Quaderni di Statistica
Vol. 10, 2008

Non-parametric two-stage active control testing method for non-inferiority tests

Rosa Arboretti Giancristofaro Stefano Bonnini
Department of Mathematics, University of Ferrara
E-mail: rbrrso@unife.it; bnnsfn@unife.it

Francesca Solmi
Department of Statistics, University of Padova
E-mail: checcasolmi@libero.it

Summary: One of statistics' most important application fields in medicine is the comparison of different populations, and in particular the evaluation of the differences between the effects of two medical treatments. In this work we deal with a specific issue directly related to this application field, i.e. the non-inferiority test. Placebo-controlled trials are in fact ideal to evaluate medical treatment effectiveness, but they are ethically justified only if no standard treatment exists. In these cases active-controlled trials are generally more appropriate, and in particular the non-inferiority trial. The Two-Stage Active Control Testing (TACT) method is suitable for evaluating differences between a new treatment and the control. Here we propose a permutation version of this technique that may be used when usual distributional assumptions do not hold.

Keywords: Non-inferiority Test, Permutation Test, Two-sample test.

1. Introduction

Randomised clinical trials (RCTs) are one of the most important techniques in biostatistics (Fisher, 1999; Harrington, 2000; Smith, 1998). Many studies on RCTs have been conducted (Hill, 1962; Spilker, 1991; Foulds, 1958; O'Brien, 1968) and in general it has been

demonstrated that they are more suitable for controlling for the false positive problem than uncontrolled clinical trials.

Among all possible RCTs, placebo-controlled trials require a small number of patients to detect a treatment effect. When an active treatment exists, placebo-controlled trials may be questionable (Rothman *et al*., 1994; Freedman *et al*., 1996a, 1996b; Angell, 1997; World Medical Association Declaration of Helsinki, 1997).

In these cases, therefore, active-controlled trials may be used, with non-inferiority trials being particularly appropriate (D'Agostino, 2003). Their objective is to establish that the effect of a new treatment, compared to a control treatment, is not below a specific pre-stated non-inferiority margin. Often the non-inferiority principle is called the 'at least as good as' criterion, which helps understand the principal idea behind the approach.

In light of this, the null and alternative conventional hypotheses have been modified into directional hypotheses and many proposals for unilateral tests have been made (Blackwelder, 1982; Blackwelder *et al*., 1984; Makuch *et al*., 1978; Laster *et al*., 2003; Wang *et al*., 2003). In these situations we also recognize the use of confidence intervals (Makuch *et al*., 1978; Durrleman *et al*., 1990): it has been shown (Holmgren, 1999) that for such cross-trial inference, the probability of establishing equivalence or non-inferiority using the preservation test approach is higher than the probability using the confidence interval approach, with the non-inferiority margin defined as the lower bound of the control treatment effect's confidence interval.

Another interesting issue is linked to the comparison between standard tests and confidence interval methods (Laster *et al*., 2003; Hung *et al*., 2003); indeed the choice of one method over another can lead to differences in the reliability of the results, depending on the invariance of the treatment effect over time (constancy condition) in historical studies. If this condition can be held, standard tests are a reasonable choice, otherwise they may lead to incorrect results. Therefore, the choice of historical studies to consider and the analysis of the presence of the constancy condition may constitute a crucial part of the researcher's work (Temple, 1983; Fleming, 1987; Pledger *et al*., 1990; Fleming, 2000; Jones *et al*., 1996).

Because of this duality between the standard test statistic and the confidence interval method, a new approach has been developed to control these issues (Wang *et al*., 2003): the Two-stage Active Control Testing (TACT) method attempts to solve this problem by anticipating the application of the decision method with a preventive analysis of whether or not the control treatment effect's constancy condition holds over time.

Section 2 of this work is devoted to the definition of the non-inferiority test and discussion of some proposals from the literature, such as the Blackwelder test (Blackwelder and Chan, 1984) and the TACT method. We also propose a permutation solution for the application of the TACT method. In section 3 the results of a simulation study will be reported in order to compare the behaviour of the proposed non-parametric procedure and of the parametric solutions. In section 4 a real-case application will be illustrated. Section 5 is devoted to some final remarks.

2. The non-inferiority issue

The concept of non-inferiority and the need to develop a theory to define the non-inferiority issue originates in the context of the comparison of different medical treatments, particularly in the pharmaceutical field.

Let us consider, for example, the case of marketing a new drug: in this field it is easy to understand the delicate nature of the situation, and the need to find alternative solutions to those already present for the purposes of comparison. Indeed, a new drug is presented onto the market, there is the need to show that the new product is not only better than the standard treatment (if one exists), but also that it is "at least as good as" the best drug already on the market (i.e. the control treatment).

2.1. The non-inferiority margin

In a clinical trial in order to show that a new experimental therapy T is "at least as good as" a standard therapy (control treatment) C, we need a hypotheses testing approach. The term "at least as good as" implies equivalent but not necessary superior effectiveness. In a non-inferiority test the non-inferiority of the new treatment is proved if the null hypothesis that the control treatment is better than the new treatment, for a given non-inferiority margin d, is rejected. Formally:

$$H_0 : m_C - m_T \geq d \quad (C \text{ is superior to } T) \tag{1}$$

against

$$H_1 : m_C - m_T < d \quad (T \text{ is not inferior to } C) \tag{2}$$

where m_C and m_T represent the mean values of the response variables for the control treatment and the new treatment respectively and constant d (>0) indicates the non-inferiority margin. The acceptance of the null hypothesis asserts the superiority of the control treatment over the new treatment. Instead, the alternative hypothesis establishes the non-inferiority of the experimental drug compared to the control drug.

The hypotheses can be tested through a one-sided test with a significance level a , using the upper bound of a symmetric $(1-2a)$ confidence interval (D'Agostino *et al*., 2003). Alternatively a one-sided $(1-a)$ confidence interval for the difference m_C - m_T can be used. In this case the null hypothesis is rejected when the upper bound of the interval is lower than d. For example, for a given d, the traditional t test for comparing the means of two independent populations rejects the null hypothesis (superiority of C) if $TS_1 = \left(\overline{X}_C - \overline{X}_T - d\right) / \hat{s}_{C-T} < -t_{m;a}$, where $\overline{X}_C$ and $\overline{X}_T$ are the sample means of the response variable for C and T respectively, $\hat{s}_{C-T}$ is the standard error of $\overline{X}_C - \overline{X}_T$ and $t_{m;a}$ is the quantile of the Student's t distribution with $m=n_C+n_T-2$ degrees of freedom, which corresponds to a cumulative probability of $1-a$ and n_C

and n_T are the sample sizes for the control group and the treated group respectively. Under H_0, the equality:

$$\text{a} = \Pr\{TS_1 < -t_{m;a}\} = \Pr\{\overline{X}_C - \overline{X}_T + t_{m;a}\hat{s}_{C-T} < d\}$$

is true and it implies that the one-sided $(1-a)$ confidence interval for $m_C - m_T$ is $(-\infty, \overline{X}_C - \overline{X}_T + t_{m;a}\hat{s}_{C-T})$ and the equality:

$$1-2a = \Pr\{|TS_1| < t_{m;a}\} = \\ = \Pr\{\overline{X}_C - \overline{X}_T - t_{m;a}\hat{s}_{C-T} < d < \overline{X}_C - \overline{X}_T + t_{m;a}\hat{s}_{C-T}\}$$

implies that the two-sided $(1-2a)$ confidence interval for $m_C - m_T$ is $(\overline{X}_C - \overline{X}_T - t_{m;a}\hat{s}_{C-T}, \overline{X}_C - \overline{X}_T + t_{m;a}\hat{s}_{C-T})$. Hence, the rejection rule of the one-sided *t* test is equivalent to the condition that the upper bound of the one-sided $(1-a)$ confidence interval is lower than d. Of course this upper bound is equal to the upper bound of the two-sided $(1-2a)$ confidence interval.

The definition of the non-inferiority margin d must follow strict rules, set and supervised by EMEA[1] (see EMEA, 2004). The non-inferiority margin d should be defined as a fraction or percentage of the standard treatment's effectiveness and this effectiveness should be documented by historical placebo-controlled trials. Hence the determination of d is based on both statistical reasoning and clinical judgements and it depends on how much the active control effect exceeds the placebo effect, calculated in historical studies. The non-inferiority margin cannot be greater than the smallest effect size from the comparison between control treatment and placebo and it reflects uncertainty due to the variability of the estimated effectiveness of *C* in the historical placebo-controlled trials. This uncertainty must be taken into consideration when the probability of the non-inferiority test's type I error is calculated.

---

[1] The EMEA (European Medicines Agency) was founded in 1995 by the European Union: its main responsibility is the evaluation and supervision of medicines for human and veterinary use.

The value of the non-inferiority margin can be defined in the following way (D'Agostino *et al*., 2003):

- test the superiority of control treatment C over placebo P through other studies in the literature, and make the assumption of constancy of the control effect over time;
- calculate d as a percentage of the difference between active control effect m_C and placebo effect m_P. In general, this percentage is taken between 20% and 50%, according to clinical evaluations.

Let us now adopt a relative risk notation letting p_T, p_C and p_P be the incidence ratios of a clinical event in the population from which patients treated respectively with experimental drug T, control treatment C and placebo P have been sampled. The non-inferiority test is then formalized as follows:

$$H_0 : p_T / p_C \geq d \quad (C \text{ is superior to } T) \tag{3}$$

against

$$H_1 : p_T / p_C < d \quad (T \text{ is not inferior to } C). \tag{4}$$

More specifically, the non-inferiority margin can be calculated as percentage $100g$ of the control effect:

$$p_P - p_T > g(p_P - p_C) \Leftrightarrow 1 - p_T / p_P > g(1 - p_C / p_P)$$

$$\Leftrightarrow p_T / p_C < 1 + (1-g)(p_P / p_C - 1)$$

where constant g and p_T, p_C, p_P take value in the interval [0, 1]. The non-inferiority margin is therefore given by $d = 1 + (1-g)(p_P / p_C - 1)$. Working in terms of logarithms of the treatment effect, the test's hypothesis becomes:

$$H_0' : \log(p_T) - \log(p_C) \geq d \tag{5}$$

against

$$H_1' : \log(p_T) - \log(p_C) < d \tag{6}$$

or equivalently:

$$H_0' : [\log(p_P) - \log(p_T)]/[\log(p_P) - \log(p_C)] \leq g \tag{7}$$

against

$$H_1' : [\log(p_P) - \log(p_T)]/[\log(p_P) - \log(p_C)] > g \tag{8}$$

and therefore the non-inferiority margin is given by $d = (1-g)[\log(p_P) - \log(p_C)]$.

Following both of the above described procedures, to determine the non-inferiority margin, an estimate for the difference between the effectiveness of the active control and the effectiveness of the placebo must be calculated. It is often better, therefore, not to use the simple point estimate of such difference, but instead, in a more conservative way, the lower bound of the confidence interval obtained from other studies in the literature.

2.2 Methodological solutions for the non-inferiority test

Several approaches have been discussed in the literature to test the non-inferiority hypothesis. An interesting testing procedure was proposed by Blackwelder (see e.g. Blackwelder, 1982) - it considers null and alternative hypotheses in the case of higher mean values indicating greater effectiveness of the drug, as in (1) and (2).

Here we will consider a more powerful version (Hung *et al*., 2003) of this test, which is defined in terms of ratio of means. In the case of increments in the means indicating an improvement in the effectiveness of the drug, the ratio version of the null and alternative hypothesis is:

$$H_0 : m_T / m_C \leq R \tag{9}$$

against

$$H_1 : m_T / m_C > R \tag{10}$$

where R (<1) indicates the non-inferiority margin, expressed in terms of lower bound for the percentage of the control drug effectiveness that we wish to preserve using the new drug to maintain its non-inferiority in relation to the control treatment (Hung *et al*., 2003).

The condition R<1 is related to the fact that T is considered non-inferior to C if the new treatment effect is greater than a percentage 100R of the active control effect. Two decision methods can be used, i.e. the standard test statistic or the confidence interval. With the former, an unbiased, uniformly most powerful test can be constructed by means of the following test statistic (Laster *et al*., 2003):

$$TS_2 = \left(\overline{X}_T - R\overline{X}_C\right) / \left[S^2\left(1 + R^2\right) / n\right]^{1/2}$$

where S^2 is an unbiased estimator of the variance of the two samples. It can be shown that this test statistic follows a Student's t distribution with $(2n - 2)$ degrees of freedom.

Alternatively, a confidence interval for the quantity of interest can be constructed. The interval is based on the following pivotal quantity:

$$\left(\hat{R} - R\right) / \left[S^2\left(1 + R^2\right) / n\left(\overline{X}_C\right)^2\right]^{1/2}.$$

For large sample sizes the approximated confidence interval for the real value of R is:

$$\hat{R} \pm z_a \left[S^2\left(1 + \hat{R}^2\right) / n\left(\overline{X}_C\right)^2\right]^{1/2}.$$

2.3 Standard test statistics against confidence intervals

Generally, uncertainty regarding the non-inferiority margin estimate has an effect on the type I error probability. This effect increases dangerously when using the standard test statistic if the control drug effect's condition of constancy over time cannot be assumed, but can be controlled using confidence intervals. In any case, if the constancy condition holds, the use of the test statistic allows us to work with an exact value of the type I error probability. Let us consider the relative risk formalisation for the null and alternative hypotheses shown in (5) and (6).

Let $\hat{p}_C$ and $\hat{p}_P$ indicate the point estimates of incidence ratios in the populations treated with C and P respectively, and $\hat{p}_{C_0}$ and $\hat{p}_{P_0}$ the point estimates from studies in the literature. Moreover, let $s_{T\text{-}C}$ and $s_{P\text{-}C0}$ denote the standard errors of $\log(\hat{p}_T) - \log(\hat{p}_C)$ and $\log(\hat{p}_{P_0}) - \log(\hat{p}_{C_0})$ respectively.

Working with the confidence interval method, it may be that:

- the null hypothesis (5) is rejected when the upper bound of the confidence interval for $\log(p_T)$-$\log(p_C)$ is less than the lower bound of the confidence interval for δ, i.e. when:

$$\log(\hat{p}_T) - \log(\hat{p}_C) + z_a s_{T-C} < (1-g)\left|\log(\hat{p}_{P_0}) - \log(\hat{p}_{C_0}) - z_a s_{P-C0}\right|.$$

In this case the maximum type I error probability is given by:

$$a' = \Pr\left\{\frac{\log(\hat{p}_T) - \log(\hat{p}_C) - (1-g)\left[\log(\hat{p}_{P_0}) - \log(\hat{p}_{C_0})\right]}{\sqrt{s^2_{T-C} + (1-g)^2 s^2_{P-C0}}} < -z_a a\right\} =$$

$$= \Phi(-z_a a),$$

where $\Phi(\bullet)$ represents the distribution function of a standard normal variable, and parameter a is given by

$a = \left[s_{T-C} + (1-g)s_{P-C0}\right] / \left[s^2_{T-C} + (1-g)^2 s^2_{P-C0}\right]^{1/2}$ which is always greater than one. Therefore we have $a = \Phi(-z_a) > \Phi(-z_a a) = a'$. This means that the value of the type I error probability calculated taking into consideration the uncertainty regarding the non-inferiority margin estimate, is always below the nominal value a;

- by comparing the confidence interval for $\log(p_T)$-$\log(p_C)$ with the point estimate of δ, the rejection rule becomes:

$$\log(\hat{p}_T) - \log(\hat{p}_C) + z_a s_{T-C} < (1-g)\left[\log(\hat{p}_{P_0}) - \log(\hat{p}_{C_0})\right]$$

and so the maximum type I error probability in this case is given by:

$$a' = \Phi(-z_a b),$$

where $b = s_{T-C} / \left[s^2_{T-C} + (1-g)^2 s^2_{P-C0}\right]^{1/2}$ and b is always positive and less than one. Therefore we have $a = \Phi(-z_a) > \Phi(-z_a b) = a'$. This means that here also the actual value of the type I error probability, calculated taking into consideration the uncertainty regarding the non-inferiority margin estimate, is always lower than the nominal a-level. Moreover, it can be noted that $a > b$, hence $\Phi(-z_a a) < \Phi(-z_a b)$, i.e. the first approach is more conservative than the second;

- the value of the non-inferiority margin that leads to a value of the type I error probability equal to the nominal value a, can be calculated as:

$$d^* = -z_a \left\{\left[s^2_{T-C} + (1-g)^2 s^2_{P-C0}\right]^{1/2} - s_{T-C}\right\} + (1-g)\left[\log(\hat{p}_{P_0}) - \log(\hat{p}_{C_0})\right],$$

since in this case the maximum type I error probability is given by:

$$a' = \Pr\left\{\log(\hat{p}_T) - \log(\hat{p}_C) + z_a s_{T-C} < d^*\right\} =$$

$$= \Pr\left\{ \frac{\log(\hat{p}_T) - \log(\hat{p}_C) - (1-g)\left[\log(\hat{p}_{P_0}) - \log(\hat{p}_{C_0})\right]}{\sqrt{s^2_{T-C} + (1-g)^2 s^2_{P-C0}}} < -z_a \right\}$$

$$= \Phi(-z_a) = a \ .$$

Note that this value for the non-inferiority margin depends on the quantity $s_{T\text{-}C}$, which is the standard error of the new treatment effect compared to the control drug, hence this quantity depends only on the current study and, in particular, is strongly related to the sample size in the present experiment. This clearly contradicts the ethical rules that should be followed in a clinical study since the non-inferiority margin can be decided a priori.

If the constancy condition of the active control effect holds, the so-called "preservation test" (Laster *et al*., 2003) can be used. The test statistic is:

$$Z_{pv} = \frac{\log(\hat{p}_T) - \log(\hat{p}_C) - (1-g)\left[\log(\hat{p}_{P_0}) - \log(\hat{p}_{C_0})\right]}{\sqrt{s^2_{T-C} + (1-g)^2 s^2_{P-C0}}},$$

and the null hypothesis is rejected when $Z_{pv} < -z_a$.

Note that this rejection rule is mathematically equivalent to the rule from the confidence interval with value d^* chosen for the non-inferiority margin. Indeed, in this case the maximum type I error probability is exactly equal to a, but the interpretation of the non-inferiority margin is different from the case of confidence intervals. Here the non-inferiority margin is considered to be a fixed quantity, and the problem related to the uncertainty of its estimate is not taken into consideration. Thus, this method controls the type I error probability only if the constancy condition holds, otherwise the actual value of a exceeds the nominal level.

2.4 The TACT method

The Two-stage Active Control Testing (TACT) method (Wang *et al*., 2003) was developed to solve the problem of controlling the type I error probability, taking into account the condition of constancy over time of the control treatment effect. The procedure consists in applying the preservation test or the confidence interval method depending on possible validity of the constancy condition of the active control effect. Let us consider the following model for the new treatment effect, expressed in terms of logarithm of the relative risk:

$$\log(event-incidence) = m + b_{C_0} Y_{C0} + b_C Y_C + b_T Y_T ,$$

where $m = \log(p_P) = \log(p_{P_0})$ is the placebo effect; $b_{C_0} = \log(p_{C_0}/p_P)$ is the control effect in relation to the placebo effect from the literature; $b_C = \log(p_C/p_P)$ is the control effect in relation to the placebo effect in the present study; $b_T = \log(p_T/p_P)$ is the new drug effect in relation to the placebo effect; Y_{C0}, Y_C and Y_T are indicator variables of the treatments.

According to (5) and (6), the hypotheses of the non-inferiority test can be formalised as:

$$H_{0g} : b_T - b_C \geq d \tag{11}$$

against

$$H_{1g} : b_T - b_C < d \tag{12}$$

where $d = -(1-g)b_C$ is the non-inferiority margin and g is the percentage of the control effect that we wish the new drug to preserve.
The TACT method consists of two steps. Firstly, the control treatment's superiority over the placebo must be proved through data from the literature. Secondly, if the active control is shown to be superior to the placebo from the collective evidence of the historical placebo controlled

trials, data from the literature and from the current study must be compared to evaluate the presence or absence of the condition of constancy over time of the control treatment effect. In other words, the next phase consists of testing whether the control effect is much less in the active controlled trial population than in the historical trial populations (that is $b_C >> b_{C_0}$).

Let us suppose that t historical studies are available, p_{C_t} indicates the control effect at time t ($t = 0, 1,\ldots, t-1$) and $p_{Ct} = p_C$ denotes the control effect in the present study. The test on the constancy condition consists of a sequential test that must be carried out using a set of test statistics such as the following:

$$Z_{Ct} = \frac{\log(\hat{p}_{C_t}) - \log(\hat{p}_{C_0})}{s_t} \quad (t = 1, 2,\ldots, t),$$

where $\hat{p}_{C_t}$ is the estimated effect of the control treatment at generic time t and s_t is the standard error of the quantity at the numerator. The constancy condition is rejected for time t if $Z_{Ct} > U_t$, where U_t is a fixed value which is chosen in line with clinical and statistical considerations. In this paper, we decided to use t tests for testing for the constancy condition hypothesis. In order to adjust the p-values for multiplicity, the Bonferroni-Holm method was used (see section 3).

Let us consider two different significance levels, $\alpha^* < \alpha^{**}$, to distinguish the situations of strong and weak significance. According to the result of the testing procedure three different decisions can be taken:

a) if the test leads to the rejection of the presence of the constancy condition at any time t (p-value $< \alpha^*$, strong significance) in favour of the hypothesis $b_C >> b_{C_0}$, the experiment must be abandoned because of the above mentioned considerations regarding the control of the type I error probability;
b) if the hypothesis of constancy is weakly rejected, i.e. the p-value of the test is less than α^{**} but not less than α^*, we can "weakly" assess that the constancy over time condition holds provided that the more

conservative confidence interval method is applied for the non-inferiority test;

c) if the constancy condition hypothesis cannot be rejected, the preservation testing method can be applied.

2.5 A non-parametric solution

As is known, the normality assumption cannot always be made in our context, and in such cases the application of parametric procedures may be inappropriate, even approximately. For this reason we are proposing a non-parametric permutation alternative for the non-inferiority issue.

The idea is to consider both permutation versions and permutation confidence intervals (Pesarin, 2001) for all the proposals from the literature that have already been presented (i.e. standard test statistics, confidence intervals and TACT method). Independently of how the effectiveness of the treatment is measured (mean values, incidence ratios of clinical events, logarithmic transformation of incidence ratios, etc.), let us consider a general notation in which parameters q_T, q_C and q_P represent the effectiveness of the new treatment, control treatment and placebo. Let us assume that greater parameter values are linked to higher effectiveness. Null and alternative hypotheses of the non-inferiority test can be formulated as:

$$H_{0g} : q_C - q_T \geq d \tag{13}$$

against

$$H_{1g} : q_C - q_T < d \ . \tag{14}$$

The permutation approach is based on the concept of exchangeability of data between the two samples under H_0 (Pesarin, 2001). When standard test statistics are used, the permutation solution can be summarized as follows:

a) firstly we calculate the value of the test statistic with the original sample data:

$$Z_{pv}^{oss} = \hat{q}_C - \hat{q}_T - (1-g)\left(\hat{q}_{C_0} - \hat{q}_P\right);$$

b) we perform B Monte Carlo iterations from the permutation sample space of the two samples and for each permutation we calculate the test statistic:

$$Z_{pv}^{*} = \hat{q}_C^{*} - \hat{q}_T^{*} - (1-g)\left(\hat{q}_{C_0} - \hat{q}_P\right);$$

c) we estimate the p-value of the test as:

$$\hat{l} = \#\left(Z_{pv}^{*} \le Z_{pv}^{oss}\right)/B\,.$$

If we consider the construction of confidence intervals, the non-parametric approach consists in calculating and comparing two permutation confidence bounds. In order to calculate the upper bound for the quantity $q_C - q_T$, we compute the permutation distribution of the corresponding standardized statistic. Then we identify the $(1-a)\cdot 100$ quantile u^* of the distribution and we estimate the upper bound of the confidence interval as $\hat{q}_C - \hat{q}_T + u^* s_{C-T}$, where $s_{C\text{-}T}$ indicates the standard error of $\hat{q}_C - \hat{q}_T$. In a similar way, we obtain the lower bound as $\left(\hat{q}_{C_0} - \hat{q}_P\right) - l^* s_{C0-P}$ for the quantity $\left(q_{C_0} - q_P\right)$, where l^* is the $a \cdot 100$ quantile of the standardised permutation distribution of $\left(\hat{q}_{C_0}^{*} - \hat{q}_P^{*}\right)$ and $s_{C0\text{-}P}$ is the standard deviation. The permutation confidence interval will lead to the rejection of the null hypothesis H_{0g} when $\hat{q}_C - \hat{q}_T + u^* s_{C-T} < (1-g)\left(\hat{q}_{C_0} - \hat{q}_P - l^* s_{C0-P}\right)$.

With the TACT method, we also propose a non-parametric approach that consists in the application of the permutation version of the two-sample t test in the first two steps of the TACT procedure, and the combination of the p-values to obtain a multiple test for the hypothesis

of consistency of the control effect over time. In case of rejection of the null hypothesis (i.e. the control effect is constant) it is possible to adjust *p*-values for multiplicity in order to detect which one of the *t* statistics provides a non significant *p*-value. In the third step of the procedure the permutation TACT method is performed using the above described permutation procedure.

3. Simulation study

A simulation study was carried out to evaluate the performances of the three considered procedures. The study was organized in the following way: three historical studies were simulated to compare the control effect and the placebo effect, and to evaluate the presence of the control drug effect's condition of constancy over time. A fourth study was simulated to represent the current study. Simulated data came firstly from normal distributions and then from exponential distributions, in order to check a possible difference in performances between parametric and non-parametric solutions.

3.1 Simulations from normal distributions

According to conditions illustrated in section 2.5 and in order to test (13) against (14), let us use X_T, X_P and X_{Ct} to indicate the response variables of the population treated with the new drug, the placebo and the control at time t (t = 0,…,3) respectively. Let us assume that $X_T \sim N(q_T,0.5)$, $X_P \sim N(q_P,0.5)$ and $X_{Ct} \sim N(q_{Ct},0.5)$, t=0,…,3. Two alternative settings for mean values have been considered and reported in Table 1.

Note that $t = 3$ represents the current study, and that we considered a value of 0.80 for γ. We worked with 1000 Monte Carlo iterations and 500 permutations of the data for the non-parametric permutation procedures. The simulations were performed using R code which is available on request from the authors. The point estimates for the mean parameters are the usual sample means.

Table 1. Simulation settings

	setting 1						setting 2					
q_{C0}	5	5	5	5	5	5	4.7	4.7	4.7	4.7	4.7	4.7
q_{C1}	5	5	5	5	5	5	4.8	4.8	4.8	4.8	4.8	4.8
q_{C2}	5	5	5	5	5	5	4.9	4.9	4.9	4.9	4.9	4.9
q_{C3}	5	5	5	5	5	5	5	5	5	5	5	5
q_P	3	3	3	3	3	3	3	3	3	3	3	3
q_T	5.6	5.2	5	4.8	4.6	4.4	5.51	5.17	5	4.83	4.66	4.49
$q_T - q_C$	0.6	0.2	0	-0.2	-0.4	-0.6	0.51	0.17	0	-0.17	-0.34	-0.51
$q_P - q_{C0}$	-2	-2	-2	-2	-2	-2	-1.7	-1.7	-1.7	-1.7	-1.7	-1.7
δ	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.34	-0.34	-0.34	-0.34	-0.34	-0.34
hypothesis	H_1	H_1	H_1	H_1	H_1	H_0	H_1	H_1	H_1	H_1	H_1	H_0

Some parametric and permutation results for a=0.5 with simulations from the normal distribution are shown in Table 2. As regards the behaviour of the estimated type I error probability in the case of testing and confidence interval methods, under the condition of constancy (first setting) the standard testing method controls the type I error probability at the nominal level, while the confidence interval method always provides a lower value of the estimated type I error probability (under H_0) and of the estimated power (under H_1). Instead, when the constancy condition does not hold (second setting), the testing method does not control the type I error probability at its nominal value, while the confidence interval method does. For the TACT method, we also report the number of abandoned studies ("Left studies"). Furthermore, the confidence interval method in this case is less powerful than the standard testing method. The TACT method results show how it represents a good compromise between the two standard approaches, reporting good power and controlling the type I error probability whether or not the constancy condition holds.

Table 2. Results of simulations from normal distribution: rejection rates (a=0.05).

n=20

Setting 1

	Parametric procedures				Permutation procedures			
$q_T - q_C$ (% of $q_P - q_{C0}$)	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.4 (-20%) H0	0.047	0.032	0.043	51	0.050	0.034	0.052	30
-0.3 (-15%) H1	0.169	0.125	0.161	47	0.160	0.123	0.160	32
-0.2 (-10%) H1	0.341	0.288	0.338	55	0.345	0.281	0.350	36
0 (0%) H1	0.800	0.744	0.804	48	0.803	0.737	0.803	37
0.2 (+10%) H1	0.986	0.975	0.987	50	0.985	0.973	0.984	47
0.6 (+30%) H1	1.000	1.000	1.000	52	1.000	1.000	1.000	45

Setting 2

$q_T - q_C$ (% of $q_P - q_{C0}$)	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.34 (-20%) H0	0.061	0.044	0.058	169	0.061	0.042	0.065	340
-0.255 (-15%) H1	0.185	0.139	0.171	204	0.187	0.135	0.202	347
-0.17 (-10%) H1	0.317	0.261	0.315	182	0.308	0.245	0.334	339
0 (0%) H1	0.739	0.672	0.712	208	0.728	0.665	0.785	340
0.17 (+10%) H1	0.949	0.928	0.943	212	0.982	0.917	0.955	350
0.51 (+30%) H1	1.000	1.000	1.000	197	1.000	1.000	1.000	342

n=30

Setting 1

	Parametric procedures				Permutation procedures			
$q_T - q_C$ (% of $q_P - q_{C0}$)	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.4 (-20%) H0	0.044	0.035	0.042	41	0.047	0.031	0.052	32
-0.3 (-15%) H1	0.201	0.153	0.198	47	0.202	0.140	0.217	29
-0.2 (-10%) H1	0.463	0.396	0.461	56	0.462	0.367	0.482	46
0 (0%) H1	0.915	0.888	0.920	47	0.918	0.872	0.921	36
0.2 (+10%) H1	0.997	0.997	0.998	46	0.998	0.996	0.997	30
0.6 (+30%) H1	1.000	1.000	1.000	51	1.000	1.000	1.000	42

Setting 2

$q_T - q_C$ (% of $q_P - q_{C0}$)	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.34 (-20%) H0	0.076	0.053	0.076	276	0.071	0.050	0.090	466
-0.255 (-15%) H1	0.197	0.157	0.183	261	0.203	0.146	0.237	455
-0.17 (-10%) H1	0.431	0.361	0.416	265	0.434	0.358	0.476	468
0 (0%) H1	0.860	0.811	0.825	280	0.853	0.787	0.869	472
0.17 (+10%) H1	0.988	0.984	0.985	287	0.983	0.981	0.993	447
0.51 (+30%) H1	1.000	1.000	1.000	273	1.000	1.000	1.000	450

n=100

Setting 1

	Parametric procedures				Permutation procedures			
$q_T - q_C$ (% of $q_P - q_{C0}$)	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.4 (-20%) H0	0.055	0.036	0.055	51	0.057	0.033	0.058	32
-0.3 (-15%) H1	0.406	0.341	0.403	43	0.431	0.319	0.428	17
-0.2 (-10%) H1	0.883	0.850	0.879	47	0.899	0.825	0.900	41
0 (0%) H1	1.000	1.000	0.995	51	1.000	1.000	1.000	39
0.2 (+10%) H1	1.000	1.000	1.000	48	1.000	1.000	1.000	28
0.6 (+30%) H1	1.000	1.000	1.000	50	1.000	1.000	1.000	45

Setting 2

$q_T - q_C$ (% of $q_P - q_{C0}$)	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.34 (-20%) H0	0.101	0.073	0.099	718	0.108	0.065	0.161	944
-0.255 (-15%) H1	0.427	0.349	0.429	741	0.447	0.335	0.625	952
-0.17 (-10%) H1	0.862	0.809	0.833	743	0.862	0.793	0.977	957
0 (0%) H1	1.000	0.999	1.000	752	1.000	1.000	1.000	956
0.17 (+10%) H1	1.000	1.000	1.000	738	1.000	1.000	1.000	953
0.51 (+30%) H1	1.000	1.000	1.000	723	1.000	1.000	1.000	950

Table 3. Results of simulations from exponential distribution: rejection rates (a=0.05).

n=20

Setting 1

$q_T - q_C$ (% of $q_P - q_{C0}$)	Parametric procedures				Permutation procedures			
	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.4 (-20%) H0	0.063	0.050	0.060	45	0.070	0.043	0.070	39
-0.3 (-15%) H1	0.069	0.052	0.071	51	0.080	0.050	0.082	36
-0.2 (-10%) H1	0.073	0.051	0.069	44	0.076	0.041	0.078	40
0 (0%) H1	0.095	0.071	0.092	53	0.100	0.067	0.104	87
0.2 (+10%) H1	0.111	0.087	0.112	38	0.119	0.072	0.119	29
0.6 (+30%) H1	0.162	0.127	0.158	48	0.165	0.143	0.167	47

Setting 2

$q_T - q_C$ (% of $q_P - q_{C0}$)								
	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.34 (-20%) H0	0.069	0.048	0.069	43	0.077	0.048	0.079	46
-0.255 (-15%) H1	0.069	0.051	0.069	35	0.075	0.044	0.073	43
-0.17 (-10%) H1	0.067	0.047	0.066	39	0.077	0.039	0.078	46
0 (0%) H1	0.089	0.057	0.087	53	0.097	0.048	0.100	48
0.17 (+10%) H1	0.121	0.094	0.119	43	0.128	0.082	0.128	51
0.51 (+30%) H1	0.160	0.121	0.160	49	0.172	0.150	0.174	48

n=30

Setting 1

$q_T - q_C$ (% of $q_P - q_{C0}$)	Parametric procedures				Permutation procedures			
	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.4 (-20%) H0	0.057	0.040	0.054	39	0.061	0.038	0.061	38
-0.3 (-15%) H1	0.068	0.047	0.063	47	0.071	0.041	0.072	45
-0.2 (-10%) H1	0.063	0.038	0.062	44	0.070	0.034	0.069	33
0 (0%) H1	0.090	0.070	0.084	55	0.093	0.063	0.089	77
0.2 (+10%) H1	0.121	0.100	0.118	46	0.128	0.085	0.124	34
0.6 (+30%) H1	0.189	0.144	0.178	53	0.184	0.169	0.180	47

Setting 2

$q_T - q_C$ (% of $q_P - q_{C0}$)								
	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.34 (-20%) H0	0.056	0.040	0.057	42	0.064	0.038	0.066	48
-0.255 (-15%) H1	0.069	0.052	0.065	32	0.079	0.048	0.078	39
-0.17 (-10%) H1	0.055	0.044	0.052	41	0.065	0.038	0.065	46
0 (0%) H1	0.081	0.059	0.081	46	0.089	0.045	0.091	46
0.17 (+10%) H1	0.096	0.072	0.095	51	0.100	0.065	0.100	50
0.51 (+30%) H1	0.150	0.115	0.150	48	0.154	0.132	0.155	47

n=100

Setting 1

$q_T - q_C$ (% of $q_P - q_{C0}$)	Parametric procedures				Permutation procedures			
	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.4 (-20%) H0	0.044	0.033	0.042	54	0.049	0.033	0.049	32
-0.3 (-15%) H1	0.061	0.042	0.059	35	0.069	0.036	0.071	24
-0.2 (-10%) H1	0.092	0.069	0.087	45	0.098	0.056	0.099	33
0 (0%) H1	0.135	0.113	0.129	40	0.149	0.102	0.149	65
0.2 (+10%) H1	0.204	0.165	0.208	41	0.223	0.151	0.226	25
0.6 (+30%) H1	0.359	0.306	0.359	49	0.363	0.311	0.361	27

Setting 2

$q_T - q_C$ (% of $q_P - q_{C0}$)								
	TEST	CI	TACT	Left studies	TEST	CI	TACT	Left studies
-0.34 (-20%) H0	0.050	0.038	0.045	54	0.059	0.035	0.057	53
-0.255 (-15%) H1	0.065	0.047	0.061	47	0.081	0.048	0.079	56
-0.17 (-10%) H1	0.081	0.059	0.080	68	0.088	0.053	0.090	71
0 (0%) H1	0.122	0.095	0.122	58	0.132	0.088	0.136	66
0.17 (+10%) H1	0.180	0.143	0.176	51	0.186	0.142	0.185	58
0.51 (+30%) H1	0.323	0.281	0.318	57	0.332	0.319	0.331	54

Both parametric and permutation results confirm that the number of left studies with the non-parametric approach is lower than the number using a parametric method in setting 1 and greater in setting 2. This emphasizes the permutation method's higher sensitivity to the condition of constancy over time, and it can be seen that the permutation approach works in a very similar way to the parametric approach, reporting power values that are very close to those of the parametric approach.

3.2 Simulations from exponential distribution

In the exponential case we simulated data by considering the same settings for the means used for the normal case. Note that the parameter of the exponential distribution is the inverse of the mean. Parametric and permutation results with simulations from the exponential distribution are reported in Table 3. Note that the power of the different methods is lower here than in the results with normal data. This is due to the particular relationship between the mean and the variance of the exponential distribution. In fact, the standard deviation is equal to the mean and for the means' chosen settings, the variances are larger than in the normal case. In any case, general better behaviour of non-parametric approaches can be seen in the reported results.

3.3 Graphical comparisons

To better display possible differences between the three considered methods, Figures 1 and 2 report the empirical power curves for TACT, testing and confidence interval methods in respectively their parametric and non-parametric versions (see Figure 1). Note that testing and TACT methods work better than the confidence interval method for both parametric and permutation.

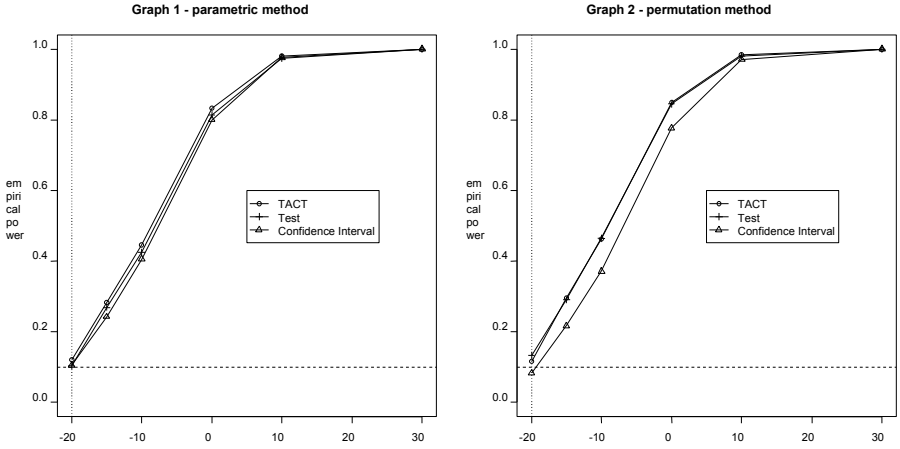

Figure 1. Empirical power of the three methods with normal data (n=20, α=0,10)

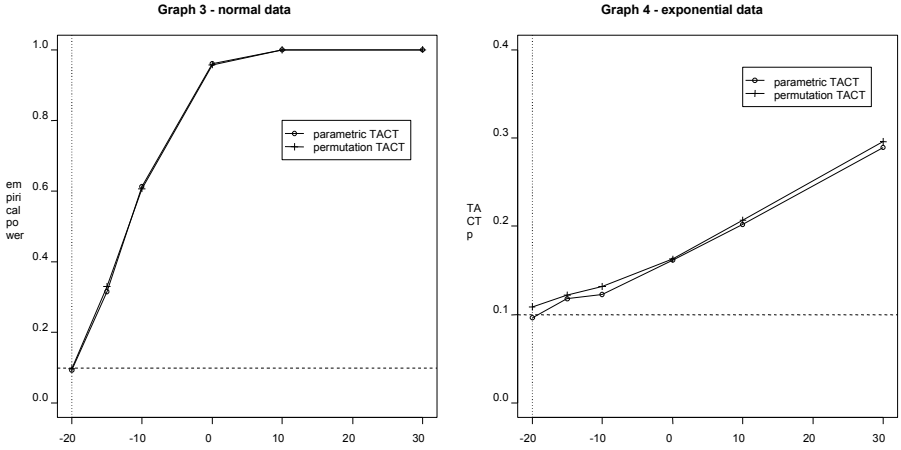

Figure 2. Empirical power of the TACT method with normal and exponential data (n=30, α=0,10)

Figures 3 and 4 display the differences between parametric and permutation solutions in the case of normally distributed or exponential data respectively (see Figure 2). Both graphs confirm the good behaviour of both versions of the TACT method, and show the similarity of behaviour of the parametric version and its permutation counterpart.

4. Real case application

A real case application of the parametric and non-parametric TACT method was carried out with reference to a randomized clinical trial. The data refers to a new treatment for temporomandibular disorders (TMD). In this field, because of medical preference for conservative treatment procedures, rather than non-reversible and invasive therapies, temporomandibular joint (TMJ) surgical interventions are limited to a small proportion of cases, and infiltrations are usually used with caution.

Arthrocentesis is the simplest and least invasive of these treatments and its use seems to be primarily indicated for disc displacement without reduction. Specifically, our data refers to a technique involving the injection of hyaluronic acid at the end of the articular lavage. In particular, we wish to compare a new treatment that consists in the use of two needles for the injection of hyaluronic acid to the standard control technique that uses only one. For the placebo treatment we considered the standard injection of physiological solution instead of hyaluronic acid. We observed the masticatory capability at the first control visit after a cycle of five infiltrations as our endpoint of interest; a score of 1 to 10 was given, with high values representing better responses to the treatment.

In our experiment we considered a historical study for the comparison of the placebo and control treatments. The present study collected data regarding the control treatment and the new technique. Our samples were unbalanced - in the present study we observed 12 patients treated with the new technique (denoted by T in Figure 3) and 16 treated with the standard method (denoted by C_1 in Figure 3); in

the historical study 13 patients treated with physiological solution were observed (denoted by P in Figure 3) and 20 using the control method (denoted by C_0 in Figure 3).

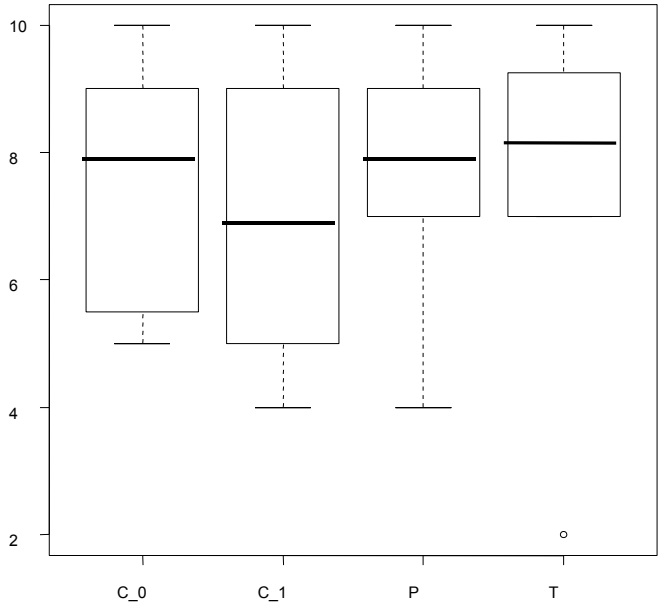

Figure 3. Frequency distributions of response variable

Note that the four box plots in Figure 3 show no apparent differences among the distributions, and in particular the placebo effect seems to be quite good. This implies an estimate of the non-inferiority margin close to zero.

We fixed a significance level a = 0.05, and the permutation results were computed with 10,000 permutations of the data. In this case, because of the presence of only one historical study, the tests for the evaluation of the constancy condition over time are reduced to only one test that compares the control treatment performances in the present study and in the historical study. Therefore, the decision as to which

method should be used, either the standard test statistic or the confidence interval method, is based on the *p*-value of this test. With the parametric and non-parametric approach, this test provided a *p*-value of 0.5877 and 0.6014 respectively. Both procedures therefore favour the use of the test statistic method, and the final *p*-values are respectively 0.1831 for the parametric approach and 0.1828 for the permutation approach. Both approaches lead to the acceptance of the null hypothesis of superiority of the control treatment over the new treatment. Note that the results of the two approaches are very close to each other.

5. Conclusions

The TACT method is a good solution for the evaluation of the non-inferiority of a new treatment over a control in active-controlled trials. Indeed, the use of the test statistic method permits a higher probability of establishing equivalence or non-inferiority than the confidence interval approach when the non-inferiority margin is defined as the lower bound of the control treatment effect's confidence interval when the control treatment effect's condition of constancy over time does not hold. This solution does not control the probability of false positives (type I error probability) at its nominal level. The permutation version of the TACT method proposed in this paper is a good solution especially when normal distribution of the data cannot be assumed. The simulation study has shown that the non-parametric approach works in a similar fashion to the parametric version in the case of normally distributed data, with little loss of power, but the permutation version is more powerful in the case of exponentially distributed data.

It is worth noting that similar results have been obtained with other non-normal distributions.

References

Angell M. (1997), The ethics of clinical research in the Third Word (editorial), *New England Journal of Medicine*, 337, 847-849.

Blackwelder W.C. (1982), 'Proving the null hypothesis' in clinical trials, *Controlled Clinical Trials*, 3, 345-353.

Blackwelder W.C., Chan M.A. (1984), Sample size graphs: for 'proving the null hypothesis', *Controlled Clinical Trials*, 5, 97-105.

D'Agostino R.B. (2003), Non-inferiority trials: advances in concepts and methodology, *Statistics in Medicine*, 22, 165-167.

D'Agostino R.B., Massaro J.M., Sullivan L.M. (2003), Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics, *Statistics in Medicine*, 22, 169-186.

Durrleman S., Simon R. (1990), Planning and monitoring of equivalence studies, *Biometrics*, 46(2), 329-336.

EMEA (2004), Points to consider on the choice of non-inferiority margin, London.

Fisher L.D. (1999), Advances in clinical trials in twentieth century, *Annual Review of Public Health*, 20, 109-124.

Fleming T.R. (1987), Treatment evaluation in active control studies, *Cancer Treatment Reports*, 71, 1061-1064.

Fleming T.R. (2000), Design and interpretation of equivalence trials, *American Heart Journal,* 139, 171-176.

Foulds G.A. (1958), Clinical research in psychiatry, *Journal of Mental Science*, 104, 259-265).

Freedman B., Weijer C., Glass K.C. (1996a), Placebo orthodoxy in clinical research I: Empirical and methodological myths, *Journal of Law and Medicine Ethics,* 24, 243-251.

Freedman B., Weijer C., Glass K.C. (1996b), Placebo orthodoxy in clinical research II: Ethical, legal and regulatory myths, *Journal of Law and Medicine Ethics,* 24, 252-259.

Harrington D.P. (2000), The randomised clinical trial, *Journal of the American Statistical Association*, 95, 312-315.

Hill A.B. (1962), *Statistical methods in clinical and preventive medicine*, Oxford University Press, New York.

Holmgren E.B. (1999), Establishing equivalence by showing that a prespecified percentage of the effect of the active control over placebo is maintained, *Journal of Biopharmaceutical Statistics*, 9(4), 651-659.

Hung H.M.J., Wang S., Tsong Y., Lawrence J., O'Neil R.T. (2003), Some fundamental issues with non-inferiority testing in active controlled trials, *Statistics in Medicine*, 219.

Jones B., Jarvis P. Lewis J.A., Ebbutt A.F. (1996), Trial to assess equivalence: the importance of rigorous methods, *British Medical Journal*, 313, 36-39.

Laster L.L., Johnson M.F. (2003), Non-inferiority trials: the 'at least as good as' criterion, *Statistics in Medicine*, 22, 187-200.

Makuch R., Simon R. (1978), Sample size requirements for evaluating a conservative therapy, *Cancer Treatment Reports*, 62, 1037-1040.

O'Brien W.M. (1968), Indomethacin: a survey of clinical trials, *Clinical Pharmacological Therapy*, 9, 94-107.

Pesarin F. (2001), *Multivariate permutation tests: with application in biostatistics*, Wiley, New York.

Pledger G., Hall D.B. (1990), Active control equivalence studies: do they address statistical issues?, S*tatistical Issues in Drug Research and Development,* 226-238, Marcel Dekker, New York.

Rothman K.L., Michels K.B. (1994), The continued unethical use of placebo controls, *New England Journal of Medicine*, 331, 393-398.

Smith R. (1998), Fifty years of randomised controlled trials, *British Medical Journal*, 317, 1166.

Spilker B. (1991), *Guide to Clinical Trials*, Wiley, New York.

Temple R. (1983), Difficulties in evaluating positive control trials, *Proceedings of th Biopharmaceutical Section of the American Statistical Association,* 1-7.

Wang S., Hung H.M.J. (2003), TACT method for non-inferiority testing in active controlled trials, *Statistics in Medicine*, 229-231.

World Medical Association Declaration of Helsinki (1997), Recommendations guiding physicians in biomedical research involving human subject, *Journal of the American Statistical Association*, 277, 925-926.