product e.g., variability, or expected (dis)similarity with the reference product, even though they may not be the best options for other cases like highly variable drugs.

With this work, we would like to contribute towards a global harmonization and convergence of generic drug developments.

8. FUTURE AREAS OF RESEARCH

8.1. Population and individual bioequivalence approaches

The average bioequivalence (ABE) approach focuses only on the comparison of the mean bioavailabilities of *R* and *T* (equation 1). However, this has been pointed to be insufficient for assessing switchability between formulations. This is because ABE does not consider the variance of bioavailability values, and does not guarantee whether or not *R* and *T* show the same therapeutic effects in each individual subject (13,61).

The *population bioequivalence* (PBE) approach guarantees prescribability by assessing bioequivalence adding the difference in the total variance of bioavailability values of the *T* and *R* formulations ($\sigma_{TT}^2 - \sigma_{TR}^2$) to the difference in mean bioavailabilities of the *T* and *R* formulations $(\mu_T - \mu_R)^2$, see equation 2.

Individual bioequivalence (IBE) guarantees switchability by assessing bioequivalence by adding three elements: the difference in mean bioavailability of *T* and *R* formulations as determined in the bioequivalence approach, the subject-by-formulation interaction, σ_D^2, and the difference in intrasubject variances of bioavailability values of *T* and *R,* $\sigma_{WT}^2 - \sigma_{WR}^2$, see equation 3. Note that IBE requires replicate crossover designs.

Because pharmaceutical products have different therapeutic ranges (narrow to wide) and variances in bioavailabilities values (large to small), these characteristics are to be considered to determine PBE and IBE acceptance ranges, δ_{PBE} and δ_{IBE}.

Average bioequivalence: Evaluated with the difference of mean bioavailability (BA) (in the logarithmic scale).

$$|\mu_T - \mu_R| < \delta_{ABE} \text{ (equation 1).}$$

Population bioequivalence: Evaluated with the difference of mean BA and total variance of bioavailability (in the logarithmic scale).

$$[(\mu_T - \mu_R)^2] + [\sigma_{TT}^2 - \sigma_{TR}^2] < \delta_{PBE} \text{ (equation 2).}$$

Individual bioequivalence: Estimated with the difference of mean BA, subject-by-formulation interaction, and intrasubject variance (in the logarithmic scale).

$$[(\mu_T - \mu_R)^2] + \sigma_D^2 + [\sigma_{WT}^2 - \sigma_{WR}^2] < \delta_{IBE} \text{ (equation 3).}$$

Where, μ_T: Average of the *T* formulation, μ_R: Average of the *R* formulation, σ_{TT}^2: Total variance of the *T* formulation, σ_{TR}^2:Total variance of the *R* formulation, σ_D^2: Subject-by-formulation interaction $\left[Var(\mu_{Tj} - \mu_{Rj})\right]$, σ_{WT}^2: Intrasubject variance of the *T* formulation, σ_{WR}^2: Intrasubject variance of the *R* formulation, δ: bioequivalence acceptance range.

See the example in the Figure 11 (61) with differences among average, population, and individual bioequivalence. Case 1 meets average, population, and individual bioequivalence because the means and total variances of BA values of *T* and *R* are equal, and the bioavailabilities of *T* and *R* in individual subjects are almost identical. In Case 2 individual bioequivalence is not met because bioavailabilities values in *T* and *R* in the individual subjects differ, but average and population bioequivalence are met. In Case 3, only bioequivalence is met.

Figure 11. Differences among average, population, and individual bioequivalence

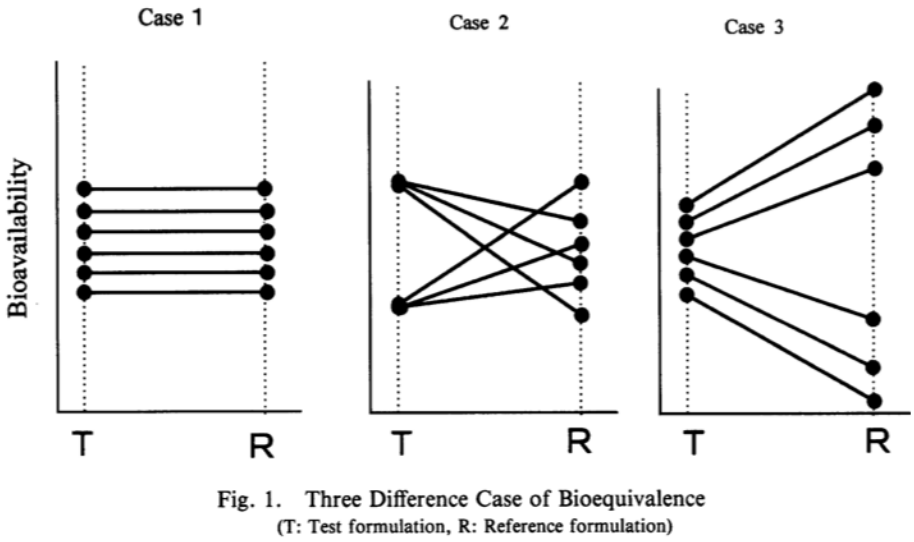


Fig. 1. Three Difference Case of Bioequivalence
(T: Test formulation, R: Reference formulation)

Source: Nakai *et al.* (61)

Some replicate design studies have demonstrated that the scaled acceptance range offered by the aggregate criterion may result in clinically unacceptable decisions in favor of IBE, although bioequivalence does not hold.

Other concerns not yet resolved are that aggregate hypotheses on the logarithmic scale have no obvious translation into the original scale; also, scaling corresponds to a modification of the bioequivalence acceptance limits, but is handled differently for IBE

and PBE than for bioequivalence; and, the proposed criteria do not consider a hierarchical testing (first means, then variances, lastly subject-by-formulation interaction). If, IBE is further pursued, statistical research should focus on disaggregate criteria that allow exact stepwise procedures for evaluating untransformed parameters (62).

Because of these limitations, PBE and IBE were put on hold early in 2000, and the final guidance calls for ABE to remain as the primary criterion by which new formulations may be judged ready for access to the marketplace.

Proposal

As previously shown, there is a limitation of declaring average bioequivalence (ABE) just using the common bioequivalence definition based on comparing the mean ratios of *R* and *T*, because in this definition the variance of *R* and *T* is not considered. This is the reason why the regulatory agencies introduced the idea of *PBE* and *IBE*, although they have some inconvenient and were put on hold early in 2020.

We propose a new line of research to study bioequivalence. Bioequivalence might be claimed based on a combined assessment: Based on the usual ABE approach (maybe re-considering the bioequivalence margins, currently fixed at 0.80-1.25 for common drugs) in order to facilitate declaring bioequivalence in case of products whose variability in the generic *T* is non-superior to *R*, i.e. based on a non-superiority test with a certain non-superior limit between *R* and *T*.

Alternatively, using the current definition of ABE to declare bioequivalence, might be discussed during the pricing and reimbursement negotiations (at local regulations) about the variability of the generic drug *T*, considering lower prices (for consumers) in case of less variable generic products.

8.2 Biosimilars

Biological medicines (‘biologicals’) contain active substances from a biological source, such as living cells or organisms. Biological medicines are well established in clinical practice and in many cases they are indispensable for the treatment of serious and chronic conditions such as diabetes, autoimmune diseases and cancers (63-67).

Biosimilars can be used as safely and effectively in all their approved indications as other biological medicines. So, a biosimilar is a biological medicine highly similar to another biological medicine already approved (called 'reference medicine').

Because biosimilars are made in living organisms there may be some minor differences from the reference medicine. These minor differences are not clinically meaningful, i.e. no differences are expected in safety and efficacy. Natural variability is inherent to all biological medicines and strict controls are always in place to ensure that it does not affect the way the medicine works or its safety.

The aim of biosimilar development is to demonstrate biosimilarity, i.e. high similarity in terms of structure, biological activity and efficacy, safety and immunogenicity profile.

By demonstrating biosimilarity, a biosimilar can rely on the safety and efficacy experience gained with the reference medicine. This avoids unnecessary repetition of clinical trials already carried out with the reference medicine.

Food and Drug Administration approved Zarxio, the first biosimilar product approved in the U.S., which is biosimilar to Neupogen (filgrastim), originally licensed in 1991 (68).

Biosimilars are not generics

Unlike generics, biosimilars are not identical to the reference biological product. Because biological products are made using living cells and processes, they may have minor differences from the reference product. For approval, biosimilars must demonstrate that these differences are not clinically meaningful. Nonetheless, certain factors should be considered when a switch from the reference product to a biosimilar is contemplated.

Unlike traditional small-molecule medications, which have standard production methods and well-defined structures, biological products have a sophisticated manufacturing process that involves the use of cell cultures. This process can result in heterogeneous products with slight variations in manufacturing (69-73).

Also, biological products have transformed therapy in several fields; however, their prohibitive prices, caused by the costs of research and development and manufacturing, are a concern (74-76).

Following a period of market exclusivity for reference products, to approve biological products, they must show the same primary amino acid sequence and mechanism of action as the reference product and there are no clinically meaningful differences between the reference product and the biosimilar (77). This is in contrast to generic medications, which are identical to brand medications (74). Therefore, although there are some similarities between generic and biosimilar medications, biosimilars are not considered generic versions of biological products (77-84). The main differences are:

- Generic medicines are chemically synthesized while biosimilars are grown in complex living systems.
- Biologic medicines are large, complex molecules or mixtures of molecules that may be composed of living material as such, biosimilars are unlikely to be exact copies of their reference products.
- Unlike generic medicines, the FDA requires a biosimilar to be highly similar, but not identical to the existing biologic medicine or “reference product”.
- A biosimilar also must demonstrate no clinically meaningful differences in efficacy, safety, and potency (safety and effectiveness) with its reference product.
- Per FDA guidance, agencies review the totality of evidence and do not necessarily focus on one type of study to evaluate a manufacturer's application for demonstration of biosimilarity.
- The manufacturer of a biosimilar demonstrates biosimilarity primarily from nonclinical analyses in a stepwise approach that includes examining the structure and functional nature of the biosimilar molecule.

What data are required for approval of a biosimilar or interchangeable product?

A biosimilar product application must include data demonstrating biosimilarity to the reference product (63,85-87). This usually includes data from:

- Analytical studies demonstrating that the biological product is highly similar to the reference product, notwithstanding minor differences in clinically inactive components;
- Animal studies, including an assessment of toxicity; and
- A clinical study or studies sufficient to demonstrate safety, purity, and potency of the proposed biosimilar product in one or more of the indications for which the reference product is licensed. This typically includes assessing immunogenicity, pharmacokinetics (PK), and, in some cases, pharmacodynamics (PD) and may also include a comparative clinical study.

In addition to the data listed above, an application for an interchangeable product (88) must also include information or data demonstrating that:

- The proposed interchangeable product is expected to produce the same clinical result as the reference product in any given patient; and,
- For a product administered more than once to an individual, switching between the proposed interchangeable product and the reference product does not increase safety risks or decrease effectiveness compared to using the reference product without such switching between products.

When considering licensure of a biosimilar product, the agencies review the totality of the data and information, including the foundation of detailed analytical (structural and functional) characterization, animal studies if necessary, then moving on to clinical pharmacology studies and, as needed, other comparative clinical studies.

FDA evaluates each biosimilar product on a case-specific basis to determine what data are needed to demonstrate biosimilarity and which data elements can be waived if deemed scientifically appropriate.

Proposal

A biosimilar product is a biological product that is approved based on a showing that it is highly similar to an already approved biological product, which is known as a reference product, and that there are no clinically meaningful differences between the biologic product and the reference product in terms of safety, purity, and potency of

the product (measure of drug activity expressed in terms of the amount required to produce an effect of given intensity).

Because no biosimilar can be scientifically or technically identical to the originator's product additional efforts are needed to compile and review all relevant published information on biosimilars (89). This would provide a visualization of the essential steps that are required to be taken for global biosimilar acceptance.

The higher complexity of biologics and the limited experience with biosimilars raise doubts about whether patients taking the reference product can be switched to the approved biosimilar. Biosimilarity relies on no expected differences in safety and efficacy so, for the approval of biosimilars, it is in most cases necessary to conduct large phase III clinical trials in patients to convince the regulatory authorities that the product is comparable in terms of efficacy and safety to the originator product.

Mielke J. *et al.* (90) proposed estimating the effect of switching from a reference medicine to a biosimilar and vice versa by means of testing a null hypothesis (switching influences the efficacy) versus an alternative hypothesis (switching has no influence on the efficacy) using semi-replicate or replicate crossover designs (with reasonable sample sizes) using linear mixed effects ANOVA models. We propose also considering using adaptive designs with interim analysis/es to assess efficacy and safety signals which would allow making decisions at earlier stages to stop the trial for futility, or alternatively, to increase the sample size, modify inclusion criteria or primary endpoints.

9. REFERENCES

1. U.S. Department of Health and Human Services, Food and Drug Administration. Frequently Asked Questions on Patents and Exclusivity. https://www.fda.gov/drugs/development-approval-process-drugs/frequently-asked-questions-patents-and-exclusivity. Content current Feb 2020. Accessed November 18, 2020.
2. Raines KW. A Primer on Generic Drugs and A Primer on Generic Drugs and bioequivalence: an overview of the bioequivalence: an overview of the generic drug approval process generic drug approval process. https://www.fda.gov/media/89135/download. Accessed November 18, 2020.
3. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for industry: Bioequivalence studies with pharmacokinetic endpoints for drugs submitted under an ANDA. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/bioequivalence-studies-pharmacokinetic-endpoints-drugs-submitted-under-abbreviated-new-drug. Published Dec 2013. Accessed November 18, 2020.
4. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for Industry: ANDAs Pharmaceutical Solid Polymorphism Chemistry, Manufacturing, and Controls Information. https://www.fda.gov/media/71375/download. Published Jul 2017. Accessed November 18, 2020.
5. Chow, S., Wang, H. On Sample Size Calculation in bioequivalence Trials. *J Pharmacokinet Pharmacodyn* 2001;28, 155-169.
6. Chow SC, Liu JP. Design and Analysis of Bioavailability and bioequivalence Studies, Third Edition. Boca Raton: Chapman & Hall/CRC Press; 2009.
7. European Medicines Agency. Guideline on the investigation of bioequivalence. CPMP/EWP/QWP/1401/98 Rev. 1.

https://www.ema.europa.eu/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf. Published Jan 2010. Accessed November 18, 2020.

8. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for Industry. Bioavailability Studies Submitted in NDAs or INDs - General Considerations. https://www.fda.gov/media/121311/download. Published Feb 2019. Accessed November 18, 2020.
9. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J of Pharmacokinet and Biopharm* 1987;15:657-680.
10. Schütz H. Two-stage designs in bioequivalence trials. *Eur J Clin Pharmacol* 2015;71(3):271-281.
11. European Medicines Agency. Questions & Answers: Positions on specific questions addressed to the pharmacokinetics working party. EMA/618604/2008 Rev. 13. https://www.ema.europa.eu/en/documents/scientific-guideline/questions-answers-positions-specific-questions-addressed-pharmacokinetics-working-party_en.pdf. Published Nov 2015. Accessed November 18, 2020.
12. Davit BM, Conner DP, Fabian-Fritsch B, et al. Highly variable drugs: Observations from bioequivalence data submitted to the FDA for new generic drug applications. *AAPS J* 2008;10:148-156
13. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for Industry. Statistical Approaches to Establishing Bioequivalence. https://www.fda.gov/media/70958/download. Published Jan 2001. Accessed November 18, 2020.
14. U.S. Department of Health and Human Services, Food and Drug Administration. Draft guidance on progesterone.

https://www.accessdata.fda.gov/drugsatfda_docs/psg/Progesterone_insertvag_22057_RC09-12.pdf. Published Sep 2012. Accessed November 18, 2020.

15. Tothfalusi L, Endrenyi L, Garcia Arieta A. Evaluation of bioequivalence for highly variable drugs with scaled average bioequivalence. *Clin Pharmacokinet* 2009;48(11):725-743.
16. Chow SC, Liu JP. On assessment of bioequivalence under a higher-order crossover design. Journal of Biopharmaceutical Statistics. *J Biopharm Stat* 1992;2(2):239-256.
17. Chen KW, Chow SC, Li G. A Note on sample Size Determination for Bioequivalence Studies with Higher-Order Crossover Designs. *J Pharmacokinet Biopharm* 1997 Dec;25(6):753-765.
18. Ocaña J, Muñoz J. Controlling type I error in the reference-scaled bioequivalence evaluation of highly variable drugs. *Pharmaceutical Statistics* 2019;18:583–599.
19. Howe WG. Approximate confidence limits on the mean of X + Y where X and Y are two tabled independent random variables. *J Am Stat Assoc* 1974;69(347):789-794.
20. Hyslop T, Hsuan F, Holder DJ. A small sample confidence interval approach to assess individual bioequivalence. *Stat Med* 2000; 19(20):2885-2897.
21. Molins E, Cobo E, Ocaña J. Two-stage designs versus European scaled average designs in bioequivalence studies for highly variable drugs: Which to choose? *Stat Med* 2017;36(30):4777-4788.
22. Molins E, Labes D, Schütz H, Cobo E, Ocaña J. An iterative method to protect the type I error rate in bioequivalence studies under two-stage adaptive 2×2 crossover designs. *Biom J* 2020;1-12.
23. Kaza M, Sokolovskyi A, Rudzki PJ. 10th Anniversary of a Two-Stage Design in bioequivalence. Why Has it Still Not Been Implemented? *Pharm Res* 2020;37(7):140.
24. Mistry P, Dunn JA, Marshall A. A literature review of applied adaptive design methodology within the field of oncology in randomised controlled trials and a

proposed extension to the CONSORT guidelines. *BMC Med Res Methodol* 2017;17:108.

25. Bandyopadhyay N, Dragalin V. Implementation of an adaptive group sequential design in a bioequivalence study. *Pharm Stat* 2007;6 (2):115-122.
26. Coffey CS, Levin B, Clark C, et al. Overview, hurdles, and future work in adaptive designs: Perspectives from a National Institutes of Health-funded workshop. *Clinical Trials* 2012;9, 671-680.
27. Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ, Smith RA. Sequential design approaches for bioequivalence studies with crossover designs. *Pharm Stat* 2008;7(4):245-262.
28. Muñoz J, Alcaide D, Ocaña J. Consumer's risk in the EMA and FDA regulatory approaches for bioequivalence in highly variable drugs. *Stat Med* 2016;35(12):1933-1943.
29. Labes D, Schütz H. Inflation of type I error in the evaluation of scaled average bioequivalence, and a method for its control. *Pharm Res* 2016;33(11):1-10.
30. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977;64(2):191-199.
31. Montague TH, Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ. Additional results for "sequential design approaches for bioequivalence studies with crossover designs". *Pharm Stat* 2012;11(1):8-13.
32. Karalis V, Macheras P. On the statistical model of the two-stage designs in bioequivalence assessment. *J Pharm Pharmacol* 2014;66(1):48-52.
33. 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. *AAPS J* 2013;15(4):974-990.
34. Kieser M, Rauch G. Two-stage designs for cross-over bioequivalence trials. *Stat Med* 2015;34(16):2403-2416.

35. Haidar SH, Makhlouf F, Schuirmann DJ, et al. Evaluation of a scaling approach for the bioequivalence of highly variable drugs. *AAPS J* 2008;10(3):450-454.

36. Davit BM, Chen ML, Conner DP, et al. Implementation of a reference-scaled average bioequivalence approach for highly variable generic drug products by the us food and drug administration. *AAPS J* 2012;14(4):915-921.

37. Davit BM, Patel DT. Bioequivalence of Highly Variable Drugs. In: Yu LX, Li BV, eds. FDA Bioequivalence Standards, AAPS Advances in the Pharmaceutical Sciences, vol. 13. New York: aapspress - Springer; 2014:139-164.

38. Karalis V, Macheras P. An insight into the properties of a two-stage design in bioequivalence studies. *Pharm Res* 2013;30(7):1824-1835.

39. Xu J, Audet C, DiLiberti CE, et al. Optimal adaptive sequential designs for crossover bioequivalence studies. *Pharm Stat* 2016;15(1):15-27.

40. Labes D, Lang B and Schütz H. Power2Stage: Power and Sample Size Distribution of 2-Stage bioequivalence Studies. R package version 0.5.2. https://CRAN.R-project.org/package=Power2Stage. Published Apr 2019. Accessed November 18, 2020.

41. 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.

42. Fuglsang A. Futility rules in bioequivalence trials with sequential designs. *AAPS J* 2014;16(1):79-82.

43. Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential clinical trials. *Biometrics* 1999;55(3):853-857.

44. Karalis V. The role of the upper sample size limit in two-stage bioequivalence designs. Int J Pharm 2013;456(1):87-94.

45. Golkowski D, Friede T, Kieser M. Blinded sample size re-estimation in crossover bioequivalence trials. *Pharm Stat* 2014;13(3):157-162.

46. European Generic Medicines Association. Revised EMA bioequivalence Guideline: Questions and Answers. Summary of the discussions held at the 3rd symposium

on bioequivalence. https://www.medicinesforeurope.com/wp-content/uploads/2016/03/EGA_BEQ_QA_WEB_QA_1_32.pdf. Published Jun 2010. Accessed November 18, 2020.

47. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Draft Guidance for Industry. Adaptive Design Clinical Trials of Drugs and Biologics. https://www.fda.gov/media/78495/download. Published Nov 2019. Accessed November 18, 2020.

48. Health Canada. Guidance Document: Conduct and Analysis of Comparative Bioavailability Studies. https://www.canada.ca/content/dam/hc-sc/documents/services/drugs-health-products/drug-products/applications-submissions/guidance-documents/bioavailability-bioequivalence/conduct-analysis-comparative.pdf. Published Sep 2018. Accessed November 18, 2020.

49. Knahl SIE, Lang B, Fleischer F, Kieser M. A comparison of group sequential and fixed sample size designs for bioequivalence trials with highly variable drugs. *Eur J Clin Pharmacol* 2018;74(5):549-559.

50. Thorlund K, Haggstrom J, Park JJH, Mills EJ. Key design considerations for adaptive clinical trials: a primer for clinicians. *BMJ* 2018;360:k698.

51. Fuglsang A. Controlling type I errors for two-stage bioequivalence study designs. *Clin Res Regul Aff* 2011;28(4):100-105.

52. Maurer W, Jones B, Chen Y. Controlling the type 1 error rate in two-stage sequential designs when testing for average bioequivalence. *Stat Med* 2018;37(10):1587-1607.

53. European Medicines Agency. Overview of comments received on draft guideline on the investigation of bioequivalence. EMA/CHMP/EWP/26817/2010. https://www.ema.europa.eu/en/documents/other/overview-comments-received-draft-guideline-investigation-bioequivalence-cpmp/ewp/qwp/1401/98-rev-1_en.pdf. Published Jan 2010. Accessed November 18, 2020.

54. O'Brien PC, Fleming TR. A Multiple Testing Procedure for Clinical Trials. *Biometrics* 1979;35(3):549-556.

55. Tothfalusi L, Endrenyi L, Midha KK, Rawson MJ, Hubbard JW. Evaluation of the bioequivalence of highly-variable drugs and drug products. *Pharmaceutical Research* 2001;18:728-733.

56. Endrenyi L, Tothfalusi L. Regulatory Conditions for the Determination of bioequivalence of Highly Variable Drugs. *J Pharm Pharm Sci* 2009;12(1):138-149.

57. Tothfalusi L, Endrenyi L. Sample Sizes for Designing bioequivalence Studies for Highly Variable Drugs. *J Pharm Pharm Sci* 2011;15(1):73-84.

58. Karalis V, Symillides M, Macheras P. Bioequivalence of highly variable drugs: A comparison of the newly proposed regulatory approaches by FDA and EMA. *Pharm Res* 2012;29(4):1066-1077.

59. Kang Q, Vahl CI. Testing for bioequivalence of highly variable drugs from TR-RT crossover designs with heterogeneous residual variances. *Pharm Stat* 2017;16(5):361-377.

60. Fuglsang A. Sequential bioequivalence Trial Designs with Increased Power and Controlled Type I Error Rates. *AAPS J* 2013; 15(3):659-661.

61. Nakai K, Fujita M, Ogata H. New bioequivalence studies: Individual bioequivalence and Population bioequivalence. *Yakugaku Zasshi* 2000;120(11):1201-1208.

62. Steinijans, VW. Some Conceptual Issues *in the Evaluation of Average, Population, and Individual bioequivalence. Ther Innov Regul Sci* 2001;35:893–899.

63. European Medicines Agency. Biosimilars in the EU. Information guide for healthcare professionals. https://www.ema.europa.eu/en/documents/leaflet/biosimilars-eu-information-guide-healthcare-professionals_en.pdf. Published Feb 2019. Accessed November 18, 2020.

64. Hung A, Vu Q, Mostovoy L. A systematic review of U.S. biosimilar approvals: what evidence does the FDA require and how are manufacturers responding? *J Manag Care Spec Pharm* 2017;23(12):1234-1244.

65. U.S. Department of Health and Human Services, Food and Drug Administration. Biosimilars Action Plan: Balancing Innovation and Competition. https://www.fda.gov/media/114574/download. Published Jul 2018. Accessed November 18, 2020.

66. Rugo HS, Linton KM, Cervi P, Rosenberg JA, Jacobs I. A clinician's guide to biosimilars in oncology. *Cancer Treat Rev* 2016;46:73-79.

67. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Guidance for Industry: Scientific considerations in demonstrating biosimilarity to a reference product. https://www.fda.gov/media/82647/download. Published Apr 2015. Accessed November 18, 2020.

68. Colwell J. FDA approves first biosimilar, Zarxio. *Cancer Discov* 2015;5(5):460.

69. Pagani E. Why are biosimilars much more complex than generics? *Einstein (São Paulo)* 2019;17(1):eED4836.

70. Stevenson JG, Popovian R, Jacobs I, Hurst S, Shane LG. Biosimilars: practical considerations for pharmacists. *Ann Pharmacother* 2017;51(7):590-602.

71. Lucio SD, Stevenson JG, Hoffman JM. Biosimilars: implications for health-system pharmacists. *Am J Health Syst Pharm* 2013;70(22):2004-2017.

72. Griffith N, McBride A, Stevenson JG, Green L. Formulary selection criteria for biosimilars: considerations for US health-system pharmacists. *Hosp Pharm* 2014;49(9):813-825.

73. Li E, Ramanan S, Green L. Pharmacist substitution of biological products: issues and considerations. *J Manag Care Spec Pharm* 2015;21(7):532-539.

74. Lyman GH, Balaban E, Diaz M, et al. American Society of Clinical Oncology statement: biosimilars in oncology. *J Clin Oncol* 2018;36(12):1260-1265.

75. Campen CJ. Integrating biosimilars into oncology practice: implications for the advanced practitioner. *J Adv Pract Oncol* 2017;8(7):688-699.

76. Boyle RM. The use of biologics in cancer therapy. *US Pharm* 2010;35(3)(Oncology suppl):4-7.

77. U.S. Department of Health and Human Services, Food and Drug Administration. Biosimilar and interchangeable products. https://www.fda.gov/drugs/biosimilars/biosimilar-and-interchangeable-products. Content current Oct 2017. Accessed November 18, 2020.

78. U.S. Department of Health and Human Services, Food and Drug Administration. Generic drugs: Questions & Answers. https://www.fda.gov/drugs/questions-answers/generic-drugs-questions-answers. Content current Jun 2018. Accessed November 18, 2020.

79. Dolinar R, Lavernia F, Edelman S. A guide to follow-on biologics and biosimilars with a focus on insulin. *Endocr Pract* 2018;24(2):195-204.

80. U.S. Department of Health and Human Services, Food and Drug Administration. Biosimilar development, review, and approval. https://www.fda.gov/drugs/biosimilars/biosimilar-development-review-and-approval. Content current Oct 2017. Accessed November 18, 2020.

81. U.S Department of Health and Human Services, Food and Drug Administration, Office of Medical Products and Tobacco, Center for Drug Evaluation and Research, Office of Generic Drugs, Office of Generic Drug Policy. Approved drug products with therapeutic equivalence evaluations, 39th ed. www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/UCM071436.pdf. Published 2020. Accessed November 18, 2020.

82. Vivian JC. Generic-substitution laws. *US Pharm* 2008;33(6):30-34.

83. Pope ND. Generic substitution of narrow therapeutic index drugs. *US Pharm* 2019;34(6):12-19.

84. U.S. Department of Health and Human Services, Food and Drug Administration. Purple Book: Lists of licensed biological products with reference product exclusivity and biosimilarity or interchangeability evaluations. https://www.fda.gov/drugs/therapeutic-biologics-applications-bla/purple-book-

lists-licensed-biological-products-reference-product-exclusivity-and-biosimilarity-or. Content current Aug 2020. Accessed November 18, 2020.

85. U.S. Department of Health and Human Services, Food and Drug Administration. What are generic drugs? https://www.fda.gov/drugs/generic-drugs/what-are-generic-drugs. Content current Aug 2017. Accessed November 18, 2020.

86. U.S. Department of Health and Human Services, Food and Drug Administration. Frequently asked questions about therapeutic biological products. https://www.fda.gov/drugs/therapeutic-biologics-applications-bla/frequently-asked-questions-about-therapeutic-biological-products. Content current Jul 2015. Accessed November 18, 2020.

87. McCamish M, Woollett G. The state of the art in the development of biosimilars. *Clin Pharmacol Ther* 2012;91(3):405-417.

88. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Guidance for Industry: Labeling for Biosimilar Products Guidance for Industry. https://www.fda.gov/media/96894/download. Published Jul 2018. Accessed November 18, 2020.

89. Kabir ER, Moreino SS, Sharif Siam MK. The Breakthrough of Biosimilars: A Twist in the Narrative of Biological Therapy. *Biomolecules* 2019;9(9):410.

90. Mielke J, Woehling H, Jones B. Longitudinal assessment of the impact of multiple switches between a biosimilar and its reference product on efficacy parameters. *Pharm Stat* 2018;17(3):231-247.

Appendix 1: Package 'betsd'

The package 'betsd' was created by means of *R* packages 'devtools' and 'roxygen2'. The package 'devtools' helps on working on the skeleton of the description file, and 'roxygen2' generates the manual which includes the arguments of the function.

This package is available in the repository *Gib, GitHub, https://github.com/eduard-molins/betsd*. The integration between *R Studio* and *GitHub* allows:

- Version control (a unique site to work with the code)

- Backup (old versions of the code are available)

- Share our project/code

The package 'betsd' can be downloaded from *GitHub* following these instructions:

> install.packages("devtools") # if not yet installed

> require(devtools)

> install_github("eduard-molins/betsd")

> require(betsd)

> help(t1e.tsd)

This package was validated during the Reproducible Research review (see Appendix 3).

Package 'betsd'

July 5, 2019

Encoding UTF-8

Version 0.1.5

Date 2019-07-03

Title Adjusting significance levels in two-stage adaptive 2x2 crossover designs

Author Eduard Molins [aut, cre], Detlew Labes [ctb], Helmut Schütz [ctb], Jordi Ocaña [ctb]

Maintainer Eduard Molins <molins.eduard@gmail.com>

Depends R (>= 3.5.0)

Imports stats, Power2Stage

Description Iteratively adjusts significance levels at each stage in order to control the overall type I error.

License GPL (>=3)

LazyData true

URL https://github.com/eduard-molins/betsd

BugReports https://github.com/eduard-molins/betsd/issues

R topics documented:

`t1e.tsd` *Significance level adjustment for bioequivalence studies using adaptive two-stage 2x2 crossover designs*

Description

This function calculates, by iterative search, the adjusted significance levels to be used in each stage of adaptive two-stage 2×2 crossover designs, to ensure an overall type I error below a specified significance level.

Usage

```
t1e.tsd(n1, CV, GMR = 0.95, Nmax = 150, min.n2 = n1/2, type = 1,
        alpha = 0.05, alpha1, alpha2, targetpower = 0.8,
        setseed = TRUE, theta1, theta2,
        details = TRUE, print = TRUE, ...)
```

Arguments

`n1`	Sample size of stage 1.
`CV`	Within subject coeficient of variation (use *e.g.*, 0.3 for 30%).
`GMR`	Expected geometric mean ratio to be used in decision scheme (power calculations in stage 1 and sample size estimation for stage 2). By default 0.95.
`Nmax`	Overall maximum number of subjects (sum of sample sizes in both stages, *i.e.*, N = n1 + n2). By default 150, see the details section.
`min.n2`	Minimum number of subjects at second stage. By default n1/2. Set min.n2 = 0 to cancel any limitation on the sample size at the second stage, N2. See the details section for more information.
`type`	Type 1 or 2 methodology. See the details section. By default 1.
`alpha`	Target overall significance level (both stages). By default 0.05.
`alpha1`	Initial significance level at stage 1. By default 0.0294.
`alpha2`	Initial significance level at stage 2. By default 0.0294.
`targetpower`	Power threshold in the power monitoring steps and power to achieve in the sample size estimation step. By default 0.8.
`setseed`	Simulations are dependent on the starting point of the (pseudo) random number generator. To avoid differences in power for different runs a setseed(1234567) is issued if set to TRUE, the default. Set this argument to FALSE for a random seed.
`theta1`	Lower limit of the bioequivalence range. By default 0.80.
`theta2`	Upper limit of the bioequivalence range. By default 1.25.
`details`	If set to TRUE (default) shows intermediate results in the console. Set this argument to FALSE to supress intermediate results.
`print`	If set to TRUE (default) shows final results in the console. Set this argument to FALSE to return a list of final results.
`...`	Optional additional arguments. See package Power2Stage, function power.tsd.

Details

The type 1 method uses the same adjusted alpha1 and alpha2 in stages 1 and 2, respectively.
The type 2 method uses an unadjusted alpha if interim power is at least the targetpower, or adjusted alpha1 and alpha2 in stages 1 and 2 otherwise. The terminology follows Schütz (2015).

By default, Nmax = 150 and min.n2 = n1/2. Its present implementation is based on controlling the total sample size: After computing the required sample size in the second step, say N2 (to ensure the required power targetpower), the number of additional observations is computed as N2 = max(min.n2, N2), but if n1 + n2 > Nmax, the study is terminated. These criteria and default values are based on Molins *et al.* (2017).

n1 and CV can be scalars or vectors. As vectors, the result might loose some power for a particular combination.
Global significance level is ensured by returning the adjusted alpha1 and alpha2 of the worst case scenario for all combinations of n1 and CV. To ensure that the overall type I error is below alpha, this code prevents from a misspecifications of CV (which is unknown at this stage) considering CV ±0.05.

By default alpha = 0.05 and alpha1 = alpha2 = 0.0294. However, alpha1 and alpha2 can be different. In this case, set alpha1 and alpha2, and alpha1 will be fixed whilst alpha2 adjusted.

This function uses the package 'Power2Stage' developed by Labes *et al.* and is based on the Potvin *et al.* (2008) methods.

Value

Adjusted significance levels, alpha1 and alpha2.

References

Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ, Smith RA. *Sequential design approaches for bioequivalence studies with crossover designs.*
Pharm Stat. 2008; 7(4):245–62. doi: 10.1002/pst.294

Montague TH, Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ. *Additional results for 'Sequential design approaches for bioequivalence studies with crossover designs'.*
Pharm Stat. 2011; 11(1):8–13. doi: 10.1002/pst.483

Fuglsang A. *Controlling type I errors for two-stage bioequivalence study designs.*
Clin Res Reg Aff. 2011; 28(4):100–5. doi: 10.3109/10601333.2011.631547

Schütz H. *Two-stage designs in bioequivalence trials.*
Eur J Clin Pharmacol. 2015; 71(3):271–81. doi: 10.1007/s0022801518062

Molins E, Cobo E, Ocaña J. *Two-stage designs versus European scaled average designs in bioequivalence studies for highly variable drugs: Which to choose?*
Stat Med. 2017; 36(30):4777–88. doi: 10.1002/sim.7452

Examples

```
## Not run:
t1e.tsd(n1 = 24, CV = c(0.3, 0.4, 0.5, 0.6), GMR = 0.95, type = 1)
# should give adjusted alpha 0.0307 at both stages and
# a maximum empirical type I error 0.04998
# Note: runtime ~10 minutes on a Xeon E3-1245 Quadcore 3.4 GHz
t1e.tsd(n1 = 12, CV = 0.2, GMR = 0.9, alpha1 = 0.0280, alpha2 = 0.0280, type = 2)
# should give adjusted alpha 0.0268 at both stages and
# a maximum empirical type I error 0.049591
# Note: runtime ~30 minutes
# List of results
x <- t1e.tsd(n1 = 12, CV = 0.10, type = 2, print = FALSE)
print(as.data.frame(x), row.names = FALSE)
## End(Not run)
```

Source: Biometrical Journal supporting information: bimj2181-sup-0001-SuppMat.docx.

Appendix 2: *R* code

Function t1e.tsd

The function 't1e.tsd' included within the package 'betsd' serves to preserve the overall type I error. The function provides with the adjusted significance levels to be used in each stage, α_1 and α_2, the probability to jump to a stage 2, the sample size at the stage 2, and the power at stages 1 and 2.

```
# Function to adjust the significance levels for bioequivalence studies using
# adaptive two-stage 2x2 crossover designs.
#
# Authors: E. Molins, D. Labes, H. Schuetz, J. Ocaña
#ADJUST ALPHA1 AND ALPHA2 ON A TSD 2x2 CXO DESIGN, CONTROLLING T1E
bioequivalenceLOW ALPHA
t1e.tsd <- function(n1, CV, GMR = 0.95, Nmax = 150, min.n2 = n1/2, type = 1,
                    alpha = 0.05, alpha1, alpha2, targetpower = 0.8,
                    setseed = TRUE, theta1, theta2,
                    details = TRUE, print = TRUE, ...) {

  #Non-modified parameters => Nmax == Inf & min.n2 == 0
  #Modified CRITERIA ACCORDING TO DOI: 10.1002/sim.7452
  min.n2 <- sapply(min.n2, function(y) if (y %% 2 != 0) y+y%%2 else y)

  #DEBUG AT FIRST ITERATION
  if (missing(n1))
    stop("Number of subjects in stage 1 must be given")
  if (any(n1 < 12))
    stop("Number of subjects in stage 1 must be at least 12")
  if (missing(CV))
    stop("CV must be given")
  if (any(CV <= 0.05)) #This is 0.05 because the algorithm will look at the
                       #given CV +/- 0.05, and CV must be > 0
    stop("CV must be > 0.05 beause the algorithm will look at CV \u00B10.05")
  if (length(alpha) > 1)
    stop("alpha must be of length 1")
  if (missing(alpha1) & missing(alpha2)) {
    alpha1 <- 0.0294
    alpha2 <- alpha1
  }
  if (missing(alpha1) & !missing(alpha2)) alpha1 <- alpha2
  if (!missing(alpha1) & missing(alpha2)) alpha2 <- alpha1
  if (alpha1 && alpha2 > 0.05)
    stop("alpha values must be <0.05 (default at 0.0294)")
  if (length(type) > 1)
    stop("type must be of length 1")
  if (type!=1 & type!=2)
    stop("type must be 1 or 2")
  if (type == 1) {type <- "B"} else {type <- "C"} #type 1: Potvin "B"; type 2:
                                                  #Potvin "C or D"
  if (missing(theta1) & missing(theta2))  theta1 <- 0.8
  if (!missing(theta1) & missing(theta2)) theta2 <- 1/theta1
  if (missing(theta1) & !missing(theta2)) theta1 <- 1/theta2
  if (GMR <= theta1 | GMR >= theta2)
    stop("GMR must be within acceptance range")

  #SETSEED
  if (is.numeric(setseed)) {
    setseed <- TRUE
  } else if (is.character(setseed)) {
```

```
    stop("setseed should be TRUE for setseed = 1234567 or FALSE for a random
          setseed.")
  }
  if (setseed) {
    seed <- 1234567 #This corresponds to set.seed(1234567) used in power.tsd
if setseed = TRUE
  } else {
    seed <- runif(1, max=1E7)
    set.seed(seed)
  }

  max_iter <- FALSE
  iter <- -1
  current_alpha <- matrix(c(alpha1, alpha2), nrow = 1, ncol = 2)
  # ------------------------------------------------------------------------
  # burn in: make an evaluation with 30 000 sims and choose the points
  # with T1E > 90% percentile
  # dataframe with the entire grid
  d <- cbind(expand.grid(n1 = n1,
                         CV = CV,
                         GMR = GMR,
                         alpha1 = alpha1,
                         alpha2 = alpha2),
                         min.n2 = min.n2)
  res_T1E <- potvin(type = type,
                    d = d,
                    Nmax,
                    targetpower,
                    setseed,
                    nsims = 30000,
                    pmethod = "nct",
                    theta0 = theta2,
                    theta1,
                    theta2, ...)
  T1E  <- res_T1E["pbioequivalence",] #This gives the T1E
  m.n2 <- res_T1E["min.n2",] #This gives min.n2
  T1E  <- t(matrix(as.numeric(T1E),
                   nrow = length(CV),
                   ncol = length(n1),
                   byrow = TRUE))
  m.n2 <- t(matrix(as.numeric(m.n2),
            nrow = length(CV),
            ncol = length(n1),
            byrow = TRUE))
  rownames(T1E) <- n1
  colnames(T1E) <- CV
  # now entries with T1E >= 90% percentile
  index <- which(T1E >= quantile(T1E, 0.90), arr.ind = TRUE)
  n1.quant <- as.numeric(rownames(T1E)[index[, 1]])
  m.n2 <- m.n2[index]
  CV.quant <- as.numeric(colnames(T1E)[index[, 2]])
  d90 <- data.frame(n1 = n1.quant,
                    min.n2 = m.n2,
                    CV = CV.quant,
                    GMR = GMR,
                    alpha1 = alpha1, alpha2 = alpha2)

  # LOOP until current_alpha and alpha_adj are different
  # ------------------------------------------------------------------------
  #browser()
  repeat{
    iter <- iter + 1
    # evaluate the T1E grid with 1E6 sims at theta0=theta2 with current_alpha
    d90$alpha1 <- current_alpha[1,1]
    d90$alpha2 <- current_alpha[1,2]
    res_T1E90 <- potvin(type = type,
                        d = d90,
                        Nmax,
```

```
                        targetpower,
                        setsee
                        nsims = 1E+06,
                        pmethod = "nct",
                        theta0 = theta2,
                        theta1,
                        theta2, ...)
    T1E_high <- data.frame(cbind(d90[, "n1"],
                                 d90[, "CV"],
                                 d90[, "GMR"],
                                 d90[, "min.n2"],
                                 d90[, "alpha1"],
                                 d90[, "alpha2"],
                                 as.numeric(res_T1E90["pbioequivalence", ])))
    colnames(T1E_high) <- c("n1",
                            "CV",
                            "GMR",
                            "min.n2",
                            "alpha1",
                            "alpha2",
                            "pbioequivalence")
    # ------------------------------------------------------------------------
    # choose the max T1E
    max_T1E <- T1E_high[T1E_high[, "pbioequivalence"] ==
max(T1E_high["pbioequivalence"]), ]
    m <- which.max(max_T1E$CV)
    max_T1E <- max_T1E[m,]
    nperc90 <- cbind(n1 = res_T1E90["n1", ],
                     CV = res_T1E90["CV", ],
                     GMR = res_T1E90["GMR", ],
                     min.n2 = res_T1E90["min.n2", ],
                     pbioequivalence_s1 = res_T1E90["pbioequivalence_s1", ],
                     pct_s2 = res_T1E90["pct_s2", ],
                     Nperc = res_T1E90["nperc", ])
    max_T1E <- merge(max_T1E,
                     nperc90,
                     all.x = TRUE,
                     by = c("n1", "CV", "GMR", "min.n2"))
    if (details) {
      if (iter < 1) {
        cat("Run-in with alphas", paste(sprintf("%.4f",
            as.numeric(current_alpha)), collapse=", "), "\n")
      } else {
        cat("Iteration", iter, "with alphas", paste(sprintf("%.5f",
            as.numeric(floor(current_alpha*1e5)/1e5)), collapse=", "), "\n")
      }
        cat("- max.TIE", sprintf("%.6f", max_T1E$pbioequivalence), "at n1 =",
max_T1E$n1,
            "and CV =" , max_T1E$CV, "\n")
    }
    # ------------------------------------------------------------------------
    # make a grid around the old_alphas
    new_alpha1 <- seq(current_alpha[1, 1] - 0.0005,
                      current_alpha[1, 1] + 0.0005,
                      length.out = 7)
    new_alpha2 <- seq(current_alpha[1, 2] - 0.0005,
                      current_alpha[1, 2] + 0.0005,
                      length.out = 7)
    step_size <- median(diff(seq(current_alpha[1, 2] - 0.0005,
                                 current_alpha[1, 2] + 0.0005,
                                 length.out = 7)))

    if (alpha1 == alpha2) {
      n_d <- cbind(n1 = max_T1E["n1"],
                   CV = max_T1E["CV"],
                   GMR = max_T1E["GMR"],
                   min.n2 = max_T1E["min.n2"],
                   alpha1 = new_alpha1,
```

```
                  alpha2 = new_alpha2,
                  row.names = NULL)
    } else {
      n_d <- cbind(n1 = max_T1E["n1"],
                  CV = max_T1E["CV"],
                  GMR = max_T1E["GMR"],
                  min.n2 = max_T1E["min.n2"],
                  expand.grid(alpha1 = alpha1, alpha2 = new_alpha2),
                  row.names = NULL)
    }
    colnames(n_d) <- c("n1", "CV", "GMR", "min.n2", "alpha1", "alpha2")

    # evaluate and use inverse regression to obtain new alphas (alpha.adj)
    res_new_T1E <- potvin(type = type,
                          d = n_d,
                          Nmax,
                          targetpower,
                          setseed,
                          nsims = 1E+06,
                          pmethod = "nct",
                          theta0 = theta2,
                          theta1,
                          theta2, ...) #Takes much time if alpha1 <> alpha2
    res_new_d_T1E <- data.frame(CV = unlist(res_new_T1E["CV", ]),
                                n1 = unlist(res_new_T1E["n1", ]),
                                GMR = unlist(res_new_T1E["GMR", ]),
                                min.n2 = unlist(res_new_T1E["min.n2", ]),
                                alpha1 = sapply(1:ncol(res_new_T1E),
                                function(x) res_new_T1E[["alpha", x]][1]),
                                alpha2 = sapply(1:ncol(res_new_T1E),
                                function(x) res_new_T1E[["alpha", x]][2]),
                                T1E =
unlist(res_new_T1E["pbioequivalence", ]))

    #INVERSE REGRESSION
    alpha_adj <- inv.reg(alpha, alpha1, alpha2, res_new_d_T1E)
    colnames(alpha_adj) <- c("alpha1", "alpha2")
    #browser()

    # Emergency brake
    if(iter > 10) {
      max_iter <- TRUE
      warning("Max. iterations reached.")
      current_alpha <- floor(current_alpha*1e4)/1e4
      repeat{
        max_T1E$alpha1 <- current_alpha[1,1]
        max_T1E$alpha2 <- current_alpha[1,2]
        check_T1E <- potvin(type = type,
                            d = max_T1E,
                            Nmax,
                            targetpower,
                            setseed,
                            nsims = 1E+06,
                            pmethod = "nct",
                            theta0 = theta2,
                            theta1,
                            theta2, ...) #Takes much time if alpha1 <> alpha
        #Punctual T1E for CV +/- 0.05 is evaluated following the article of Xu
        #et al.
        lo.power.tsd <- power.tsd(alpha=c(max_T1E$alpha1, max_T1E$alpha2),
                                  CV = CV[1] - 0.05,
                                  n1 = max_T1E$n1,
                                  GMR = GMR,
                                  min.n2 = max_T1E$min.n2,
                                  Nmax = Nmax,
                                  targetpower = targetpower,
                                  setseed = setseed,
                                  pmethod = "nct",
```

```
                                theta0 = 1.25,
                                theta1 = theta1,
                                theta2 = theta2,
                                method = type)
        hi.power.tsd <- power.tsd(alpha=c(max_T1E$alpha1, max_T1E$alpha2),
                                CV = tail(CV, n=1) + 0.05,
                                n1 = max_T1E$n1,
                                GMR = GMR,
                                min.n2 = max_T1E$min.n2,
                                Nmax = Nmax,
                                targetpower = targetpower,
                                setseed = setseed,
                                pmethod = "nct",
                                theta0 = 1.25,
                                theta1 = theta1,
                                theta2 = theta2,
                                method = type)
        if (unlist(check_T1E["pbioequivalence", ]) < alpha &&
lo.power.tsd$pbioequivalence < alpha &&
            hi.power.tsd$pbioequivalence < alpha) {
            max_T1E$pbioequivalence <- check_T1E["pbioequivalence",]
            break
        } else if (alpha_adj[1] == alpha_adj[2]) {
          current_alpha <- current_alpha - 1E-4
        } else {
          current_alpha[1] <- floor(alpha1*1e4)/1e4 #In case of alpha 1
          #different than alpha 2, alph1 is fixed
          current_alpha[2] <- current_alpha[2] - 1E-4
        }
      } #end repeat
      break
    } #end if

    #check that the significance level is below alpha, and the new alphas
    #against the old_one   current_alpha[2] because this is the only one we
    #adjust. If alpha1=alpha2, then the alpha2 adjusted is copied in alpha1.
    #If alpha1 and alpha are different, the only that we adjust is alpha2
    diff <- abs(current_alpha[2] - alpha_adj[2])
    #SCENARIO 1/4: T1E < 0.05 and Diff <= 2E-4
    if (max_T1E["pbioequivalence"] < alpha & diff <= 2E-4) {
      current_alpha <- floor(current_alpha*1e4)/1e4
      #The following repeat serves to find a solution when the algorithm
      #enters into a cycle
      repeat{
        max_T1E$alpha1 <- current_alpha[1,1]
        max_T1E$alpha2 <- current_alpha[1,2]
        check_T1E <- potvin(type = type,
                            d = max_T1E,
                            Nmax,
                            targetpower,
                            setseed,
                            nsims = 1E+06,
                            pmethod = "nct",
                            theta0 = theta2,
                            theta1,
                            theta2, ...) #Takes much time if alpha1 <> alpha2
        #Punctual T1E for CV +/- 0.05 is evaluated following the article of Xu
        #et al.
        lo.power.tsd <- power.tsd(alpha=c(max_T1E$alpha1,
                                max_T1E$alpha2),
                                CV = CV[1] - 0.05,
                                n1 = max_T1E$n1,
                                GMR = GMR,
                                min.n2 = max_T1E$min.n2,
                                Nmax = Nmax,
                                targetpower = targetpower,
                                setseed = setseed,
                                pmethod = "nct",
```

```
                                theta0 = 1.25,
                                theta1 = theta1,
                                theta2 = theta2,
                                method = type)
        hi.power.tsd <- power.tsd(alpha=c(max_T1E$alpha1,
                                max_T1E$alpha2),
                                CV = tail(CV, n=1) + 0.05,
                                n1 = max_T1E$n1,
                                GMR = GMR,
                                min.n2 = max_T1E$min.n2,
                                Nmax = Nmax,
                                targetpower = targetpower,
                                setseed = setseed,
                                pmethod = "nct",
                                theta0 = 1.25,
                                theta1 = theta1,
                                theta2 = theta2,
                                method = type)
        if (unlist(check_T1E["pbioequivalence", ]) < alpha &&
lo.power.tsd$pbioequivalence < alpha &&
            hi.power.tsd$pbioequivalence < alpha) {
          max_T1E$pbioequivalence <- check_T1E["pbioequivalence",]
          break
        } else if (alpha_adj[1] == alpha_adj[2]) {
          current_alpha <- current_alpha - 1E-4
        } else {
          current_alpha[1] <- floor(alpha1*1e4)/1e4 #In case of alpha 1
          #different than alpha 2, alph1 is fixed
          current_alpha[2] <- current_alpha[2] - 1E-4
        }
      } #end repeat
    break
    #SCENARIO 2/4: T1E > 0.05 and Diff <= 2E-4
    } else if (max_T1E["pbioequivalence"] > alpha & diff <= 2E-4) { #Scenario
2/4: T1E >
      #0.05 and Diff <= 1E-4
      if (alpha_adj[2] > current_alpha[2]) {
        if (alpha_adj[1] == alpha_adj[2]) {
          current_alpha <- current_alpha - step_size*(iter+1)
        } else {
          current_alpha[1] <- alpha1 #In case of alpha 1 different than alpha
          #2, alph1 is fixed
          current_alpha[2] <- current_alpha[2] - step_size*(iter+1)
        }
      } else {
        if (alpha_adj[1] == alpha_adj[2]) {
          current_alpha <- alpha_adj
        } else {
          current_alpha[1] <- alpha1 #In case of alpha 1 different than alpha
          #2, alph1 is fixed
          current_alpha[2] <- alpha_adj[2]
        }
      }
    #SCENARIO 3/4: Diff > 2E-04 and T1E < 0.05
    } else if (max_T1E["pbioequivalence"] < alpha & diff > 2E-4) {
      if (max(current_alpha[2], alpha_adj[2]) > alpha) { # Scenario 3 (first
      #option): T1E < 0.05 but either current_alpha or alpha_adj is > alpha
        if (alpha_adj[1] == alpha_adj[2]) {
          current_alpha <- matrix(c(min(current_alpha - step_size*(iter+1),
                                  alpha_adj),
                                  min(current_alpha - step_size*(iter+1),
                                  alpha_adj)),
                                  nrow = 1,
                                  ncol =2)
          if (current_alpha[2] < 0.025)
              current_alpha <- matrix(c(alpha/2 + step_size*(iter+1),
                                    alpha2/2+ step_size*(iter+1)),
                                    nrow = 1,
```

```
                                    ncol = 2)
      } else {
        current_alpha[1] <- alpha1
        current_alpha[2] <- min(current_alpha[2] - step_size*(iter+1),
                                alpha_adj[2])
        if (current_alpha[2] < 0.025)
            current_alpha[2] <- alpha/2 + step_size*(iter+1)
      }
    } else { # Secenario 3 (second option): T1E < 0.05 and current_alpha <
             #alpha
      if (alpha_adj[1] == alpha_adj[2]) {
          current_alpha <- matrix(c(max(current_alpha + step_size*(iter+1),
                                    alpha_adj),
                                    max(current_alpha + step_size*(iter+1),
                                    alpha_adj)),
                                    nrow = 1,
                                    ncol =2)
      } else {
          current_alpha[1] <- alpha1 #In case of alpha 1 different than
          #alpha 2, alph1 is fixed
          current_alpha[2] <- max(current_alpha[2] + step_size*(iter+1),
                                    alpha_adj[2])
      }
    }
  #SCENARIO 4/4: Diff > 2E-04 and T1E > 0.05
  } else {
    if (min(current_alpha[2], alpha_adj[2]) > alpha) { #Both are > alpha
      if (alpha_adj[1] == alpha_adj[2]) {
        current_alpha <- matrix(c(alpha1 - step_size*(iter+1),
                                  alpha2 - step_size*(iter+1)),
                                  nrow = 1,
                                  ncol =2)
      } else {
        current_alpha[1] <- alpha1 #In case of alpha 1 different than alpha
          #2, alph1 is fixed
        current_alpha[2] <- alpha2 - step_size*(iter+1)
      }
    } else {
      if (alpha_adj[1] == alpha_adj[2]) {
        current_alpha <- matrix(c(min(current_alpha - step_size*(iter+1),
                                  alpha_adj),
                                  min(current_alpha - step_size*(iter+1),
                                  alpha_adj)),
                                  nrow = 1,
                                  ncol =2)
        if (current_alpha[2] < 0.025)
            current_alpha <- matrix(c(alpha/2 + step_size*(iter+1),
                                      alpha2/2 + step_size*(iter+1)),
                                      nrow = 1, ncol = 2)
      } else {
        current_alpha[1] <- alpha1 #In case of alpha 1 different than alpha
        #2, alph1 is fixed
        current_alpha[2] <- min(current_alpha[2] - step_size*(iter+1),
                                  alpha_adj[2])
        if (current_alpha[2] < 0.025)
            current_alpha[2] <- alpha/2 + step_size*(iter+1)
      }
    }
  }
} # end repeat loop current_alpha different from alpha_adj
# ------------------------------------------------------------------------
-
# output
mt <- if (type == "B") "1 (B)" else "2 (C/D)"
#browser()
p <- potvin(type = type,
            d = max_T1E,
            Nmax,
```

```
              targetpower,
              setseed,
              nsims = 1E+05,
              pmethod = "nct",
              theta0 = max_T1E[["GMR"]],
              theta1 = theta1,
              theta2 = theta2, ...)
  if (print) {
    cat("\nMethod type =", mt, "\n",
        "setseed =", seed, "\n",
        "Bioequivalence acceptance range (theta1, theta2) =",
        paste(c(round(unlist(res_T1E["theta1", ][1]), 4),
              round(unlist(res_T1E["theta2", ][1]), 4)), collapse=" ... "),
"\n",
        "Power calculation method =", unlist(p["pmethod",][1]), "\n",
        "n1 =", n1, "\n",
        "GMR =", GMR, "\n",
        "CV =", paste(CV, collapse=", "), "\n",
        "Nmax =", Nmax, "\n",
        "min.n2 =", min.n2, "\n",
        "Adjusted alpha at stage 1 =",
        sprintf("%.4f", (max_T1E$alpha1*1e4)/1e4), "and alpha at stage 2 =",
        sprintf("%.4f", (max_T1E$alpha2*1e4)/1e4), "\n",
        "Maximum empirical type I error =", sprintf("%.6f",
max_T1E$pbioequivalence[[1]]),
        "at n1 =", max_T1E$n1, "and CV =" , max_T1E$CV, "\n",
        "Power: Overall probability of bioequivalence =",
unlist(p["pbioequivalence", ]),"\n",
        "Power: Probability of bioequivalence at stage 1 =",
unlist(p["pbioequivalence_s1", ]),"\n",
        "Studies in stage 2 =", paste0(unlist(p["pct_s2", ]),"%\n"),
        names(p["nperc",][[1]]), "percentiles of N =",
        paste(unlist(p["nperc", ]), collapse=", "),"\n",
        "max.iter =", max_iter, "\n\n"
    )
  }
  if (print == FALSE) {
    pct <- as.numeric(unlist(p["nperc", ], use.names=FALSE))
    res <- list(type = mt, alpha = alpha, CV = CV, n1 = n1, GMR = GMR,
        min.n2 = min.n2, Nmax = Nmax, targetpower = targetpower,
        alpha1 = max_T1E$alpha1, alpha2 = max_T1E$alpha2,
        theta1 = round(unlist(res_T1E["theta1", ][1]), 4),
        theta2 = round(unlist(res_T1E["theta2", ][1]), 4),
        pmethod = unlist(p["pmethod", ]),
        TIE = max_T1E$pbioequivalence[1],
        loc.CV = max_T1E$CV,loc.n1 = max_T1E$n1,
        pbioequivalence = as.numeric(unlist(p["pbioequivalence", ])),
        pbioequivalence_s1 = as.numeric(unlist(p["pbioequivalence_s1", ])),
        pct_s2 = as.numeric(unlist(p["pct_s2", ])),
        nperc5 = pct[1], median = pct[2], nperc95 = pct[3],
        max.iter = max_iter)
    return(res)
  }
} # end of t1e.tsd()
```

Function Potvin

Potvin function invokes the function 'power.tsd' (from repository CRAN) (40). This function assesses the probability of declaring bioequivalence. This probability can sometimes return a 'power' or a 'type I error'. If theta0 <= 0.8 o theta0 >= 1.25, the function returns the probability of a type I error. If 0.8 < theta0 < 1.25 each bioequivalence claim is true and the result corresponds to the power. So, 'power.tsd' allows assessing the power curve, i.e. the likelihood of rejecting H0 (i.e. bioinequivalence) as a function of theta0. This probability is estimated through simulations (by means of generating statistics, not subjects) as the proportion of bioequivalence claims in 'nsims' simulations

```
# Power.tsd function
# d is a dataframe with alpha1, alpha2, n1, GMR, CV, min.n2
potvin <- function(type, d, Nmax, targetpower, setseed, nsims, pmethod
= "nct", theta0=theta0, theta1=theta1, theta2=theta2, npct = c(0.05,
0.5, 0.95)) { # uses FUNCTION power.tsd
  # BY LBioequivalenceS D., LANG B., SCHUETZ H.
  sapply(1:nrow(d), function(x) {
    return(power.tsd(method = type,
                     alpha = c(d[x, "alpha1"], d[x, "alpha2"]),
                     n1  = d[x, "n1"],
                     GMR = d[x, "GMR"],
                     CV  = d[x, "CV"],
                     targetpower = targetpower,
                     pmethod = pmethod,
                     Nmax = Nmax,
                     min.n2 = d[x, "min.n2"],
                     # theta0: True unknown GMR (t_effect);
                     # theta0=theta2 for T1E; theta0=GMR for power
                     theta0 = theta0,
                     theta1 = theta1,
                     theta2 = theta2,
                     npct = npct,
                     nsims = nsims,
                     setseed = setseed))
  }
  )
}
```

Function inv.reg

Type I errors are simulated/obtained for each proposed significance levels (α_1, α_2) by means of the function 'power.tsd' (from repository CRAN) (40). Then, we fit linear and quadratic regression models where the independent variable is the simulated type I error which is adjusted by (α_1, α_2) pairs, and we choose the 'best' model based on the minimum Akaike Information Criterion (*AIC*). To obtain the adjusted alpha 'adj_alpha', we isolate (α_1, α_2) for *T1E* = 5%. For example, if *e.g.* if $\alpha_1 = \alpha_2$ and the min(*AIC*) is obtained from the linear regression model, then 'alpha.adj' is obtained as:

$$\alpha_{adj1} = \alpha_{adj2} = \frac{(0.05 - \widehat{\beta}_0)}{\widehat{\beta}_1}$$

```
#LOOKING FOR A ALPHA AT EACH STAGE WHOSE T1E < alpha (BASED ON
MIN(AIC)) USING INVERSE REGRESSION
inv.reg <- function(alpha, alpha1, alpha2, res_new_d_T1E) {
  mod1 <- lm(T1E ~ alpha2, data = res_new_d_T1E)   # linear
  mod2 <- lm(T1E ~ alpha2 + I(alpha2^2), data = res_new_d_T1E)
#quadratic
  if (extractAIC(mod1, k=2)[2] <= extractAIC(mod2, k=2)[2]) { # select
  #the better model
    #cat("B0 =", coef(mod1)[[1]], "B1 =", coef(mod1)[[2]], "\n")
    alpha.adj <- (alpha-coef(mod1)[[1]])/coef(mod1)[[2]]
  } else {
    #cat("B0 =", coef(mod2)[[1]], "B1 =", coef(mod2)[[2]],
         "B2 =", coef(mod2)[[3]], "\n")
    det <- (coef(mod2)[[2]]/2/coef(mod2)[[3]])^2-(coef(mod2)[[1]]-
           alpha)/coef(mod2)[[3]]
    #cat("det =", det, "sqrt.det =", sqrt(det), "\n")
    if (det > 0) {
      if (coef(mod2)[[3]] < 0) {
        alpha.adj <- -(coef(mod2)[[2]]/2/coef(mod2)[[3]]+sqrt(det))
      } else {
        alpha.adj <- -(coef(mod2)[[2]]/2/coef(mod2)[[3]]-sqrt(det))
      }
    } else {
      alpha.adj <- (alpha-coef(mod1)[[1]])/coef(mod1)[[2]]
    }
  } #end else
  #if (alpha.adj > alpha) alpha.adj <- alpha/2
  if (alpha.adj < 0) alpha.adj <- alpha/2
  if (alpha1 == alpha2) {
    return(matrix(c(alpha.adj, alpha.adj), nrow = 1, ncol = 2))
  } else {
    return(matrix(c(alpha1, alpha.adj), nrow = 1, ncol = 2))
  }
}
```

Examples

```
#Examples:
t1e.tsd(n1 = 24, CV = c(0.3, 0.4, 0.5, 0.6), GMR = 0.95, type = 1)
t1e.tsd(n1 = 24, CV = c(0.3, 0.4, 0.5, 0.6), GMR = 0.95, type = 1,
setseed=FALSE)
t1e.tsd(n1 = 24, CV = c(0.3, 0.4, 0.5, 0.6), GMR = 0.95, type = 1,
setseed=FALSE)
t1e.tsd(n1 = 24, CV = 0.6, GMR = 0.95, Nmax = 4000, type = 1,
setseed=FALSE)
t1e.tsd(n1 = 36, CV = 0.4, GMR = 0.95, Nmax = 4000, type = 1,
setseed=FALSE)
t1e.tsd(n1 = 12, CV = 0.8, GMR = 0.95, Nmax = 4000, targetpower = 0.8,
type = 1, alpha = 0.05, setseed=FALSE)
```

Appendix 3: Reproducible Research (RR)

We conducted a validation process to ensure that the code and results shown in the article published in the Biometrical Journal are absolutely reproducible. So, this article earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in the article could fully be reproduced.

We followed Wiley’s Guidelines for Code and Data submissions Specific on Reproducible research (RR):

https://onlinelibrary.wiley.com/pb-assets/assets/15214036/RR_Guideline-1509621643000.pdf

We submitted a README.txt document with the following content:

Source code for manuscript "An iterative method to protect the type I error rate in bioequivalence studies under two-stage adaptive 2x2 crossover designs"
by Eduard Molins, Detlew Labes, Helmut Schutz, Erik Cobo, Jordi Ocaña.

The code has been written by by Eduard Molins, Detlew Labes, Helmut Schutz, Jordi Ocaña. Please, contact molins.eduard@gmail.com for any comment.

R version 3.6.0 (2019-04-26)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS 10.15.4

Matrix products: default
BLAS:
/System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/libBLAS.dylib
LAPACK:
/Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

Random number generation:
RNG: Mersenne-Twister
Normal: Inversion
Sample: Rounding

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats graphics grDevices utils datasets methods base

other attached packages:
[1] reshape2_1.4.3 reshape_0.8.8 Power2Stage_0.5.2

loaded via a namespace (and not attached):
[1] cubature_2.0.3 compiler_3.6.0 magrittr_1.5 plyr_1.8.4
[5] tools_3.6.0 PowerTOST_1.4-7 Rcpp_1.0.1 TeachingDemos_2.10
[9] mvtnorm_1.0-10 stringi_1.4.3 stringr_1.4.0

In addition, functions for Monte Carlo simulations (t1e.tsd.R, potvin.R, inv.reg.R) are provided. These functions are placed in code\functions\...

The working directory should be set to '~/code'. To reproduce the results presented in the manuscript, just run the main analysis file 'Analyses_tsd_simulation.R'. All tables and Figure 3 (in this thesis Figure 10) are stored in the \results subfolder.

Information on the data set can be found in data\README_tsd_t1e_bimj.docx.

In addition, a *data* subfolder was submitted with a 'README_tsd_t1e_bimj.docx' file containing the following information:

This function calculates the 'empiric' type 1 error and power of 2-stage bioequivalence studies according to a modified Potvin et al. with max(N = N_1+ N_2) = 150, and $N_2 \geq N_1/2$, via simulations. But instead of simulating subject data, the statistics point estimate at stage 1 (pe1), mean square error at stage 1 (MSE1) or intra-subject residual variance calculated from CV, and point estimate at stage 2, and sum of square at stage 2 (SS2) are simulated via their associated distributions (normal and $\chi 2$ distributions).

Reproducible Research (RR) - Analyses_tsd_simulation.R

This is the *R* code approved for Reproducible Research (RR) in the Biometrical Journal.

Note that below there are the Table and Figure numbers as appearing in the journal.

But, Table and Figure numbers have been accommodated in this thesis as follow:

Correspondence between Table and Figure numbers published in the Biometrical Journal and those used in this thesis

Article number	Thesis number
Table 1	Table 6
Table 2	Table 7
Table 3	Table 8
Table4	Table 9
Table 5	Table 10
Table 6	Table 11
Figure 1	Figure 8
Figure 2	Figure 9
Figure 3	Figure 10

```
#########################################################################
#
#   Filename    :         Analyses_tsd_simulation.R
#   Project     :         BiomJ article "An iterative method to
protect the type I error rate in bioequivalence studies under two-
stage adaptive 2×2 crossover designs"
#   Authors     :         E. Molins, D. Labes, H. Schütz, J. Ocaña
#   Date        :         2020-08-24
#   Purpose     :         As described in BiomJ article
#   R Version   :         R version 3.6.0 (2019-04-26)#
#   Input data files  :   Simulation-based (see:
data\README_tsd_t1e_Biomj.txt)
#   Output data files :   Table1, Table2, Table3, Table4, Table5,
Table6, and Figure3 (see: results\)#
#   Required R packages : Power2Stage, t1e.tsd, inv.reg, potvin
#
#########################################################################

rm(list = ls())
#########################################################################
##
## Code to obtain the resutls of Tables 1, 2, 3, 4, 5, 6 and Figure 3
##
#########################################################################
## Working directory
## The current working directory is the folder code/

## source function definitions:
library(Power2Stage)  # Function power.tsd() is loaded

source("functions/t1e.tsd.R")   # Iterative method to calculate
significance levels of two-stage designs (Figure 2)
source("functions/inv.reg.R")   # Inverse regression for selection
linear or quadric regression models
```

```
source("functions/potvin.R")    # Function power.tsd() calculate type
I error or power

require(reshape)  # For functions melt() and cast()
require(reshape2) # For functions melt() and cast()

###################
###   Table 1   ###
###################
# Adjusted α1 and α2 in both stages preserving the overall T1E below
5%

type <- c(1,2)
n1 <- c(12, 24)
CV <- c(1,2,3,4)
CV_range <- rbind(seq(0.1,0.19,0.01), seq(0.2,0.29,0.01),
seq(0.3,0.39,0.01), seq(0.4,0.49,0.01))
d <- cbind(expand.grid(CV = CV, n1 = n1, type = type), CV_range)

pBioequivalence <- sapply(1:nrow(d), function(x) {
        return(t1e.tsd(n1 = d[x, "n1"], CV = as.numeric(d[x,4:13]),
                       GMR = 0.95, targetpower = 0.8, type = d[x,
"type"], print=FALSE))
})

res <- pBioequivalence[c("n1","CV","alpha1",
"alpha2","TIE","pbioequivalence_s1","pct_s2","pbioequivalence","nperc5
","median","nperc95"),]
#save(res, file = "../results/Table1.RDa")

load("../results/Table1.RDa")

res_table1 <- data.frame(t(res))
res_table11 <- cbind(CV_id = rep(1:4,4), res_table1)
res_table12 <- res_table11[,c(2,1,4:length(res_table11))]
sapply(res_table12, mode)

res_table13 <- data.frame(matrix(unlist(res_table12),
ncol=length(res_table12), byrow=F))
colnames(res_table13) <- c("n1", "CV_id", "alpha1", "alpha2", "TIE",
"pbioequivalence_s1", "pct_s2", "pbioequivalence", "nperc5", "median",
"nperc95")
res_table13[c("pbioequivalence_s1", "pbioequivalence")] <-
round(res_table13[c("pbioequivalence_s1", "pbioequivalence")]*100, 2)
res_table13[c("pct_s2")] <- round(res_table13[c("pct_s2")], 2)

res_table13$CV_id <- sapply(res_table13$CV_id, function(y) if (y == 1)
                            { "0.10-0.19"} else if (y == 2)
                            { "0.20-0.29"} else if (y == 3)
                            { "0.30-0.39"} else { "0.40-0.49"})

res_table13

###################
###   Table 2   ###
###################
# Type 1 method to adjust α2 for a fixed α1 preserving the overall T1E
below 5%

n1 <- c(12, 24)
CV <- seq(0.2,0.29,0.01)
```

```
alpha <- rbind(c(0.0294, 0.0300), c(0.0320, 0.0294))
type <- 1
d <- merge(n1, alpha)
colnames(d) <- c("n1", "alpha1", "alpha2")

pBioequivalence <- sapply(1:nrow(d), function(x) {
  return(t1e.tsd(n1 = d[x, "n1"],
                 CV = CV,
                 alpha1 = d[x, "alpha1"],
                 alpha2 = d[x, "alpha2"],
                 GMR = 0.95,
                 targetpower = 0.8,
                 type = 1,
                 print=FALSE))
})

res <- pBioequivalence[c("n1",
            "CV",
            "alpha1",
            "alpha2",
            "TIE",
            "pbioequivalence_s1",
            "pct_s2",
            "pbioequivalence",
            "nperc5",
            "median",
            "nperc95"),]
#save(res, file = "../results/Table2.RDa")

load("../results/Table2.RDa")

res_table2 <- data.frame(t(res))

res_table21 <- cbind(CV_id = rep(1,4), res_table2)
res_table22 <- res_table21[,c(2,1,4:length(res_table21))]
sapply(res_table22, mode)

res_table23 <- data.frame(matrix(unlist(res_table22),
ncol=length(res_table22), byrow=F))
colnames(res_table23) <- c("n1", "CV_id", "alpha1", "alpha2", "TIE",
                           "pbioequivalence_s1", "pct_s2",
"pbioequivalence", "nperc5",
                           "median", "nperc95")
res_table23[c("pbioequivalence_s1", "pbioequivalence")] <-
round(res_table23[c("pbioequivalence_s1",
                                        "pbioequivalence")]*100, 2)
res_table23[c("pct_s2")] <- round(res_table23[c("pct_s2")], 2)

res_table23$CV_id <- sapply(res_table23$CV_id,
                            function(y) if (y == 1) "0.20-0.29")

res_table23

####################
###   Table 3   ###
####################
# Empiric type 1 error and power for CV_w at 0.05 below and above LB
and UB

CV <- rep(c(0.10 - 0.05, 0.19 + 0.05, 0.20 - 0.05, 0.29 + 0.05, 0.30 -
            0.05,0.39 + 0.05, 0.40 - 0.05, 0.49 + 0.05),4)
```

```
alpha1 <- c(rep(0.0299,2), rep(0.0307,2), rep(0.0303,2),
            rep(0.0377,2), rep(0.0381,2), rep(0.0306,2),
            rep(0.0302,2), rep(0.0306,2), rep(0.0280,2),
            rep(0.0280,2), rep(0.0295,2), rep(0.0377,2),
            rep(0.0314,2), rep(0.0301,2), rep(0.0303,2),
            rep(0.0306,2))
alpha2 <- alpha1
method <- c(rep("B",16),rep("C",16))
n1 <- rep(c(rep(12,8), rep(24,8)),2)
Nmax <- rep(150, length(CV))
min.n2 = n1/2
theta0_1.25 <- rep(1.25, length(CV))
theta0_0.95 <- rep(0.95, length(CV))

d_1.25 <- data.frame(method, n1 = n1, alpha1 = alpha1, alpha2 =
                     alpha2, GMR = 0.95, CV = CV, Nmax = Nmax,
                     min.n2 = min.n2, theta0 = theta0_1.25)
d_0.95 <- data.frame(method, n1 = n1, alpha1 = alpha1,
                     alpha2 = alpha2, GMR = 0.95, CV = CV,
                     Nmax = Nmax, min.n2 = min.n2,
                     theta0 = theta0_0.95)

pBioequivalence_1.25 <- sapply(1:nrow(d_1.25),
                    function(x) {
                    return(power.tsd(method = as.character(d_1.25[x,
                    "method"]), alpha = c(d_1.25[x, "alpha1"],
                    d_1.25[x, "alpha2"]),
                    n1  = d_1.25[x, "n1"], GMR = d_1.25[x, "GMR"],
                    CV  = d_1.25[x, "CV"],
                   targetpower = 0.8, pmethod = "nct", Nmax =
d_1.25[x, "Nmax"], min.n2 = d_1.25[x, "min.n2"],
                   theta0 = d_1.25[x, "theta0"], theta1 = 0.8,
                   theta2 = 1.25, npct = c(0.05, 0.5, 0.95),
                   nsims = 10^6, setseed = 1234567))
})
Bioequivalence_1.25 <- pBioequivalence_1.25
Bioequivalence_1.25 <- unlist(Bioequivalence_1.25["pbioequivalence",])
t1e.table3 <- cbind(d_1.25,Bioequivalence_1.25)

pBioequivalence_0.95 <- sapply(1:nrow(d_0.95), function(x) {
  return(power.tsd(method = as.character(d_0.95[x, "method"]),
                   alpha = c(d_0.95[x, "alpha1"],
                   d_0.95[x, "alpha2"]),
                   n1  = d_0.95[x, "n1"],
                   GMR = d_0.95[x, "GMR"],
                   CV  = d_0.95[x, "CV"],
                   targetpower = 0.8,
                   pmethod = "nct",
                   Nmax = d_0.95[x, "Nmax"],
                   min.n2 = d_0.95[x, "min.n2"],
                   theta0 = d_0.95[x, "theta0"],
                   theta1 = 0.8, theta2 = 1.25,
                   npct = c(0.05, 0.5, 0.95),
                   nsims = 10^5, setseed = 1234567))
})

Bioequivalence_0.95 <- pBioequivalence_0.95
Bioequivalence_0.95 <- unlist(Bioequivalence_0.95["pbioequivalence",])
pow.table3 <- cbind(d_0.95,Bioequivalence_0.95)
```

```
t3 <- cbind(t1e.table3, bioequivalence_0.95 =
pow.table3$Bioequivalence_0.95)
#save(t3, file = "../results/Table3.RDa")

load("../results/Table3.RDa")

res_table3 <- data.frame(t3)
res_table31 <- subset(res_table3, select = c(n1, alpha1, alpha2,
                      method))
res_table31$method <- sapply(res_table31$method,
                      function(y) if (y == "B") "Type 1"
                      else "Type 2")
res_table31$id <- sort(rep(1:16,2))
res_table31 <- unique(res_table31)

# Ordering Type I error and Power
res_table32 <- cbind(id = sort(rep(1:16,2)), CV_id = rep(1:2,16),
subset(res_table3, select = c(CV, bioequivalence_1.25,
bioequivalence_0.95)))
res_table33 <- melt(res_table32, id = c("id","CV_id", "CV"))
res_table34 <- as.data.frame(cast(res_table33, id ~ variable + CV_id))
colnames(res_table34) <- c("id", "t1e_CV_LB-
0.05","t1e_CV_UB+0.05","power_CV_LB-0.05","power_CV_UB+0.05")
res_table34[c("t1e_CV_LB-0.05","t1e_CV_UB+0.05")] <-
round(res_table34[c("t1e_CV_LB-0.05","t1e_CV_UB+0.05")], 4)
res_table34[c("power_CV_LB-0.05","power_CV_UB+0.05")] <-
round(res_table34[c("power_CV_LB-0.05", "power_CV_UB+0.05")]*100, 2)

# Ordering CV
res_table35 <- cbind(id = sort(rep(1:16,2)), CV_id = rep(1:2,16),
res_table3["CV"])
res_table36 <- melt(res_table35, id = c("id","CV_id"))
res_table37 <- as.data.frame(cast(res_table36, id ~ variable + CV_id))
c <- 0.05 # According to CV at 0.05 above and below the upper and
lower CV bounds (see the article)
res_table37[c("CV_1","CV_2")] <- c(res_table37["CV_1"] + c,
res_table37["CV_2"] - c)
colnames(res_table37) <- c("id","CV_LB","CV_UB")

# Merging data.frames
res_table38 <- merge(res_table37, res_table34, by = "id")
res_table39 <- merge(res_table31, res_table38, by = "id")
res_table39 <- res_table39[, c(5,2,6,7,3,4,8,9,10,11)]

res_table39

###################
###   Table 4   ###
###################
# Xu et al. Optimal TSD designs of methods E and F and our methodology
(type 1 and 2 methods)

t1.n1.18 <- t1e.tsd(n1 = 18, CV = c(0.1, 0.15, 0.20, 0.25, 0.30),
                    GMR = 0.95, targetpower = 0.8, type = 1,
                    alpha = 0.05, print=FALSE)
t1.n1.48 <- t1e.tsd(n1 = 48,
                    CV = c(0.30, 0.35, 0.40, 0.45, 0.50, 0.55),
                    GMR = 0.95, targetpower = 0.8, type = 1,
                    alpha = 0.05, print=FALSE)
t2.n1.18 <- t1e.tsd(n1 = 18, CV = c(0.1, 0.15, 0.20, 0.25, 0.30),
                    GMR = 0.95, targetpower = 0.8, type = 2,
```

```
                    alpha = 0.05, print=FALSE)
t2.n1.48 <- t1e.tsd(n1 = 48,
                    CV = c(0.30, 0.35, 0.40, 0.45, 0.50, 0.55),
                    GMR = 0.95, targetpower = 0.8, type = 2,
                    alpha = 0.05, print=FALSE)
t4 <- cbind(t1.n1.18, t1.n1.48, t2.n1.18, t2.n1.48)
#save(t4, file = "../results/Table4.RDa")

load("../results/Table4.RDa")

# Taken from Table I - Article: Xu, J., Audet, C., DiLiberti, C. E.,
Hauck, W. W., Montague, T. H., Parr, A. F., Potvin, D., and
Schuirmann, D. J.  (2016).
# Optimal adaptive sequential designs for crossover bioequivalence
studies.
# Pharmaceutical Statistics 15, 15-27.
Xu_E_18_alpha1 <- 0.0249
Xu_E_18_alpha2 <- 0.0363
Xu_E_48_alpha1 <- 0.0254
Xu_E_48_alpha2 <- 0.0357
Xu_F_18_alpha1 <- 0.0248
Xu_F_18_alpha2 <- 0.0364
Xu_F_48_alpha1 <- 0.0259
Xu_F_48_alpha2 <- 0.0349
Xu_E_18_f <- "93.74 - 106.67"
Xu_E_48_f <- "93.05 - 107.47"
Xu_F_18_f <- "94.92 - 105.35"
Xu_F_48_f <- "93.50 - 106.95"

Xu <- rbind(c(Xu_E_18_alpha1, Xu_E_48_alpha1),
            c(Xu_E_18_alpha2, Xu_E_48_alpha2),
            c(Xu_E_18_f, Xu_E_48_f),
            c(Xu_F_18_alpha1, Xu_F_48_alpha1),
            c(Xu_F_18_alpha2, Xu_F_48_alpha2),
            c(Xu_F_18_f, Xu_F_48_f))

# Molins et. all: Obtained using our methodology
type1_18_alpha1 <- t4["alpha1",][1]
type1_18_alpha2 <- t4["alpha2",][1]
type1_48_alpha1 <- t4["alpha1",][2]
type1_48_alpha2 <- t4["alpha2",][2]
type2_18_alpha1 <- t4["alpha1",][3]
type2_18_alpha2 <- t4["alpha2",][3]
type2_48_alpha1 <- t4["alpha1",][4]
type2_48_alpha2 <- t4["alpha2",][4]

Molins <- rbind(c(type1_18_alpha1, type1_48_alpha1),
                c(type2_18_alpha1, type2_48_alpha1))

# This is to check the CV values used in Molins et al.: From 0.1 to
0.3 and from 0.3 to 0.55, for n1 = 18 and n2 = 48, respectively
t4_CV <- as.data.frame(unlist(t4["CV", ]))

# Merging Xu et al. and Molins et al. results
t4_Xu_Molins <- as.data.frame(rbind(Xu, Molins))
colnames(t4_Xu_Molins) <- c("CVw 0.1-0.3 & N1=18", "CVw 0.3-0.55 &
N1=48")
rownames(t4_Xu_Molins) <- c("Xu_E_alpha1",
                            "Xu_E_alpha2",
                            "Xu_E_f",
                            "Xu_F_alpha1",
```

```
                              "Xu_F_alpha2",
                              "Xu_F_f",
                              "Molins_Type1",
                              "Molins_Type2")

t4_Xu_Molins

###################
###   Table 5   ###
###################
# Percentiles of N (5th, 50th, 95th) and % of Studies in Stage 2

#Percentiles N and % of Studies in Stage 2
CV <- c(seq(0.10,0.30,0.05), seq(0.30,0.55,0.05))
method <- c(rep("B",length(CV)),rep("C",length(CV)))
n1 <- c(rep(18,5),rep(48,6))
min.n2 <- sapply(n1/2, function(y) if (y %% 2 != 0) y+y%%2 else y)
alpha <- c(rep(0.0303,5), rep(0.0305,6), rep(0.0331,5), rep(0.0331,6))
d <- data.frame(method = method, CV = CV, GMR = 0.95, n1 = n1,
                alpha1 = alpha, alpha2 = alpha,
                Nmax = 150, min.n2 = min.n2, theta0 = 0.95)

pBioequivalence <- sapply(1:nrow(d),
               function(x) {
               return(power.tsd(method = as.character(d[x, "method"]),
                      alpha = c(d[x, "alpha1"], d[x, "alpha2"]),
                      n1 = d[x, "n1"], GMR = d[x, "GMR"],
                      CV  = d[x, "CV"],
                      targetpower = 0.8, pmethod = "nct",
                      Nmax = d[x, "Nmax"], min.n2 = d[x, "min.n2"],
                      theta0 = d[x, "theta0"], theta1 = 0.8,
                      theta2 = 1.25, npct = c(0.05, 0.5, 0.95),
                      nsims = 10^6, setseed = 1234567))
})
Bioequivalence_0.95 <- pBioequivalence
Bioequivalence_nperc <-
data.frame(matrix(unlist(Bioequivalence_0.95["nperc",]),
nrow=length(Bioequivalence_0.95["nperc",]), byrow=T))
Bioequivalence_st2 <- unlist(Bioequivalence_0.95["pct_s2",])
pow.table5 <- data.frame(d,Bioequivalence_nperc, bioequivalence_st2)
#save(pow.table5, file = "../results/Table5.RDa")

load("../results/Table5.RDa")

# Taken from Table II - Article: Xu, J., Audet, C., DiLiberti, C. E.,
#Hauck, W. W., Montague, T. H., Parr, A. F., Potvin, D., and
#Schuirmann, D. J.  (2016).
# Optimal adaptive sequential designs for crossover bioequivalence
#studies.
# Pharmaceutical Statistics 15, 15-27.
Xu  <- as.data.frame(rbind(cbind(n1 = 18, CV = 0.10,
                     MethodE = '(18,18,18) 0%',
                     MethodF = '(18,18,18) 0%'),
                     cbind(n1 = 18, CV = 0.15,
                     MethodE = '(18,18,18) 2.4%',
                     MethodF = '(18,18,18) 1.3%'),
                     cbind(n1 = 18, CV = 0.20,
                     MethodE = '(18,18,32) 24.1%',
                     MethodF = '(18,18,32) 21.8%'),
                     cbind(n1 = 18, CV = 0.25,
                     MethodE = '(18,24,42) 54.2%',
```

```
                      MethodF = '(18,24,42) 53.7%'),
                      cbind(n1 = 18, CV = 0.30,
                      MethodE = '(18,42,42) 75.8%',
                      MethodF = '(18,42,42) 76.9%'),
                      cbind(n1 = 48, CV = 0.30,
                      MethodE = '(48,48,52) 7.6%',
                      MethodF = '(48,48,48) 3.6%'),
                      cbind(n1 = 48, CV = 0.35,
                      MethodE = '(48,48,74) 28.2%',
                      MethodF = '(48,48,74) 22.8%'),
                      cbind(n1 = 48, CV = 0.40,
                      MethodE = '(48,48,98) 46.2%',
                      MethodF = '(48,48,98) 44.0%'),
                      cbind(n1 = 48, CV = 0.45,
                      MethodE = '(48,80,124) 61.3%',
                      MethodF = '(48,80,124) 60.5%'),
                      cbind(n1 = 48, CV = 0.50,
                      MethodE = '(48,104,150) 74.3%',
                       MethodF = '(48,104,152) 73.6%'),
                       cbind(n1 = 48, CV = 0.55,
                       MethodE = '(48,128,176) 85.2%',
                       MethodF = '(48,128,180) 84.3%')))

# Molins et. all: Obtained using our methodology
pow.table5$method <- as.integer(sapply(pow.table5$method, function(y)
if (y == "B") 1 else 2))
names(pow.table5)[1]<-paste("id")

pow.table51 <- cbind(CV_id = rep(1:11,2), pow.table5)
pow.table52 <- subset(pow.table51, select = c(n1, id, CV_id, CV))
pow.table53 <- subset(pow.table51, select = c(id, CV_id, X1, X2, X3,
bioequivalence_st2))

pow.table54 <- melt(pow.table53, id = c("id", "CV_id"))
pow.table54 <- pow.table54[order(pow.table54$id, pow.table54$CV_id),]
pow.table55 <- cast(pow.table54, id + CV_id ~ variable)

pow.table56 <- merge(pow.table52, pow.table55, by = c("id", "CV_id"))
pow.table56 <- pow.table56[order(pow.table56$id, pow.table56$CV_id),]
pow.table57 <-pow.table56[,c(1,3:8)]
row.names(pow.table57) <- NULL
colnames(pow.table57) <- c("method_type", "n1", "CV",
                           "P_N_0.05", "P_N_0.5", "P_N_0.95",
                           "per_ST2")
Molins <- pow.table57

Xu
Molins

###################
###   Table 6   ###
###################
# Power and mean sample size with constraint N ≤ 4000 for HVD

t1e.tsd(n1 = c(12,24,36), CV = seq(0.4,0.8,0.01), GMR = 0.95, Nmax =
4000, targetpower = 0.8, type = 1, alpha = 0.05)

res <- power.tsd(method = "B", alpha0 = 0.05,
                 alpha = c(0.0302, 0.0302),
                 n1 = 36, GMR = 0.95, CV = 0.4,
                 targetpower = 0.8, pmethod = "nct",
```

```
                Nmax = 4000, min.n2 = 18, theta0 = 0.95,
                theta1 = 0.8, theta2 = 1.25,
                npct = c(0.05, 0.5, 0.95), nsims = 10^6)
res_1 <- c(n1=36, CV=0.4, unlist(res[c("pbioequivalence","nmean")]))

res <- power.tsd(method = "B", alpha0 = 0.05,
                 alpha = c(0.0302, 0.0302),
                 n1 = 24, GMR = 0.95, CV = 0.6,
                 targetpower = 0.8, pmethod = "nct",
                 Nmax = 4000, min.n2 = 12, theta0 = 0.95,
                 theta1 = 0.8, theta2 = 1.25,
                 npct = c(0.05, 0.5, 0.95), nsims = 10^6)
res_2 <- c(n1=24, CV=0.6, unlist(res[c("pbioequivalence","nmean")]))

res <- power.tsd(method = "B", alpha0 = 0.05,
                 alpha = c(0.0302, 0.0302),
                 n1 = 12, GMR = 0.95, CV = 0.8, targetpower = 0.8,
                 pmethod = "nct", Nmax = 4000, min.n2 = 6,
                 theta0 = 0.95, theta1 = 0.8, theta2 = 1.25,
                 npct = c(0.05, 0.5, 0.95), nsims = 10^6)
res_3 <- c(n1=12, CV=0.8, unlist(res[c("pbioequivalence","nmean")]))

t6 <- rbind(res_1,res_2,res_3)
#save(t6, file = "../results/Table6.RDa")

load("../results/Table6.RDa")

# Taken from Article: Maurer, W., Jones, B., and Chen, Y. (2018).
#Controlling the type 1 error rate in two-stage sequential designs
#when testing for average bioequivalence.
# Statistics in Medicine 37, 1587-1607.
# See Table 8 - Potvin et al. Method B, and Maurer, Jones, and Chen
Maximum Combination Test (MCT) (w, w*): (0.5, 0.25)
method <- c("Potvin", "Potvin", "Potvin")
n1 <- c(36, 24, 12)
CV <- c(0.4, 0.6, 0.8)
power <- c(82, 77, 72)
nmean <- c(67, 161, 257)
Potvin <- data.frame(method, n1, CV, power, nmean,
stringsAsFactors=FALSE)
rm(method, n1, CV, power, nmean)

method <- c("Maurer", "Maurer", "Maurer")
n1 <- c(36, 24, 12)
CV <- c(0.4, 0.6, 0.8)
power <- c(81, 80, 76)
nmean <- c(67, 180, 325)
Maurer <- data.frame(method, n1, CV, power, nmean,
stringsAsFactors=FALSE)
rm(method, n1, CV, power, nmean)

# Molins et al.: : Obtained using our methodology
Molins <- as.data.frame(t6)
Molins$pbioequivalence <- Molins$pbioequivalence*100
names(Molins)[3] <- "power"
Molins[c("power","nmean")] <- round(Molins[c("power","nmean")], 0)
row.names(Molins) <- NULL
Molins <- cbind(method = 'Molins', Molins)

t62 <- merge(Potvin, Maurer, by = c("n1", "CV"))
t63 <- merge(t62, Molins, by = c("n1", "CV"))
```

```
t63 <- t63[order(-t63$n1),]
colnames(t63) <- c("n1", "CV",
                   "Pot.method", "Pot.power", "Pot.nmean",
                   "Mau.method", "Mau.power", "Mau.nmean",
                   "Mol.method", "Mol.power", "Mol.nmean")
row.names(t63) <- NULL

t63

###################
###   Figure 1  ###
###################
# R Studio was not used

###################
###   Figure 2  ###
###################
# R Studio was not used

###################
###   Figure 3  ###
###################
# Power assessment based on true GMR and CV_w with N1 = 12 and type 1
methodology
# The seed for simulations is the same, i.e. 1234567.

t_e <- seq(0.81,1.24,0.01)
cv_w <- seq(.1,.49,.01)
type <- "B"
n1 <- 12
min.n2 <- n1/2

d <- cbind(expand.grid(n1 = n1, t_e = t_e, cv_w = cv_w), min.n2 =
min.n2)

potvin <- function(type, alpha1, alpha2, d) {
  sapply (1:nrow(d), function(x) {
    return(power.tsd(method = type, alpha = c(alpha1, alpha2),
                     n1  = d[x, "n1"], GMR = d[x,"t_e"],
                     CV  = d[x,"cv_w"],
                     targetpower = 0.8, pmethod = "nct",
                     Nmax = 150, min.n2 = d[x, "min.n2"],
                     npct = c(0.05, 0.5, 0.95), nsims = 10^5,
                     setseed = 1234567))
  }
  )
}

#0.0299
result_0.0299 <- potvin(type = type, alpha1 = 0.0299,
                        alpha2 = 0.0299, d)
res_0.0299 <- cbind(d, alpha1 = 0.0299, alpha2 = 0.0299, power =
unlist(result_0.0299["pbioequivalence",]))
z.0.0299 <- matrix(res_0.0299$power, nrow = length(t_e),
                   ncol = length(cv_w))
save(z.0.0299, file = "../results/Figure3_0.0299.RDa")

pdf("../results/Figure3_0.0299.pdf", height = 5, width = 5)
contour(x = seq(0.81,1.24, length.out = nrow(z.0.0299)),
        y = seq(0.1,0.49, length.out = ncol(z.0.0299)),
        z.0.0299,
```

```
        xlab = "GMR",
        ylab = "CVw",
        lwd = 1)
title(main = "Figure3_0.0299")
dev.off()

#0.0307
result_0.0307 <- potvin(type = type, alpha1 = 0.0307,
                        alpha2 = 0.0307, d)
res_0.0307 <- cbind(d, alpha1 = 0.0307, alpha2 = 0.0307,
                    power = unlist(result_0.0307["pbioequivalence",]))
z.0.0307 <- matrix(res_0.0307$power, nrow = length(t_e),
                   ncol = length(cv_w))
save(z.0.0307, file = "../results/Figure3_0.0307.RDa")

pdf("../results/Figure3_0.0307.pdf", height = 5, width = 5)
contour(x = seq(0.81,1.24, length.out = nrow(z.0.0307)),
        y = seq(0.1,0.49, length.out = ncol(z.0.0307)),
        z.0.0307,
        xlab = "GMR",
        ylab = "CVw",
        lwd = 1)
title(main = "Figure3_0.0307")
dev.off()

#0.0294 and 0.0310
result_0.0294_0.0310 <- potvin(type = type, alpha1 = 0.0294,
                               alpha2 = 0.0310, d)
res_0.0294_0.0310 <- cbind(d, alpha1 = 0.0294, alpha2 = 0.0310,
                     power =
unlist(result_0.0294_0.0310["pbioequivalence",]))
z.0.0294_0.0310 <- matrix(res_0.0294_0.0310$power, nrow = length(t_e),
ncol = length(cv_w))
save(z.0.0294_0.0310, file = "../results/Figure3_0.0294_0.0310.RDa")

pdf("../results/Figure3_0.0294_0.0310.pdf", height = 5, width = 5)
contour( x = seq(0.81,1.24, length.out = nrow(z.0.0294_0.0310)),
         y = seq(0.1,0.49, length.out = ncol(z.0.0294_0.0310)),
         z.0.0294_0.0310,
         xlab = "GMR",
         ylab = "CVw",
         lwd = 1)
title(main = "Figure3_0.0294_0.0310")
dev.off()

#0.0320 and 0.0279
result_0.0320_0.0279 <- potvin(type = type, alpha1 = 0.0320,
                               alpha2 = 0.0279, d)
res_0.0320_0.0279 <- cbind(d, alpha1 = 0.0320, alpha2 = 0.0279,
                           power =
unlist(result_0.0320_0.0279["pbioequivalence",]))
z.0.0320_0.0279 <- matrix(res_0.0320_0.0279$power,
                          nrow = length(t_e),
                          ncol = length(cv_w))
save(z.0.0320_0.0279, file = "../results/Figure3_0.0320_0.0279.RDa")

pdf("../results/Figure3_0.0320_0.0279.pdf", height = 5, width = 5)
contour(x = seq(0.81,1.24, length.out = nrow(z.0.0320_0.0279)),
        y = seq(0.1,0.49, length.out = ncol(z.0.0320_0.0279)),
        z.0.0320_0.0279,
        xlab = "GMR",
```

```
        ylab = "CVw",
        lwd = 1)
title(main = "Figure3_0.0320_0.0279")
dev.off()

#Reading 4 individual plots
load("../results/Figure3_0.0299.RDa")
load("../results/Figure3_0.0320_0.0279.RDa")
load("../results/Figure3_0.0307.RDa")
load("../results/Figure3_0.0294_0.0310.RDa")

#Individual plots
# z.0.0299
system2('open', args = c('-a Preview.app',
'../results/Figure3_0.0299.pdf'), wait = FALSE)
# z.0.0320_0.0279
system2('open', args = c('-a Preview.app',
'../results/Figure3_0.0320_0.0279.pdf'), wait = FALSE)
# z.0.0307
system2('open', args = c('-a Preview.app',
'../results/Figure3_0.0307.pdf'), wait = FALSE)
# z.0.0294_0.0310
system2('open', args = c('-a Preview.app',
'../results/Figure3_0.0294_0.0310.pdf'), wait = FALSE)

#Combined plot (Figure 3)
pdf("../results/Figure3.pdf")
par(mfrow=c(2,2))
contour( x = seq(0.81,1.24, length.out = nrow(z.0.0299)),
         y = seq(0.1,0.49, length.out = ncol(z.0.0299)),
         z.0.0299,
         xlab = "GMR",
         ylab = "CVw",
         lwd = 1)
title(main = "Figure3_0.0299")
contour( x = seq(0.81,1.24, length.out = nrow(z.0.0307)),
         y = seq(0.1,0.49, length.out = ncol(z.0.0307)),
         z.0.0307,
         xlab = "GMR",
         ylab = "CVw",
         lwd = 1)
title(main = "Figure3_0.0307")
contour( x = seq(0.81,1.24, length.out = nrow(z.0.0294_0.0310)),
         y = seq(0.1,0.49, length.out = ncol(z.0.0294_0.0310)),
         z.0.0294_0.0310,
         xlab = "GMR",
         ylab = "CVw",
         lwd = 1)
title(main = "Figure3_0.0294_0.0310")
contour( x = seq(0.81,1.24, length.out = nrow(z.0.0320_0.0279)),
         y = seq(0.1,0.49, length.out = ncol(z.0.0320_0.0279)),
         z.0.0320_0.0279,
         xlab = "GMR",
         ylab = "CVw",
         lwd = 1)
title(main = "Figure3_0.0320_0.0279")
dev.off()

system2('open', args = c('-a Preview.app', '../results/Figure3.pdf'),
wait = FALSE)
```

```
#########################################################################
## End Code to obtain the resutls of Tables 1, 2, 3, 4, 5, 6 and
Figure 3 ##
#########################################################################
sessionInfo()
```