The therapeutic Delta-equivalence. The large sample case

Jean-Luc Bosson, Claudine Robert

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

The therapeutic Δ-equivalence. The large sample case.

J.L. Bosson , C. Robert

N 2532

Avril 1995

PROGRAMME 5

ISSN 0249-6399

INRIA
RHÔNE-ALPES

The therapeutic Δ-equivalence. The large sample case.

J.L. Bosson *, C. Robert **

Programme 5 — Traitement du signal, automatique et productique
Projet Action SYSTOL

Rapport de recherche n° 2532 — Avril 1995 — 23 pages

Abstract: Several approaches are given in the literature for therapeutic equivalence (the two one-sided test approach, the power approach, the confidence and Westlake intervals approaches, the UMP test approach) yielding distinct definitions. This paper is aimed at passing beyond this heterogeneity in order to propose a standard definition of therapeutic equivalence. We firstly show that there can be no absolute definition, and that we have to define Δ-equivalence, where Δ is a threshold that has to be chosen from clinical considerations. Then we study the links and inconsistencies of the different approaches in the case of large samples (which allows us to come back to calculations with normal distributions). From the obtained results, we are able to propose the two one-sided test definition as a standard definition. We replace the p-value computation by that of an a posteriori limit value of Δ. This standard definition is simple, it is identical to the UMP test definition in case of one-sided equivalence, and it is close to it in case of two-sided equivalence. We finally propose a strategy for equivalence trials and we initiate the discussion on the medical areas where equivalence should be investigated.

Key-words: clinical trials, equivalence, UMP test, confidence intervals.

(Résumé : tsvp)

*SIIM Centre Hospitalier Universitaire de Grenoble.
**Université Joseph Fourier et INRIA Rhône-Alpes.

La Δ-equivalence thérapeuthique. Le cas des grands échantillons.

Résumé : Le but de cet article est de voir si une définition standard de l'équivalence thérapeutique peut être adoptée. Nous montrons d'abord qu'il n'existe pas de notion absolue d'équivalence, mais uniquement une notion de Δ-équivalence pour un seuil Δ choisi a priori en fonction de critères cliniques. Dans le cas de grands échantillons, nous étudions les liens et les contradictions entre les différentes approches théoriques, à savoir celle de deux tests unilatéraux, celle de contraintes de puissance, les intervalles de confiance et plus particulièrement les intervalles de Westlake, le test UPP à hypothèses composées. Il apparaît alors que l'approche par deux tests unilatéraux est à la fois simple à mettre en œuvre, numériquement proche d'une approche optimale (au sens de la théorie des tests) au plan théorique, qu'elle s'adapte aussi bien au concept d'équivalence bilatérale ou unilatérale. Par ailleurs, dans le cadre de tests d'équivalence, nous remplaçons la notion de p-value par celle de seuil d'équivalence limite. Enfin, nous proposons une stratégie pour les essais d'équivalence et amorçons une réflexion sur les domaines d'applications possibles en médecine.

Mots-clé : essais cliniques, équivalence, test UMP, intervalles de confiance.

1 Introduction

In medicine, therapeutic equivalence between two treatments is an issue that should be more and more often investigated. For example, in the treatment of arterial hypertension, all the current molecules are efficient and one of the main goals is actually to obtain "equivalent" treatments better tolerated and easier to use in order to improve therapeutic observance. Therefore, to prevent cardiovascular complications for patients at high risk (such as hypertension), one could adopt, taking into account an equivalent effectiveness, this new medication given these theoretical advantages. The treatment of hypercholesterolemia shows exactly the same kind of problem. However, in spite of an increasing need, few equivalence trials are done, as one can see when consulting the Medline data taken over the period of the last fifteen years : one finds less than twenty equivalence trials published per year over this period in spite of the fact that there have been thousands of efficiency trials referenced during this same period. This very modest recourse to equivalence trials is explained, in part, by the coexistence of several definitions for equivalence with apparently different theoretical approaches. Moreover, one finds, at present, more methodological publications on therapeutic equivalence than publications on real applications.

We intend to work on the therapeutic equivalence of two treatments , A and B, the effects of which are measured by two continuous variables. Let X_A and X_B be the variables that measure the effects of the treatments A and B, with theoretical means and variances μ_A, μ_B, σ_A^2, σ_B^2. For (global) therapeutic equivalence, we are interested in $d = \mu_A - \mu_B$. With the assumption of large samples, the sample means $\bar{X}_A$ and $\bar{X}_B$ can be considered as normally distributed. We will see that therapeutic equivalence can be restricted to situations were the theoretical variances σ_A^2 and σ_B^2 can be assigned the same common value σ^2. With large samples of respective sizes n_A and n_B, $(\bar{X}_A - \bar{X}_B)/s(1/n_A + 1/n_B)^{1/2}$ where s is an estimate of σ , has a Student distribution that can be approximated by a gaussian distribution $N((\mu_A - \mu_B), s(1/n_A + 1/n_B))$. Therefore the calculations are identical, up to classical approximations, to thoses where $\bar{X}_A$ and $\bar{X}_B$ have the respective $N(\mu_A, \sigma^2/n_A)$ and $N(\mu_B, \sigma^2/n_B)$ with known common variance. Thus we come back to the most simple situation of gaussian distributions with known variances.

Statistics in current practice is done today essentially through dedicated software; before reaching the stage where the computer integrates different methods for treating therapeutic equivalence when small or large samples are available, paired or not, we feel it is useful to clarify the notions linked to equivalence in the most simple case, that is the case of normal distributions with known variance. Furthermore, we shall see that the notion of equivalence is interesting in practice, especially in this setting, and it concerns large medical domain such as cardio-vascular disease, diabetes or infectious disease.

Having defined the statistical setting for which we consider the equivalence, we first come back to general considerations on hypotheses testing. Then we define comparability of treatments and clarify the concept of Δ equivalence for which we review the most classical approaches and their links. Fortunately, it yields a choice of a stan-

dard definition for which several explicit formulas are established (we favor, in this article, working out formulas ; they complete the simulations done for small samples in the literature [Phillips 90, Metzler 91, Diletti 90]). We furthermore propose, in the first section, to replace the calculation of the p-value with that of a limit equivalence threshold.

Let us stress the fact that we do not consider here the problem of individual equivalence of two treatments: two treatments can be globally equivalent on a population but may not be efficient for the same subjects (the problem is therefore to characterize and to differentiate their target population); also, we do not consider here the situation of changing a treatment where, for a given patient usually treated with an A molecule, the physician wonders if he can use a new B molecule, cheaper for example, without modifying the therapeutic effect. Also, we do not deal with bioequivalence in the pharmacological sense : pharmacological bioequivalence between two treatments (generic medication for example) is defined by measuring pharmacological parameters on a small number of healthy subjects (area under curve for the distribution of a molecule in the body, plasma concentrations...). If one can prove that the two treatments perform similarly in the organism then one assess that they are pharmacologically equivalent and therefore, one makes the assumption that they have the same therapeutic effect (one can see Hauck and Anderson to make clear the different types of equivalence [Hauck 92, Anderson 90]).

In the second section, we briefly consider the case where the parameter of interest for measuring the effect of the treatments is a proportion and, to conclude, we give strategy for realizing an equivalence trial with large samples.

2 Different approaches for equivalence

We wish to formulate a definition of equivalence of the two treatments A and B for which two independent samples are available.

2.1 Remarks about a classical situation

We here assume that the model adopted is that of two Gaussian distributions $N(\mu_A, \sigma_A^2)$, $N(\mu_B, \sigma_B^2)$, whose means and variances are unknown.

Let us start by considering the usual efficacy trial situation for comparing the treatments. We proceed in two stages. In the first, we perform a test for comparing the variances, which is written :

$$(1)\ H_0 : \sigma_A^2 = \sigma_B^2\ ;\ H_1 : \sigma_A^2 \neq \sigma_B^2\ \text{, level } \alpha.$$

Assume that we encounter the most pleasant situation, meaning here, that the hypothesis H_0 is accepted. We hasten to conclude that the *variances are equal* in order to move on to the second stage for comparing the means. We then use the following test:

$$(2)\ H_0 : \mu_A = \mu_B\ ;\ H_1 : \mu_A^2 \neq \mu_B\ \text{, level } \alpha.$$

If H_0 is accepted, we say that the difference between the observed empirical means is not significant at the level α, but that we haven't "proven that the means were equal". We sometimes add, if the samples are not judged to be very large, that we cannot deduce from this that treatments A and B are therapeutically equivalent, the test not being powerful enough. In short, we stress the fact that *the equality of the theoretical means μ_A and μ_B is neither rejected nor proven*, which is of course awkward.

So, for the first test we conclude on the equality of the variances and for the second, we emphasize the fact that we haven't proven that the means are equal. Would the policy of "two weights, two measures" be used that often ? This merits some comments on the most simple situation of comparison of two parameters through a standard hypothesis test :

$$(3)\ H_0 : \Theta_A = \Theta_B\ ;\ H_1 : \Theta_A \neq \Theta_B\ \text{, level } \alpha.$$

i) Accepting H_0 tells us that the model where $\theta_A = \theta_B$ agrees with the experimental data, at level α. This does not imply that models where $\theta_A \neq \theta_B$ are not also compatible with the available data.

ii) Rejecting H_0 means that the model where $\theta_A = \theta_B$ is not compatible with the experimental data, at level α. We therefore prove that all models compatible with the data verify $\theta_A \neq \theta_B$, but no particular model is proposed, meaning that no particular value of $(\theta_A - \theta_B)$ is proved to agree with the experimental data.

Let us note here that giving a logical meaning to the expression "to prove that $\theta_A = \theta_B$ " yields a paradox. Indeed, "to prove that $\theta_A = \theta_B$ " would therefore be logically equivalent to "to prove that θ_A and θ_B are not different". But the parameters θ_A and θ_B are *theoretical parameters* used to identify a model within a parametric family of models, and there is no unicity of values for such parameters : whatever the amount of available data, if they agree with a model where $\theta_A = \theta_B$, we can always find a number small enough so that a model with $(\theta_A - \theta_B) = \epsilon$ could as well agree with these same data.

In other words, the usual terminology of "true value" to designate the theoretical parameters θ_A or θ_B is deceptive, in that it infers that there exists somewhere a unique value for each theoretical parameter which would therefore give meaning to the question of the equality of these "true values". In fact, everything happens as if, in test (2), we suddenly started to believe in the existence of "true values"; the rejection of H_0 would therefore prove that they are unequal, accepting H_0 would neither prove nor disprove their equality.

We propose to keep the terminology "true value" in case we have simulated data. For example, if we simulate a sample from the distribution $N(\sqrt{0.13} + \sqrt{145}, 1)$, we call $\sqrt{0.13} + \sqrt{145}$ the true mean; note that no statistical method allows us to find

this true value in its form as a sum of two square roots and that a theoretical mean will be just one approximation of this true value.

In both cases i) and ii), we get some results concerning the possible models for the data available. If we perform test (3) having as a goal the definition of a model for the experimental data, we wish to prove the compatibility of a precise model with these data, in other words, we wish to accept H_0 . If, on the other hand, we want to take a practical decision at the outcome of the test, we want to exclude one of two possibilities and we wish to reject H_0.

Let us now come back to tests (1) and (2) and set the following definition:

Definition 0: Treatments A and B are comparable at level α if their variances and their means are comparable, in other words, if in the two tests (1) and (2) we accept H_0 .

In other words, two treatments are comparable if we can choose the same model for both.

For example, in efficacy trials where two groups are randomly selected, we often verify that these groups are comparable, with respect to age, weight or any other covariate : if the distributions related to these covariates in the samples are Gaussian, we verify that we accept the null hypothesis in tests (1) and (2).

We want a definition of therapeutic equivalence of two treatments such that if the two treatments are declared equivalent we can, within the whole population, prescribe either one. The decision of prescribing treatment A or treatment B will then depend on other criteria (better tolerance, easier observance, cost...).

If two treatments A and B are comparable at level α , it might be that a model with $\mu_A - \mu_B = \pm\Delta$ is compatible with the data even though such a difference Δ is clinically inadmissible (in other words, the power of the performed test at the point Δ is too weak cf. [Lee 91, Blackwelder 82]). Two treatments could therefore be declared comparable by an efficacy test without being declared therapeutically equivalent.

Conversely, it could be that we accept H_0 with test (1) and that we reject H_0 with test (2) even though we would accept the null hypothesis with test (2') :

$$(2')\ H_0 :\mid \mu_A - \mu_B \mid \leq \Delta\ ,\ H_1 :\mid \mu_A - \mu_B \mid > \Delta\ ,\ \text{level } \alpha.$$

where Δ is a difference judged clinically unimportant (we would, therefore have shown a significant difference having no clinical consequences cf. [Rocke 84, Westlake 72, Westlake 79]). In this case, the two treatments are declared non-comparable but their therapeutic equivalence must not be excluded.

These two classical arguments show that comparability of two treatments cannot be used as a definition for therapeutic equivalence; nevertheless it allows to set the classical statistical background for the equivalence :

- Without ambiguity it is a test problem where equivalence will be associated with the rejection of a null hypothesis.

- The notion of equivalence is not absolute, but it is tied to a threshold Δ. One can only define the notion of Δ-equivalence at a level α. (It may be that the impossibility to define a notion of equivalence that does not involve a threshold Δ, together with the belief in "true values" for the means of the distributions of the two treatments that could be proven to be equal, causes the idea that the notion of equivalence is simply a last resort).

- In order to define the equivalence using theoretical means, we will assume first that one accepts H_0 with the test (1) concerning the variances. In fact, we feel that it would not be judicious to talk about equivalence using the theoretical means for two treatments for which the associated models cannot have the same variance. For example, if for two hypertension drugs therapy we had samples of size 100 whose empirical means were equal to about 0.0001 mm Hg, but for which the empirical standard deviations had a ratio of about 1 to 10, it would be most regrettable to have a definition of equivalence which would allow them to be declared equivalent, for example, up to 1 mm Hg. If with the test (1), one rejects the null hypothesis, one can still consider the equivalence, by taking as a measure for the treatments effect, the proportion of individuals for which the results are within some interval judged however to be clinically pertinent. (For example, in the case of treatment for arterial hypertension, the therapeutic goal is to control the risk factor, that is, to bring down a population at risk (diastolic arterial blood pressure superior to 90 mm Hg) within a blood pressure range considered to be normal (60 to 90 mm Hg)) .

The first problem is the choice of Δ. This is obviously a clinical choice which involves the responsibility of the experts in the domain. Nevertheless, a sensible choice of Δ is tied, unconsciously or not, to the value of the standard deviation σ, common to the models of the two treatments [Kirshner 91]. One can, in practice, help the physicians by giving them an order of magnitude obtained with considering Δ/σ ; in fact, values of Δ/σ smaller than 0.1 seem to us, from practical experience, too restrictive in the sense that few treatments will reach Δ-equivalence, whereas values larger than 1 seem to us really large, leading too easily to a Δ-equivalence. The part of the statistician could be here to advise the physicians to choose of Δ in the interval $[\sigma/10, \sigma]$ where σ is firstly roughly estimated. We think that if the notion of Δ-equivalence is spread sufficiently within a precise domain, it is probable that standard values of Δ will be proposed for this domain. By analogy, in the neighbouring field of bioequivalence, where A designates a new treatment and B a well known standard treatment, it seems that the choice of Δ is made from a value of Δ/μ_B [Phillipps 90, Metzler 91]. But the considered choices of Δ/μ_B generally lead to values of $\Delta/\sigma = (\Delta/\mu_B)(\mu_b/\sigma)$ within [0.1; 1]. One can find in Welleck and Michaelis considerations for the choice of Δ in case of paired samples [Welleck 91].

3 Various approaches for the definition of equivalence

As far as the variances are comparable, an assumption that we made in paragraph 1, we are going to consider the situation where the sum n of the sizes of the two samples is large enough to be able to identify a Student distribution with $n-2$ degrees of freedom to a standard normal distribution ($n \geq 100$ will do). This means as quoted in the introduction that the difference of the empirical means can be modelled by a Gaussian distribution with a theoretical mean d and a known variance σ_n^2 (if σ^2 is the variance common to the two treatments and if the two samples are of the same size $n/2$, then $\sigma_n^2 = 4\sigma^2/n$, otherwise $\sigma_n^2 = \sigma^2(1/n_A + 1/n_B)$).

3.1 The optimal definition

It is well known (in [Lehmann 70, Neymann 50] or in many other statistical textbooks that there exists a uniformly most powerful test (UMP) for the following hypotheses :

$$(4)\ H_0 :\mid d \mid \geq \Delta\ ,\ H_1 :\mid d \mid < \Delta\ ,\ \text{level } \alpha.$$

The definition of the equivalence which is optimal from a statistical point of view is therefore the following :

Definition 1: The treatments A and B are Δ -equivalent at level α if, with the UMP test (4) above, the hypothesis H_0 is rejected.

This definition is proposed for example in Hauschke 90, Diletti 1990, Wellek 91, Rocke 84.

The critical region of the UMP test (4) is an interval $\mathcal{I}_n$, symmetrical about the origin and which could be written in the form :

$$\mathcal{I}_n = [-\Delta + \alpha_n\sigma_n; \Delta - \alpha_n\sigma_n]$$

where α_n verifies :

$$P(-2\Delta/\sigma_n + \alpha_n \leq U \leq -\alpha_n) = \alpha.$$

with U denoting a standard normal random variable.

It seems that this definition of equivalence has been often judged too restricting, in that $\mathcal{I}_n$ is often small and so that few treatements will be declared Δ-equivalent.

Let us try now to approximate α_n. Let u_α be such that

$$P(U \leq -u_\alpha) = \alpha.$$

We have :

$$P(-2\Delta/\sigma_n + \alpha_n \leq U \leq -\alpha_n) < P(U \leq -\alpha_n)$$

thence,

$$-\alpha_n \geq -u_\alpha\ ,\ \text{that is } \alpha_n \leq u_\alpha.$$

If $-2\Delta/\sigma_n + \alpha_n$ is small, $P(U \leq -2\Delta/\sigma_n + \alpha_n)$ is negligible and one would have :

$$P(-2\frac{\Delta}{\sigma_n} + \alpha_n \leq U \leq -\alpha_n) \approx P(U \leq -\alpha_n) \text{ and } \alpha_n \approx u_\alpha.$$

In practice, for $\alpha > 0.01$, since $P(U \leq -3) \leq 0.002$ and $-2\Delta/\sigma_n + \alpha_n \leq -2\Delta/\sigma_n + u_\alpha$:

$$-2\Delta/\sigma + u_\alpha \leq -3 \Rightarrow \alpha_n \approx u_n.$$

Table 1 gives, for different values of Δ/σ and for samples of equal sizes $n/2$, the minimum value of $n/2$ which enables the approximation α_n by u_n . The critical region can therefore be written as :

$$\mathcal{I}_n \approx [-\Delta + u_\alpha\sigma_n; \Delta - u_\alpha\sigma_n]$$

However this region can be obtained by using the definition given in the following subsection.

Δ/σ	1/10	1/5	1/4	1/3	1/2	1
n/2 $\alpha = 0.05$	1078	270	173	48	43	11
n/2 $\alpha = 0.01$	1420	355	228	128	57	14

Table 1: Values of $n/2$ such that definitions 1 and 2 give approximately the same critical region (samples of equal size).

3.2 An approach by means of two one sided tests.

Definition 2: The treatments A and B are Δ -equivalent at level α if in the two one-sided tests (5) and (6) defined below, the hypothesis H_0 is rejected.

$$(5)\ H_0 : d \geq \Delta\ ,\ H_1 : d < \Delta\ \text{, level } \alpha.$$

$$(6)\ H_0 : d \leq -\Delta\ ,\ H_1 : d > -\Delta\ \text{, level } \alpha.$$

The region where equivalence will be reached for this definition is the intersection of the critical regions of these two tests, that is :

$$\mathcal{J}_n = [-\Delta + u_\alpha\sigma_n\ ;\ \Delta - u_\alpha\sigma_n]$$

Table 2 gives the minimal values of $n/2$ such that two samples of size $n/2$ lead to an interval $\mathcal{J}_n$ which is not empty.

Since $\alpha_n \leq u_\alpha$, one always has $\mathcal{J}_n \subseteq \mathcal{I}_n$, which is normal since test 1 is uniformly the most powerful. In other words, although definition 2 is easier to use than definition 1, it is slightly more restrictive. However, for large values of n (see Table 1), $\mathcal{J}_n \approx \mathcal{I}_n$ and the two definitions are therefore identical at the considered numerical precision level. This definition is proposed in [Hauck 92, Shuirman 87 and Phillips 90].

Remarks :

Δ/σ	1/10	1/5	1/4	1/3	1/2	1
n/2	540	135	86	49	21	6

Table 2: Values of $n/2$ such that the critical region is not empty (α=0.05). (samples of equal size).

1. In practice, it is enough to perform just one of these two tests: if the observed empirical mean difference $\bar{d}_n$ is ≥ 0 (resp. $\bar{d}_n \leq 0$), it is sufficient to apply test (5) (resp. (6)) since, if the null hypothesis is rejected, it will be automatically rejected by the other test, and if it is not rejected, no equivalence can be assessed.

2. We have noted that two treatments for which relevant mean differences can be taken equal up to a level so small that it is clinically unimportant, may not be comparable at level α (and often would not be if there are enough subjects in each samples) but these treatments should however, be considered equivalent. In fact, if n is sufficiently large so that $\Delta \geq (u_\alpha + u_{\alpha/2})\sigma_n$, for all $\bar{d}_n$ verifying $u_{\alpha/2}\sigma_n \leq \bar{d}_n \leq \Delta - u_\alpha\sigma_n$, one effectively concludes that the treatments are not comparable ($\bar{d}_n$ is in the critical region of the comparability test), but are Δ-equivalent at level α ($\bar{d}_n$ is in $\mathcal{J}_n$).

3. For the unilateral equivalence [Ferner 92] where for example A is a new treatment and B a control treatment, if we consider that the mean should not be lowered "too" much, we require that $\mu_A \geq \mu_B - \Delta$, and the definitions 1 and 2 coincide, leading us to consider only test (6) , whatever the sign of $\bar{d}_n$.

3.3 Approach using the notion of power

The use of tests where the null hypothesis is a simple hypothesis has lead to a definition which, in its general form, involves two levels α and α' (cf. [Schuirmann 87] who compares the definition 3 given below with the definition 2).

Definition 3: Treatments A and B are Δ -equivalent if they are comparable with a level α' and if in the test (2) for comparison of means, that is in :

$$(2)\ H_0 : d = 0\ ;\ H_1 : d \neq 0\ \text{, level } \alpha'.$$

The power function $1 - \beta_\Delta$ is larger than or equal to $1 - \alpha$.

The region where equivalence is concluded for this definition is that where H_0 is accepted, that is to say :

$$\mathcal{K}_{n,\alpha'} = [-u_{\alpha'/2}\sigma_n; +u_{\alpha'/2}\sigma_n]$$

But the power condition can then be written $\mathcal{K}_{n,\alpha'} \subseteq \mathcal{I}_{n,\alpha}$, where $\mathcal{I}_{n,\alpha}$ is, here, the region where Δ-equivalence is concluded at level α.

This definition *is incorrect* insofar as we have seen that two treatments could be declared non comparable due to a statistically significant difference, clinically unimportant, whereas the question of their equivalence is nevertheless posed (see remark 2, 1.2.2). Furthermore, the approach using the notion of power is based on a test that is not UMP whereas in definition 1, one uses a UMP test.

The condition $\Delta \geq (\alpha_n + u_{\alpha'/2})\sigma_n$ is also the one that is obtained by taking definition 1 and by adding the condition that the power at 0 is $1-\alpha'$. We therefore replace definition 3 with definition 3' :

Definition 3': A and B are Δ -equivalent at the levels α, α' if their variances are comparable at level α, and if, in the above UMP test (4) with level α, the hypothesis H_0 is rejected, the test (4) having a power larger than $1-\alpha'$ at 0.

Definition 3' therefore appears as a particular case of definition 1 derived by adding a power condition, that is to say, by adding a constraint on the sample sizes. The domain of equivalence is therefore $\mathcal{I}_{n,\alpha}$, whereas for definition 3, it is a smaller area, that is $\mathcal{K}_{n,\alpha'}$. In other words, definition 3' is less strict than definition 3 but makes use of exactly the same number of cases.

We have seen that for $\alpha \geq 0.01$, definitions 1 and 2 lead to the same interval as soon as $\Delta/\sigma_n \geq (u_\alpha + 3)/2$. Although in theory this is not automatic, the condition $\Delta/\sigma_n \geq (\alpha_n + u_{\alpha'/2})$ of definition 3' is in most cases associated to $\Delta/\sigma_n \geq (u_\alpha + 3)/2$. For $\alpha = \alpha' = 0.05$ for example, definition 3 necessitates a number $n/2$ of subjects per sample larger than $2[(u_{\alpha/2} + u_\alpha)\sigma/\Delta]^2 = 2[(1.96 + 1.65)\sigma/\Delta]^2$. One can see in Table 3 the sample sizes necessary to consider the equivalence of two treatments for definition 3' with standard values of α and α'.

Δ/σ	1/10	1/5	1/4	1/3	1/2	1
n/2	2600	650	416	234	104	26

Table 3: Values of n/2 such that the power at 0 is 0.95. (samples of equal size).

3.4 Procedures with confidence intervals

A classical definition of the bioequivalence [Kirkwood 81, Kirshner 91, Metzler 91, Rocke 84, Schuirmann 87] and now a standard one since being recommended by the FDA in 1988, [Report support FDA 88, Food and Drug 1986] states that the two treatments are Δ -equivalent at level α when the confidence interval $[\bar{d}_n - u_{\alpha/2}\sigma_n; \bar{d}_n + u_{\alpha/2}\sigma_n]$ at level $1-2\alpha$ is included in $[-\Delta, +\Delta]$. This means that there is a Δ-equivalence at level $\alpha/2$ when $-\Delta + u_{\alpha/2}\sigma_n \leq \bar{d}_n \leq \Delta - u_{\alpha/2}\sigma_n$, which is exactly the definition 2 of the Δ-equivalence at level α. Let us note here that Hauschke and Steinijans suggest replacing the tests in definition 2 with the Mann-Whitney-Wilcoxon non parametric test for which they then compute the confidence intervals [Hauschke 90].

Westlake [Westlake 72, Westlake 76] then suggested replacing this classical confidence interval with another interval centred not about the empirical mean $\bar{d}_n$, but about 0, whence the following definition:

Definition 4: Two treatments are Δ -equivalent at level α if Westlake's interval $\mathcal{W}(\bar{d}_n)$is included in $[-\Delta, +\Delta]$, where Westlake's interval, defined by :

$$\mathcal{W}(\bar{d}_n) = [-w_n(\bar{d}_n)\sigma_n; +w_n(\bar{d}_n)\sigma_n]$$

verifies the condition (W):

$$(\mathrm{W}) : P[-(\bar{d}_n)/\sigma_n - w_n(\bar{d}_n) \leq U \leq -(\bar{d}_n)/\sigma_n + w_n(\bar{d}_n)] \simeq 1-\alpha$$

where U denotes a standard normal random variable.

Definition 4 consists in replacing, in definition 1, the UMP test with a test based on the statistic $w_n(\bar{d}_n)$, with critical region the set $\mathcal{W}_n$ of values of $\bar{d}_n$ such that the function $w_n(\bar{d}_n)$, positive valued, defined by the condition (W) here above, verifies $w_n(\bar{d}_n) \leq \Delta/\sigma_n$, that is :

$$\mathcal{W}_n = \{\bar{d}_n; w_n(\bar{d}_n) \leq \Delta/\sigma_n\}.$$

Let us note that the condition (W) implies that :

$$-\bar{d}_n/\sigma_n + w_n(\bar{d}_n) \geq u_\alpha \text{ and } -\bar{d}_n/\sigma_n - w_n(\bar{d}_n) \leq -u_\alpha\,,$$

which can be written as :

$$(u_\alpha - w_n(\bar{d}_n))\sigma_n \leq \bar{d}_n \leq (-u_\alpha + w_n(\bar{d}_n))\sigma_n.$$

If definition 4 is satisfied, in other words if $w_n(\bar{d}_n)\sigma_n \leq \Delta$, one then has:

$$-\Delta + u_\alpha\sigma_n \leq (u_\alpha - w_n(\bar{d}_n))\sigma_n \leq \bar{d}_n \leq (-u_\alpha + w_n(\bar{d}_n))\sigma_n \leq +\Delta - u_\alpha\sigma_n$$

and definitions 1 and 2 are also satisfied (which is normal for definition 1, being optimal since it uses a UMP test); we finally have :

$$\mathcal{W}_n \subseteq \mathcal{J}_n \subseteq \mathcal{I}_n$$

Definition 4 is not only more complicated to set up but it is also more restrictive than definition 2. Let us also note that if $\bar{d}_n \leq 0$, we have $\bar{d}_n + w_n(\bar{d}_n)\sigma_n \leq u_{\alpha/2}\sigma_n$ and, if $\bar{d}_n \geq 0$, $\bar{d}_n - w_n(\bar{d}_n)\sigma_n \geq -u_{\alpha/2}\sigma_n$, thence finally $w_n(\bar{d}_n)\sigma_n \leq u_{\alpha/2}\sigma_n + \mid \bar{d}_n \mid$. One can deduce that:

$$\mathcal{J}_{n,\alpha/2} \subseteq \mathcal{W}_{n,\alpha} \subseteq \mathcal{J}_{n,\alpha}$$

where $\mathcal{J}_{n,\alpha/2}, \mathcal{J}_{n,\alpha}$ denote the equivalence intervals at levels $\alpha/2$ and α. Thus, the Westlake definition is more strict than the definition at level α and weaker than the definition at level $\alpha/2$. Nevertheless, if n is large, we have : $\mathcal{J}_{n,\alpha/2} \approx \mathcal{W}_{n,\alpha} \approx \mathcal{J}_{n,\alpha}$.

We particularly think, that since this is a decision problem and not one of estimation, it should be formulated as such, and this is also the aim of [Kirkwood 81] with

whom we completely agree. First, if the experimental data lead to a confidence interval, centred at $\bar{d}_n$ or at 0, and not included within $[-\Delta, +\Delta]$, this does not signify, contrary to that which is often suggested in the definition with confidence intervals, that the treatments are inequivalent. Basing the equivalence definition on confidence intervals tends to neglect power considerations [Metzler 91] in a decision problem and is inappropriate. Kirshner argues that confidence interval have the advantage that they do not depend on a threshold Δ [Kirshner 91]. Let us nevertheless remark that a conclusion on equivalence will necessitate to define Δ . This choice is of great clinical importance, and we think it is better to consider it first and not after the experiments. It is therefore advisable to reset the notion of confidence interval, with respect to that of the rejection region of a null hypothesis. In the case of the classical confidence interval centred at $\bar{d}_n$, this can be done very simply, the rejection region is $\mathcal{J}_n$ (it seems that at the origin of the Westlake procedure there has been some confusion between the classical confidence interval, not centred at 0, and that of the rejection region, which is symmetric) and the power at 0 is easily computed. On the other hand, the Westlake procedure leads to a complicated equivalence region $\mathcal{W}_n$ and we do not know if the computation of the power at 0 is feasible. We haven't actually found an argument, theoretical or practical, which could justify the use of definition 4.

4 The calculation of a limit threshold

In efficiency tests, one a posteriori computes the limit value of α , which is called the p-value. This p-value is the realization of a random variable, whose distribution, under the null hypothesis, is uniform. In the case of the Δ-equivalence, one can either look at the p_Δ-value for a fixed Δ threshold or at a limit Δ_α threshold for a fixed value of α, such that the equivalence be, a posteriori, accepted at level α for all $\Delta \geq \Delta_\alpha$. For example, in the study for two hypertension drugs therapy, one can conclude that the treatments are a posteriori equivalent up to 5 mm Hg with a p-value of 0.035, or that they are equivalent at level 0.05, for any value of Δ clinically acceptable and $\geq$ 3.5 mm Hg. The equivalence threshold limit seems to us more interesting than the p_Δ-value in many cases.

Given an observation $\bar{d}_n$ and a level α, the value of Δ_α such that the Δ-equivalence at level α is proven for all $\Delta \geq \Delta_\alpha$, for definition 1, is given by:

$$\Delta_\alpha = \mid \bar{d}_n \mid + \alpha_n \sigma_n,$$

and for definition 2 by:

$$\Delta'_\alpha = \mid \bar{d}_n \mid + u_\alpha \sigma_n.$$

Note that we always have :

$$\Delta_\alpha \leq \Delta'_\alpha.$$

Let d_n be the realization of a random variable with distribution $N(d, \sigma_n^2)$. The cumulative distribution function of the random variable, whose Δ'_α is a realization,

is given by :

$$F(\Delta) = P(\Delta'_\alpha \leq \Delta) = \Phi[(\Delta - d)/\sigma_n - u_\alpha] - \Phi[u_\alpha - (\Delta + d)/\sigma_n],$$

where Φ is the cumulative distribution function of $N(0,1)$.

The calculation of the a posteriori limit values of α or Δ tempers the a priori arbitrarily choice of α or Δ . In other words, the interest lies in the comparison of Δ_α (or of p_Δ) to a range of values that could have, a priori, been reasonably done. One cannot simultaneously calculate a limit value for α and Δ , and in order to balance the choice for the calculation of Δ_α or of Δ'_α, one could calculate Δ_α and/or Δ'_α for several classical values of α (for example 0.1, 0.05, 0.01).

4.1 The choice of a definition

We have studied four definitions; the second is an approximation of the first, the third gives, after correction, a particular case of the first one and the fourth has been eliminated. We do not talk here, about other definitions specific to the bioequivalence (rule of the 75/75 [Food and Drug 88]). That leaves the choice between the two first definitions (one will find in Rocke 84 the comparison of the equivalence intervals for definition 1 at level α, definition 2 with the levels α and $\alpha/2$ and definition 5 when one does not approximate the Student distribution with a normal distribution).

Although slightly more restrictive than the first, we propose to take as a standard the second definition. This choice is justified by its practical simplicity and also because this definition generalizes to some other parameters for which the statistic of test (4) does not have a symmetrical distribution. This definition can also be generalized to the case of a proportion and allows, very easily, the determination of a limit threshold Δ'_α. Note that, in bioequivalence, one also considers a non symmetric notion of equivalence [Stein] where, in test (4), one replaces the hypotheses with:

$$H'_0 : d \geq \Delta_2 \text{ or } d \leq -\Delta_1 \ ; \ H'_1 : -\Delta_1 \leq d \leq \Delta_2 \text{ , level } \alpha.$$

The two tests for definition 2 should therefore be adapted [Schuirmann 87, Hauck 84].

Such a situation of non symmetric equivalence doesn't ocuur in therapeutic equivalence when the variable of interest has a symmetric distribution.

Lastly, the postulated definitions and their ties remain sensible in the case of normal distributions, when the sum n of the sample sizes is too small to adopt the Gaussian approximation of the Student distribution. However the notion of the therapeutic Δ-equivalence for reasonable values of Δ concerns, in our opinion, the treatments for which one can obtain large size samples. In fact, the region $\mathcal{J}_n$ is very small and sometimes even empty for small samples and for values of Δ such that $\Delta/\sigma \leq 1$; for example, for $n \leq 40$, with samples of the same size, $\mathcal{J}_n$ is empty for $\Delta/\sigma \leq 1/2$ (see Table 2), and in such cases, I_n will be empty or very small.

5 Calculating the necessary number of subjects

In the majority of cases, using definition 2, the number of subjects necessary in order to have a power larger than or equal to $1 - \beta_0$ at 0 is given, when the two samples are of the same size (i.e. $n/2$ with the previous notation) by:

$$n/2 \approx 2[(u_\alpha + u_{\beta_0/2})(\sigma/\Delta)]^2$$

One can see that this value of n insures a power larger than $1 - \beta_0$ at 0 for test (4) of definition 1.

This formula is also given in Ferner and Neumann [Ferner 92], as well as similar formulas in the case of one sided equivalence. Power plots are given in Diletti for typical situations in bioequivalence, as well as in Phillips and in Metzler for small samples [Diletti 91, Phillips 90, Metzler 91].

Note that the number n' of subjects necessary in an efficiency test at level α, if one wants to have a power of $1 - \beta_\Delta$ at $\pm\Delta$, is given for samples of the same size $n'/2$, by:

$$n'/2 \approx 2[(u_{\alpha/2} + u_{\beta_\Delta})(\sigma/\Delta)]^2$$

To compare the cost of a classical efficiency test (test 2) to that of an equivalence test for the same threshold with similar power restrictions ($\beta_0 = \beta_\Delta = \beta$), one can calculate the ratio n/n' for the number of subjects necessary in the two situations. One therefore has:

$$n/n' = [(u_\alpha + u_{\beta/2})/(u_{\alpha/2} + u_\beta)]^2$$

Hence, for $\alpha = \beta$, $n = n'$ and for $\alpha = 0.05$ and $\beta_0 = \beta_\Delta = 0.1$, $n/n' = 1.03$. Contrary to the belief (Metzler 91), the required numbers of subjects are rather close to each other.

Example 1. A quantitative case.

This concerns a therapeutic trial in the treatment for high blood pressure. The goal is to show the therapeutic equivalence between two forms of the same anti hypertension therapy with, on the one hand, a conventional treatment (immediate liberation (IL) given twice a day) and on the other hand, a new low dosage formulation with prolonged liberation (PL) given once a day. One hopes to reduce the number of doses taken daily, and thus the tolerance would be greater and the observance of the therapy would be better. It is a double blinded multicentrical study with two independent groups. The main criterion to consider is the diastolic blood pressure (DBP) after three months of treatment. In the IL group, out of 205 patients, the mean effect observed (DBP M0-M3) is a drop of 11.1 ± 7.9 mm de Hg. In the PL group, out of 200 patients, the mean effect observed (DBP M0-M3) drop is of 10.7 ± 7.4 mm of Hg. In our opinion, one should choose $\Delta = 5$ mm Hg.

Given that the comparison test for the two variances is not significant, one can assume a common variance $\sigma^2 = 58,63; \sigma = 7,66$, therefore $\sigma_n = 7.66(1/205 +$

$1/200)^{1/2} = 0.76$. With $\Delta = 5$, $\Delta/\sigma \approx 1/2$ and $2\Delta/\sigma_n \approx 13 \gg 3$, definitions 1 and 2 give the same equivalence interval:

$$\mathcal{J}_n = [-5 + 1.645 \times 0.76; +5 - 1.645 \times 0.76] = [-3.75 + 3.75].$$

The observed difference between the two groups of treatment $\bar{d}_n = 11.1 - 10.7 = 0.4$ lies within this interval, and therefore one can conclude to the equivalence (p-value $\ll 0.001$). One has:

$$\Delta_{0.05} = 0.4 + 1.645 \times 0.76 = 1.65 \text{ and } \Delta_{0.01} = 0.4 + 2,3 \times 0.76 = 2.15.$$

For illustration purposes, the Westlake interval at level 0.05 is:

$$\mathcal{W}(\bar{d}_n) = [-w_n(\bar{d}_n)\sigma_n; +w_n(\bar{d}_n)\sigma_n] = [-1.675; +1.675] \ (w_n = 2,204).$$

This interval is included in [-5,+5] for $\Delta = 5$. Thus, the two treatments are also equivalent according to the Westlake procedure. The limit threshold calculated for the Westlake method is obviously here $w_n(\bar{d}_n)\sigma_n$, i.e. 1,675, thus slightly larger than the limit threshold $\Delta_{0,05}$ calculated for definitions 1 and 2 (1,67 vs 1,65).

But in any case, in this example for arterial hypertension, a limit smaller than 2 mm of Hg would not make sense since, even under the best conditions, the precision for measuring the DBP is of the order of $\pm$ 1 mm of Hg. Therefore, for any reasonable choice of the limit Δ , the treatments are therapeutically equivalent at level 0.05. The estimated tolerance on biological criteria is significantly better for the group PL (19recommend the treatment with prolonged liberation, whose effectiveness is equivalent, which has less secondary effects and is more simple to use. In this example, like in many others, there is no great difference between all the considered approaches. Nevertheless, we claim that definition 2 is both the simplest and the best justified one.

6 The case of a proportion

The average therapeutic equivalence for two treatments is considered here, in the case of effects measured with variables for which one adopts the Bernouilli model with respective parameters p_A and p_B.

It would seem that there is a general agreement for the choice of $\Delta = 0.1$ but we have not found any medical discussions concerning this. We can only say that insofar as p_A and p_B are both between 0.1 and 0.9, the standard deviation of the empirical frequency of each of them will be between 0.3 and 0.5, from where it follows that $1/5 \leq \Delta/\sigma \leq 1/3$, which is in conformity with the mathematical restrictions imposed in the first paragraph, independent of any clinical consideration. A much smaller choice of Δ , for example 0.01, would lead to a definition that few pairs of treatments could attain.

Let us point out, however, an important one sided situation, concerning the new treatments that do not reduce the success rate by "to much" while allowing a notable

reduction in toxicity (for example in treatments for cancer where the parameters p_A and p_B are the survival rates [Rodary 89]). It appears delicate, from the ethical point of view, to fix Δ with regard to even the lowest toxicity of the new treatment ; on the other hand, one generally only has small samples available; in such a case, the comparability and the consideration of the p-value seems to us to be better adapted than the therapeutic equivalence, for a decision that will rely in fine on a qualitative appreciation of the secondary effects of the two treatments.

We use here definition 2 for the equivalence, but, of course, we can no longer assume that the variances relative to treatment distributions are equal. We still assume that the samples are large and that one can approximate the binomial distributions with normal distributions. The problem is therefore to estimate the variance of the difference between empirical frequencies under the null hypothesis of the test to be performed. Two solutions are usually envisaged, notably in Dunnet and Gent [Dunnet 77]:

Let $\nu_A = n_A/N_A$ and $\nu_B = n_B/N_B$ be the observed empirical frequencies on the samples of size N_A and N_B (here $n = N_A + N_B$)

1. one estimates the variance by:

$$s_n^2 = \nu_A(1-\nu_A)/N_A + \nu_B(1-\nu_B)/N_B$$

2. if $\nu_A \geq \nu_B$, one estimates the variance by:

$$s'^2_n = \pi_A(1-\pi_A)/N_A + \pi_B(1-\pi_B)/N_B$$

where $\pi_A = (n_A+n_B+N_B\Delta)/(N_A+N_B)$ and $\pi_B = (n_A+n_B-N_A\Delta)/(N_A+N_B)$ verify

$$N_A\pi_A + N_B\pi_B = n_A + n_B \text{ and } \pi_A - \pi_B = \Delta.$$

In any case, we can see that, since it is always possible to write that s_n^2 and s'^2_n are smaller than $(1/N_A + 1/N_B)/4$, the equivalence interval $\mathcal{J}_n = [-\Delta + u_\alpha\sigma_n; \Delta - u_\alpha\sigma_n]$ always contains the interval :

$$min\mathcal{J}_n = [-\Delta + u_\alpha(1/N_A + 1/N_B)^{1/2}/2\,;\ \Delta - u_\alpha(1/N_A + 1/N_B)^{1/2}/2].$$

Thence, it suffices that $\bar{d}_n$ be in $min\mathcal{J}_n$ in order to prove Δ-equivalence at level α.

Lastly ,we always have :

$$\Delta_\alpha \leq max\Delta_\alpha = \bar{d}_n + u_\alpha(1/N_A + 1/N_B)^{1/2}/2.$$

We know however, that, in this case, the right reference parameter is the odds ratio, and that by fixing the bounds for $p_A(1-p_B)/p_B(1-p_A)$, one would no longer have any nuisance parameters [Dunnet 77]. Dunnet and Gent show, when comparing with the Gart method, that a direct calculation of Dunnet and Gent's chi square, whose value here is $(n_A - N_A\pi_A)^2(1/N_A\pi_A(1-\pi_A) + 1/N_B\pi_B(1-\pi_B))$, produces better p-values, but less than about 0.001, than the calculation of the p-values by

means of a normal approximation with either one of the two above methods. We therefore use the results from Dunnet and Gent to confirm the choice of definition 2 in the case of proportions, with method 2 for approximating the variance of the difference between empirical frequencies.

In practice, in order to calculate the number of subjects necessary for having a given power at 0, we suggest replacing, in the formula of section 1.5 concerning the efficiency trials, the number σ by $\sigma' = (\nu'(1-\nu'))^{1/2}$, where ν' is an a priori rough approximation of the frequencies ν_A and ν_B. Thus, if one knows that ν_A and ν_B will be between 0.8 and 0.9, one would take $\nu' = 0.8$ and if one knows that ν_A and ν_B will be between 0.2 and 0.3, one would take $\nu' = 0.3$, that is, we sort out an approximation of an upper bound for the larger of the two proportions (Remember that, similarly, when calculating the number of subjects necessary in an efficiency trial, we are often led to estimate, a priori, the common variance for the models of the two treatments). Finally, one finds in [Lee 91 and Shuirmann 1990] some excellent studies of equivalence for the case where the parameter of interest is a proportion, but where the two samples are paired (the "cross-over trials" case). Also, one finds in [Mehta] a generalization of the work done by Dunnet and Gent for ordinal data, based on a numerical algorithm of the p-value of the Wilcoxon test for this case. Now let us consider two examples: in the first, ν_A and ν_B are very close to $1/2$; in the second, they are smaller than 0.05.

Example 2. C. Carlier *et al* propose an equivalence trial concerning the use of β-carotene instead of retinyl palmitate in the case of vitamin A deficiency [Carlier 93]. The trial is a double blind trial with random samples. The individuals treated are Senegalese children (2 to 15 years old) with xerophthalmia or with cytological signs of vitamin A deficiency. Seven weeks after absorbing doses of retinyl palmitate or β-carotene, one looks at the rate of children who have been cured, that is to say, that do no longer present signs of vitamin A deficiency. The trial is formulated as a one sided equivalence trial with $\Delta = 0.1$ and $\alpha = 0.05$. If A denotes the treatment with retinyl palmitate and B the treatment with β-carotene, one wants to show that B is not worse than A. Therefore it is sufficient to perform the test (5). The data are the following:

$$n_A = 124,\ N_A = 242,\ n_B = 123,\ N_B = 246.$$

We then have $\nu_A = 0.512$ and $\nu_B = 0.5$. The variances calculated with 1) or 2) are approximately equal to $max(s_n^2)$, that is to $(0.045)^2$. The equivalence interval is approximately :

$$min\mathcal{J}_n = [-0.026; +0.026] \text{ , where } 0.026 = 0.1 - 1.645 \times 0.045$$

and contains the difference 0.012=0.512-0.500. The two treatments are therefore 0.1 equivalent at level 0.05. One can note that here limit threshold $\Delta_{0.05} = 0.012+0.074 = 0.086$ and that $\Delta/\sigma \approx 0.1/0.5 = 1/5$. The limit value 0.086 of $\Delta_{0.05}$ is approximately equal to the limit $\Delta = 0,1$ that was a priori chosen by the authors of the paper. One can thus consider that with the same protocol for the two treatments, Δ-equivalence for $\Delta = 0.05$ or $\Delta = 0.01$ is far from being proved. Finally $\Delta = 0.1$ appears both

experimentally sensible and justified in such a public issue concerning prevention in a country where any preventive policy based on drug administration meets many impediments.

Example 3. Two treatments, one with low molecular weight heparin (LMWH) and the other with standard heparin (SH) are studied. They are known to be effective in the prevention of postoperative venous thrombo-embolism, and their safety is under investigation in this study [Kakkar 93]. In a multicenter randomized trial with 3809 patients undergoing major abdominal surgery, heparin was given preoperatively and continued at least 5 postoperative days (1894 LMWH and 1915 SH). Major bleeding event, which was here the event of interest, appeared in 69 patients (3,6groups appear to be comparable with respect to numerous covariates. Let us consider now the 0.03-equivalence with $\alpha = 0.05$ for the safety of the two heparins with regard to major bleeding events. The two methods for computing the variance yield approximately equal equivalence intervals :

$$\mathcal{J}_n = [-0.019; +0.019].$$

As $\bar{d}_n = 0.048 - 0.038 = 0.012$, one concludes that the safety is equivalent up to 0.02 (here $\Delta_0.05 = 0.023$). Finally, treatments with LMWH need a unique daily injection (instead of 2 or 3 for SH treatments). LMWH and SH treatments are equivalent with a clinically sensible choice of Δ ($\Delta = 0.02$) with regard to the risk of haemorrhage (which are in most cases minor haemorrhages). LMWH treatments are at least as effective as SH treatments to prevent deep venous thrombosis. Theses results indicate that LMWH should be preferred to SH treatments.

7 Conclusion

To conclude, we suggest a simple strategy for an equivalence trial.

1. Firstly, it is necessary to consider whether the situation is one for which there is a question of equivalence. In particular, there must be no ethical obstacle for officially defining an equivalence threshold and one must think about having relatively large samples available. Also, a decision must be made with respect to the question being asked (for example, when one randomly picks samples in a population, it is sufficient in our opinion, to look at their comparability with respect to the covariates, which does not necessitate the choice of a threshold Δ).

2. If it is a problem of equivalence, is it a one sided or a symmetric problem? A one sided problem is a problem associated with only one of the two following tests:

$$H_0 : d \geq \Delta \ , \ H_1 : d < \Delta \ , \text{ level } \alpha,$$

or:

$$H_0 : d \leq -\Delta \ , \ H_1 : d > -\Delta \ , \text{ level } \alpha,$$

and a symmetric problem is one associated with the test:

$$H_0 :\mid d \mid \geq \Delta \ , \ H_1 :\mid d \mid < \Delta.$$

For example, if we study a new treatment for high blood pressure, the aim is to regulate blood pressure which must be neither too high nor too low. It is therefore a symmetric problem. On the other hand, in the study of a new treatment, less toxic than the standard one used for a certain illness, if the parameter of interest is the percentage of patients who have been either cured or have improved using this treatment, one would be in a one sided situation where the advantages of the new treatment are such that one would tolerate, with respect to the standard treatment, a decrease at most equal to Δ for cured or improved patients.

3. One chooses therefore, an equivalence value Δ and a level α. The choice of Δ depends on clinical considerations, on the precision of the available measurements and on a rough a priori estimation of the standard deviations that could be observed.

4. Given an a priori estimation of σ, if it is possible, one will take equal size samples and one should choose the total number n of subjects in order that:

   - the equivalence region for definition 2 is not empty, that is, for the symmetric case:
     $$\Delta/\sigma_n \geq u_\alpha.$$
     That is, for samples of equal size $n/2$:
     $$n/2 \geq 2[u_\alpha(\sigma/\Delta)]^2$$
     (cf. Table 2 for numerical values). In the one sided case, this condition has to be maintained if one wishes that the equivalence region lies entirely on the same side of 0.
   - the power at 0 is $1-\beta$, which implies that (cf. Table 3) :
     $$\Delta/\sigma_n \geq u_\alpha + u_{\beta/2}.$$
     That is, for samples of equal size $n/2$:
     $$n/2 \geq 2[(u_\alpha + u_{\beta_0/2})(\sigma/\Delta)]^2$$

5. After having made the trial and gathered the data, one could:

   - check that the variances are comparable at level α.
   - look if there is a Δ-equivalence at level α.
   - calculate the limit Δ_α threshold. One can compare it to the precision of the measurements, compute the ratio Δ_α/σ and compare it to 1.

Finally, considering large samples clarifies the notion of Δ equivalence and yield to a simple strategy for many equivalence trials (which do not necessitate much larger samples than efficacy trials). Furthermore, we think that the public health domain could have a great use of the Δ-equivalence notion in the next years, since it often deals with well known frequent pathologies where life is not directly at risk, where reference treatments are effective and where the optimization of such treatments is the issue to investigate. The above strategy might also be useful in situations where a null hypothesis has to be accepted contrary to a strong a priori conviction. Then, if the issue is formulated in terms of equivalence, one might come back to rejecting a null (new) hypothesis.

References

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

Blackwelder W.C. (1982). Proving The Null Hypothesis In Clinical Trials. *Controlled Clinical Trials*, **3**, 345-353.

Carlier C., Coste J., Etchepare M. *et al.* [1993]. A Randomised Controlled Trial to Test Equivalence Between Retinyl Palmitate and Beta Carotene for Vitamine A Deficiency. *British Medical Journal*, **307**, 1106-10

Diletti E., Hauschke D., and Steinijans V. (1991). Sample Size Determination for Bioequivalence Assessment by Means of Confidence Intervals. *International Journal of Clinical Pharmacology, Therapy and Toxicology*, **29**, 1-8.

Dunnet C.W. and Gent M. (1977). Significance Testing to Establish Equivalence Between Treatments, with Special Reference to Data in the Form of 2*2 Tables. *Biometrics*, **33**, 593-602.

Ferner U. and Neumann N. (1992). Active Control Equivalence Trials : Some Methodological Aspects. *Psychopharmacology*, **106**, S93-S95.

Food and Drug administration. (1988). *Report by the Bioequivalence Task Force on recommendations from the bioequivalence hearing, conducted by the Food and Drug Administration*, September 29 October 1, 1986.

Hauck W.W. and Anderson S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **12**, 83-91.

Hauck W.W. and Anderson S. (1992). Types of Bioequivalence and Related Statistical Considerations. International Journal of Clinical Pharmacology. *Therapy and Toxicology*, **30**, 181-7.

Hauschke D., Steinijans V.W., and Diletti E. (1990). A distribution-free procedure for the statistical analysis of bioequivalence studies. *International Journal for Clinical Pharmacology, Therapy and Toxicology*, **28**, 72-78.

Kakkar V.V., Cohen A. T., Edmonson R.A. *et al.* (1993). Low Molecular Weight versus Standard Heparin for Prevention of Venous Thromboembolism After Major Abdominal Surgery.*Lancet*, **341**, 259-265.

Kirkwood T.B.L. (1981). Bioequivalence Testing-A Need to Rethink. *Biometrics*, **37**, 589-94.

Kirshner B. (1991). Methodological Standards for Assessing Therapeutic Equivalence. *Journal of Clinical Epidemiology*, **44**, 839-49.

Lee M.L. (1991). The problem of therapeutic equivalence with paired qualitative data : an example from a clinical trial using haemophiliacs with an inhibitor to factor VIII.*Statistics In Medicine*, **10**, 433-441.

Lehmann E.L (1959). *Testing Statistical Hypotheses*. Wiley.

Metzler C.M. (1991). Sample Sizes for Bioequivalence Studies.*Statistics in Medicine*, **10**, 961-970.

Mehta C., Patel N. R., and Tsiatis A. (1984). Exact Significance Testing to Establish Treatement Equivalence with Ordered Categorical Data. *Biometrics*, **40**, 819-825.

Neyman J. (1950). *First Course in Probability and Statistics*. Ed H. Holt and Company. New-York.

Phillips K.F. (1990). Power of the Two One Sided Tests Procedure in Bioequivalence.*Journal of Pharmacokinetics and Biopharmaceutics*, **18**, 137-144.

Report supports FDA's bioequivalence evaluation criteria, advocates more stringent statistical testing. *Clinical Pharmacy*, 1988, **7**, 336-337.

Rodary C., Com-Nougue C., and Tournade M.F. (1989). How To Establish Equivalence Between Treatments : A One Sided Clinical Trial in Pediatric Oncology. *Statistics In Medicine*, **8**, 593-8.

Rocke D.M. (1984). On Testing for Bioequivalence. *Biometrics*, **40**, 225-230.

Schuirmann D. J. (1987). A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing in the Equivalence of Average Bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657-680.

Stein.S., Koeleman H.A. Gouws E., and Ritschel W.A. (1991). An approache to select the appropriate statistical method for testing bioequivalence. *International Journal for Clinical Pharmacology, Therapy and Toxicology*, **29**, 156-160.

Wellek S. and Michaelis J. (1991). Elements of Significance Testing with Equivalence Problems. *Methods of Information in Medicine*, **30**, 194-8.

Westlake W.J. (1972). Use of Confidence Intervals in Analysis of Comparative Bioavailability Trials. *Journal of Pharmaceutical Sciences*, **61**, 1340-1341.

Westlake W.J. (1976). Symmetrical Confidence Intervals for Bioequivalence Trials. *Biometrics*, **32**, 741-744.

Westlake W.J. (1979). Statistical Aspects of Comparative Bioavailibility Trials. *Biometrics*, **35**, 273-280.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399