Carryover negligibility and relevance in bioequivalence studies

Jordi Ocaña[a*], María P. Sánchez O[b] and Josep L. Carrasco[c]

Abstract
The carryover effect is a recurring issue in the pharmaceutical field. It may strongly influence the final outcome of an average bioequivalence study. Testing a null hypothesis of zero carryover is useless: not rejecting it does not guarantee the non-existence of carryover, and rejecting it is not informative of the true degree of carryover and its influence on the validity of the final outcome of the bioequivalence study. We propose a more consistent approach: even if some carryover is present, is it enough to seriously distort the study conclusions or is it negligible? This is the central aim of this paper, which focuses on average bioequivalence studies based on 2×2 crossover designs and on the main problem associated with carryover: type I error inflation. We propose an equivalence testing approach to these questions and suggest reasonable negligibility or relevance limits for carryover. Finally, we illustrate this approach on some real datasets.
Keywords: carryover, crossover design, average bioequivalence, bioequivalence study.

1. INTRODUCTION

Average bioequivalence (ABE) studies are performed to demonstrate that the ratio of geometric mean bioavailabilities (BA) of a brand or reference (*R*) drug and a generic or test drug (*T*) lies within pre-specified limits of equivalence. In the original scale of measurements, these limits are typically 0.80 and 1.25 [1]. Bioavailability is measured in terms of specific variables like "area under the curve until time *t*", $AUC_{0\text{-}t}$, or maximum concentration, Cmax.

Normally, a logarithmic transformation of data is recommended. In the transformed scale, these limits become ±0.2231 and the difference of mean log-bioavailabilities, the formulation effect, must lie between them. Most regulatory agencies recommend that ABE studies be based on a 2×2, RT/TR, crossover design (two treatments, two periods and two sequences) and inference on the TOST (two one-sided tests) procedure. The α level TOST is operationally equivalent to the interval inclusion principle, say, to declare ABE if the usual parametric normal $1 - 2\alpha$ "shortest" confidence interval for the formulation effect lies within the bioequivalence limits.

[a] *Universitat de Barcelona, Department of Statistics, Faculty of Biology*
[b] *Universitat de Barcelona*
[c] *Universitat de Barcelona, Department of Public Health, Faculty of Medicine*
** Correspondence to: Jordi Ocaña, Departament d'Estadística, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal, 643, 08028 Barcelona, Spain.*
E-mail: jocana@ub.edu

Crossover designs allow within-subject comparison, but, as each subject receives a sequence of treatments, a carryover (or residual) effect may occur in the second (and any subsequent) administration period of the assay [1]. One of the assumptions underlying the standard ABE methods based on crossover trials is that carryover effects are absent [1]. In theory, we can avoid, minimise or rule out these effects if there is a presumed sufficient washout time between drug administrations. It is recommended that washout periods exceed five drug elimination half-lives [1], [2].

Given the possibility of disturbing carryover effects, Grizzle [3] proposed a two-stage procedure for the analysis of data from 2×2 crossover studies. First, to test the null hypothesis of non-existence of carryover at a significance level of $\alpha = 0.1$, or even 0.15, to ensure there is enough power. In case of non-rejection of the null hypothesis, he recommended proceeding with the standard analysis under no carryover. Otherwise, the recommendation was to use only the data from the first period, like data obtained in a fully randomised parallel trial. This strategy has been recommended in the past by the FDA [1], and is widely used in practice despite much criticism ([4], [5], [6], [7], [8]). The two-stage procedure is not mentioned in recent regulations (e.g. [2]).

Opponents of the two-stage procedure state that the best policy is not to test for carryover beforehand (or not to use this test as a basis for any further decisions on the analysis course) and to proceed as if it were absent. In well-performed experiments, carryover will commonly be absent, as the washout will normally succeed in eliminating it. This opinion seems to be confirmed by D'Angelo *et al*. [9] in their review of 324 two-way and 96 three-way crossover studies. Only a small proportion of these studies, compatible with the common significance level at which they were performed, resulted in a significant carryover. Moreover, for the subset of studies reporting the *p*-value, its empirical distribution was very close to the uniform. With these data, this distributional null hypothesis is never rejected by the Kolmogorov-Smirnov (KS) test [8]. These results are contested in [10] and [11], with simulations that suggest the lack of power of these KS tests. Senn *et al*. in [12] rebut these arguments, arguing the irrelevance of power calculations to interpret observational data. However, a presumed proper washout time doesn't always guarantee that carryover effects are removed, as is suggested, for example, in [13] and [14] (contested by [15]). Mills *et al*. in [16] review the methodological aspects of 116 crossover studies and conclude that carryover may likely be present in some of them. Their arguments mainly concern the design, including the lack of washout, and not the outcome of a carryover significance test. In a 71% of papers, the possibility of carryover is not taken into consideration in the methods section. Similar conclusions are reported in [17].

In recent years, a growing body of pharmacogenetics evidence also suggests that avoiding carryover in bioequivalence studies may pose problems. Peiró *et al*. in [18] identify a SNP polymorphism associated with cytochrome P-450 (CYP2C9*3), directly related to the pharmacokinetics of Tenoxicam. It may affect a bioequivalence study if, by chance, different proportions of each genotype are assigned to each sequence, as it is related to low drug clearance and high $AUC_{0-\infty}$ and $t_{1/2}$ (high-life time) values. The study was developed in 18 healthy volunteers. A detectable plasma drug concentration before the second administration (and after a presumed adequate washout period of 21 days) was observed in five volunteers. This situation could strongly influence the existence of carryover. Bioequivalence is declared when all volunteers are considered, but no bioequivalence is declared if only the volunteers with a particular variant of the polymorphism (CYP2C9) are considered. Wu *et al*. in [19] describe three different types of pharmacokinetic behaviour related to individual genotypes,

the so-called extensive, high and early metabolisers. The above results seem to reinforce the experimental grounds of the simulation studies in [20], where differences in pharmacokinetic behaviour between individuals may induce some carryover. It seems unquestionable that the genetic characteristics associated with the metabolising ability (high, medium or slow) of the volunteers in a bioequivalence study directly affect the concentration of a drug in the second period, and that, despite a presumed adequate washout period, in some cases a percentage of the drug is left over from the first period.

Carryover considerations aside, in more general statistical terms any pre-testing strategy like Grizzle's two-stage procedure should be avoided, as it leads to invalid tests which do not respect the nominal *global* test size [6, 21]. On the other hand, if used as a complementary diagnostic instead of a pre-test, it provides some insight on possible carryover, which seems desirable in any crossover study. But testing a null hypothesis of zero carryover is useless: not rejecting it does not guarantee the non-existence of carryover, and rejecting it is not informative of the true degree of carryover and its influence on the validity of the main conclusions of the study, e.g. to conclude bioequivalence (or not). In other words, statistical significance is not synonymous of relevance.

A more reliable approach would be equivalence testing: even if some carryover is present, is it enough to seriously distort the study conclusions or is it negligible? This is the point of view taken in this paper, with average bioequivalence studies based on 2×2 crossover designs as the main goal.

In the next section, we summarise some results and notation. In section 3, an approach for establishing the equivalence or negligibility limits (and their complementary relevance limits) for carryover in ABE studies is proposed. Section 4 introduces an equivalence testing procedure based on these limits. Section 5 is devoted to some illustrative examples. The paper concludes with a short discussion and some conclusions.

2. BASIC RESULTS AND NOTATION

In a 2×2 crossover design, each experimental subject receives a single dose of both formulations, *R* and *T*, in only one of two possible orders or treatment sequences, *RT* or *TR*. A sample of $N = n_1 + n_2$ subjects are randomly allocated, n_1 to sequence *RT* and n_2 to sequence *TR*. For a given variable *Y in the logarithmic scale*, say, *Y* = log Cmax or *Y* = log $AUC_{0\text{-}t}$, Y_{ijk} will designate an observation made on the *i*–th individual, in the *j*–th period and the *k*–th sequence, $i = 1, \ldots, n_k$, $j = 1,2$ and $k = 1,2$.
We consider the following underlying linear model:

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

where μ is a global mean, P_j is the fixed effect of the administration period *j*, $F_{(j,k)}$ is the fixed effect of the formulation administered on the *k*-th sequence and *j*-th period, and $C_{(j\text{-}1,k)}$ corresponds to the fixed effect of carryover. The possible carryover effect of the reference formulation from the first period to the second period in sequence 1 is denoted by C_R, while the equivalent effect of the test formulation in sequence 2 is denoted by C_T. Therefore:

$$C_{(j-1,k)} = \begin{cases} C_R & \text{if} \quad j = 2 \text{ and } k = 1 \\ C_T & \text{if} \quad j = 2 \text{ and } k = 2 \\ 0 & \text{otherwise} \end{cases}$$

with $C_R = -C_T = C$. Similarly,

$$F_{(j,k)} = \begin{cases} F_R & \text{if} \quad j = k \\ F_T & \text{if} \quad j \neq k \end{cases}$$

with $F_R = -F_T = F$, and $P_1 = -P_2 = P$ as we consider $\sum_{j=1}^{2} P_j = 0$.

We will designate the formulation effect as $\phi = F_T - F_R = -2F$, the carryover effect as $\kappa = C_T - C_R = -2C$ and the period effect as $\pi = P_2 - P_1$. $S_{i(k)} \sim N\left(0, \sigma_S^2\right)$ represents the random effect of the i-th subject nested in the k-th sequence. σ_S^2 is the inter-subject variance. $e_{ijk} \sim N(0, \sigma^2)$ is the random error, residual or disturbance term. Additionally, we assume independence between all $S_{i(k)}$, all e_{ijk}, and mutual independence between the $\{S_{i(k)}\}$ and the $\{e_{ijk}\}$.

For simplicity we assume constant residual (or within or intrasubject) variance, σ^2.

The inference on the formulation effect is based on the period difference contrasts for each subject i within each sequence k, $d_{ik} = 0.5\,(Y_{i2k} - Y_{i1k})$. Its expectation and variance are:

$$E\left(d_{ik}\right) = \begin{cases} \frac{1}{2}\left(\pi + \phi + C_R\right) \text{ if } k = 1 \\ \frac{1}{2}\left(\pi - \phi + C_T\right) \text{ if } k = 2 \end{cases} \qquad (2)$$

$$\operatorname{var}\left(d_{ik}\right) = \frac{1}{2}\sigma^2.$$

If $\bar{d}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} d_{ik}$ are the sample means of the period differences, its difference:

$$\bar{D} = \bar{d}_1 - \bar{d}_2 \qquad (3)$$

is an unbiased estimate of the formulation effect ϕ, provided that no carryover is present, *i.e.* if $\kappa = 0$. But in general, $\bar{D}$ is a biased estimator of ϕ:

$$E\left(\bar{D}\right) = \phi - \frac{1}{2}\kappa. \qquad (4)$$

The variance of the semidifference contrasts d_{ik} may be estimated as:

$$\hat{\sigma}_d^2 = \frac{1}{N-2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(d_{ik} - \bar{d}_k\right)^2 = \frac{1}{2}\hat{\sigma}^2 \qquad (5)$$

and then the standard error of $\bar{D}$ can be independently estimated by

$$\widehat{\operatorname{se}}_{\bar{D}} = \hat{\sigma}_d\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \hat{\sigma}_d\sqrt{\frac{N}{n_1 n_2}} = \hat{\sigma}\sqrt{\frac{N}{2n_1 n_2}}. \qquad (6)$$

According to the confidence interval inclusion principle, ABE is declared if the $1 - 2\alpha$ "shortest" confidence interval:

$$\bar{D} \pm t_{N-2}^{1-\alpha}\,\widehat{\operatorname{se}}_{\bar{D}} \qquad (7)$$

lies within the bioequivalence limits. In (7), $t_{N-2}^{1-\alpha}$ corresponds to the $1 - \alpha$ quantile of a Student's t distribution with $N - 2$ degrees of freedom.

While inference on the formulation effect is typically based on the difference contrasts, the inference on the carryover may be based on the sums of observations within each subject along all periods. Using the common "dot" notation, writing $Y_{i\bullet k} = Y_{i1k} + Y_{i2k}$ we have:

$$E\left(Y_{i\bullet k}\right) = \begin{cases} 2\mu + C_R & \text{if } k = 1 \text{ (in sequence 1)} \\ 2\mu + C_T & \text{if } k = 2 \text{ (in sequence 2)} \end{cases} \quad (8)$$

and

$$\sigma_+^2 = \text{var}\left(Y_{i\bullet k}\right) = 4\sigma_S^2 + 2\sigma^2 \quad (9)$$

that may be estimated as:

$$\hat{\sigma}_+^2 = \frac{\sum_{i=1}^{n_1}\left(Y_{i\bullet 1} - \bar{Y}_{\bullet\bullet 1}\right)^2 + \sum_{i=1}^{n_2}\left(Y_{i\bullet 2} - \bar{Y}_{\bullet\bullet 2}\right)^2}{N-2}. \quad (10)$$

From the above results, the usual estimator of the carryover effect may be expressed as:

$$\hat{\kappa} = \bar{Y}_{\bullet\bullet 2} - \bar{Y}_{\bullet\bullet 1} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{i\bullet 2} - \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{i\bullet 1} \quad (11)$$

with variance:

$$\text{var}\left(\hat{\kappa}\right) = \sigma_+^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \quad (12)$$

For a more in-depth introduction to these matters see, for example, [22] or [23].

3. ESTABLISHING CARRYOVER NEGLIGIBILITY (OR RELEVANCE) LIMITS

The numerical specification of the equivalence limits depends on each field of application, e.g. as a consensus among experts in the field. This is the origin of the 0.80/1.25 or ±0.2231 limits used in ABE [1]. Many studies on the impact of carryover in crossover assays refer to the case where the end goal of these assays is establishing difference and the main magnitude under consideration is the test power. For example, in this context, Willan and Patter [24] obtained a threshold for the relative carryover, κ/ϕ, in order to determine which strategy (either analysing the full set of data or only data from the first period) is better in terms of power.

In a previous paper [25] Sanchez *et al.* established that the most disturbing effect of carryover in bioequivalence studies is the considerable increase in the probability of type I error or consumer risk, that is, of inappropriately declaring bioequivalence. This inflation occurs when the carryover effect and the formulation effect both have the same sign (and then the relative carryover κ/ϕ is positive), in accordance with the fact that the expectation of the usual estimator of the formulation effect is $\phi - \kappa/2$. Then, in a scenario of true non-bioequivalence (e.g. positive ϕ, to the right of the bioequivalence limit), if the true carryover effect has the same sign as the formulation effect (e.g. it is positive), the estimated values of the formulation effect will more frequently tend to be within the bioequivalence limits (e.g. left-deviated with respect to ϕ). On the other hand, when the carryover effect and the formulation effect have different signs, the size of the usual bioequivalence test is only slightly reduced. Thus, it seems appropriate to establish carryover negligibility limits in terms of its tolerable impact on the true test size, say α^*. With a fixed nominal ABE significance level α (e.g. the usual 0.05),

our proposed strategy will be to determine the maximum tolerable value of α^* over α (e.g. two times α) and then to determine the level of carryover in which this level of true type I error is reached.

In Appendix I, we conclude that the crucial parameter in establishing carryover negligibility should be based on the scaled carryover, κ/σ. Specifically, we recommend the parameter θ defined as:

$$\theta = (\kappa/\sigma)\sqrt{n_1 n_2/(2N)} \tag{13}$$

A good, simple approximation to the negligibility limit in terms of this parameter is:

$$\theta_0 = \Phi^{-1}(\alpha^*) + z_{1-\alpha} \tag{14}$$

where Φ corresponds to the $N(0,1)$ distribution function and $z_{1-\alpha}$ to its $1-\alpha$ quantile. On the other hand, this negligibility limit may be computed more exactly (resulting in slightly more permissive negligibility limits), without a great deal of computational effort. In any case, irrespective of the origin of the limits, the carryover negligibility problem should be stated as an equivalence problem:

$$H_0: \sqrt{\frac{n_1 n_2}{2N}}\frac{\kappa}{\sigma} \le -\theta_0 \quad \text{or} \quad \sqrt{\frac{n_1 n_2}{2N}}\frac{\kappa}{\sigma} \ge \theta_0 \quad \text{vs} \quad H_1: -\theta_0 < \sqrt{\frac{n_1 n_2}{2N}}\frac{\kappa}{\sigma} < \theta_0. \tag{15}$$

Alternatively, a carryover relevance test, to prove the existence of a very disturbing level of carryover (out of a given threshold θ_0 associated with a given unacceptable level of consumer risk, α^*) should be stated as the complementary problem:

$$H_0: -\theta_0 \le \sqrt{\frac{n_1 n_2}{2N}}\frac{\kappa}{\sigma} \le \theta_0 \quad \text{vs} \quad H_1: \sqrt{\frac{n_1 n_2}{2N}}\frac{\kappa}{\sigma} < -\theta_0 \quad \text{or} \quad \sqrt{\frac{n_1 n_2}{2N}}\frac{\kappa}{\sigma} > \theta_0. \tag{16}$$

Note that greater sample sizes and/or lesser residual variabilities will tend to make $\theta = (\kappa/\sigma)\sqrt{n_1 n_2/(2N)}$ greater. In other words (and perhaps counter-intuitively at first sight), the same level of carryover will affect type I error to a greater extent than with smaller sample sizes and/or greater variability. This tendency and the validity of the above limits was confirmed in the simulations in [25] and in the simulations presented below. Note also that Wellek's test of carryover negligibility ([26], p. 284) is not directly applicable to (15), as its scaling variance is $\sigma_+^2 = 4\sigma_S^2 + 2\sigma^2$ while the scaling considered here is based on the residual variance σ^2.

Figure 1

4. TESTING CARRYOVER NEGLIGIBILITY AND RELEVANCE

4. 1. Carryover negligibility

The testing problem (15) for carryover negligibility may be rewritten as:

$$H_0: \frac{n_1 n_2}{2N}\left(\frac{\kappa}{\sigma}\right)^2 \ge \theta_0^2 \quad \text{vs} \quad H_1: \frac{n_1 n_2}{2N}\left(\frac{\kappa}{\sigma}\right)^2 < \theta_0^2$$

or, equivalently:

$$H_0 : \kappa^2 - \left(\frac{2N}{n_1 n_2} \theta_0^2 \right) \sigma^2 \geq 0 \quad \text{vs} \quad H_1 : \kappa^2 - \left(\frac{2N}{n_1 n_2} \theta_0^2 \right) \sigma^2 < 0. \qquad (17)$$

Note in advance that there may be some confusion because we are concerned with three "alpha" values: the nominal significance level α of the BE test, the limit of permissibility for its true BE test size, α^*, and the significance level at which we are testing if carryover is negligible, test (17). From now, this last significance level will be designated as α'.

Let U_η be the upper limit of a $1-\alpha'$ confidence interval $\left(-\infty, U_\eta\right]$ for $\eta = \kappa^2 - \left(\frac{2N}{n_1 n_2} \theta_0^2 \right) \sigma^2$.

According to the interval inclusion principle, to reject H_0 if $U_\eta < 0$ defines a test of size α'. This upper limit may be derived using the Howe's method, [27], adapted to a bioequivalence context in [28] and [29]. For a linear combination of parameters $\sum c_j \theta_j$, like η with $c_1 = 1$, $\theta_1 = \kappa^2$, $c_2 = -\left(2N/\left(n_1 n_2\right)\right)\theta_0^2$ and $\theta_2 = \sigma^2$, let E_j be independent point estimators for each summand $c_j\theta_j$ and U_j be the corresponding upper limits of $1-\alpha'$ one-sided confidence intervals for $c_j\theta_j$. If $D_j = \left(U_j - E_j\right)^2$ then:

$$U_\eta = \sum E_j + \sqrt{\sum D_j}, \qquad (18)$$

is the upper limit of an approximate $1-\alpha'$ one-sided confidence interval for η.

Unfortunately, the variance of the usual estimator of κ is σ_+^2, which depends on the intra and inter subject variation and is usually large. The variance of $\hat{\kappa}/\hat{\sigma}$ is even larger due to the random denominator. As a consequence, the test for (17) based on (18) tends to be biased for the most reasonable α^* values, like 0.06 (a 20% increase over 0.05), 0.1 or 0.15 for a nominal α = 0.05, even using "permissive" values $\alpha' = 0.10$ or 0.15 in the same line suggested by Grizzle in [3]. Their power properties improve for more extreme values like α^*= 0.50, but the statement that carryover "is negligible" because "the risk of inadequately declaring ABE is not over 0.50" lacks any interest.

So, for the moment, the problem of carryover negligibility must remain in a descriptive but not inferential status: the estimate of the scaled carryover (13) may only suggest lack of alarming carryover levels.On the other hand, limited but possibly more interesting results may be obtained for the reciprocal problem of carryover relevance.

4.2. Carryover relevance

A test of carryover relevance for the problem (16) may be of interest for "large" values α^* like 0.10 or 0.20, as an a posteriori diagnostic of extreme carryover. An interesting value is α^*= 0.50; then, rejecting the null hypothesis of carryover negligibility will suggest that the BE study under consideration has a user's risk control not better than simply tossing a coin and deciding to declare BE or not, ignoring data. If $\tau = 1 / \theta$, that is, $\tau = \left(\sigma / \kappa\right)\sqrt{2N / \left(n_1 n_2\right)}$, the above problem reduces to an equivalence or negligibility problem:

$$H_0 : \sigma^2 - \left(\frac{n_1 n_2}{2N} \left(\frac{1}{\theta_0} \right)^2 \right) \kappa^2 \geq 0 \quad \text{vs} \quad H_1 : \sigma^2 - \left(\frac{n_1 n_2}{2N} \left(\frac{1}{\theta_0} \right)^2 \right) \kappa^2 < 0. \qquad (19)$$

According to Howe's method, we can obtain an upper confidence interval limit U_η for the parameter

$$\eta = \sigma^2 - \frac{n_1 n_2}{2N} \left(\frac{1}{\theta_0} \right)^2 \kappa^2 \qquad (20)$$

from the estimators and upper confidence interval limits summarized in Table I:

Table I. Point estimators and confidence intervals to construct a confidence interval for the parameter η defined in (20).

Parameter	**Point estimator**	**Upper confidence interval limit**
σ^2	$E_1 = \hat{\sigma}^2$	$U_1 = \dfrac{\hat{\sigma}^2 (N-2)}{\chi^2_{\alpha', (N-2)}}$
$\left(-\dfrac{n_1 n_2}{2N} \left(\dfrac{1}{\theta_0} \right)^2 \right) \kappa^2$	$E_2 = -\dfrac{n_1 n_2}{2N} \left(\dfrac{1}{\theta_0} \right)^2 \hat{\kappa}^2$	$U_2 = -\dfrac{n_1 n_2}{2N} \left(\dfrac{1}{\theta_0} \right)^2 \max \left\{ 0, \left(\lvert\hat{\kappa}\rvert + t^{\alpha'}_{N-2} \hat{\sigma}_+ \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}} \right)^2 \right\}$ based on [30])

where $\chi^2_{\alpha', (N-2)}$ corresponds to the α' quantile of a chi-square distribution with $N - 2$ degrees of freedom, and $t^{\alpha'}_{N-2}$ corresponds to the α' quantile of a Student's t distribution with $N - 2$ degrees of freedom.

If $U_\eta < 0$ then the null hypothesis in (19) may be rejected, concluding that there is a relevant carryover, perhaps questioning the validity of a previous bioequivalence study declaring bioequivalence. This test is approximately valid provided that the intersubject variance σ_S^2 is not much larger than the residual variance, or more precisely, provided that the intraclass correlation $\rho_I = \sigma_S^2 / \left(\sigma_S^2 + \sigma^2 \right)$ is not too large. Once fixed an upper bound for the maximum degree of true type I error level for the relevance test, the maximum allowable ρ_I is a growing function of α^*. Figure 2 displays the maximum allowable intraclass correlation for which the true type I error probability of the negligibility test is sufficiently closer (±20%) to a nominal size $\alpha' = 0.05$, in a balanced 2×2 design for sample sizes n = 12, 24 and 36.

Figure 2

These results were obtained in a simulation study whose complete results and R code are available at www.ub.edu/stat/recerca/materials/Carryover_negligibility_and_relevance.htm. Figure 3 displays a subset of the more interesting simulation results. It corresponds to the power curve of the relevance test (say the probability of declaring carryover relevance) when "relevance" is set at $\alpha^* = 0.50$ and the test is performed at three possible significance levels, $\alpha' = 0.05$, 0.1 or 0.15, for a balanced sample size n = 12. The probability of declaring carryover relevance is displayed in function of the parameter θ defined in (13), in terms of a fraction of the relevance limit $\theta_0(\alpha^*)$. Each probability line corresponds to a given proportion between the "intra" and the "inter" subject variances, σ^2 and σ_S^2, expressed in terms of a given value of intraclass correlation, ρ_I. Ideally, the probability of declaring carryover relevance should be below α' (horizontal thick line) for fractions at left of 1 in the abscises axis, it should be exactly 0.05 for a unit fraction and should be above this reference value for fractions at right of 1. This behavior is acceptably displayed in all situations except when the intraclass correlation is too high.

Figure 3

5. NUMERICAL EXAMPLES

5.1 Example 1

We illustrate the above procedures using a dataset, which is accessible through the FDA website. It corresponds to dataset 29, "Cholinesterase inhibitor", in Section II, which is devoted to non-replicate designs, at:
http://www.fda.gov/downloads/Drugs/ScienceResearch/UCM301914.txt.
These data correspond to a balanced 2×2 crossover design for a total of N = 28 subjects, n = 14 in each sequence. The measured variables were the area under the curve until time t, $AUC_{0\text{-}t}$, and the peak plasma concentration of a drug after oral administration, Cmax.
Table II shows the main results of a standard bioequivalence and ANOVA analysis. ANOVA is performed through a parameterisation that allows for estimation of the overall mean, period effects, treatment effects and carryover effects, assuming that no sequence effects exist. The drug has low within-subject variability and the study has adequate power, provided that the number of healthy volunteers included in the protocol is sufficient. For both variables, the standard ANOVA or Student's t test procedures reject the null hypothesis of null carryover effect, $\kappa = 0$, but do not give any idea of the magnitude of these non-null carryovers and their possible impact on a bioequivalence study.

Table II. ABE and ANOVA analysis

Bioavailability measure	**90% IC shortest*** (regulatory)	**TOST*** (p-value superior and inferior)	**carryover effect p-value (ANOVA)****
$AUC_{0\text{-}t}$	[96.60; 106.79] *Bioequivalent*	<0.0001 <0.0001	0.0236
Cmax	[96.87; 111.00] *Bioequivalent*	<0.0001 <0.0001	0.0318

** Procedure TTEST and TOST calculated with SAS v. 9.2 ** Calculated with STATA v. 11.00, by parameterisation #1*

Bioavailability measure	**Within-subject coefficient of variation [%]**	**Minimum required sample size, *n* for sequence, *N* total***	**Calculated power of the BE test**
$AUC_{0\text{-}t}$	10.03	$n_1 = n_2 = 8$, $N = 16$	91.41%
Cmax	14.21	$n_1 = n_2 = 12$, $N = 24$	86.82%

** Calculated with R library PowerTOST*

For the logarithmically transformed $AUC_{0\text{-}t}$,, the estimated carryover (expression (11)) is $\hat{\kappa} = -0.7568$ and the residual standard deviation $\hat{\sigma} = 0.1166$. This makes $\hat{\kappa}/\hat{\sigma} = -6.4878$ and the estimated θ parameter becomes -12.1376. Considering a limit for type I error $\alpha^* = 0.50$, and the standard α = 0.05 for the bioequivalence test, the associated relevance limits for standardised carryover become $\pm\theta_0 = \pm 1.6889$, so the estimated θ is more than seven times this limit. This may be interpreted as a suggestion of a possibly highly relevant carryover. In fact, the test for carryover relevance proposed in section 4 gives a significant result at a standard significance level $\alpha' = 0.05$ as the upper limit of the one-sided confidence interval for the parameter (20) is negative: -0.0457. Following Grizzle's recommendation of testing carryover with more permissive significance levels, like 0.10 or 0.15, the preceding result is still clearer, for example for $\alpha' = 0.15$ the confidence interval upper limit becomes -0.2969.
These results must be taken with care as the estimated intraclass correlation is very high, 0.9245, which makes the relevance test too permissive, according to Figures 2 and 3. They

may suggest the convenience of revising the experimental protocols, but should not be taken as a full evidence of distorting carryover.

For the logarithmically transformed Cmax, the corresponding values are $\hat{\kappa} = -0.4782$ and $\hat{\sigma} = 0.1453$. Then $\hat{\theta} = (\hat{\kappa}/\hat{\sigma})\sqrt{n_1 n_2/(2N)} = -6.1585$ suggests a carryover level of nearly four times the relevance limit $\pm\theta_0 = \pm 1.6889$. Again, relevant carryover may be suspected. This is not corroborated at $\alpha' = 0.05$ (upper limit of the confidence interval 0.0042) but for $\alpha' = 0.15$ the confidence interval limit is negative, −0.0588, which conducts to the rejection of the null hypothesis of irrelevant carryover. This result is more reliable as the estimated intraclass correlation, 0.7608, lies within the validity range of the test.

It is worth to say that for both bioavailability variables, the formulation effect estimates $\bar{D}$ were positive (and so with a different sign than the carryover estimate), although very close to zero: 0.0051 and 0.0363 respectively for AUC and Cmax. Only when the formulation effect and the carryover effect have the same sign, carryover is potentially dangerous with respect to type I error distortion. On the other hand, the evidences of strong carryover, under expression (4) suggest that the formulation effect estimate may be strongly biased towards positive values. The negative values of alternative formulation effect estimators like the one based only on the first period data, say $\hat{\phi}_1 = \bar{Y}_{\cdot 12} - \bar{Y}_{\cdot 11}$ (= −0.3733 and −0.2028, for AUC and Cmax respectively) where $\bar{Y}_{\cdot 1k} = n_1^{-1}\sum_{i=1}^{n_j} Y_{i1k}$ stands for the mean of all observations in period 1 and sequence *k*, and the synthetic estimator of Longford [31] (−0.2161 and −0.0787, respectively) which is based on a weighted average of $\bar{D}$ and $\hat{\phi}_1$, give some credibility to the possibility of a truly negative formulation effect and thus to the formulation effect and the carryover effect having the same sign.

Table III. Carryover relevance analysis

Parameter	Estimated carryover	Estimated θ	Carryover limits $\pm\theta_0$	Upper limit U_η. Negative value ⇒ carryover relevance at α'=0.05 (α'=0.15)
$AUC_{0\text{-}t}$	−0.7568	−12.1376	±1.6889 (θ is 7.2 times θ_0)	−0.0457 (−0.2969)
Cmax	−0.4782	−6.1585	±1.6889 (θ is 3.6 times θ_0)	0.0042 (−0.0588)

5.2 Example 2

As a second example, we use the results of a true but unrecognisable bioequivalence study available at www.ub.edu/stat/recerca/materials/Example2Carryover.pdf.

In short, for a balanced sample size of 12 in each sequence, for all three pharmacokinetic parameters, the ANOVA for carryover is non-significant at a 0.05 level. The p-values are 0.1859, 0.2077 and 0.1123 for the logarithms of $AUC_{0\text{-}t}$, $AUC_{0\text{-}\infty}$ and Cmax, respectively. The null hypothesis of zero carryover may be rejected for Cmax using the more permissive level 0.15. But in any case these results do not provide any indication on the true distorting effect of carryover on the BE study, if present. For example, one may question if these carryovers may put the probability of erroneously declaring BE at an unacceptable α^* = 0.50 level.

For Cmax, the estimated carryover is 1.096 and the estimated within-subject σ is 0.333 which gives an estimated θ value of 5.699. Testing relevance at a 0.05 level, 5.699 corresponds to more than 3 times the relevance limit θ_0 = 1.6977. These results seem to suggest carryover relevance, but the upper confidence interval limit U_η is 0.114 so no significant carryover relevance may be declared. On the other hand, if relevance is also tested at a 0.15 level, the upper confidence interval limit U_η becomes –0.0485 and carryover relevance is declared, thus suggesting evidence for an unacceptable user's risk of 0.50 of incorrectly declaring bioequivalence. The intraclass correlation for Cmax is 0.8446, in the limit but still supporting a credible carryover relevance test.
The same results (evidence of relevant carryover at a 0.15 level but not at 0.05, when the possibility of reaching a true type I error probability 0.5 is considered) are obtained for the other two pharmacokinetic parameters, but at very high intraclass correlation values, 0.9538 and 0.9566 for $AUC_{0\text{-}t}$ and $AUC_{0\text{-}\infty}$ respectively, which decrease the credibility of the corresponding relevance results.

6 DISCUSSION

In our opinion, there is enough evidence to state that some factors may directly affect the concentration of a drug in the second period of a crossover study, and that, despite a presumed adequate washout period, sometimes a certain degree of carryover may be present. Among these factors, there is the possible presence of phenotypes associated with metabolising ability (e.g. extensive, intermediate, poor and ultra-rapid metabolizer, [32]) in the volunteers who participate in a bioequivalence study, a factor that may directly affect the concentration of a drug in the second period. Provided that the frequencies of the alleles associated to these phenotypes may vary across human groups, these considerations also pose some doubt in the automatic transportability of bioequivalence studies between countries or ethnical groups. Obviously, these considerations are only relevant (for carryover) if the differences in metabolizing ability are translated in some way to differences associated with the different formulations.
Our first example was chosen quite deliberately to illustrate a case where high carryover was suspected in advance due to heterogeneity with respect to gender. These effects of subgroup heterogeneity (gender, phenotypes, age, etc.) that induce some subject-by-formulation interaction (confounded with the carryover effect) have been emphasized by the regulators. Chapter III of [33], "Methods to document BA and BE, Part A, Pharmacokinetic Studies, item 5. Study Population" recommends that "in vivo BE studies be conducted in individuals representative of the general population, taking into account age, sex, and race. We recommend that if the drug product is intended for use in both sexes, the sponsor attempt to include similar proportions of males and females in the study". Here we are dealing with subject-by-formulation interaction and not with representativeness, but it is worth pointing that there is also some scepticism among specialists concerning these recommendations, as BE studies are performed on healthy volunteers and not in patients, and presumably representativeness will be not an issue.
Therefore, ignoring the carryover issue in bioequivalence studies may not be the best strategy, especially given that carryover may severely affect the type I error, that is to say, the user risk associated with wrongly declaring bioequivalence. We suggest that bioequivalence studies should be accompanied by some analysis exploring the possible presence of disturbing degrees of carryover, as a way of reinforcing its credibility or lack thereof. In its present status, a positive result of the testing procedure for carryover relevance may not be presented

as a feasible proof of inadequacy of the BE study (and even more clearly, the test for carryover negligibility may not be presented as a proof of its adequacy), but perhaps should be taken by a regulatory authority as a suggestion for the convenience of requiring more information about the experiment to the applicant laboratory.

The above comment suggests where our method may be of main interest: when analysing data coming from an external source, with limited control on the amount of information available by the analyser (e.g. a journal reviewer or a regulatory agency examining a generic application). There are other possible ways of evaluating carryover, e.g. using baseline measurements before the second administration; our method may conduct to conclude that such complementary information is necessary and to seek for it.

In any case we are not promoting a two-stage approach to BE determination. Our point of view is strictly one-stage, always assuming null or negligible carryover and thus the correctness of the decision on BE based on the confidence interval (7). But in the same way that, for example, a look to the residuals is always advisable, a look to any suspected trace of possible disturbing carryover may be a good policy, possibly for asking for supplementary information on the experiment, with the desirable end goal of finally stablishing its correctness.

Acknowledgments
This research was supported by research projects MTM2008-00642, Ministerio de Ciencia e Innovación, and 2005SGR00871, Generalitat de Catalunya, and partially supported by Spanish MEyC grant CGL2008-05448-C02-02/BOS. In its last stage, it was also supported by grant 2014 SGR 464, Generalitat de Catalunya.

REFERENCES

[1] FDA U.S. Food and Drug Administration., "Guidance for industry: Statistical approaches to establishing bioequivalence". CDER, Department of Health and Human Services. Rockville, MD., 2001. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070244.pdf (accessed 8 August 2014)

[2] Committee or Medicinal Products for Human Use (CHMP). "EMEA European Medicines Agency. Guideline on the investigation of Bioequivalence". Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf (accessed 8 August 2014).

[3] Grizzle JE. The two-period change-over design and its use in clinical trials. *Biometrics* 1965; **21**:467-480, http://www.jstor.org/pss/2528104 (accessed 8 August 2014).

[4] Brown B. The crossover experiment for clinical trials. *Biometrics* 1980; **36**:69-79. Available at: ftp://www.biostat.wisc.edu/pub/chappell/641/papers/paper18.pdf (accessed 8 August 2014)

[5] Senn S. Cross-over trials, carry-over effects and the art of self-delusion. *Statistics in Medicine* 1988; **7**(10):1099-1101. DOI:10.1002/sim.4780071010. Available at: http://onlinelibrary.wiley.com/doi/10.1002/sim.4780071010/abstract (accessed 8 August 2014)

[6] Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Statistics in Medicine* 1989; **8**(12):1421-1432. DOI:10.1002/sim.4780081202. Available at: http://onlinelibrary.wiley.com/doi/10.1002/sim.4780081202/abstract (accessed 8 August 2014)

[7] Senn S. The AB/BA Cross-over: How to perform the two-stage analysis if you can't be persuaded that you shouldn't. Published in:Hansen and de Ridder Eds, Liber Amicorum Roel van Strik, Rotterdam, Erasmus University, 1996; pp.93-100. Available at: ftp://ftp.biostat.wisc.edu/pub/chappell/641/notes.week5and6/senn1.pdf (accessed 8 August 2014)

[8] Senn S, D'Angelo G and Potvin D. Carry-over in cross-over trials in bioequivalence: theoretical concerns and empirical evidence. *Pharmaceutical Statistics* 2004; **3**(2):133-142. DOI:10.1002/pst.111. Available at: http://onlinelibrary.wiley.com/doi/10.1002/pst.111/abstract (accessed 8 August 2014)

[9] D'Angelo G, Potvin D and Turgeon J. Carry-over effects in bioequivalence studies. *Journal of Biopharmaceutical Statistics* 2001; **11**(1-2):35-43. DOI:10.1081/BIP-100104196. http://www.tandfonline.com/doi/abs/10.1081/BIP-100104196 (accessed 8 August 2014)

[10] Putt M. Comment on "Carry-over in cross-over trials in bioequivalence: theoretical concerns and empirical evidence". Senn S, D'Angelo G, Potvin D. *Pharmaceutical Statistics* 2005; **4**(3):215-216. DOI: 10.1002/pst.174. Available at: http://onlinelibrary.wiley.com/doi/10.1002/pst.174/abstract (accessed 8 August 2014)

[11] Putt M. Power to detect clinically relevant carry-over in a series of cross-over studies. *Statistics in Medicine* 2006; **25**(15):2567-2586. DOI:10.1002/sim.2275. Available at: http://onlinelibrary.wiley.com/doi/10.1002/sim.2275/abstract (accessed 8 August 2014)

[12] Senn S and D'Angelo G. Rejoinder: Dr. Putt's analysis. *Pharmaceutical Statistics* 2005; **4**(3):216-219. DOI: 10.1002/pst.181. Available at: http://onlinelibrary.wiley.com/doi/10.1002/pst.181/abstract (accessed 8 August 2014)

[13] Bellamy MF, W. McDowell, Ramsey M, Brownlee M, Newcombe R and Lewis M. Oral folate enhances endothelial function in hyperhomocysteinaemic subjects. *European Journal of Clinical Investigation* 1999; **29**:659–662. DOI:10.1046/j.1365-2362.1999.00527.x. Available at: http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2362.1999.00527.x/abstract Responses: Bellamy MF, McDowell IFW, Lewis MJ. *Circulation* 2000; DOI: 10.1161/01.CIR.102.11.e92. Available at: http://circ.ahajournals.org/content/102/11/e92.full (accessed 8 August 2014)

[14] Blakesley V, Awni W, Locke C, Ludden T, Granneman GR, Braverman LE. Are Bioequivalence Studies of Levothyroxine Sodium Formulations in Euthyroid Volunteers Reliable?. *Thyroid* 2004; **14**(3):191-200. DOI:10.1089/105072504773297867. Available on line at: http://online.liebertpub.com/doi/abs/10.1089/105072504773297867 (accessed 8 August 2014)

[15] Bolton S. Bioequivalence Studies for Levothyroxine. *The AAPS Journal* 2005; **7**(1):E47-E53. DOI: 10.1208/aapsj070106. Available at: http://www.pharmagateway.net/ArticlePage.aspx?DOI=10.1208/aapsj070106 (accessed 8 August 2014)

[16] Mills EJ, Chan A-W, Wu P, Vail A, Guyatt GH, Altman DG. Design, analysis, and presentation of crossover trials. *Trials* 2009; **10**:27. DOI:10.1186/1745-6215-10-27. Available on line at: http://www.trialsjournal.com/content/10/1/27 (accessed 8 August 2014).

[17] Diaz-Uriarte R. Incorrect analysis of crossover trials in animal behaviour research. *Animal Behaviour* 2002; 63:815–822. DOI:10.1006/anbe.2001.1950. Available on line at: http://ligarto.org/rdiaz/Papers/cross-over-animal-behaviour.pdf (accessed 8 August 2014).

[18] Peiró AM, Novalbos J, Zapater P, Moreu R, López-Rodríguez R, Rodríguez V, Abad-Santos F, Horga JF. Pharmacogenetic relevance of the CYP2C9*3 allele in a tenoxicam bioequivalence study performed on Spaniards. *Pharmacological Research* 2009; **59**:62–68. DOI:10.1016/j.phrs.2008.09.018. Available on line at: http://www.sciencedirect.com/science/article/pii/S1043661808001813 (accessed 8 August 2014)

[19] Wu R, Tong Ch, Wang Z, Mauger D, Tansitira K, Szefler SJ, Chinchilli V, Israel E. A conceptual framework for pharmacodynamics genome-wide associated studies in pharmacogenomics. *Drug Discovery Today* 2011, **16**(19/20):884-890. Available on line at: http://www.sciencedirect.com/science/article/pii/S1359644611002911 (accessed 8 August 2014)

[20] Dhariwal K, Jackson A. Effect of Length of Sampling Schedule and Washout Interval on Magnitude of Drug Carryover From Period 1 to Period 2 in Two-Period, Two-Treatment Bioequivalence Studies and Its Attendant Effects on Determination of Bioequivalence. *Biopharmaceutics & Drug Disposition* 2003, **24**: 219–228. DOI: 10.1002/bdd.359. Available at: http://onlinelibrary.wiley.com/doi/10.1002/bdd.359/abstract? (accessed 8 August 2014)

[21] Senn, S. **The AB/BA Cross-over: How to perform the two-stage analysis if you can't be persuaded that you shouldn't.** *Liber Amicorum Roel van Strik*. B. Hansen and M. de Ridder. Rotterdam, Erasmus University: 93-100. Available on line at: ftp://ftp.biostat.wisc.edu/pub/chappell/641/notes.week5and6/senn1.pdf (accessed 23 February 2015).

[22] Chow S-C, J-P Liu. Design and Analysis of Bioavailability and Bioequivalence Studies (3rd edn), Chapman & Hall/CRC, Boca Raton, 2009.

[23] Patterson S, Jones B. Bioequivalence and Statistics in Clinical Pharmacology,: Chapman & Hall/CRC, Boca Raton, 2006.

[24] Willan AR, Pater JL. Carryover and the two-period crossover clinical trial. *Biometrics* 1986;

42(3):593-599. Available on line at: http://www.jstor.org/stable/2531209 (accessed 8 August 2014)

[25] Sánchez MP, Ocaña J, Carrasco JL. The effect of variability and carryover on average bioequivalence assessment: a simulation study,» *Pharmaceutical Statistics* 2010; **10**(2):135-142. DOI:10.1002/pst.431. Available on line at: http://onlinelibrary.wiley.com/doi/10.1002/pst.431/full (accessed 8 August 2014)

[26] Wellek S. Testing Statistical Hypotheses of Equivalence and Noninferiority (2nd ed). Chapman and Hall/CRC Press, Boca Raton, 2010.

[27] Howe WG. Approximate confidence limits on the mean of X + Y where X and Y are two tabled independent random variables. *Journal of the American Statistical Association* 1974; **69**(347):789-794. Available on line at: http://www.jstor.org/pss/2286019 (accessed 8 August 2014)

[28] Hyslop T, Hsuan F, Holder DJ. A small sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine* 2000; **19**(20):2885-2897. DOI: 10.1002/1097-0258(20001030)19:20<2885::AID-SIM553>3.0.CO;2-H. Available on line at: http://onlinelibrary.wiley.com/doi/10.1002/1097-0258%2820001030%2919:20%3C2885::AID-SIM553%3E3.0.CO;2-H/abstract (accessed 8 August 2014)

[29] Tothfalusi L and Endrenyi, L. Limits for the scaled bioequivalence of highly variable drugs and drugs products. *Pharmaceutical Research* 2003, **20**(3): 382–389. Available on line at: http://www.pharmagateway.net/ArticlePage.aspx?DOI=10.1023/A:1022695819135 (accessed 9 August 2014)

[30] J. Hsu, J. Hwang, H.-K. Liu and S. Ruberg. Confidence intervals associated with tests of bioequivalence. Biometrika 1994, **81**(), 103-114. Available on line at: http://biomet.oxfordjournals.org/content/81/1/103.full.pdf (accessed 9 August 2014)

[31] Longford N. Synthetic estimators with moderating influence: the carry-over in cross-over trials revisited. *Statistics in Medicine* 2001; **20**(21):3189–3203. DOI:10.1002/sim.926. Available on line at: http://onlinelibrary.wiley.com/doi/10.1002/sim.926/pdf (accessed 9 August 2014)

[32] Evans WE, McLeod HL. Pharmacogenomics — Drug Disposition, Drug Targets, and Side Effects. New England Journal of Medicine. 2003, 348(6):538- 549. Available on line at: http://www.nejm.org/doi/full/10.1056/NEJMra020526 (accessed 9 August 2014)

[33] FDA in Guidance for Industry: Bioavailability and Bioequivalence, Studies for Orally Administered Drug Products – "General Considerations" in chapter III "Methods to document BA and BE, Part A. Pharmacokinetic Studies, item 5. Study Population. Available on line at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070124.pdf (accessed 9 August 2014)

Appendix I. Determining the carryover negligibility limits

Consider an ABE assay where the usual test (i.e. TOST/Interval inclusion principle for the "shortest" confidence interval) is unbiased. In other words, with a fixed significance level α, the probability of rejecting the null hypothesis of bioinequivalence will be $\leq \alpha$ if this hypothesis is true and $\geq \alpha$ otherwise. This will exclude some possible cases from our consideration, such as the study of a high variability drug under insufficient sample size. Let the null hypothesis of bioinequivalence be true. Then, the true unknown formulation effect will verify $\phi \geq \phi_0$ (= 0.2231, usually) or $\phi \leq -\phi_0$. We will wrongly reject H_0 and declare ABE

if the confidence interval (7) is fully included in the interval $[-\phi_0, +\phi_0]$. The probability of this event, i.e. the probability of type I error, is:

$$\begin{aligned}&\Pr\left\{\left[\bar{D}-t_{N-2}^{1-\alpha}\hat{\sigma}\sqrt{\frac{N}{2n_1n_2}}>-\phi_0\right]\cap\left[\bar{D}+t_{N-2}^{1-\alpha}\hat{\sigma}\sqrt{\frac{N}{2n_1n_2}}<\phi_0\right]\right\}\\&=\Pr\left\{-\phi_0+t_{N-2}^{1-\alpha}\hat{\sigma}\sqrt{\frac{N}{2n_1n_2}}<\bar{D}<\phi_0-t_{N-2}^{1-\alpha}\hat{\sigma}\sqrt{\frac{N}{2n_1n_2}}\right\}.\end{aligned}\qquad(21)$$

The above quantity may be approximated by simulation, generating a large number *m* (e.g. 10^6) of independent $\bar{D}\sim N\left(\phi-\frac{1}{2}\kappa,\ \sigma_{\bar{D}}=\sigma\sqrt{\frac{N}{2n_1n_2}}\right)$ and, independently, *m* $\hat{\sigma}^2\sim\frac{\sigma^2}{N-2}\chi^2(N-2)$ values and computing the frequency of $-\phi_0+t_{N-2}^{1-\alpha}\hat{\sigma}\sqrt{\frac{N}{2n_1n_2}}<\bar{D}<\phi_0-t_{N-2}^{1-\alpha}\hat{\sigma}\sqrt{\frac{N}{2n_1n_2}}$.

A simple approximation to (21) may be derived by substituting the estimate of the residual standard deviation by the corresponding population value σ:

$$\Pr\left\{-\phi_0+t_{N-2}^{1-\alpha}\sigma\sqrt{\frac{N}{2n_1n_2}}<\bar{D}<\phi_0-t_{N-2}^{1-\alpha}\sigma\sqrt{\frac{N}{2n_1n_2}}\right\}.\qquad(22)$$

Then, this probability depends on the fact that:

$$\bar{D}\sim N\left(\phi-\frac{1}{2}\kappa,\ \sigma_{\bar{D}}=\sigma\sqrt{\frac{N}{2n_1n_2}}\right).$$

Standardising the above variable, we have:

$$\Phi\left(\frac{\phi_0-t_{N-2}^{1-\alpha}\sigma\sqrt{\frac{N}{2n_1n_2}}-\left(\phi-\frac{1}{2}\kappa\right)}{\sigma\sqrt{\frac{N}{2n_1n_2}}}\right)-\Phi\left(\frac{-\phi_0+t_{N-2}^{1-\alpha}\sigma\sqrt{\frac{N}{2n_1n_2}}-\left(\phi-\frac{1}{2}\kappa\right)}{\sigma\sqrt{\frac{N}{2n_1n_2}}}\right)=$$

$$\Phi\left(\sqrt{\frac{n_1n_2}{2N}}\left[\frac{2(\phi_0-\phi)}{\sigma}+\frac{\kappa}{\sigma}\right]-t_{N-2}^{1-\alpha}\right)-\Phi\left(\sqrt{\frac{n_1n_2}{2N}}\left[\frac{2(-\phi_0-\phi)}{\sigma}+\frac{\kappa}{\sigma}\right]+t_{N-2}^{1-\alpha}\right)$$

where Φ corresponds to the $N(0,1)$ distribution function. The worst case, the maximum type I error probability corresponding to the bioequivalence test size, is reached in the bioequivalence limit, when $\phi_0=|\phi|$:

$$\alpha^*\cong\Phi\left(\sqrt{\frac{n_1n_2}{2N}}\frac{\kappa}{\sigma}-t_{N-2}^{1-\alpha}\right)-\Phi\left(\sqrt{\frac{n_1n_2}{2N}}\frac{\kappa}{\sigma}+t_{N-2}^{1-\alpha}-2\frac{\phi_0}{\sigma}\sqrt{\frac{2n_1n_2}{N}}\right).\qquad(23)$$

Note that the crucial parameter is the scaled carryover: $\theta=(\kappa/\sigma)\sqrt{n_1n_2/(2N)}$. The expression above provides values of the true test size that approximate the exact (21) values. Provided that when $\kappa=0$, α^* should equate α, this approximation may be improved by substituting the critical *t* value by the normal quantile, $z_{1-\alpha}$:

$$\alpha^*\cong\Phi\left(\sqrt{\frac{n_1n_2}{2N}}\frac{\kappa}{\sigma}-z_{1-\alpha}\right)-\Phi\left(\sqrt{\frac{n_1n_2}{2N}}\frac{\kappa}{\sigma}+z_{1-\alpha}-2\frac{\phi_0}{\sigma}\sqrt{\frac{2n_1n_2}{N}}\right).\qquad(24)$$

Numerically inverting function (24) –or inverting a table of pairs (θ, α^*) coming from the exact expression (22)– the equivalence (negligibility) limits $\pm\theta_0$ for the standardised carryover $\theta = (\kappa/\sigma)\sqrt{n_1 n_2/(2N)}$ may be determined. In fact, α^* also depends on σ, unknown, in the denominator of ϕ_0 / σ. But even for small sample sizes like $n_1 = n_2 = 6$,

$$\Phi\left(\sqrt{\frac{n_1 n_2}{2N}}\frac{\kappa}{\sigma} - z_{1-\alpha}\right) \gg \Phi\left(\sqrt{\frac{n_1 n_2}{2N}}\frac{\kappa}{\sigma} + z_{1-\alpha} - 2\frac{\phi_0}{\sigma}\sqrt{\frac{2n_1 n_2}{N}}\right) \simeq 0$$

and then

$$\alpha^* \simeq \Phi\left(\sqrt{\frac{n_1 n_2}{2N}}\frac{\kappa}{\sigma} - z_{1-\alpha}\right) = \Phi(\theta - z_{1-\alpha}) \qquad (25)$$

that may be inverted in exact form:

$$\tilde{\theta}_0 = \Phi^{-1}(\alpha^*) + z_{1-\alpha}. \qquad (26)$$

Once α^* and α are fixed, $\tilde{\theta}_0$ in (26) corresponds to the limit when $\max\{n_1,n_2\} \to \infty$ of θ_0 obtained from the exact probabilities (22). Provided that $\tilde{\theta}_0 < \theta_0$ (though they are always very similar quantities), $\tilde{\theta}_0$ may be considered a simple, pessimistic approximation to the true negligibility limit θ_0, although this is not difficult to precisely approximate by simulation. For example, if $\alpha = 0.05$ and $\alpha^* = 0.1$, for a sample size of $n_1 = n_2 = 12$, then the negligibility limit of the scaled carryover is $\theta_0 = 0.3742$ and $\tilde{\theta}_0 = 0.3633$.

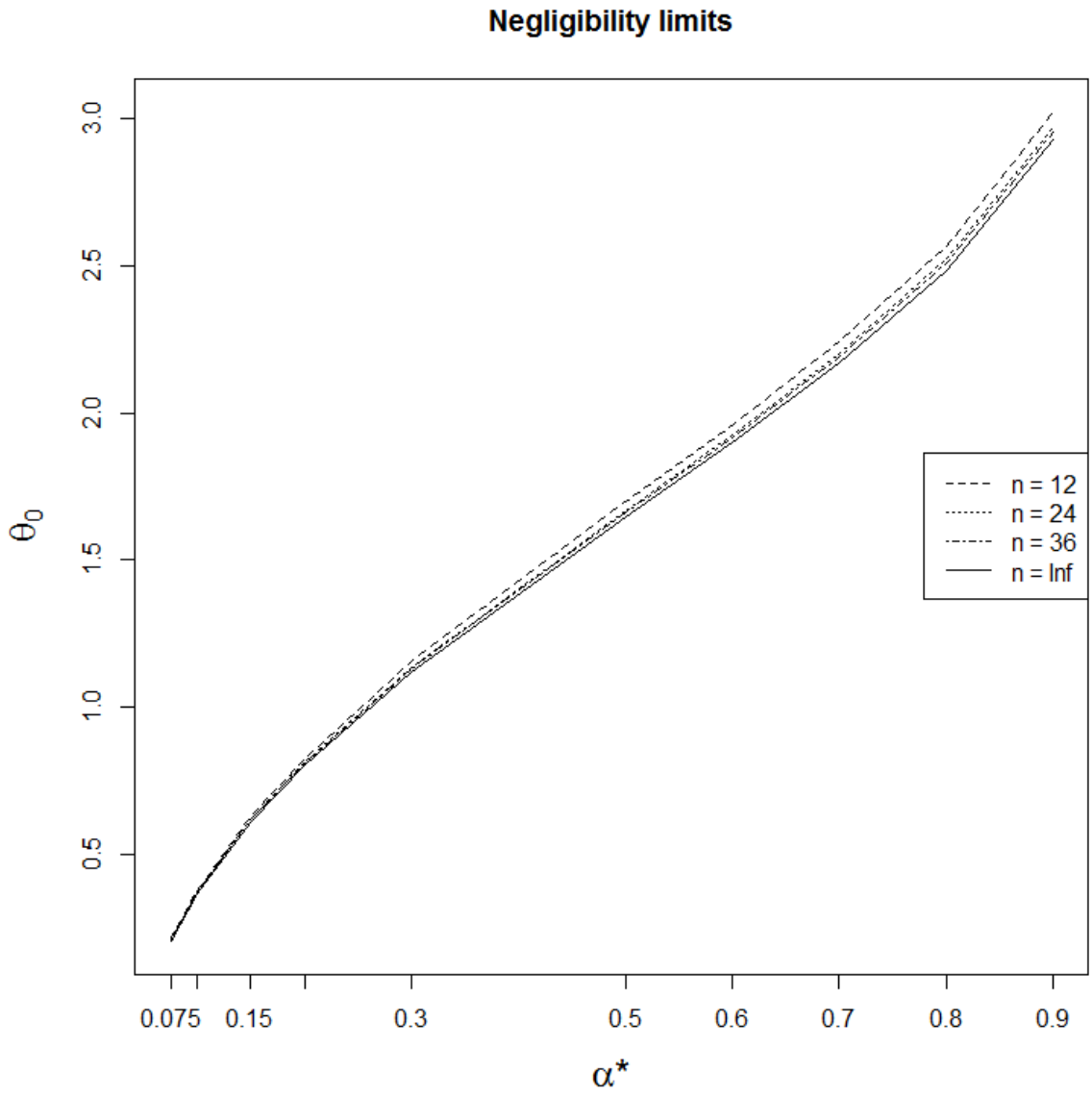

Figure 1. **Negligibility limit for scaled carryover**, $\theta = (\kappa/\sigma)\sqrt{n_1 n_2/(2N)}$, in function of the maximum allowable type I error, α^*, for diverse sample sizes ($n = n_1 = n_2$ corresponds to each sequence size in a balanced design, n = Inf corresponds to the approximation (13)).

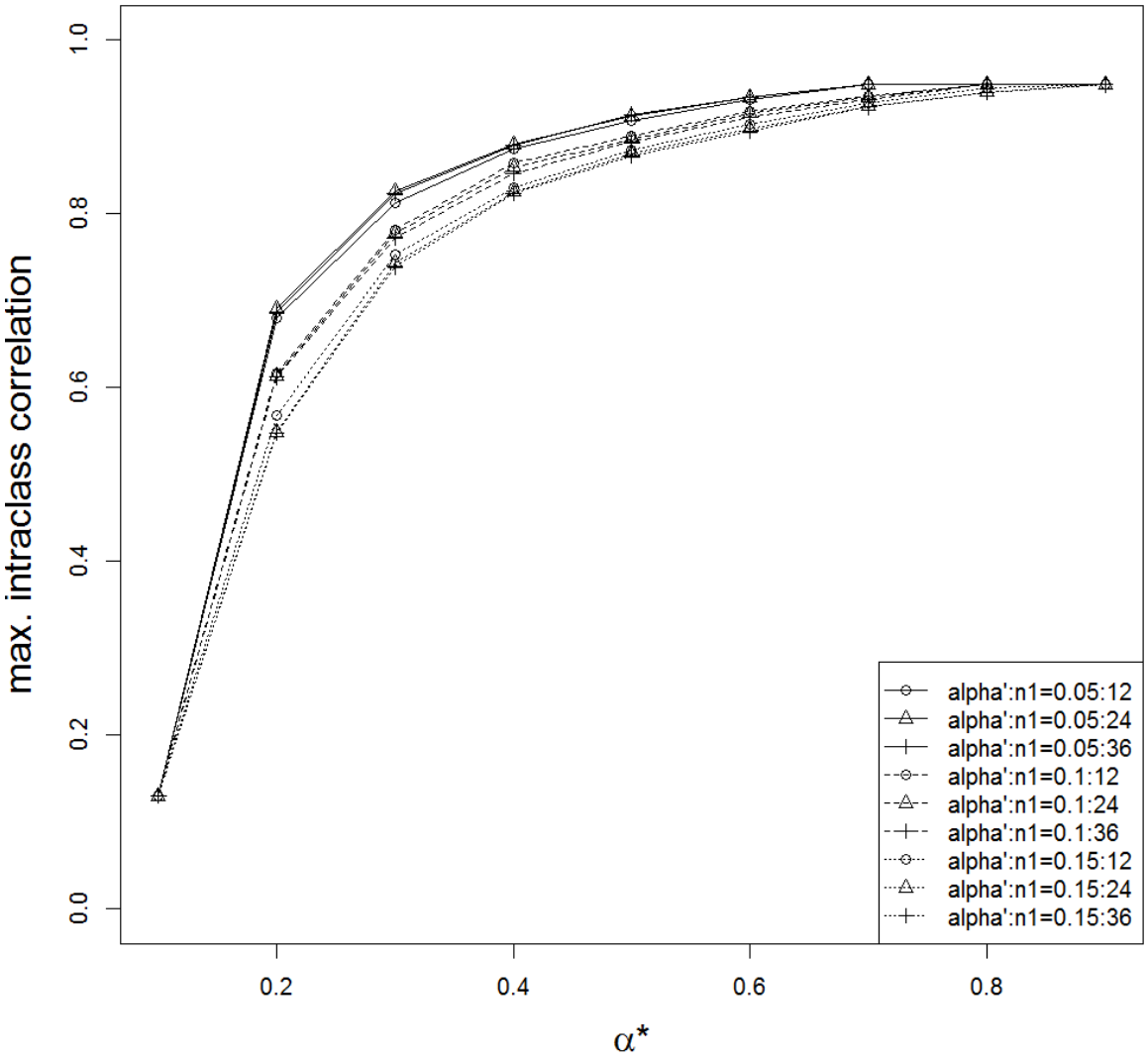

Figure 2. **Maximum allowable intraclass correlation to ensure approximate validity of the test for carryover relevance.** Too high intraclass correlations may seriously distort the test for carryover relevance. In order to assure a true type I error probability of the relevance test sufficiently closer (±20%) to a nominal size $\alpha' = 0.05$, in a balanced 2×2 design for sample sizes n = 12, 24 and 36, the intraclass correlation should not be greater than these values.

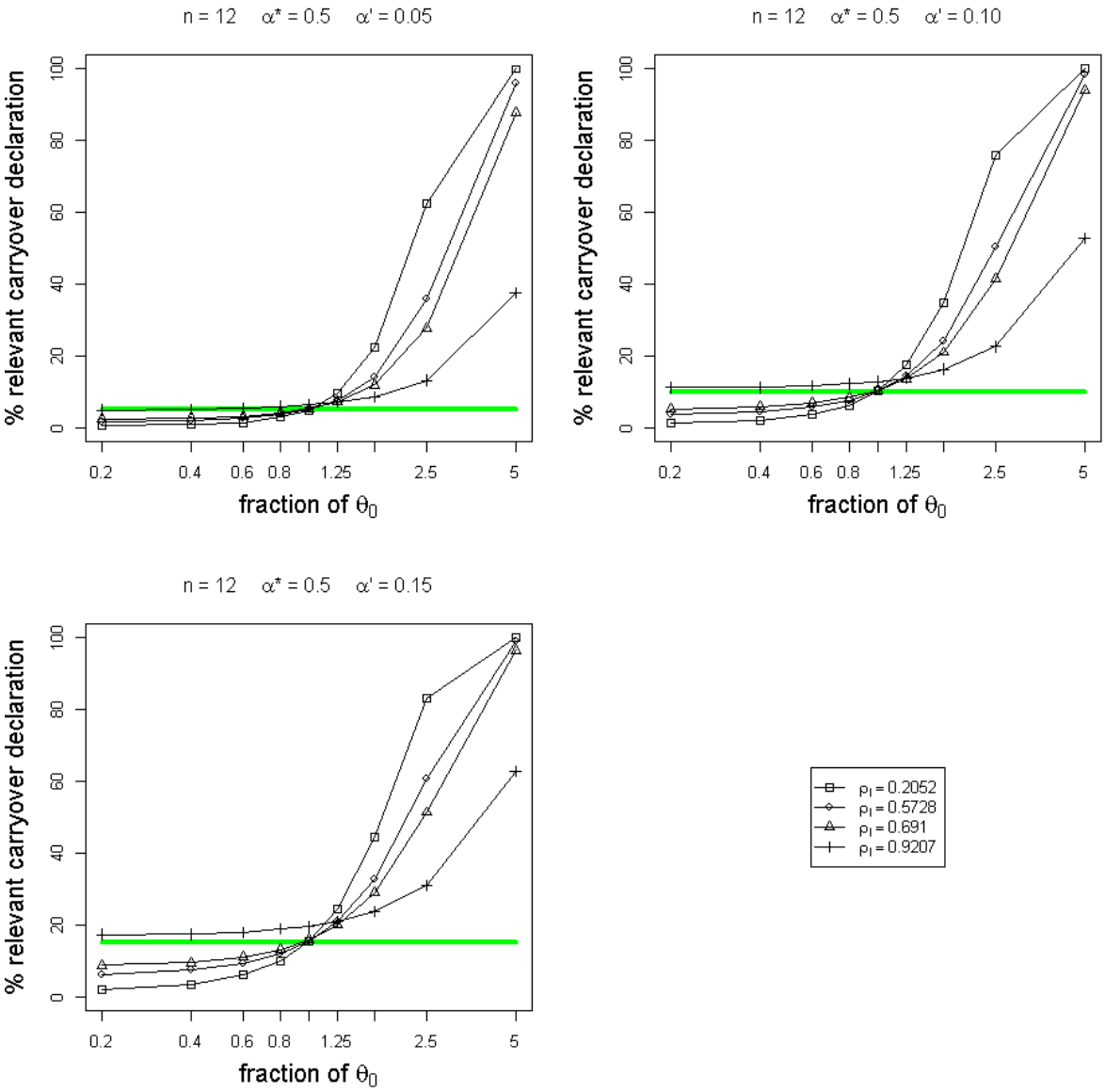

Figure 3. **Probability of declaring relevant carryover.** When "relevance" is set at $\alpha^* = 0.50$ and the test is performed at three possible significance levels, $\alpha' = 0.05,\ 0.1$ or 0.15, for a balanced sample size $n = 12$ in a 2×2 crossover design. The probability is displayed in function of the carryover relevance parameter θ defined in (13), in terms of a fraction of the relevance limit $\theta_0(\alpha^*), \theta / \theta_0$. Fraction values below 1 reflect non-relevant carryover degrees, fraction values above 1 reflect relevant carryovers, able to put the true user risk at a too high $\alpha^* = 0.50$ in a BE study. Each probability line corresponds to a given proportion between the "intra" and the "inter" subject variances, σ^2 and σ_S^2, expressed in terms of a given value of intraclass correlation, ρ_I.