SORT 32 (2) July-December 2008, 151-176

On equivalence and bioequivalence testing

Jordi Ocaña[1], M. Pilar Sánchez O.[2], Alex Sánchez[1] and Josep Lluís Carrasco[1]

[1] *Universitat de Barcelona and* [2] *Universidad de Chile*

Abstract

Equivalence testing is the natural approach to many statistical problems. First, its main application, bioequivalence testing, is reviewed. The basic concepts of bioequivalence testing (2×2 crossover designs, TOST, interval inclusion principle, etc.) and its problems (TOST biased character, the carry-over problem, etc.) are considered. Next, equivalence testing is discussed more generally. Some applications and methods are reviewed and the relation of equivalence testing and distance-based inference is highlighted. A new distance-based method to determine whether two gene lists are equivalent in terms of their annotations in the Gene Ontology illustrates these ideas. We end with a general discussion and some suggestions for future research.

MSC: 62F03, 62P10, 62F12, 62F15.

Keywords: crossover designs, TOST, intersection-union, distance-based inference, validation of simulation models, Gene Ontology.

1 Introduction and motivation

Consider the following situation: an experimenter has obtained some data under each of two distinct experimental conditions (e.g. control and treated patients), with the objective of demonstrating the *non-existence* (or more properly, the *near non-existence*) of differences between the two experimental conditions –e.g. the absence of an adverse drug reaction. In the subsequent statistical analysis, a known (but still frequent in practice) error is to test a null hypothesis stating equality (e.g. the corresponding means are equal) vs. an alternative hypothesis stating the existence of differences.

[1] Universitat de Barcelona.
[2] Universidad de Chile.

Received: February 2008
Accepted: May 2008

Suppose that the above test has been properly applied. If the computed p-value is greater than the previously stated significance level (e.g. a p-value of 0.12 when the significance level was 0.05), the null hypothesis will not be rejected. As is well known (but not always taken into account in practice), non-rejection of the null hypothesis is not "proof" of its validity. In such a situation, the p-value may be accompanied by some post-experiment or post-hoc power calculations, in order to give more "credibility" to the null hypothesis. Such observed power computations consist of calculating the probability of rejecting the null hypothesis under a distributional setting compatible with the test assumptions, and under parameter values defined by appropriate summary statistics obtained from the data. For example, under the typical Student's t-test for comparing the means of two separate groups, the post-experiment power calculations will assume normality and a nature state compatible with the alternative hypothesis, characterised by the estimated difference of means and the value of the pooled variance estimate. The assumed rationale of these power calculations is that a high observed power (e.g. greater than 0.9 or 0.95) "reinforces the credibility" of the null hypothesis, which could not be rejected "even" under such high power or low type II error probability. In fact, observed power calculations do not have any evidential value: see, for example, Hoenig and Heisey (2001).

It might also be possible to find a situation where the p-value leads to significant results with a very high power, possibly due to a very large sample size. The null hypothesis of effect *non-existence* would, therefore, be rejected, even if such an effect was negligible, i.e. the effect is statistically but not practically (e.g. clinically) significant.

More tenable approaches are based on improving the evidential value of p-values, for example using some sort of p-value calibration (preferably under a Bayesian point of view), as in Sellke *et al.* (2001) or in Girón *et al.* (2006), or, in a fully Bayesian setting, using Bayes factors and posterior probabilities, as in Moreno and Girón (2006). Under a frequentist approach, the best policy would be to recognise the inherent asymmetry of the risks associated with both hypotheses, and to invert their roles. If the end goal is to "demonstrate" the non-existence of effect (or more generally, of differences), a more dependable approach would be to establish an alternative hypothesis of "near equality" (not of strict equality) *versus* a complementary null hypothesis of "sufficiently large difference". This approach, where the alternative hypothesis defines the effect *non-existence* as equivalence of parameters rather than strict equality, is taken in equivalence testing. It is also compatible with a Bayesian point of view.

The rest of the paper is organised as follows. The second section is devoted to bioequivalence (BE) testing, by far the most common equivalence situation. This will provide a reference case sufficient to establish the main ideas and problems. The third section is devoted to evaluating the potentialities of this approach in more general terms, to establish their relations with a distance-based approach to statistics and to illustrate these ideas with a new distance-based equivalence test in bioinformatics. The last section brings together these ideas in a final discussion.

2 Bioequivalence testing

2.1 Statement of the bioequivalence problem

When the patent period of a drug is going to expire, the company that developed the brand-name "innovator" product based on this drug may try to develop a new formulation or dosage form with the same active ingredient, in order to extend its market exclusivity. Concurrently, other companies may try to develop generic forms based on the same active principle as the innovator product. To obtain approval of these alternatives, most regulatory agencies, including the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMEA), require proof of "equivalence of average bioavailabilities" or average bioequivalence (ABE), between the brand (innovator) product, commonly referred to as the "reference" (R) product, and the new dosage form or generic copy, commonly referred to as the "test" (T) product. An equivalence trial is much less expensive and easier to perform than a clinical trial for the development of a brand new drug. The former is based on sample sizes usually of 24 to 36 healthy volunteers as experimental subjects, in comparison with the sample sizes of thousands of patients commonly required in the latter.

The concept of bioavailability refers to the rate and extent by which the drug is available at its site of action. This is a complex and multidimensional concept. Quantitatively, it is expressed by several measures obtained from the curve of the concentration of drug in blood or plasma *versus* time, observed in each subject after a single-dose administration. The main bioavailability measures are t_{max}, the time until the maximum concentration is reached, C_{max}, the maximum concentration, AUC_{0-t}, the area under the curve from the dose administration to the last observation time, and $AUC_{0-\infty}$, the area under the curve until infinity. The underlying assumption in the requirement of equivalence of average bioavailability is that, as the drug or active principle is the same in all formulations under comparison, its therapeutic effect depends mainly on its concentration at the site of action, which should be similar for all products.

As summarised in Chow and Liu (2000), several criteria of ABE have been used since the 1970s, according to several regulatory recommendations. These include what are known as the 80/20 rule, the 75/75 rule, the ±20% rule and the (currently most widely used) 80%/125% rule.

The 80/20 rule states that, to declare bioequivalence, these two conditions must be fulfilled:

1. The test and reference means should not be statistically significantly different (commonly at a 5% level).
2. There should be at least an 80% power of detecting differences if the true difference were at least as large as 20% of the *observed* reference average.

Note that condition (ii) is an example of "observed power" computation.

The 75/75 states that at least 75% of the subjects must show a bioavailability value for the new formulation that is at least 75% of the corresponding bioavailability measurement for the reference formulation.

The ±20% rule concludes bioequivalence if the mean bioavailability of the test formulation, μ_T, is within ±20% of the mean of the reference formulation, μ_R, i.e., in terms of a ratio of means, if $0.8 < \mu_T / \mu_R < 1.20$.

Most regulatory agencies (e.g. CDER, 2001) recommend making all analyses for log-transformed data. The 80%/125% rule adapts the preceding criterion to analyses made at a logarithmic scale and, at the same time, enables bioequivalence to be stated in terms of a difference rather than a ratio. If, at the original bioavailability scale, a geometric means ratio between 0.8 and 1.25 is admissible, i.e. $0.8 < m_T / m_R < 1.25 = 1 / 0.8$, assuming that the means of the log-transformed variables correspond to the log-transformed geometric means, this inequality becomes $-0.22314 = \log(0.8) < \mu_T - \mu_R < \log(1.25) = +0.22314$. This is the basis of the ±0.223 rule on the logarithmic scale, equivalent to the 80%/125% rule on the original scale.

The ±20% and the 80%/125% criteria are used in conjunction with inferential procedures to ensure type I and type II error control. The first one requires inferential methods on the ratio of means; and the second one, on the difference.

Metzler (1974) was possibly the first author to recognise the inadequacy of the classical testing approach in bioavailability studies and the need for an equivalence approach, though the need for such an approach in a more general context can be traced back to Lehmann (1959). The reviews by Senn (2001) and Zapater and Horga (1999) are to some extent complementary to the present paper. In the next subsections we review bioequivalence in its most common setting: under a fixed sample size crossover design and for normal log-transformed data.

2.2 Average Bioequivalence. Design and basic statistical analysis

The commonest experimental design in bioequivalence studies is a 2×2 crossover design. In it, each experimental subject receives a single dose of both formulations, *R* and *T*, in one and only one of two possible orders or treatment sequences, *RT* or *TR*. There is always a "washout period" between dose administrations, in order to avoid "carry-over" effects, a possible influence or interaction of the first dose on the second. 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 on the logarithmic scale*, say $Y = \log C_{max}$ or $Y = \log AUC_{0-t}$, 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 $k = 1,2$.

With slight variants, all authors follow the linear model and basic analysis for 2×2 crossover trials proposed by Grizzle (1965). We consider the following underlying linear

model:

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

where μ is an overall 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 $C_{(j-1,k)}$ corresponds to the fixed effect of carry-over. It can only occur during the second period.

The possible carry-over 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:

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

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

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

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 period effect as $\pi = P_2 - P_1 = -2P$ and the carry-over effect as $\kappa = C_T - C_R = -2C$.

$S_{i(k)} \sim N(0, \sigma_S^2)$ represents the random effect of the i-th subject nested in the k-th sequence and $e_{ijk} \sim N(0, \sigma^2_{\tau(j,k)})$ is the random error or residual, or disturbance term. Additionally, we assume independence between all $S_{i(k)}$ and all e_{ijk}, and mutual independence between $\{S_{i(k)}\}$ and $\{e_{ijk}\}$.

Subindex $\tau(j,k)$ in the residual variance indicates a possible dependence on the experimental conditions. Obviously one possibility is constant variance, $\sigma^2 = \sigma^2_{\tau(j,k)}$. We will assume a slightly more general model, with possible dependence on the administered formulation:

$$\sigma^2_{\tau(j,k)} = \begin{cases} \sigma_R^2 \text{ if } j = k \\ \sigma_T^2 \text{ if } j \neq k. \end{cases} \tag{2}$$

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:

$$\begin{aligned} E(d_{ik}) &= \begin{cases} \frac{1}{2}(\pi + \phi + C_R) \text{ if } k = 1 \\ \frac{1}{2}(\pi - \phi + C_T) \text{ if } k = 2 \end{cases} \\ var(d_{ik}) &= \tfrac{1}{4}(\sigma_R^2 + \sigma_T^2). \end{aligned} \tag{3}$$

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

$$\bar{D} = \bar{d}_1 - \bar{d}_2 \tag{4}$$

is an unbiased estimate of the formulation effect ϕ, provided that no carry-over is present, *i.e.* if $\kappa = 0$.

The standard error of $\bar{D}$ can be independently estimated by

$$\widehat{se}_{\bar{D}} = \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \hat{\sigma}_d \sqrt{\frac{N}{n_1 n_2}} \tag{5}$$

where

$$\hat{\sigma}_d^2 = \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left(d_{ik} - \bar{d}_k \right)^2 \tag{6}$$

estimates the variance in (3). Alternatively, one may consider that $\hat{\sigma}_d^2$ corresponds to half the ANOVA estimate of the disturbance terms variance:

$$\sigma^2 = \frac{1}{2} \left(\sigma_R^2 + \sigma_T^2 \right), \tag{7}$$

sometimes denoted $\hat{\sigma}_{Res}^2$(for "residual") or $\hat{\sigma}_W^2$ (for "within" subjects).

An alternative way of defining $\bar{D}$ is based on the "least squares" means of the test and reference formulation:

$$\bar{Y}_R = \frac{1}{2} \left(\bar{Y}_{\cdot 11} + \bar{Y}_{\cdot 22} \right) \quad \text{and} \quad \bar{Y}_T = \frac{1}{2} \left(\bar{Y}_{\cdot 21} + \bar{Y}_{\cdot 12} \right),$$

where $\bar{Y}_{\cdot jk} = (1/n_k) \sum_{i=1}^{n_k} Y_{ijk}$ is the average of all observations in the j-th period and k-th sequence. Its difference coincides with (4):

$$\bar{D} = \bar{Y}_T - \bar{Y}_R. \tag{8}$$

The establishment of ABE is stated in terms of an equivalence test for the formulation effect ϕ:

$$\begin{aligned} H_0 &: \phi \leq \theta_1 \quad \text{and} \quad \phi \geq \theta_2 \\ H_1 &: \theta_1 < \phi < \theta_2 \end{aligned} \tag{9}$$

where normally $-\theta_1 = \theta_2 = \theta = 0.223$. In the following, if nothing more is specified, we will assume symmetrical equivalence limits, $\pm\theta = \pm 0.223$ for data on the logarithmic scale.

Schuirmann (1987) suggested decomposing the above hypothesis testing problem in two one-tail hypothesis testing problems:

$$\begin{array}{ll} H_{01} : \phi \leq \theta_1 & \\ H_{11} : \phi > \theta_1 & \end{array} \quad \text{and} \quad \begin{array}{ll} H_{02} : \phi \geq \theta_2 & \\ H_{12} : \phi < \theta_2 & \end{array} \tag{10}$$

and to conclude ABE, if and only if both H_{01} and H_{02} were rejected at a chosen α nominal level of significance (e.g. 0.05). The one-sided tests are easily implemented, since the statistic

$$T = \frac{\bar{D} - \phi}{\widehat{se}_{\bar{D}}} \tag{11}$$

follows a Student's central distribution, with $N - 2$ degrees of freedom. This provides an α level test, as a direct consequence of the intersection-union principle: see Berger (1982) and Berger and Hsu (1996).

The above procedure, known as the Two One-Sided Test (TOST) procedure, is operationally equivalent to the "confidence interval inclusion principle", say, to declare ABE if the usual $1 - 2\alpha$ shortest confidence interval:

$$\bar{D} \pm t_{(\alpha,N-2)}\ \widehat{se}_{\bar{D}} \tag{12}$$

where $t_{(\alpha,N-2)}$ is the $1 - \alpha$ quantile of a Student's t distribution with $N - 2$ degrees of freedom, is fully included in the bioequivalence limits, $[\theta_1, \theta_2]$. This principle was first pointed out by Westlake (1972). See Wellek (2003) for a discussion in more general terms. Declaring ABE if the 90 % interval (12) for log-transformed data is included between the limits ±0.233 is the current methodological mainstream in bioequivalence studies.

The use of a $1 - 2\alpha$ interval for a test of size α may seem counter-intuitive. As is shown in Munk and Pflüger (1999) in more general terms, this relation between confidence and test size requires the fulfilment of two conditions: convexity of the parametric region associated with the alternative hypothesis and an equivariance property of the confidence interval. If $I_{1-2\alpha}$ stands for a $1–2\alpha$ confidence interval, particularising the equivariance condition to the inference problem considered here, it may be stated as $I_{1-2\alpha}\left(d_\phi\left(\bar{D}\right), \widehat{se}_{\bar{D}}\right) = 2\phi - I_{1-2\alpha}\left(\bar{D}, \widehat{se}_{\bar{D}}\right)$ with respect to the transformation $d_\phi(x) = 2\phi - x$. This equivariance condition is fulfilled by (12) but relaxing this requirement in other confidence intervals leads to $1 - \alpha$ confidence intervals associated to α size tests. This is the case for the three confidence intervals described below.

Westlake (1976), from controversial considerations on the need of symmetry for any bioequivalence decision rule, introduced the following confidence interval:

$$\left[\bar{D} - t_2\ \widehat{se}_{\bar{D}}, \quad \bar{D} - t_1\ \widehat{se}_{\bar{D}}\right] \tag{13}$$

where t_1 and t_2 must satisfy the equations

$$\begin{aligned} &\Pr\{t_1 < T < t_2\} = 1 - \alpha \\ &(t_1 + t_2)\,\widehat{se}_{\bar{D}} = 2\bar{D} \end{aligned} \tag{14}$$

in order to define a symmetric around-zero interval, with 100 % coverage if the true formulation effect (in the logarithmic scale) is null, $\phi = 0$, and coverage tends to $1 - \alpha$ as ϕ tends to infinity. It ensures a bioequivalence test of size α when the interval inclusion rule is applied. The computation of (13) needs a trial-and-error iteration.

Hsu *et al.* (1994) introduced the following intervals not requiring any trial and error step, also with confidence level $1 - \alpha$ and associated with bioequivalence tests of size α:

$$\pm\left(\left|\bar{D}\right| + t_{(\alpha, n_1+n_2-2)}\widehat{se}_{\bar{D}}\right) \tag{15}$$

and

$$\left[\min\left(0,\ \bar{D} - t_{(\alpha, n_1+n_2-2)}\widehat{se}_{\bar{D}}\right),\quad \max\left(0,\ \bar{D} + t_{(\alpha, n_1+n_2-2)}\widehat{se}_{\bar{D}}\right)\right]. \tag{16}$$

Interval (15) is symmetrical and both have asymptotic confidence level $1 - \alpha$, and 100 % coverage if $\phi = 0$. By construction, there is an inclusion relation in the sense of (16) $\subset$ (15) $\subset$ (13). Thus, from (13) to (16), the power of the corresponding α level tests is not decreasing, and possibly increases.

The properties of the above intervals, and their relation to α level tests, are summarized in Chow and Shao (2002).

Rodda and Davis (1980) interpreted the confidence interval inclusion principle from a Bayes point of view. Under model (1) and no carry-over effect, the statistics $\bar{d}_1$, $\bar{d}_2$ and $(N-2)\,\hat{\sigma}_d^2$ are independently distributed, $\bar{d}_k \sim N\left(\xi_k, \frac{\sigma^2}{2}\right)$ with $\xi_1 = 0.5\,(\pi + \phi)$, $\xi_2 = 0.5\,(\pi - \phi)$ and $(N-2)\,\hat{\sigma}_d^2 \sim \frac{\sigma^2}{2}\chi^2_{N-2}$. Assuming independent and locally uniform non-informative priors for ξ_1, ξ_2 and σ^2, it is finally found that the posterior distribution of

$$\frac{\phi - \bar{D}}{\widehat{se}_{\bar{D}}} \tag{17}$$

is a central Student's t with $N - 2$ degrees of freedom. This allows a probabilistic interpretation in terms of credible intervals. For example, the $1 - 2\alpha$ highest density interval is computationally identical to the shortest confidence interval (12). But now, declaring bioequivalence when it is included within the bioequivalence limits $\pm\theta$ may be interpreted as imposing the condition that the posterior probability of $-\theta < \phi < +\theta$ must be no less than $1 - 2\alpha$.

2.3 An ABE study example

We illustrate the preceding basic bioequivalence analyses with the results of Al Mohizea *et al.* (2007), a bioequivalence study on two forms (the new form gemifloxacin 320 mg/tablet vs. the reference form factive 320 mg/tablet) of the antibiotic Gemifloxacin. The study was performed in 24 healthy volunteers under a 2×2 crossover design.

		AUC_{0-t}		$AUC_{0-\infty}$		C_{max}	
		Interval limits					
		Lower	Upper	Lower	Upper	Lower	Upper
Interval type	"Shortest" (12)	87.48	107.83	88.72	108.19	92.08	113.47
	Westlake (13)	87.12	114.79	88.15	113.44	87.55	114.22
	Symmetric (15)	87.48	114.32	88.71	112.72	88.13	113.47
	"Optimal" (16)	87.48	107.83	88.71	108.19	92.08	113.47

The $\bar{D}$ values for log AUC_{0-t}, log $AUC_{0-\infty}$ and log C_{max} were –0.0292, –0.0205 and 0.0220, respectively. The standard errors $\widehat{se}_{\bar{D}}$ for the same pharmacokinetic measures were 0.0609, 0.0578 and 0.0608. These results lead to the following standard "shortest" 90 % confidence interval for log AUC_{0-t}:

$$\bar{D} \pm t_{(\alpha,N-2)}\ \widehat{se}_{\bar{D}} = -0.0292\ \pm\ 1.7171\ \times\ 0.0609 = [-0.1338,\ 0.0754]$$

which leads to the following confidence interval for the ratio in the original scale: $[\exp(-0.1338), \exp(0.0754)] = [0.8748, 1.0783]$ or [87.48, 107.83] in percentage terms.

As this interval is included in the bioequivalence limits [80, 125], bioequivalence should be declared. Similarly, the confidence intervals for the other variables are indicated in the table above, with a similar conclusion of bioequivalence.

Equivalently, the p-value for the upper and lower Schuirmann's TOST is less than 0.0001 in all cases.

The limits of the alternative confidence intervals discussed in the previous section are also shown in the table. In all cases bioequivalence should be declared. Note that to compute the 95 % Westlake confidence interval, we first need the limits satisfying equations (14). For example, these values are $t_1 = -2.7442$ and $t_2 = 1.7845$ for AUC_{0-t}.

2.4 The power of the TOST procedure and scaling methods

The power of the TOST test, or its interval inclusion equivalent, can be computed as:

$$\beta(\phi,\sigma) = \int_0^{\nu(\theta/\sigma;n_1,n_2,\alpha)} \left[\Phi\left(\sqrt{\frac{n_1 n_2}{N}}\frac{(\theta-\phi)}{\sigma} - t_{(\alpha,N-2)}\nu\right) - \Phi\left(\sqrt{\frac{n_1 n_2}{N}}\frac{(\theta+\phi)}{\sigma} - t_{(\alpha,N-2)}\nu\right)\right] \sqrt{N-2}\ g_\chi\left(\sqrt{N-2}\nu\right) d\nu \tag{18}$$

where g_χ stands for the χ-distribution with $N - 2$ degrees of freedom and

$$\nu\left(\theta/\sigma; n_1, n_2, \alpha\right) = \sqrt{\frac{n_1 n_2}{N}}\frac{\theta}{\sigma}$$

(see Wellek, 2003, p. 211).

The most obvious consequence of (18) is that, for a fixed sample size, $\beta(\phi, \sigma) \to 0$ as $\sigma \to \infty$, for any value of the formulation effect, ϕ. This means that there are alternatives with $\beta(\phi,\ \sigma) < \alpha$. This biased character of the TOST procedure has great practical importance in "high-variability" (HV) drugs or drug products. These are products containing drugs of poor pharmaceutical quality as a cause of their high variability.

A drug is assumed to be HV when the observed coefficient of variation CV (on the original scale) associated with the ANOVA estimate $\hat{\sigma}^2_{Res} = 2\hat{\sigma}^2_d$ exceeds 30 % (Blume and Midha, 1993). Sometimes this threshold is put at 25 %. The coefficient of variation on the original scale is related to variance on the logarithmic scale by means of the relation:

$$CV\left(\sigma^2\right) = \sqrt{\exp\left(\sigma^2\right) - 1}. \tag{19}$$

A general discussion on HV drugs analysis can be found in Shah *et al.* (1996). The main problem with HV drugs is the low power of the TOST procedure when used with the usual sample sizes in bioequivalence trials: at most, a few dozen subjects. Bioequivalence trials with hundreds or even thousands of individuals, which would be required for some HV drugs, are generally considered unfeasible.

Anderson and Hauck (1983) proposed a procedure that is more powerful than TOST, but does not adequately control Type I error probability. The test of Berger and Hsu (1996) is nearly unbiased and uniformly more powerful than TOST, but it is not widely used in practice, possibly due to its (moderate) complexity and because the rejection region includes values of $\bar{D}$ outside the limits $\pm\theta$ for large values of $\widehat{se}_{\bar{D}}$. This counter-intuitive character was pointed out by Schuirmann in the discussion accompanying Berger and Hsu (1996), and questioned in Perlman and Wu (1999) in a well-founded argument. This latter paper (which adopts a Fisherian perspective) seems to continue the debate between Fisher and Neyman (see Barnett, 1999). In fact, the approaches that are commonly taken in practice rely on widening in some way the alternative hypothesis (i.e., the bioequivalence limits), which may also seem arbitrary.

Widening the bioequivalence limits to new fixed values has been regulated by the FDA (70 %/143 % or ±0.3567 in logarithmic scale), which assumed the proposals in Shah *et al.* (1996), and by the EMEA (75 %/133 % or ±0.2877) in CPMP (2001). These proposals mainly refer to C_{max}, the bioavailability measure most frequently found to be HV. Clearly, these enlargements do not solve in a general way the power problems of the ABE testing procedures.

Boddy *et al.* (1995) suggested linearly scaling the bioequivalence limits in function of variability, jointly with deciding bioequivalence in the usual way, according to the confidence interval inclusion principle based on the classical shortest interval. Under this setting, the bioequivalence limits (*BEL*) become $\mp k\sigma_{SC}$ instead of a fixed quantity, $\pm\theta$. In 2×2 crossover studies, the most reasonable choice for the scaling variance σ^2_{SC} is the residual variance σ^2. As it is unknown, it must be replaced by an appropriate estimate. Then, the scaled bioequivalence limits become a random function of data:

$$BELsc = \mp k\hat{\sigma}_{SC}. \tag{20}$$

There are some possibly reasonable choices for the constant k (1.116 in CDER, 2003; 1.0 in Boddy *et al.*, 1995; and 0.759 in Tothfalusi and Endrenyi, 2003), but in any case the choice is somewhat arbitrary. The selection of the constant k should be drug-specific and the responsibility of regulatory agencies. Additionally, for a sufficiently large estimated variance, bioequivalence will be declared for $\bar{D}$ values far from the usual bioequivalence limits, a similar criticism to the one about the Berger and Hsu (1996) method. To try to mitigate these drawbacks, families of more flexible scaled limits were developed. Technical details, with an illustrative example, can be found in http://hdl.handle.net/2072/5456.

Apart from their possible arbitrariness, all these bioequivalence limit functions share the same problem: the size and in general the statistical properties of the decision criteria based on them are not guaranteed, as they are not based on known and constant bioequivalence limits, but on limits that are random functions of data, and no additional theoretical support is provided. As is done with individual and population bioequivalence (see below), the bootstrap method gives a possible approach, but this possibility has still not been sufficiently explored.

A more well-founded approach is to make equivalency inferences on scaled parameters for fixed limits, rather than to scale the equivalency limits. In other words, one may restate the problem as that of establishing bioequivalence from fixed limits using a scaled metric ϕ/σ_{sc}. Again, under 2×2 crossover designs, the most natural choice for scaling variability is the residual variance (7). Then $\sigma_{SC} = \sigma$ and equivalence is stated as:

$$-k < \frac{\phi}{\sigma} < +k. \tag{21}$$

An adequate criterion would be to base the final decision on an appropriate confidence interval or test procedure for this scaled parameter. A direct approach is to use the fact that, on rescaling the previous inequality as

$$-\tilde{k} < \tilde{\phi} < +\tilde{k} \tag{22}$$

with $\tilde{\phi} = (\phi/\sigma)\sqrt{2n_1n_2/N}$ and $\tilde{k} = k\sqrt{2n_1n_2/N}$, the statistic $T = \bar{D}\big/\widehat{se}_D$ has a non-central Student's t distribution with $N - 2$ degrees of freedom and a non-centrality parameter $(\phi/\sigma)\sqrt{2n_1n_2/N}$. This defines the following $1 - 2\alpha$ confidence interval:

$$t_\alpha(\lambda, N-2) \leq \tilde{\phi} \leq t_{1-\alpha}(\lambda, N-2) \tag{23}$$

where $t_\alpha(\lambda, N-2)$ corresponds to the α quantile of a non-central Student's t distribution with $N - 2$ degrees of freedom and non-centrality parameter T. Bioequivalence is declared if the above confidence interval lies within the limits $\pm\tilde{k} = \pm k\sqrt{2n_1n_2/N}$. Obviously, the choice of k still remains arbitrary; as before. Some reasonable choices may be 1.116, 1 or 0.759.

The above interval inclusion procedure is equivalent to a testing procedure with rejection region of general form

$$\{c_1 < T < c_2\}$$

which is optimal (in the sense of being most powerful unbiased) for a wide class of distributions, including the normal case discussed here, as is extensively shown in Wellek (2003).

The scaled procedure, despite its optimality, is not always accepted as the adequate approach to the bioequivalence problem, which is still primarily articulated in the scale of the means and not of the scaled means.

2.5 The carry-over controversy

As has been mentioned, under model (1), $\bar{D}$ is an unbiased estimator of the true formulation effect ϕ only in absence of carry-over effect.

The analysis of the carry-over effect is straightforward. In order to estimate it, we first form the sums inside each individual, $Y_{ij\cdot} = Y_{ij1} + Y_{ij2}$. Simple computations from model (1) lead to the following expressions:

$$\begin{aligned} &var\,(Y_{ij\cdot}) = 4\sigma_S^2 + \sigma_R^2 + \sigma_T^2 = \sigma_+^2 \\ &E\,(Y_{i1\cdot}) - E\,(Y_{i2\cdot}) = \kappa. \end{aligned} \tag{24}$$

Then, the difference:

$$\hat{\kappa} = \bar{Y}_{\cdot 1\cdot} - \bar{Y}_{\cdot 2\cdot} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{i1\cdot} - \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{i2\cdot} \tag{25}$$

is an unbiased estimator of the carry-over effect with standard error estimated by:

$$\widehat{se}_{\hat{\kappa}} = \sqrt{\frac{\sum_{i=1}^{n_1}\left(Y_{i1\cdot} - \bar{Y}_{\cdot 1\cdot}\right)^2 + \sum_{i=1}^{n_2}\left(Y_{i2\cdot} - \bar{Y}_{\cdot 2\cdot}\right)^2}{n_1 n_2 \left(N-2\right)/N}}. \tag{26}$$

According to standard results (e.g. Chow and Liu, 2000, pp. 60-61) the statistic:

$$\frac{\hat{\kappa} - \kappa}{\widehat{se}_{\hat{\kappa}}} \tag{27}$$

follows an Student's central t distribution with $N - 2$ degrees of freedom if $\kappa = 0$. Grizzle (1965) proposed testing this null hypothesis of non-existence of carry-over at a significance level of $\alpha = 0.1$, or even 0.15, in order to have enough power. In case of no rejection of the null hypothesis, he recommended proceeding with the standard analysis under no carry-over. Otherwise, the recommendation was to use only the data of the first period, like data obtained in a fully randomised parallel trial. This strategy is recommended by the FDA (CDER, 2001).

This two-stage procedure is widely used in practice, despite the criticisms of Brown (1980) in terms of cost and those of Senn (1988) and Freeman (1989) in terms of its inadequate test size and power. Additional arguments against the two-stage procedure are given in Senn (1996) and Senn *et al.* (2004).

When higher-order crossover studies are performed, another possibility is to adjust for the presence of carry-over, assuming specific models for it, as in Laird *et al.* (1992) and Putt and Chinchilli (1999). This approach is also discouraged in Senn (1992), Senn and Lambrou (1998) and Senn *et al.* (2004) in terms of the implausibility of assumptions on carry-over and the analysis' lack of robustness.

Both detractors to the two-stage procedure and opponents to adjusting for carry-over state that the best policy is to not previously test for carry-over (or not use this test to take any further decision on the analysis course) and to proceed as if it was absent. In well-performed experiments, carry-over will usually be absent as 'washout' will normally succeed in eliminating it. This opinion seems to be confirmed in D'Angelo et al. (2001) 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 carry-over. Moreover, for the subset of these 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 (Senn *et al.*, 2004). These results are contested in Putt (2005, 2006) with simulations that suggest the lack of power of these KS tests. Senn *et al.* (2005) rebut the arguments of Putt, arguing the irrelevance of power calculations to interpretation of their observational data.

Putt's analysis is not an observed power study (in the sense of our introductory section). However, as the author herself recognises, it does not prove that carry-over existed – as in Senn *et al.* (2004), results are not a proof of non-existence of carry-over.

These latter authors note the difference between significant and important (sufficient to distort the subsequent analysis) carry-over and the appropriateness of an equivalence approach. This last suggestion refers to their analyses of the possible uniformity of *p*-value data lists, but it would also be a more suitable approach to directly discarding the possible presence of sufficiently distorting carry-over in each particular crossover study.

To perform an equivalence test for *scaled* carry-over:

$$H_0 : \frac{|\kappa|}{\sigma_+} \geq \varepsilon \qquad \text{vs.} \qquad H_1 : \frac{|\kappa|}{\sigma_+} < \varepsilon \tag{28}$$

Wellek (2003, pp. 196-203) uses the fact that the statistic:

$$T = \frac{\hat{\kappa}}{\widehat{se}_{\hat{\kappa}}} \tag{29}$$

follows a non-central *t*-distribution with $N - 2$ degrees of freedom and non-centrality parameter $\varepsilon\sqrt{n_1 n_2/N}$ in the boundaries of the equivalence region. The *p*-value associated with an observed sample value, T_{obs}, may be computed as $\Pr\{F < T^2_{\text{obs}}\}$, where F stands for a random variable with non-central F distribution, with non-centrality parameter $\varepsilon^2 (n_1 n_2/N)$ and 1 and $N - 2$ degrees of freedom. Alternatively, if one wishes to test the absolute carry-over, κ, an approach like Schuirmann's test for bioequivalence can be performed with the "inside each subject sum" data $Y_{ij\cdot}$ and using results from (24) to (27). As the variance in (24) may be high, this last approach may result in a test with low power.

On the other hand, what seems to be the most interesting measurement of the possible disturbing effect of carry-over is its size in relation with the formulation effect, and not its absolute or relative value in relation with variability.

An interesting alternative to the two-stage procedure, although for the moment not sufficiently developed for practical use in bioequivalence testing, is the synthetic estimators of Longford (2001), a way to combine the with-carry-over and without-carry-over formulation effect estimators. Similarly, the Bayesian approach taken in Grieve (1985) avoids taking an all-or-nothing approach for possible carry-over. The usefulness and/or correctness of this approach is contested by some authors, but it is successfully used with real data, as shown by Racine *et al.* (1986).

2.6 Other approaches to BE

An obvious extension of the preceding techniques is to make them multivariate, i.e. to test simultaneously for all available bioavailability measures and not separately for each one. The basic approach to ABE, based on confidence interval inclusion, was generalised to the multivariate case by Wang *et al.* (1999). Ghosh and Gonen (2008) provide a semi-parametric Bayesian solution to ABE. Using Montecarlo Markov Chain

Methods (MCMC), these authors assume a realistic multivariate prior, with dependent parameters.

Pharmacokinetic measures like C_{max} or AUC are statistical summaries computed from the concentration-by-time curves, which are the true raw data in bioavailability experiments. Thus, another possible approach to bioequivalence is to directly analyse these curves, either as multivariate data or by any modelling approach that adequately describes the curves. In this context, a standard tool are mixed models, and an adequate approach would be to establish the equivalence of their parameters or, perhaps better, to compute confidence regions for the mean curves. However, up to the authors knowledge, equivalence testing in mixed models is a still unexplored field, with the exception of Rashid (2003).

Even though all these approaches to BE may be very interesting, the regulatory approach exclusively recommends testing pharmacokinetic parameters individually, which makes the multivariate approach rare in theoretical papers and nearly absent in practical work. Next we describe two (univariate) BE approaches that have merited regulatory consideration.

Complete or nearly complete similarity between the means does not imply equivalence between both formulations. If for example the bioavailability of the test formulation is much more variable than the bioavailability of the reference formulation ($\sigma_T^2 \gg \sigma_R^2$), replacing R by T will probably imply some user risks. The concept of "population bioequivalence" (PBE) refers to equivalence both in mean and in variability; and more generally, to equivalence in the general form of the distribution of the bioavailability variable. This concept tries to express the idea that a generic form is fully *prescribable* to a patient who initiates its treatment. However, even when the distributions under T or R are marginally equivalent, it is not guaranteed that R is *exchangeable* with T in a patient who started treatment with R. The concept of "individual bioequivalence" (IBE) tries to reflect this last concept of exchangeability within the same individual.

These concepts were introduced by Anderson and Hauck (1990) and formalised in Schall and Luus (1993). These authors suggested the following aggregate scaled measure of global dissimilarity to define PBE:

$$\frac{\phi^2 + \sigma_{totT}^2 - \sigma_{totR}^2}{\sigma_{totR}^2} \tag{30}$$

where

$$\begin{aligned} \sigma_{totR}^2 &= var\,(Y_{i11}) = var\,(Y_{i22}) = \sigma_S^2 + \sigma_R^2 \\ \sigma_{totT}^2 &= var\,(Y_{i12}) = var\,(Y_{i21}) = \sigma_S^2 + \sigma_T^2 \end{aligned} \tag{31}$$

are the "total" variances (that is, including both the between-subject variance, σ_S^2, and the residual variance) of the response under each treatment. The corresponding

moment-based measurement of individual bioequivalence uses the concept of subject-by-formulation interaction that requires higher-order crossover designs and will not be dealt with here.

Measure (30) combines, rather arbitrarily, a squared Euclidean distance, ϕ^2, with a difference of variances. The natural scaling factor for the first summand is residual variance (7) and not the total variance under the reference formulation. Given these and other difficulties with the above-mentioned concept of population bioequivalence, Schall (1995) proposed a criterion based on the probabilities of discrepancy between the responses under the test and the reference formulation, in relation to the same probability when both individuals receive the reference formulation. In a completely different approach, Wellek (2000) proposed a "disaggregate" test in the sense of separately testing for ϕ/σ and for $\sigma^2_{totT}/\sigma^2_{totR}$ and then combining both tests by means of the intersection-union principle.

In the FDA guidance CDER (2001), there are precise instructions for individual and population bioequivalence. But CDER (2003) seems to abandon the requirement of individual and population bioequivalence and to return to average bioequivalence exclusively, perhaps due to the difficulties in these concepts and in their implementation. Moreover, Senn (2001) points out that the concept of exchangeability of drugs is meaningless in clinical terms and only prescribability is useful when the clinician has to decide whether to prescribe a formulation.

3 Equivalence testing: a more general perspective

3.1 Some selected equivalence problems and applications

Bioequivalence is just one of the potential applications of the equivalence testing concept. There are many applications and potential applications of the equivalence concept, focusing either on statistical methodology or on specific fields of application.

Wellek (2003) reviews some common statistical problems that may be treated more adequately under an equivalence approach. These include comparing binomial variables, goodness of fit to a distribution, testing for homoscedasticity and testing for non-importance of interactions, i.e. for additivity in a linear model. Barker *et al.* (2001) perform an extensive (though not complete, as is pointed out by Martín Andrés and Herranz Tejedor, 2002) review of equivalence tests for binomial variables. Bayesian alternatives to some of these tests are discussed in Williamson (2007).

The following are some discussions of the applicability and/or concrete applications of equivalence testing in diverse areas: Stegner *et al.* (1996) in social sciences, Burns and Elswick (2001) in dental clinical trials, Barker *et al.* (2002) in epidemiology and Mecklin (2003) in educational research. Van Steen *et al.* (2005) propose an equivalence

procedure in DNA sequence comparison. This is an example of the distance-based approach to equivalence, to be treated in more detail in the next section.

A problem of central practical importance is simulation model validation. According to Sargent (2005), operational validation is "determining whether the simulation model's output behaviour has the accuracy required for the model's intended purpose over the domain of the model's intended applicability". If the modelled system is observable, the objective methods for validation are, essentially, two (or more) sample comparison methods: the data observed in the real system vs. the generated data experimenting with the model (i.e., simulating). Reynolds and Deaton (1982) and Kleijnen (1999) review hypothesis test methods for validation.

In contradiction with the preceding quoted definition (which in our opinion reflects pretty well the concept that simulation practitioners have in mind), the common approach to model validation states a null hypothesis of *exact* model validity. This strategy leads to severe methodological problems, illustrated by common recommendations (e.g. Sargent, 2005) of not using too large sample sizes (especially from the simulated data side), in order to avoid rejecting adequate (to the goals of the study) models. It is obvious that an equivalence approach would be much more dependable. The authors are not aware of any equivalence approach to simulation model validation, except Robinson and Froese (2004) and the ideas outlined in Warner (2002).

Most likely, in model validation a difficult problem will be to establish the equivalence limits, which may be very application area- and model-dependent. There are some regulations on how to construct and validate simulation models (e.g. http://cdds.ucsf.edu/research/sddgpreport.php is a best-practice document on simulation in drug development), but none of them considers the equivalence approach in depth.

3.2 Equivalence testing and distance-based Statistics

The great majority of the equivalence problems commented on above admit a distance-based representation, with general form:

$$H_0 : d(A, B) \geq d_0 \text{ vs. } H_1 : d(A, B) < d_0 \tag{32}$$

where d is a distance or dissimilarity index, A and B are two objects (distributions, models...) to be compared and d_0 is an equivalence limit. For example, admitting model (1) and in absence of carry-over and period effects, the distributions under T and R are, respectively,

$$A \equiv N\left(\mu_T = \mu + C_T; \sigma^2_{totT}\right) \text{ and } B \equiv N\left(\mu_R = \mu + C_R; \sigma^2_{totR}\right). \tag{33}$$

The ABE distance and criterion are $d(A, B) = |\mu_T - \mu_R| = |\phi| < d_0 (= \theta)$. The index d is a true distance measure under $\sigma^2_{totT} = \sigma^2_{totR}$.

PBE is based on the index:

$$d(A, B) = \frac{(\mu_T - \mu_R)^2 + \sigma^2_{totT} - \sigma^2_{totR}}{\sigma^2_{totR}}. \tag{34}$$

Note that (34) is not a metric distance, nor a reasonable dissimilarity measurement. For example, it is possible that $d(A, B) = 0$, when $\mu_T \neq \mu_R$ and $\sigma^2_{totT} \neq \sigma^2_{totR}$. Index (34) rewards a generic product with less variability than the brand product.

This distance-based approach is explicitly taken in Munk and Czado (1998), using a trimmed version of the pth Mallows distance between distributions:

$$\Gamma_{\alpha,p}(F, G) = (1 - 2\alpha)^{-1} \left\{ \int_{\alpha}^{1-\alpha} \left| F^{-1}(u) - G^{-1}(u) \right|^p du \right\}^{1/p} \quad \alpha \in \left[0, \tfrac{1}{2}\right), p \geq 1 \tag{35}$$

Their asymptotic results allow non-parametric goodness of fit testing, and average and population bioequivalence testing, in a unified way. One drawback of these tests is that their true size exceeds the nominal size, unless large sample sizes (much larger than is usual in bioequivalence testing) are employed.

Dragalin *et al.* (2003) use the squared Kullback-Leibler divergence, $d(f, g) = \Delta_1^2(f, g)$, where

$$\Delta_1(f, g) = \sqrt{I(f, g) + I(g, f)} \tag{36}$$

is the Jeffreys J-divergence based on the Kullback-Leibler information:

$$I(f, g) = E_f \left\{ \log \frac{f(X)}{g(X)} \right\} \tag{37}$$

for densities f and g. (36) is not a distance index, but has reasonable dissimilarity properties.

If f_T and f_R are the densities of (33), PBE is associated with the index:

$$d(f_T, f_R) = \frac{1}{2} \left\{ (\mu_T - \mu_R)^2 + \sigma^2_{totT} + \sigma^2_{totR} \right\} \left(\frac{1}{\sigma^2_{totT}} + \frac{1}{\sigma^2_{totR}} \right) - 2. \tag{38}$$

Equivalent results are also obtained for the exponential family of distributions and for the multivariate normal case. An obvious advantage of the distance approach is that the generalisation to the multivariate case is much more straightforward.

In the univariate normal case, when $\sigma^2_{totT} = \sigma^2_{totR} = \sigma^2_{tot}$, the preceding index defines a scaled BE criterion in relation to total variance, not in relation to residual variance, as in (21).

Approximate inference with the preceding indices is based on the interval inclusion approach. Bioequivalence is declared if the upper limit of the one-sided bootstrap

percentile interval for d falls below d_0. An advantage of using the Kullback-Leibler metric is that the FDA bioequivalence limits can be easily adapted to the corresponding d_0 values, because both are simple functions of the same moments.

3.3 Combining studies based on Gene Ontology

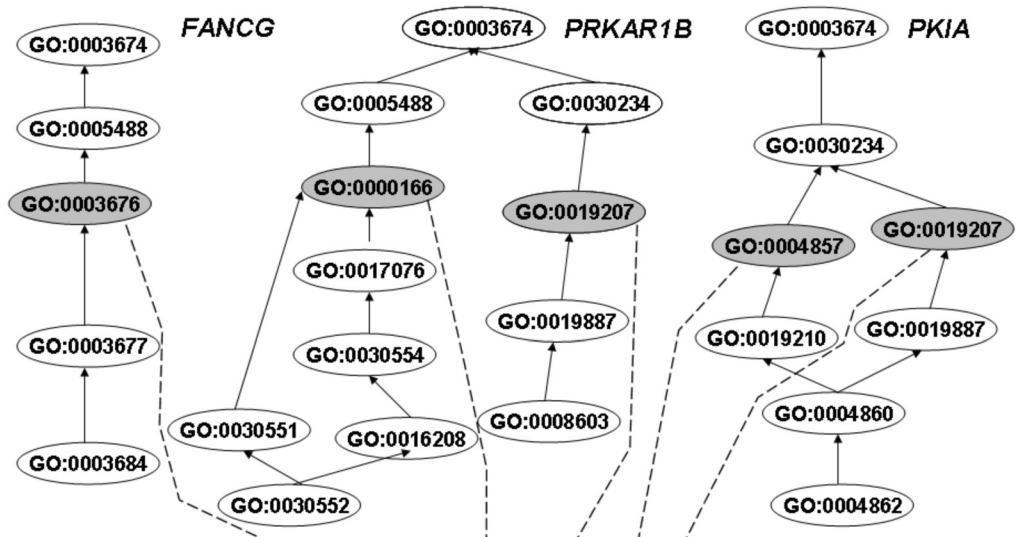
With the help of recently developed technologies like DNA microarrays, it is now possible to analyse the behaviour of thousands of genes in a single experiment. Gene Ontology (GO, www.geneontology.org) is an annotation database created and maintained by the Gene Ontology Consortium in order to systematise these huge amounts of quickly growing information. GO is organised in three basic ontologies: molecular function (MF), biological processes (BP) and cellular components (CC). Each of them can be viewed as directed acyclical graphs (DAG). The nodes in the DAG represent concepts that may help to characterise a gene (e.g. the biological processes in which it participates). The known information on a given gene is expressed as *annotations* or *hits* on one or more nodes in the GO. A way to summarise a given list of genes (e.g. those over-expressed in individuals suffering from a specific disease) is to determine its GO *profile* for a given level in one of the three GO ontologies. A level is the set of all nodes at the same distance from the origin of the ontology; it is like a cross-section in the rich DAG structure. A profile is the vector of annotation counts (or percentages or relative frequencies) in the s nodes of the chosen level, for all genes on the list: $\hat{P} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_s)$. To compute the relative frequency $\hat{p}_i$, one counts the number of annotations in node i. As a given gene may be annotated in two (or more) nodes, these relative frequencies may add more than one.

Figure 1 illustrates these concepts for a list of three human genes (FANCG, PRKAR1B and PKIA) annotated in several nodes (in grey) at level three in the MF ontology. GO nodes are solely identified with a node code, GO:nnnnnnn. Note that the sum of annotation percentages is greater than 100 %, because one of the genes is annotated in two nodes. Thus, direct use of chi-squared tests or related techniques is not adequate for overall comparison of profiles.

In Sánchez *et al.* (2007), a statistical model for GO profiles is provided. It allows a distance-based analysis, using squared Euclidean distance:

$$d\left(\hat{P}, \hat{Q}\right) = \sum_{i=1}^{s} (\hat{p}_i - \hat{q}_i)^2. \tag{39}$$

The comparison of two sample lists of genes, in terms of the squared Euclidean distance over their GO profiles, is studied in Salicrú *et al.* (2008). The comparisons can be made in terms of either a difference problem (i.e., $H_0 : d(P, Q) = 0$ vs. $H_1 : d(P, Q) > 0$) or an equivalence problem (i.e., $H_0 : d(P, Q) \geq d_0$ vs. $H_1 : d(P, Q) < d_0$).



GO term\gene	FANCG	PRKAR1B	PKIA	Profile	%
GO:0003676	1	0	0	1	33.33%
GO:0000166	0	1	0	1	33.33%
GO:0019207	0	1	1	2	66.67%
GO:0004857	0	0	1	1	33.33%

Figure 1: *Functional profile at level 3 of MF Ontology associated with a list of three genes*

In the most general case, one wishes to compare a sample of n genes with another sample of m genes, with $n = n_1 + n_0$ and $m = m_1 + n_0$. The quantity n_0 corresponds with the number of genes shared by both lists. If $n_0 = 0$, the two lists of genes are mutually excluding: e.g., a set of recessive vs. a set of dominant genes. If $m_1 = 0$, the first list completely includes the second one: e.g., all genes analysed in a microarray vs. those differentially expressed in a given pathology.

Designating as $\hat{P}_1$, $\hat{Q}_1$ and $\hat{P}_0$ the profiles associated with the n_1, m_1 and n_0 genes, respectively, and the sample profiles to be compared as:

$$\hat{P} = \frac{n_0}{n}\hat{P}_0 + \frac{n_1}{n}\hat{P}_1 \quad \text{and} \quad \hat{Q} = \frac{n_0}{m}\hat{P}_0 + \frac{m_1}{m}\hat{Q}_1. \tag{40}$$

If P and Q are the population profiles, the asymptotic distribution of the sample profiles is multivariate normal:

$$\left(\frac{n\,m}{n+m}\right)^{1/2} (\hat{P} - P, \hat{Q} - Q) \xrightarrow{d} Y \sim N(0,\ \Sigma_{PQ}) \approx N(0,\ \Sigma_{\hat{P}\hat{Q}}) \tag{41}$$

with covariance matrix of form:

$$\Sigma_{\hat{P}\hat{Q}} = \begin{pmatrix} \frac{m}{n+m}\left[\frac{n_0}{n}\Sigma_{\hat{P}_0} + \frac{n-n_0}{n}\Sigma_{\hat{P}_1}\right] & \frac{n_0}{n+m}\Sigma_{\hat{P}_0} \\ \frac{n_0}{n+m}\Sigma_{\hat{P}_0} & \frac{n}{n+m}\left[\frac{n_0}{m}\Sigma_{\hat{P}_0} + \frac{m-n_0}{m}\Sigma_{\hat{Q}_1}\right] \end{pmatrix}. \tag{42}$$

The $s \times s$ covariance matrices $\Sigma_{\hat{P}_0}$, $\Sigma_{\hat{P}_1}$ and $\Sigma_{\hat{Q}_1}$have the general form (Sánchez *et al.*, 2007) $\Sigma_{\hat{P}} = (\hat{\sigma}_{ij})$, with

$$\hat{\sigma}_{ij} = \begin{cases} \hat{p}_i(1-\hat{p}_i) & \text{if } i = j \\ \hat{p}_{ij} - \hat{p}_i\hat{p}_j & \text{if } i \neq j \end{cases} \tag{43}$$

where $\hat{p}_{ij}$ designates the relative frequency of genes simultaneously annotated in nodes i and j and possibly also annotated in other nodes.

An asymptotic solution to the equivalence problem (i.e., to test whether both GO profiles are not very dissimilar) may be obtained from:

$$\left(\frac{nm}{n+m}\right)^{1/2} \left\{ d\left(\hat{P}, \hat{Q}\right) - d\left(P, Q\right) \right\} \xrightarrow{d} Y \sim N\left(0; \omega^2\right) \tag{44}$$

where ω^2can be estimated by:

$$\hat{\omega}^2 = 4 \begin{pmatrix} \hat{P} - \hat{Q} \\ -(\hat{P} - \hat{Q}) \end{pmatrix}^{\mathsf{T}} \Sigma_{\hat{P}\hat{Q}} \begin{pmatrix} \hat{P} - \hat{Q} \\ -(\hat{P} - \hat{Q}) \end{pmatrix}. \tag{45}$$

Thus, for a given equivalence limit d_0 and according to squared Euclidean distance, we may conclude equivalence of GO profiles if

$$d\left(\hat{P}, \hat{Q}\right) - z_\alpha \hat{\omega} \sqrt{\frac{1}{n} + \frac{1}{m}} < d_0 \tag{46}$$

where z_α corresponds to the α quantile of standard normal distribution. For example, if $\alpha = 0.05$, then we have $z_\alpha = -1.64$.

A possible criterion to establish the equivalence limit d_0 is to fix a maximum allowed discrepancy in each GO node, $|p_i - q_i| < \varepsilon$. Then $d_0 = s\varepsilon^2$, where s is the number of compared nodes.

To illustrate the above ideas, a comparison between two microarray experiments performed by Welsh *et al.* (2001) and Singh *et al.* (2002) to study prostate tumors based on gene expression data is put forward. Although the studies were performed independently, they had similar characteristics in type of tumors, microarray platforms and sample size (see table).

Study	**Platform**	**Sample**
Welsh *et al.*, 2001	HGU95A	32: normal 8, tumor 24
Singh *et al.*, 2002	HGU95Av2	102: normal 50, tumor 52

The comparability of these studies has been exploited by various authors, such as Manoli *et al.* (2006), who used them to compare different microarray data analysis methods, or Moradi *et al.* (2006), who combined them in a predictive analysis (one

data set was used as training set and the other as test set). In either situation the study combination was justified simply on the basis of their common topic, but no quantitative argument was given.

The example below shows the results of the equivalence test performed on the second level of Gene Ontology. The lists of differentially expressed genes were selected using a p-value cutoff of 0.05. The analysis was performed with R package *goProfiles* (Sánchez *et al.*, 2008) available at Bioconductor 2.2 (www.bioconductor.org).

Applying the equivalence tests to the resulting profiles for each of the ontologies gives the following results:

	MF	**BP**	**CC**
Squared Euclidean distance	0.000619	0.001768	0.004081
d_0 threshold for equivalence test (computed as $d_0 = s\varepsilon^2$ with $\varepsilon = 0.05$)	0.037500	0.050000	0.032500
Upper confidence interval limit	0.001329	0.003548	0.006386
Reject null hypothesis of inequivalence	Yes	Yes	Yes

This suggests that it is appropriate to combine the two datasets, as Moradi *et al.* (2006) did.

4 Discussion

Equivalence testing is the most adequate way to address situations where the primary aim is to prove similarity. As is shown with some detail in the case of bioequivalence testing, it is not free from difficulties or controversy, but it does seem to be the most dependable approach to bioequivalence and to many other important problems.

As many of the difficulties with the equivalence approach are essentially technical in nature, solutions to them are likely to be found or, in the worst case, the non-existence of a solution proven. In practice, other questions are more problematic, such as, in our view, the adequate determination of the equivalence limits. Wellek (2003, pp. 11-13) makes some reasonable suggestions regarding the parameters and statistical problems under consideration. However, this problem still depends, to a great extent, on specific areas of application and even on specific problems.

The distance-based approach may be a natural way to include many equivalence problems under the same paradigm, and to permit a smooth path from a univariate to a multivariate approach. There are many distance or dissimilarity indexes that may be adequate. To some extent the decision as to which index to use is arbitrary. Some are adequate due to their simplicity, ease of interpretation or easy mathematical handling. This is the case with Euclidean distance. Other indexes have nice or natural statistical properties, unfortunately sometimes associated with some handling difficulties. This is the case of measurements associated with intrinsic criteria, like those discussed in

García and Oller (2006). A natural intrinsic distance is the distance based on Fisher's information metric and proposed in Rao (1945). In this setting some concepts may have a more natural treatment; for example, the determination of the equivalence limits, possibly related to concepts like the curvature of the parametric Riemannian manifold.

A final consideration: equivalence problems generally admit either a frequentist or a Bayesian approach, but frequentist solutions are more common in the literature and much more often used in practice, despite the nice properties of many Bayesian solutions. This may be due in part to the weight of regulatory agencies in bioequivalence testing, the most significant application area. There may well be a regulatory bias towards the frequentist approach, but it is also likely to be based on criteria of clarity and ease of use for the potential users of the methods.

Acknowledgements

We thank the reviewers and the editor for constructive suggestions in the previous version of this paper.

References

Al-Mohizea A. M., Kadi, A. A., Al-Bekairi, A. M., Al-Balla, S. A., Al-Yamani, M. J, Al-Khamis, K. I, Niazy, E. M., El-Sayed, Y. M. (2007). Bioequivalence evaluation of 320 mg gemifloxacin tablets in healthy volunteers. *International Journal of Clinical Pharmacology and Therapeutics*, 45, 617-622.

Anderson, S. and Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics-Theory and Methods*, 12, 2663-2692.

Anderson, S. and Hauck, W. W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, 18, 259-273.

Barker, L., Luman, E. T., Mc Cauley, M. M. and Chu, S. Y. (2002). Assessing equivalence: an alternative to the use of difference tests for measuring disparities in vaccination coverage. *American Journal of Epidemiology*, 156, 1056-1061.

Barker, L., Rolka, H., Rolka, D. and Brown, C. (2001). Equivalence testing for binomial random variables: Which test to use? *The American Statistician*, 55, 279-287.

Barnett, V. (1999). *Comparative Statistical Inference*. New York: John Wiley.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24, 295-300.

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11, 283-319.

Blume, H. H. and Midha, K. K. (1993). Bio-International '92, Conference on Bioavailability, Bioequivalence and Pharmacokinetic Studies. *Pharmaceutical Research*, 10, 1806-1811.

Boddy, A. W., Snikeris, F. C., Kringle, R. O., Wei, G. C. G., Oppermann J. A. and Midha, K. K. (1995). An approach for widening the bioequivalence acceptance limits in the case of highly variable drugs. *Pharmaceutical Research*, 12, 1865-1868.

Brown, B. W. (1980). The crossover experiment for clinical trials. *Biometrics*, 36, 69-79.

Burns, D. R. and Elswick Jr., R. K. (2001). Equivalence testing with dental clinical trials. *Journal of Dental Research*, 80, 1513-1517.

CDER, Center for Drug Evaluation and Research (2001). *Statistical Approaches to Establishing Bioequivalence*. Rockville, MD: Food and Drug Administration. www.fda.gov/cder/guidance/3616fnl.htm#P353_2704.

CDER, Center for Drug Evaluation and Research (2003). *Guidance for Industry. Bioavailability and Bioequivalence Studies for Orally Administered Drug Products-General Considerations*. Rockville, MD: Food and Drug Administration. http://www.fda.gov/cder/guidance/index.htm.

Chow, S-C. and Liu, J-P. (2000). *Design and Analysis of Bioavailability and Bioequivalence studies*. New York: Marcel Dekker.

Chow, S-C. and Shao, J. (2002). A note on statistical methods for assessing therapeutic equivalence. *Controlled Clinical Trials*, 23, 515-520.

Committee for Proprietary Medicinal Products (CPMP) (2001). *Note for Guidance on the Investigation of Bioavailability and Bioequivalence*. London: EMEA.

D'Angelo, G., Potvin, D. and Turgeon, J. (2001). Carry-over effects in bioequivalence studies. *Journal of Biopharmaceutical Statistics*, 11, 35-43.

Dragalin, V., Fedorov, V., Patterson, S. and Jones, B. (2003). Kullback-Leibler divergence for evaluating bioequivalence. *Statistics in Medicine*, 22, 913-930.

Freeman, P. R. (1989). The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Statistics in Medicine*, 8, 1421-1432.

García, G. and Oller, J. M. (2006). What does intrinsic mean in statistical estimation? *SORT*, 30, 125-170.

Girón, F. J., Martínez, M. L., Moreno, E. and Torres, F. (2006). Objective testing procedures in linear models: calibration of *p*-values. *Scandinavian Journal of Statistics*, 33, 765-784.

Ghosh, P. and Gonen, M. (2008). Bayesian modeling of multivariate average bioequivalence. *Statistics in Medicine*, 27, 2402-2419.

Grieve, A. P. (1985). A Bayesian Analysis of the Two-Period Crossover Design for Clinical Trials. *Biometrics*, 41, 979-990.

Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics*, 21, 467-480.

Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55, 19-24.

Hsu, J. C., Hwang, J. T. G., Liu, H-K. and Ruberg, S. J. (1994). Confidence intervals associated with test for bioequivalence. *Biometrika*, 81,103-114.

Kleijnen, J. P. C. (1999). Validation of models: statistical techniques and data availability. *Proceedings of the 1999 Winter Simulation Conference, P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds.*, 647-654.

Laird, N. M., Skinner, J. and Kenward, M. (1992). An analysis of two-period crossover designs with carry-over effects. *Statistics in Medicine*, 11, 1967-1979.

Lehmann, E. L. (1959). *Testing Statistical Hypothesis*. New York: Wiley.

Longford, N. T. (2001). Synthetic estimators with moderating influence: the carry-over in cross-over trials revisited. *Statistics in Medicine*, 20, 3189-3203.

Manoli, T., Gretz, N., Grone, H-J., Kenzelmann, M., Eils, R. and Brors, B. (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, 22, 2500-2506.

Martín Andrés, A. and Herranz Tejedor, I. (2002). Equivalence testing for binomial random variables: which test to use? (Letter to the Editor). *The American Statistician*, 56, 253-254.

Mecklin, C. J. (2003). The use of equivalence testing in conjunction with standard hypothesis testing and effect sizes. *Journal of Modern Applied Statistical Methods*, 2, 329-340.

Metzler, C. M. (1974). Bioavailability: a problem of equivalence. *Biometrics*, 30, 309-317.

Midha K. K., Rawson M. J. and Hubbard J. W. (2005). The bioequivalence of highly variable drugs and drug products. *International Journal of Clinical Pharmacology and Therapeutics*, 43,485–498.

Moradi, M., Mousavi, P. and Abolmaesoumi, P. (2006). Pathological distinction of prostate cancer tumors based on DNA microarray data. http://cscbc2006.cs.queensu.ca/assets/documents/Papers/paper127.pdf.

Moreno, E. and Girón, F. J. (2006). On the frequentist and Bayesian approaches to hypothesis testing. *SORT*, 30, 3–28.

Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60, 223-241.

Munk, A. and Pflüger, R. (1999). $1 - \alpha$ equivariant confidence rules for convex alternatives are $\alpha/2$–level tests –with applications to the multivariate assessment of Bioequivalence. *Journal of the American Statistical Association*, 94, 1311-1319.

Perlman, M. D. and Wu, L. (1999). The emperor's new tests. *Statistical Science*, 14, 355-381.

Putt, M. and Chinchilli, V. M. (1999). A mixed effects model for the analysis of repeated measures cross-over studies. *Statistics in Medicine*, 19, 3037-3058.

Putt, M. (2005). Comment on 'Carry-over in cross-over trials in bioequivalence: theoretical concerns and empirical evidence'. Senn S, D'Angelo G, Potvin D. *Pharmaceutical Statistics*, 4, 215-216.

Putt, M. (2006). Power to detect clinically relevant carry-over in a series of cross-over studies. *Statistics in Medicine*, 25, 2567-2586.

Racine, A., Grieve, P., Fluhler, H. A. and Smith, F. M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry. *Applied Statistics*, 35, 93-150.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 81-91.

Rashid, M. M. (2003). Rank-based tests for non-inferiority and equivalence hypotheses in multi-centre clinical trials using mixed models. *Statistics in Medicine*, 22, 291-311.

Reynolds, M. R. and Deaton, M. L. (1982). Comparisons of some tests for validation of stochastic simulation models. *Communications in Statistics – Simulation and Computation*, 11, 769-799.

Robinson, A. P. and Froese, R. E. (2004). Model validation using equivalence tests. *Ecological Modelling*, 176, 349-358.

Rodda, B. E. and Davis, R. L. (1980). Determining the probability of an important difference in bioavailability. *Clinical Pharmacology and Therapeutics*, 28, 247-252.

Salicrú, M., Ocaña, J. and Sánchez, A. (2008). Comparison of lists of genes based on functional profiles. *Submitted*.

Sánchez, A., Salicrú, M. and Ocaña, J. (2007). Statistical methods for the analysis of high-throughput data based on functional profiles derived from the Gene Ontology. *Journal of Statistical Planning and Inference*, 137, 3975-3989.

Sánchez, A., Ocaña, J. and Salicrú, M. (2008). *goProfiles: an R package for the Statistical Analysis of Functional Profiles*. http://estbioinfo.stat.ub.es/pubs/goProfiles-Usersguide.pdf

Sargent, R. G. (2005). Verification and validation of simulation models. *Proceedings of the 2005 Winter Simulation Conference, M. E. Kuhl, N. M. Steiger, F. B. Armstrong and J. A. Joines, eds.*, 130-143.

Schall, R. (1995). Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics*, 51, 615-626.

Schall, R. and Luus, H. G. (1993). On population and individual bioequivalence. *Statistics in Medicine*, 12, 1109-1124.

Schuirmann D. J. (1987). A comparison of the Two One-sided Test procedure and the Power Approach for assessing the equivalence of Average Bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657-680.

Sellke, T., Bayarri, M. J. and Berger, O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, 55, 62-71.

Senn, S. (1988). Cross-over trials, carry-over effects and the art of self-delusion. *Statistics in Medicine*, 7, 1099-1101.

Senn, S. (1992). Is the 'simple carry-over' model useful? *Statistics in Medicine*, 11, 715-726.

Senn, S. (1996). The AB/BA Cross-over: How to perform the two-stage analysis if you can't be persuaded that you shouldn't. *Hansen, B and De Ridder, M. eds. Liber Amicorum Roel van Strik*, 93-100. Rotterdam: Erasmus University.

Senn, S. (2001). Statistical issues in bioequivalence. *Statistics in Medicine*, 20, 2785–2799.

Senn, S., D'Angelo, G. and Potvin, D. (2004). Carry-over in cross-over trials in bioequivalence: theoretical concerns and empirical evidence. *Pharmaceutical Statistics*, 3, 133-142.

Senn, S., D'Angelo, G. and Potvin, D. (2005). Rejoinder: Dr. Putt's analysis. *Pharmaceutical Statistics*, 4, 217-219.

Shah, V. P., Yacobi, A., Barr, W. H., Benet, L. Z., Breimer, D., Dobrinska, M. R., Endrenyi, L., Fairweather, W., Gillespie, W., Gonzalez, M. A., Hooper, J., Jackson, A., Lesko, L., Midha, K. K., Noonan, P. K., Patnaik R. and Williams R. L. (1996). Evaluation of Orally Administered Highly Variable Drugs and Drug Formulations. *Pharmaceutical Research*, 13, 1590-1594.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1, 203–209.

Stegner, A. L., Bostrom, A. G. and Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning*, 19, 193-198.

Tothfalusi, L. and Endrenyi, L. (2003). Limits for the scaled average bioequivalence of highly variable drugs and drug products. *Pharmaceutical Research*, 20, 382-389.

Van Steen, K., Raby, B. A., Molenberghs, G., Thijs, H., De Wit, M. and Peeters M. (2005). An equivalence test for comparing DNA sequences. *Pharmaceutical statistics*, 4, 203-214.

Wang, W. W., Hwang, J. T. G. and Dasgupta, A. (1999). Statistical tests for multivariate bioequivalence. *Biometrika*, 86, 395-402.

Warner, B. (2002). Equivalence testing. *MORS Workshop "Test & Evaluation, Modeling & Simulation and VV&A: Quantifying the Relationship Between Testing and Simulation", Kirtland AFB, Albuquerque, NM*. http://www.mors.org/meetings/test_eval/presentations/C_Warner.pdf.

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Science*, 61, 1340-1341.

Wellek, S. (2001). On a reasonable disaggregate criterion of population bioequivalence admitting of resampling-free testing procedures. *Statistics in Medicine*, 19, 2755-2767.

Wellek, S. (2003). *Testing Statistical Hypotheses of Equivalence*. Boca Raton: Chapman & Hall/CRC.

Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson Jr, H. F. and Hampton, G. M. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Research*, 61, 5974–5978.

Williamson, P. P. (2007). Bayesian equivalence testing for binomial random variables. *Journal of Statistical Computation and Simulation*, 77, 739 - 755

Zapater, P., Horga, J. F. (1999). Bioequivalencia y Genéricos. Los estudios de Bioequivalencia. I. Una aproximación a sus bases teóricas, diseño y realización. *Revista de Neurología*, 29:1235-1246.