Far East J. Theo. Stat. **13**(1) (2004), 47-65
(A special volume on Biostatistics and related areas)

IMPACT OF MULTIPLE ENDPOINTS ON TYPE I ERROR RATE AND POWER OF TEST STATISTIC IN NON-SUPERIORITY CLINICAL TRIALS*

ABDUL J. SANKOH and MOHAMMAD F. HUQUE*

Global Biostatistics, Aventis Pharmaceuticals
P. O. Box 6890, 200 Crossing Blvd., BX2-300E, Bridgewater
NJ 08807, U. S. A.
e-mail: abdul.sankoh@aventis.com

Division of Biometrics III, Office of Biostatistics
Center for Drug Evaluation and Research
US Food and Drug Administration, MD, U. S. A.

Abstract

The efficacy of a new drug may be demonstrated by showing it is clinically equivalent or non-inferior to a standard active drug. To arrive at such a non-superiority efficacy conclusion, one must (ideally) first establish that both the standard and test drugs have efficacy profiles that are superior to placebo or no treatment in the prospective clinical trial. The statistical analysis approach for the demonstration of clinical equivalence or non-inferiority involves the construction of a confidence interval that rules out any possibility of the test drug being less efficacious than the standard drug by more than a pre-defined clinically irrelevant amount called an inferiority margin. In this paper we review statistical inference and methodological concepts for clinical equivalence

2000 Mathematics Subject Classification: 62P10.

Key words and phrases: clinical equivalence, multiple endpoints, non-inferiority, power, sample size, type I error rate.

*Views expressed are of the author and not necessarily of the US Food and Drug Administration (FDA).

Communicated by Dejian Lai

Received December 17, 2003; Revised February 25, 2004

and non-inferiority clinical trials. Additionally, we also show, at least for two correlated endpoints, multiplicity due to multiple endpoints could negatively impact the power of the test statistics.

1. Introduction

In a positive control clinical trial, the efficacy of a test drug can be compared to a reference without the objective of demonstrating superiority of the test to the reference drug. This class of trials could be divided into three groups: bio-equivalence, clinical equivalence, and non-inferiority trials. For bio-equivalence trials (which are mostly blood level trials in normal volunteers), the expectation is that both the test and the reference drugs studied in the prospective clinical trial are pharmacokinetically absorbed in the blood compartment at the same rate and extent to produce the same therapeutic effects. So the objective is to demonstrate that the test drug (e.g., a generic version or a new formulation of an approved drug) is bio-equivalent to the reference drug (e.g., the approved original drug itself). That is, for some class of drugs, it is possible to use bio-equivalence argument to establish therapeutic equivalence between a test and a reference drug.

However, it is not possible to establish clinical equivalence in this way for drugs that do not work through systemic absorption. Examples of such drugs are sucralfate for acute healing of duodenal ulcer, bulk laxatives, antidiarrheals, and pancreatic enzymes, which work locally within the gastro-intestinal tract and hence cannot be demonstrated to be therapeutically equivalent in the blood stream. For such non-absorbable drugs, one way to establish therapeutic equivalence between a test and a reference drug is through clinical trials with one or more appropriate clinical endpoints. This class of trials falls within the domain of clinical equivalence trials where both the inferiority and the superiority of the test drug in comparison to the reference product are undesirable (1). However, it is sometimes acceptable to design a clinical trial with the objective of showing that the efficacy of the test drug is no worse than the reference drug with some added advantages including a superior toxicity profile and/or some pharmacoeconomic benefits. The pharmacoeconomic edge may include an overall cost benefit or simplicity in administration

(e.g., intravenous versus subcutaneous or subcutaneous versus oral). This type of clinical trials is generally referred to as non-inferiority clinical trials in the efficacy E-9 ICH (International Conference on Harmonization) document.

One key consideration in the design of clinical equivalence or non-inferiority clinical trials is the need to establish that the active comparator (i.e., the reference drug) is also effective in the prospective clinical trial. This is usually achieved by incorporating a concurrent placebo, or a low dose arm of the investigational drug, or any other inferior arm of an effective treatment relevant to the indication being studied in the design of the prospective clinical trial. When there is an ethical concern for the use of non-effective treatment, an appropriate historical control patient population is often used to achieve this goal. If the active comparator is not shown to be effective in the prospective clinical trials either through internal or external validations, then the results of such clinical trials are generally difficult to interpret. These and other relevant issues regarding clinical equivalence and/or non-inferiority clinical trials have been addressed by many authors including Fleming [4], Gould [5], Lamborn [7], and Temple [9].

Another key consideration in such trials is the formulations of the null and alternate hypotheses. These formulations are different from those of the superiority trials where the null hypothesis is usually the conventional hypothesis of no treatment difference. For example, consider a simple clinical trial where n_1 patients are randomized to the test drug T and n_2 to the reference drug R. Suppose that π_T and π_R are the true unknown response rates for the test T and reference R drug, respectively. Assume that higher response rates are more desirable. Testing the conventional null hypothesis $H_0 : \pi_T = \pi_R$ against (the 2-sided or 1-sided) alternative hypothesis $H_a : \pi_T < \pi_R$, or $H_a : \pi_T > \pi_R$ at a nominal level of significance α is inappropriate for declaring similarity in efficacy profiles between T and R if H_0 is not rejected at level α. This is because the non-rejection of the null hypothesis does not necessarily imply that the null hypothesis is true. This point is articulated in Blackwelder [2]. In addition, even if there is a sufficient built in statistical power of the

test, this approach does not give much incentive to the experimenter, because a sloppy trial could have a very good chance of not rejecting the null hypothesis of no treatment effect.

We revisit in this paper the design, hypothesis formulations, and analysis approach for clinical equivalence and non-inferiority clinical trials for a single primary clinical endpoint. We then formulate the problem for such clinical trials for two primary endpoints and provide results on the impact of multiplicity due to multiple endpoints on the type I error rate and the power of the test.

2. Single Endpoint Case

2.1. Non-inferiority trials

Consider a simple clinical trial in which n_1 patients are randomized to a test drug T and n_2 patients to a reference drug R. Let π_T and π_R denote the true unknown response rates (e.g., success rates) for the test and reference drug, respectively. Assume that higher response rates are more desirable. In a clinical trial with a non-inferiority objective, the alternate hypothesis is that the test drug T is not inferior to the reference drug R by some pre-specified inferiority margin (of clinical relevance) δ_0. The corresponding null hypothesis is that T is inferior to R by at least this amount δ_0. Thus, if H_0 and H_a are, respectively, the null and alternate hypotheses, then for this simple trial we have

$$H_0 : \pi_T \leq \pi_R - \delta_0 \text{ (hypothesis of inferiority) vs.}$$
$$H_a : \pi_T > \pi_R - \delta_0 \text{ (hypothesis of non-inferiority).} \quad (1)$$

The null hypothesis H_0 is an interval hypothesis and can be tested using a z-statistic $Z = [D - (-\delta_0)]/\sigma_D$, where $D = p_T - p_R$ is the observed treatment difference, p_T and p_R being the observed response rates for T and R, respectively, $\delta = E(D) = \pi_T - \pi_R$ (the true treatment difference), and σ_D = standard error of D defined by $\sigma_D(\delta, \pi_R) = \{\pi_T(1 - \pi_T)/n_1 + \pi_R(1 - \pi_R)/n_2\}^{1/2}$. That is, assume that sample sizes are sufficiently large so that Z follows approximately a normal distribution with some

mean and standard deviation under H_0. Note that the standard error σ_D is a function of two unknown parameters δ and π_R.

While one could consider test statistics based on exact binomials, or odds ratio, or risk ratio, we consider here a simple test formulation to illustrate statistical principles and to show some results related to the power of the test and sample size for single as well as for multiple endpoints. (Lui [8] considers an exact binomial approach for the risk ratio under inverse sampling.)

The rule for establishing non-inferiority based on the above z-statistic is then to reject the hypothesis of inferiority H_0 if $Z > u_{1-\alpha}$, where $u_{1-\alpha}$ is such that $\Phi(u_{1-\alpha}) = 1 - \alpha$, and $\Phi(x)$ denotes the area under the standard normal probability curve from $-\infty$ to x.

Example. For the above 2-treatment-group clinical trial case, suppose that the observed response rates for the test T and reference R drugs are, respectively, $p_T = 55\%$ and $p_R = 60\%$, and the corresponding sample sizes are $n_1 = n_2 = 100$ patients (per treatment group). Assume pre-specified $\delta_0 = 20\%$ and $\alpha = 5\%$. Then one can calculate the test statistic in three different ways:

(i) If it is known that $\pi_R = .70$ and that $\delta = -\delta_0 = -.20$, then from above, $\sigma_D = \{.5(.5) + .7(.3)\}^{1/2} = \{.0025 + .0021\}^2 = .0678$, leading to $Z = (-.05 + .20)/.0678 = 2.21$.

(ii) If π_R is unknown but that $\delta = -\delta_0 = -.20$, one could assume $\pi_R = \pi_T = .50$ to obtain maximum $\sigma_D = .0707$, leading to a minimum Z value of 2.12.

(iii) Alternatively, one could estimate the value of σ_D from the data to obtain $\sigma_D = \{.5(.5) + .6(.4)\}^{1/2} = .070$, leading to $Z = 2.15$.

For this example, we notice that each of the three approaches for calculating σ_D gives about the same value of Z and leads to the same conclusion, namely rejection of the null hypothesis H_0 in all three cases.

The rule for rejection of the null hypothesis of inferiority $Z > u_{1-\alpha}$ implies that $D - u_{1-\alpha} \times \sigma_D > -\delta_0$, where the left-hand side of the inequality is the lower limit of the $100(1-2\alpha)\%$ 2-sided confidence interval of δ. This leads to a confidence interval based rule:

> Reject H_0 if the $1-2\alpha$ lower confidence lower limit $(\text{LCL}_{1-2\alpha})$ of the true treatment difference δ exceeds $-\delta_0$.

This confidence interval approach is more appealing to clinicians because the lower confidence limit value can easily be related to $-\delta_0$ for clinical interpretations. In the above example, we see that the minimum $\text{LCL}_{1-2\alpha} = -.05 - .0707(1.96) = -.1886 > -20 = \delta_0$. Thus we reject the null hypothesis of inferiority H_0 at the .025 level of significance (for pre-specified 2-sided .05 level).

Power function. The power function for the above test procedure is the probability of δ_0 falling in the rejection region given the expected value of D [i.e., $E(D) = \delta$]. In addition to being a function of δ, the power function also depends on δ_0, π_R, n_1, n_2 and α. For convenience, let $P(\delta; *) = P(\delta; \delta_0, \pi_R, n_1, n_2, \alpha)$ denote the power function. With the above assumptions about the test statistic $Z = [D - (-\delta_0)]/\sigma_D$, we have

$$P(\delta; *) = Pr[(p_T - p_R) - \sigma_D u_{1-\alpha} > -\delta_0 \mid E(p_T - p_R) = \delta]$$
$$= 1 - \Phi(u_{1-\alpha} - (\delta + \delta_0)/\sigma_D). \quad (2)$$

It is easy to see that $P(\delta; *) = \alpha$ at $\delta = -\delta_0$, and $P(\delta; *) < \alpha$ for $\delta < -\delta_0$. Therefore, the type I error rate for the above test procedure is $\text{Sup}_\delta\{P(\delta; *) \text{ for } \delta \le -\delta_0\} \le \alpha$, with a maximum value at the boundary point $\delta = -\delta_0$.

Sample size calculations. If $1-\beta$ is the pre-specified power of the test for sample size determination, then on equating (2) to $1-\beta$ leads to the expression

$$\Phi[u_{1-\alpha} - (\delta + \delta_0)/\sigma_D] = \beta, \text{ or } u_{1-\alpha} - (\delta + \delta_0)/\sigma_D = -u_\beta,$$

where $\Phi(-u_\beta) = \beta$. On assuming equal samples sizes of n patients per treatment group, the last expression leads to the following sample size expression:

$$n = 2(\delta + \delta_0)^{-2}(Z_\alpha + Z_\beta)^2[\pi_R(1 - \pi_R) - \delta\pi_R + \delta(1 - \delta)/2], \quad (3)$$

where Z_p denotes the critical z-value corresponding to the proportion p on the upper tail of the standard normal distribution. For example, $Z_{0.025} = 1.96$ and $Z_{0.05} = 1.645$. Upon assuming that the true treatment difference $\delta = 0$, then (3) reduces to

$$n = 2\delta_0^{-2}(Z_\alpha + Z_\beta)^2(\pi_R(1 - \pi_R)),$$

which corresponds to the conventional sample size formula of

$$n = 2\sigma^2(Z_\alpha + Z_\beta)^2/(\mu_1 - \mu_2)^2$$

for a normally distributed continuous endpoint with unknown mean μ and known standard deviation σ for testing the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$ using 1-sided alternative.

If the sample size is calculated using the 1-sided approach, for example, using the test statistic $Z_0 = D/(\text{standard error of } D)$, and assuming that Z_0 follows a standard normal distribution under the null hypothesis of no treatment difference, then the formula for n for detecting a treatment difference of $\delta = -\delta_0$, is given as follows:

$$n = 2\delta^{-2}(Z_\beta + fZ_\alpha)^2[\pi_R(1 - \pi_R) - \delta\pi_R + \delta(1 - \delta)/2], \quad (4)$$

$$f^2 = \frac{\pi_R(1 - \pi_R)}{\pi_R(1 - \pi_R) - \delta\pi_R + \delta(1 - \delta)/2}, \ \delta = -\delta_0,$$

where the correction factor f is due to the fact that the standard error of D also depends on δ and π_R, which are different under the null and alternate hypotheses for the binary endpoint case. For a normally distributed continuous endpoint with unknown mean μ and known standard deviation σ, however, the term $[-\delta\pi_R + \delta(1 - \delta)/2]$ in (4) is zero,

resulting in the standard sample size expression for a 1-sided test of $n = 2\sigma^2(Z_\alpha + Z_\beta)^2/(\mu_1 - \mu_2)^2$.

2.2. Clinical equivalence trials

For non-inferiority clinical trials as discussed above, it is not desirable that the test drug T is inferior to the standard treatment R by an amount $\delta_0^{(1)}$ or greater. However, for clinical equivalence trials, in addition to the non-inferiority condition, it is also not desirable that the test drug T is superior to R by an amount $\delta_0^{(2)}$ or greater. This is the case when a generic version of a brand has to be a duplicate with identical chemical or biological formulation. Clinical trials currently conducted for the approval of generic versions of sucralfate fall into this category. As a specific example, sucralfate was approved for Marion Laboratories in 1981 and is marketed under the brand name Carafate® for the treatment of acute duodenal ulcer. Sucralfate is a complex of sucrose octasulfate and aluminum hydroxide, not easy to identify chemically. It has various postulated modes of action some of which are adherence to the ulcer crater, increasing prostaglandin synthesis in mucosa, and bringing growth factor in greater concentration to the ulcer site. This unique chemical complexity compounds the problem of demonstrating bio-equivalence for these class of drugs, and has led to the current requirement that a clinical endpoint, e.g., healing of acute duodenal ulcers as demonstrated by endoscopy, must be used to demonstrate therapeutic equivalence instead. These sucralfate trials are at the moment being designed as 3-arm trials with sucralfate, Carafate, and placebo with the aim of demonstrating the generic sucralfate is clinically equivalent to the marketed Carafate after validating that Carafate is effective in the given trial.

Thus, in a clinical equivalence trial, one has to address the two test interval hypotheses: $H_0^{(1)} : \pi_T \leq \pi_R - \delta_0^{(1)}$ (hypothesis of inferiority), and $H_0^{(2)} : \pi_T \geq \pi_R + \delta_0^{(2)}$ (hypothesis of superiority). Both must be rejected simultaneously through an efficient test procedure on assuring that in

the test procedure the overall probability of wrongly concluding that the true treatment response π_T is within the clinical equivalence interval $(\pi_R - \delta_0^{(1)}, \pi_R + \delta_0^{(2)})$ is no more than a pre-specified α. For our simple two-arm trial, the null hypothesis $H_0^{(1)}$ can be tested against the alternative hypothesis $H_a^{(1)} : \pi_T > \pi_R - \delta_0^{(1)}$ by the criterion

$$Z^{(1)} = \sigma_D^{-1}[(p_T - p_R) - (-\delta_0^{(1)})] > u_{1-\alpha},$$

leading to the rejection of $H_0^{(1)}$. Similarly the null hypothesis $H_0^{(2)}$ can be tested against the alternative $H_a^{(2)} : \pi_T < \pi_R + \delta_0^{(2)}$ by the criterion

$$Z^{(2)} = \sigma_D^{-1}[(p_T - p_R) - (\delta_0^{(2)})] < -u_{1-\alpha},$$

leading to the rejection of $H_0^{(2)}$. Therefore, the (overall) null hypothesis of non-clinical equivalence with respect to both endpoints $H_0 : \delta \in (-\delta_0^{(1)}, \delta_0^{(2)})$ will be rejected in favor of the alternative hypothesis of clinical equivalence with respect to both endpoints if the values of the statistics $Z^{(1)}$ and $Z^{(2)}$ for the given data fall in the 'intersection' critical region C given by

$$C : (Z^{(1)} > u_{1-\alpha}) \cap (Z^{(2)} < -u_{1-\alpha}).$$

Note that the alternate hypothesis H_a is the intersection of the two alternate hypotheses $H_a^{(1)}$ and $H_a^{(2)}$, i.e., $H_a = H_a^{(1)} \cap H_a^{(2)}$. Similarly, the null hypothesis H_0 can be written as the union of the two null hypotheses $H_0^{(1)}$ and $H_0^{(2)}$, i.e., $H_0 = H_0^{(1)} \cup H_0^{(2)}$. Berger and Hsu [1] address the 'intersection-union' hypothesis testing principle in the context of bio-equivalence clinical trials.

Also note that the symmetric case is obtained when $\delta_0^{(1)} = \delta_0^{(2)} = \delta_0$. For this case, $Z^{(1)} > u_{1-\alpha}$ implies $\text{LCL}_{1-2\alpha} = (D - \sigma_D \times u_{1-\alpha}) > -\delta_0$ and $Z^{(2)} < u_{1-\alpha}$ implies $\text{UCL}_{1-2\alpha} = (D - \sigma_D \times u_{1-\alpha}) < \delta_0$. This leads to a

confidence interval rule. That is, reject the null hypothesis of non-clinical equivalence with respect to both endpoints $H_0 : \delta \in (-\delta_0^{(1)}, \delta_0^{(2)})$ in favor of the alternative if the $1 - 2\alpha$ confidence interval of the true treatment difference δ is within the clinical equivalence interval $(-\delta_0, \delta_0)$.

Power function. Again the power function of the above test procedure is the probability of both $Z^{(1)}$ and $Z^{(2)}$ falling in the critical region C given $E(D) = \delta$. Besides being a function of δ, the power function would conditionally depend on δ_0, π_R, n_1, n_2 and α. On assuming that for sufficiently large n, both $Z^{(1)}$ and $Z^{(2)}$ follow normal distributions, this probability, for the symmetric case, is given by, $P(\delta; *) = P(\delta; \delta_0, \pi_R, n_1, n_2, \alpha)$,

$$P(\delta; *) = Pr[(Z^{(1)} > u_{1-\alpha}) \cap (Z^{(2)} < -u_{1-\alpha}) | E(D) = \delta]$$

$$= Pr\left[\left(U > u_{1-\alpha} - \frac{\delta + \delta_0}{\sigma_D}\right) \cap \left(U < -u_{1-\alpha} - \frac{\delta - \delta_0}{\sigma_D}\right)\right],$$

where U is the standard normal random variate. Using normal probability integrals, the above power function expression can be equivalently written as

$$P(\delta; *) = \Phi[-u_{1-\alpha} - (\delta - \delta_0)/\sigma_D] - \Phi[u_{1-\alpha} - (\delta + \delta_0)/\sigma_D]. \qquad (5)$$

The type I error probability is $\text{Sup}_\delta\{P(\delta; *)$ for $\delta \in (-\delta_0, \delta_0)\}$, and equals the nominal α when δ is at the boundary. For example, when $\delta_0 = .20$, $\pi_R = .50$, $n_1 = n_2 = 100$, and $\alpha = .05$, we obtain from expression (5) $P(-\delta_0; *) = \Phi(-1.645 + .40/.0678) - \Phi(1.645) = \Phi(4.25) - \Phi(1.645) \approx$ nominal α. Similarly, we obtain $P(\delta_0; *) = \Phi(-1.645) - \Phi(-4.25) \approx$ nominal α. In addition, given that there is no treatment difference (i.e., $\delta = 0$), the power of the test is $\beta(0|\pi_R = .5) = \Phi(-1.645 + .20/.0707) - \Phi(1.645 - .2/.0707)$ $= \Phi(1.184) - \Phi(-1.184) = 76\%$ when the response rate for the standard treatment is $\pi_R = .5$. For $\pi_R = .7$, the power of the test increases to $\beta(0 | \pi_R = .70) = \Phi(1.475) - \Phi(-1.475) = 86\%$.

Sample size formula. We notice that at $\delta = 0$, the power function in expression (5) is of the form $\Phi(h) - \Phi(-h)$, where $h = (\delta_0/\sigma_D - u_{1-\alpha})$. Thus equating (5) to a nominal power, say $1 - \beta$, leads to the equation $\delta_0/\sigma_D - u_{1-\alpha} = u_{\beta/2}$, where $\Phi(-u_{\beta/2}) = \beta/2$. This gives the sample size expression:

$$n = 2\delta_0^{-2}(Z_\alpha + Z_{\beta/2})^2 \pi_R(1 - \pi_R). \tag{6}$$

Fleiss [3] gave a similar sample size expression $n = 2\sigma^2(Z_{\alpha/2} + Z_{\beta/2})^2/\delta_0^2$ for clinical equivalence trials at $\delta = 0$. In his expression, $\alpha/2$ is used in place of α, and $100(1 - \alpha)\%$ confidence interval criterion is used instead of the $100(1 - 2\alpha)\%$ criterion used here in our calculations. This latter formulation is consistent with the two-tailed test formulation for a superiority trial with a total significance level of α, when $\alpha/2$ is spent in assuring that the clinical efficacy of the test drug is different from or inferior to the reference drug.

The sample size expression in (6) differs from the conventional one in expression (3) obtained when $\delta = 0$ only in that $Z_{\beta/2}$ replaces Z_β. This means that if the true treatment difference $\delta = 0$ is assumed, then the increase in sample size in using a non-inferiority trial instead of a clinical equivalence trial is by a factor of $g = (Z_\alpha + Z_{\beta/2})^2/(Z_\alpha + Z_\beta)^2$. However, when δ is not zero, the power function in (5) is of the form $\Phi(h - \delta/\sigma_D) - \Phi(-h - \delta/\sigma_D)$, where for $h = \delta_0/\sigma_D - u_{1-\alpha}$, the sample size value can be obtained by solving the following non-linear equation:

$$\Phi(h - \delta/\sigma_D) - \Phi(-h - \delta/\sigma_D) = 1 - \beta,$$

$$\sigma_D^2 = 2[\pi_R(1 - \pi_R) - \delta\pi_R + \delta(1 - \delta)/2]/n. \tag{7}$$

This is a non-linear equation in n, and can easily be solved by a standard programming for desired α, $1 - \beta$, δ_0, δ, and π_R. If the sample size is calculated using the conventional 2-sided approach [with the test statistic

$Z_0 = D/se(D)$, with the usual normal distribution assumptions as in (4)], then the expression for n for detecting a treatment difference of δ is the same as in expression (4) except that Z_α is replaced by $Z_{\alpha/2}$.

3. Multiple Endpoints Case

Similar to placebo-controlled trials, active-controlled trials are often designed and conducted with multiple efficacy and safety endpoints. For example, in an acute duodenal ulcer healing trial, healing rates are generally measured at different time points such as week 4 and week 8. In this case, the clinical expectation may be that the test drug T must be non-inferior (or clinically equivalent) to a reference drug R for both time points. Similarly, in a large vaccine trial, it may be necessary to show that a new vaccine is at least as good as the old vaccine with respect to not one but a number of key efficacy and safety measurements. For anti-infective clinical trials, often the non-inferiority of a test drug in comparison to an active control is sought with respect to both a cure rate endpoint and a suitable bacteriological endpoint.

The computational results presented in this section are for binary endpoints under the assumption of large sample size to provide sufficient power for a single endpoint case. To introduce some notations, suppose that, for endpoint j $(j = 1, 2, ..., K)$, π_{jT} and π_{jR} represent the true response rates for the test and reference drug, respectively, $\delta_j = \pi_{jT} - \pi_{jR}$ represents the true treatment difference between the test and reference drug, δ_{0j} is a pre-specified minimal clinical significant difference of interest, and $Z_j = \{D_j - (-\delta_{0j})\}/\sigma_j$ is the test statistic for testing the null hypothesis of inferiority $H_{0j} : \pi_{jT} \leq \pi_{jR} - \delta_{0j}$ against the alternative hypothesis of non-inferiority $H_{aj} : \pi_{jT} > \pi_{jR} - \delta_{0j}$, where $D_j = (p_{jT} - p_{jR})$ is the observed treatment difference, and σ_j is the standard error of D_j for endpoint j. Suppose that we adopt the following rule for multiple endpoint testings.

Rule. Claim non-inferiority for the test drug T in comparison to the reference treatment R, if $Z_j > u_{1-\alpha}$ for all $j = 1, ..., K$, i.e., $\mathbf{Z} = (Z_1, ..., Z_K)$ is within the intersection region $\cap_{1\le j\le K}(Z_j > u_{1-\alpha})$.

In the following, we evaluate the impact on the type I error probability and the power of the test procedure for multiple endpoint testing. We assume that the endpoints are correlated, the sample sizes for treatments T and R in the clinical trial are sufficiently large so that the test statistic vector $\mathbf{Z} = (Z_1, ..., Z_K)$ follows a multivariate normal distribution.

3.1. Two-endpoint case

Let (δ_1, δ_2) denote the axes of a rectangular co-ordinate centered about $(-\delta_{01}, -\delta_{02})$. Then the overall null hypothesis H_0 is the union of the three quadrants C_{i0} $(i = 1, 2, 3)$ and the alternative hypothesis is the quadrant C_a as defined in the following:

$$C_{10} : \{\delta_1 \le -\delta_{01}, \delta_2 \le -\delta_{02}\}$$

$$C_{20} : \{\delta_1 \le -\delta_{01}, \delta_2 > -\delta_{02}\}$$

$$C_{30} : \{\delta_1 > -\delta_{01}, \delta_2 \le -\delta_{02}\}$$

$$C_a : \{\delta_1 > -\delta_{01}, \delta_2 > -\delta_{02}\}.$$

That is, the test drug T is inferior to the reference drug R with respect to both endpoints in quadrant C_{10}, but it is inferior to R with respect to endpoint 1 only in quadrant C_{20}, and inferior to R with respect to endpoint 2 only in quadrant C_{30}. Let $P(\delta_1, \delta_2; *)$ denote the power function of the test procedure. The critical regions for non-inferiority, clinical equivalence, and superiority are displayed in Figure 1 where $C(jk)$ stands for C_{jk}, dj stands for δ_j and $dj0$ stands for δ_{j0} $(j = k = 0, 1)$, and wrt = with respect to. Note that the critical regions for a single endpoint can be read off of any one of the two axes.

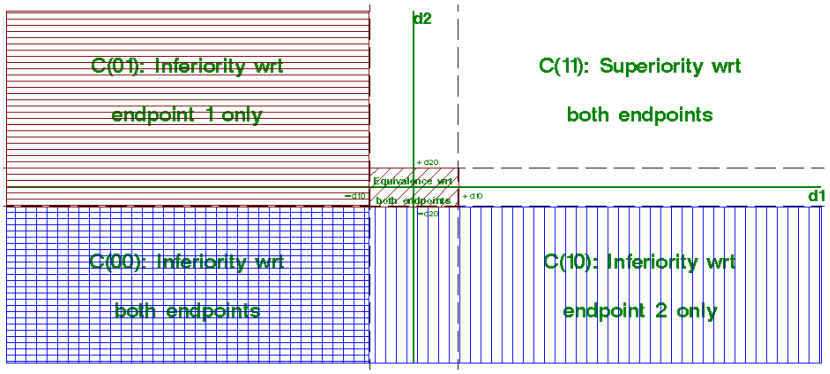

Figure 1. Critical regions $C(jk)$ for two-endpoint case, $j = k = 0, 1$

If

$$\alpha_1 = \text{Sup}\, P(\delta_1, \delta_2; *) \text{ when } (\delta_1, \delta_2) \in C_{10}$$

$$\alpha_2 = \text{Sup}\, P(\delta_1, \delta_2; *) \text{ when } (\delta_1, \delta_2) \in C_{20}$$

$$\alpha_3 = \text{Sup}\, P(\delta_1, \delta_2; *) \text{ when } (\delta_1, \delta_2) \in C_{30},$$

then the type I error rate of the test procedure is $\alpha = \max\{\alpha_1, \alpha_2, \alpha_3\}$.

Power function. The power function $P(\delta_1, \delta_2; *) = P\{(\delta_1, \delta_2); \delta_{10}, \delta_{20}, \pi_{1R}, \pi_{2R}, n_1, n_2, \alpha\}$ is the probability of C given $E(D_j) = \delta_j$, where $C : \{(Z_1 > u_{1-\alpha}) \cap (Z_2 > u_{1-\alpha})\}$. This probability can be calculated using $L(h, k; \rho, \delta_1, \delta_2)$ of the standard bivariate normal distribution as

$$P(\delta_1, \delta_2; *) = L(h, k; \rho, \delta_1, \delta_2) = \int_h^\infty \int_k^\infty (4\pi^2(1-\rho^2))^{-1/2}\, \phi(u, v) du dv,$$

where $h = u_{1-\alpha} - \xi_1$, $k = u_{1-\alpha} - \xi_2$, $\xi_1 = (\delta_1 + \delta_{10})/\sigma_1$, $\xi_2 = (\delta_2 + \delta_{20})/\sigma_2$, and ρ is the correlation coefficient between Z_1 and Z_2. It is easy to see that $\alpha_1 = L(u_{1-\alpha}, u_{1-\alpha}; \rho)$, because the supremum of $L(h, k)$ when $(\delta_1, \delta_2) \in C_{10}$ is attained at $\delta_1 = -\delta_{10}$, $\delta_2 = -\delta_{20}$, leading to $\alpha_1 = \alpha^2$

when $\rho = 0$, and $\alpha_1 < \alpha$ with α_1 approaching α in the limit as correlation ρ approaches 1. In addition,

$$\begin{aligned}\alpha_2 &= \text{Sup}\, L(h, k) \text{ when } (\delta_1, \delta_2) \in C_{20} \\ &= L(h, k; \rho) \text{ at } \delta_1 = -\delta_{10},\ \delta_2 = \max \delta_2 \le 1 - \pi_{2R} \\ &= L(u_{1-\alpha},\ u_{1-\alpha} - (\max \delta_2 + \delta_{20})/\sigma_2;\ \rho) \\ &< 1 - \Phi(u_{1-\alpha}) = \alpha.\end{aligned}$$

Similarly,

$$\begin{aligned}\alpha_3 &= \text{Sup}\, L(h, k) \text{ when } (\delta_1, \delta_2) \in C_{30} \\ &= L(h, k; \rho) \text{ at } \delta_1 = \max \delta_1 \le 1 - \pi_{1R},\ \delta_2 = -\delta_{20} \\ &= L(u_{1-\alpha} - (\max \delta_1 + \delta_{10})/\sigma_1,\ u_{1-\alpha};\ \rho) \\ &< 1 - \Phi(u_{1-\alpha}) = \alpha.\end{aligned}$$

That is, as $n \to \infty$, both α_2 and α_3 approach the nominal significance level α. To see this, let $\phi(\delta_2) = (\delta_2 + \delta_{20})/\sigma_2$ and $\pi_{2T} = \delta_2 + \pi_{2R}$ in the expression for α_2. Then

$$\sqrt{\frac{2}{n}\phi(\delta_2)} = (\delta_2 + \delta_{20}) \Big/ \sqrt{\pi_{2R}(1 - \pi_{2R}) - \pi_{2R}\delta_2 + \frac{\delta_2(1 - \delta_2)}{2}}.$$

Note that $f(\delta_2) = \pi_{2R}(1 - \pi_{2R}) - \pi_{2R}\delta_2 + \delta_2(1 - \delta_2)/2$ attains its maximum value at $\delta_2 = (.5 - \pi_{2R})$, and is strictly decreasing for all $\delta_2 \in (.5 - \pi_{2R}, 1 - \pi_{2R})$. The largest possible value of $\delta_2 = 1 - \pi_{2R}$ when $\pi_{2T} = 1$. Therefore, the maximum of $\phi(\delta_2)$ is attained at $1 - \pi_{2R}$ and is given by $\phi^* = (1 - \pi_{2R} + \delta_{20}) \Big/ [\pi_{2R}(1 - \pi_{2R})/n]^{1/2}$. Consequently, $\alpha_2 = L(u_{1-\alpha}, u_{1-\alpha} - \phi^*; \rho)$ converges to $L(u_{1-\alpha}, -\infty; \rho)$ as $n \to \infty$. But $L(u_{1-\alpha}, -\infty; \rho) = \alpha$ and hence the result for α_2. The result for α_3 can be derived similarly. Note that in evaluating the L function, both δ_1 and δ_2 are evaluated at their maximum values.

Tables 1a, 1b and 2 give values of α_2 for various values of δ_{10}, δ_{20}, $\max \delta_2$, and n; Table 3 presents the results of some power calculation results.

Table 1a. α_2 -values for two endpoints

$\delta_{10} = \delta_{20} = .10$, $n = 100$, $\max(\delta_2) = 30\%$

ρ	$\pi_R = .5$	$\pi_R = .6$	$\pi_R = .7$
0.0	.049994	.049998	.050000
0.3	.050000	.050000	.050000
0.5	.050000	.050000	.050000
≥ 0.6	.050000	.050000	.050000

Table 1b. α_2 -values for two endpoints

$\delta_{10} = \delta_{20} = .10$, $n = 200$, $\max(\delta_2) = 10\%$

ρ	$\pi_R = .5$	$\pi_R = .6$	$\pi_R = .7$
0.0	.049585	.049747	.049934
0.3	.049958	.049978	.049996
0.5	.049998	.049999	.050000
≥ 0.6	.050000	.050000	.050000

Table 2. α_2-values for two endpoints

$\delta_{10} = \delta_{20} = .15$, $n = 100\,(150)$, $\max(\delta_2) = (15\%, 20\%)$

ρ	$\pi_R = .5$ $\delta_2 = 15\%$, $\delta_2 = 20\%$	$\pi_R = .7$ $\delta_2 = 15\%$, $\delta_2 = 20\%$
0.0	.049825, .049978 (.049950, .049996)	.049989, .050000 (.049997, .050000)
0.1	.049911, .049990 (.049977, .049999)	.049995, .050000 (.049969, .050000)
0.5	.049999, .050000 (.050000, .050000)	.050000, .050000 (.049996, .050000)
0.9	.050000, .050000 (.050000, .050000)	.050000, .050000 (.050000, .050000)

Single endpoint power: 74.4%, 84.8%
(Single endpoint power = 79.7%, 88.3%)

Table 3. Power of the test for two endpoints

$\delta_{10} = \delta_{20} = .15$, $n = 100$, Power$(\delta_1 = \delta_2 = 0)$

ρ	$\pi_R = .5$	$\pi_R = .7$
0.0	63.6% (79.7%)*	78.0% (88.3%)*
0.1	64.4%	78.4%
0.5	68.4%	80.6%
0.9	74.7%	84.8%
0.99	78.2%	87.2%

*Single endpoint power

4. Discussion

In an ideal situation, a non-inferiority (or a clinical equivalence) trial would include a concurrent placebo arm for internal validation that the active control is effective in the given trial, before testing for similarity or non-inferiority between the test and the reference drug. However, this situation is often not possible for ethical reasons, thus leading to an external validation based on historical data or other external clinical evidence. Without such an internal or external validation, active control trials pose some serious challenges regarding results interpretation. These and other relevant issues about active control trials have been addressed by many authors including Temple [9], Lamborn [7], Gould [5] and Fleming [4].

A suitable testing of hypothesis framework for such trials is to make the null hypothesis an interval hypothesis of non-inferiority (or non-equivalence) and to set up the testing procedure, as discussed in this presentation, for the purpose of rejecting such a null hypothesis. This approach leads to a confidence interval rule that helps the interpretation of the clinical trial results. For clinical equivalence trials, however, the lower and the upper equivalence margins may not be the same, and thus the test procedure may not convert to a symmetric confidence interval rule.

The choice of an acceptable margin of clinical inferiority (δ_0), often called upper and lower equivalence margins, have not been addressed in this paper. For the binary-type endpoint these margins depend on the background response rate and the clinical input. If the background response rate is high, these margins could be small. One obvious restriction is that the lower and/or upper margins need to be smaller than the size of the observed treatment difference between the reference drug and placebo on the δ-scale. In particular, the size of the margin may not be greater than half the observed and documented treatment difference between the reference drug and placebo in the trial that formed the basis for the approval of the reference drug.

We have presented some sample size and power calculations both for a one and two correlated endpoints for a positive-controlled clinical trial.

In a superiority trial, multiple endpoints, depending on the strength of the correlation among the endpoints, generally lead to inflation of the type I error rate. Therefore, for non-inferiority trials one would expect the opposite of this, i.e., deflation of the type I error to a much lower value since the roles of the null and alternative hypotheses are reversed. However, our computations show that, at least for the two endpoint case, a test for non-inferiority in the hypothesis-testing framework discussed above leads to a type I error rate that remains very close to the nominal α. However, our calculations show that the power of the test procedure is adversely affected depending on the extent of the correlations between the endpoints.

References

[1] R. L. Berger and J. C. Hsu, Bio-equivalence trials, intersection-union tests, and equivalence confidence sets, Stat. Sci. 11(4) (1996), 283-319.

[2] W. C. Blackwelder, Proving the null hypothesis in clinical trials, Cont. Clinical Trials 3 (1983), 345-353.

[3] J. L. Fleiss, General design issues in efficacy, equivalency and superiority trials, J. Period. Res. 27(Special Issue) (1992), 306-313.

[4] T. R. Fleming, Evaluation of active control trials in AIDS, J. Acquired Immune Deficiency Syndromes 3(Suppl. 2) (1990), S82-S87.

[5] L. Gould, Another view of active-controlled trials, Cont. Clinical Trials 12 (1991), 474-485.

[6] M. F. Huque, S. D. Dubey and S. B. Fredd, Establishing therapeutic equivalence with clinical endpoints, Amer. Stat. Assoc. Proceedings of the Biopharmaceutical Section (1989), 46-52.

[7] K. R. Lamborn, Some practical issues and concerns in active-controlled trials, Amer. Stat. Assoc. Proceedings of the Biopharmaceutical Section (1983), 8-11.

[8] K. J. Lui, Exact equivalence test for risk ratio and its sample size determination under inverse sampling, Stat. Med. 16 (1997), 1777-1786.

[9] R. Temple, Difficulties in evaluating positive-controlled trials, Amer. Stat. Assoc. Proceedings of the Biopharmaceutical Section (1983), 1-7.

■