STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:1131–1147
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2359

Power and sample size computations in simultaneous tests for non-inferiority based on relative margins

Gemechis Dilba[1,‡], Frank Bretz[2,§], Ludwig A. Hothorn[1,¶] and Volker Guiard[3,*,†]

[1] *Bioinformatics Unit, University of Hannover, Germany*
[2] *Novartis Pharma AG, Basel, Switzerland*
[3] *Research Institute for the Biology of Farm Animals, Research Unit Genetics and Biometry, Dummerstorf, Germany*

SUMMARY

In this paper, we address the problem of calculating power and sample sizes associated with simultaneous tests for non-inferiority. We consider the case of comparing several experimental treatments with an active control. The approach is based on the ratio view, where the common non-inferiority margin is chosen to be some percentage of the mean of the control treatment. Two power definitions in multiple hypothesis testing, namely, complete power and minimal power, are used in the computations. The sample sizes associated with the ratio-based inference are also compared with that of a comparable inference based on the difference of means for various scenarios. It is found that the sample size required for ratio-based inferences is smaller than that of difference-based inferences when the relative non-inferiority margin is less than one and when large response values indicate better treatment effects. The results are illustrated with examples. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: ratio-to-control; multiple comparison; least favourable configuration; non-central *t*; non-inferiority

1. INTRODUCTION

Non-inferiority trials are becoming increasingly more popular as an alternative to placebo controlled clinical trials. If the use of a placebo group is unethical, an active competitor may be included against which non-inferiority has to be claimed. Such approach replaces the traditional superiority trials including a placebo arm if well-established active competitors are

[*] Correspondence to: Volker Guiard, Research Institute for the Biology of Farm Animals, Research Unit Genetics and Biometry, Wilhelm-Stahl-Allee 2, D-18196 Dummerstorf, Germany.
[†] E-mail: v.guiard@fbn-dummerstorf.de
[‡] E-mail: dilba@bioinf.uni-hannover.de
[§] E-mail: frank.bretz@novartis.com
[¶] E-mail: hothorn@bioinf.uni-hannover.de

Contract/grant sponsor: Deutsche Forschungsgemeinschaft; contract/grant number: BR 2202/1

Received 15 March 2005
Accepted 15 July 2005

available. In many therapeutic areas, such active competitors do exist and it may, therefore, not be sufficient to solely show superiority of the newly developed experimental treatment over placebo since its relative performance to the competitors on the market would remain uninvestigated. Another reason to conduct non-inferiority trials is the increased number of studies claiming a better safety profile of the experimental treatment over competing treatments while not being inferior in efficacy. In such instances, the new experimental treatment is shown to be safer than its competitors in which case a proven non-inferior efficacy justifies its potential release on the market.

In a special issue of Statistics in Medicine (2003, volume 22, issue 2), several statistical problems related to non-inferiority trials were discussed. Among others, D'Agostino *et al.* [1] addressed the design concepts while Rashid [2] dealt with non-parametric analysis and Laster and Johnson [3] discussed the use of ratio hypotheses. Ratio hypotheses re-formulate the standard hypotheses of differences, say for the efficacy parameters, in terms of relative effects which are particularly appealing in non-inferiority trials. The primary merit of this approach is that margins of non-inferiority (or equivalence) or superiority can be easily defined as a percentage of the unknown mean of the control treatment, particularly when the definition of the non-inferiority margin in an absolute term is difficult [4, 5]. Hauschke *et al.* [6] dealt with sample size calculations in testing of equivalence based on ratio. In this problem, a single ratio is involved, but the nature of the problem leads to the computation of percentage points of a non-central bivariate *t*-distribution. Pigeot *et al.* [7] considered the problem of comparing an experimental treatment for non-inferiority with a reference including a placebo arm. This problem is succinctly formulated as inference for a single ratio of linear combinations of the treatment means. They also derived formulas for determining power and sample size. Laster and Johnson [3] described the ratio-based inference in detail and compared it with the classical testing approach based on difference of means [8]. In terms of the sample size, they conclude that under certain conditions testing for ratio of means is more efficient than testing the associated difference of means.

Our aim is to extend the results from Laster and Johnson [3] to the case of testing multiple treatments for non-inferiority against an active competitor. The problem of multiple ratios occurs, for example when several dose levels of a certain compound are assessed for their efficacy in comparison to an established treatment. Multiple testing for non-inferiority based on ratios was first addressed by Hauschke and Kieser [9]. The related estimation problems and derivation of simultaneous confidence intervals for multiple ratios were studied by Dilba *et al.* [10]. In dose finding studies, Bretz *et al.* [11] implemented a stepwise test procedure with associated confidence intervals to identify effective and/or safe doses based on ratios. Tamhane *et al.* [12] discussed determining sample sizes in the context of estimating the maximum safe dose.

Power or sample size formulas related to simultaneous testing of non-inferiority of several treatments against a common control, without prioritizing the hypothesis, are yet not available. This paper aims at bridging this gap and derives the power formulas associated with the single-step multiple test procedure described by Hauschke and Kieser [9]. Due to the inherent multiplicity aspect, different power concepts are available. They are thoroughly discussed in the context of non-inferiority testing and advice is given on how to proceed in practical situations. The inverse problem of determining the necessary sample size for a given power is also addressed and numerical comparisons related to the different power definitions are performed. In particular, we investigate how the required sample sizes for the ratio-based inference compare with that of the inference based on differences (or absolute margin) for various scenarios.

Accordingly, the layout of the paper is as follows. In Section 2, we present two examples of pharmacological and clinical trials to motivate our discussions. In Section 3, the statistical methodologies related to simultaneous testing as well as power and sample size computations are described in detail. In Section 4, we present the results of a numerical power study and re-visit the motivating examples. Concluding remarks are given in Section 5.

2. EXAMPLES

In this section we introduce two examples to motivate the subsequent discussions. We come back to these examples in Section 4 in order to illustrate the resulting power and sample size formulas.

2.1. Example 1: Non-inferiority in a pharmacological study

Consider the problem of determining the sample size in a simultaneous non-inferiority test of three intermittent administration schedules against a continuous administration of the same total dose of ibandronate in a long-term *in vivo* study on osteoporosis [13]. The continuous administration represents the active control group and the endpoint is trabecular bone mass in tibiae (in per cent). According to previous results, it can be assumed that the coefficient of variation for the control group is 50 per cent (mean and standard deviation of 10 and 5 per cent, respectively). Suppose that the interest is to design a new confirmatory non-inferiority trial based on these previous study results with a relative non-inferiority margin of 0.70 (which was explicitly described for trabecular bone mass in osteoporosis trials [14]), a minimal power of 80 per cent and overall type I error rate $\alpha = 0.05$. In particular, it is of interest to investigate the optimum sample size allocation across the four treatment arms when the total sample size is fixed.

2.2. Example 2: Superiority in a clinical trial

Knapp *et al.* [15] described a double-blind, placebo-controlled, multi-centre trial with four arms on subjects with hypercholesterolaemia. The study involved the comparisons of placebo with two doses of Simvastatin and a third treatment group with a fixed Simvastatin/Colesevelam dose combination. The primary outcome variable was the serum LDL cholesterol level after 45 days. The goal of this study was to show superiority by a cholesterol level reduction of at least 10 per cent over placebo in at least one treatment arm. In this example, small response values indicate better treatment benefits. Assuming that a follow-up confirmatory trial is planned, we are interested in calculating the necessary sample size for the proof of efficacy due to the superiority associated with a minimal power of 80 per cent and $\alpha = 0.025$. From the previous study, it can be assumed that the coefficient of variation for the control group is 17 per cent (mean and standard deviation of 177 and 30, respectively).

3. METHODOLOGY

3.1. Simultaneous tests for non-inferiority

We consider the problem of simultaneously comparing r experimental treatments $(i=1,\dots,r)$ with one active control treatment $(i=0)$. Let Y_{ij} denote independent observations from a normal distribution with mean μ_i and common unknown variance σ^2, $i=0,1,\dots,r$, $j=1,\dots,n_i$. The primary interest is to simultaneously test the non-inferiority of the r treatments as compared with the control based on the ratios of the treatment means to the control mean. The problem of multiple testing concerning several ratios in non-inferiority trials was addressed by Hauschke and Kieser [9]. The objective is now to determine the power and sample sizes associated with these tests, and also to compare the results with that of the classical approach of formulating the hypotheses (in terms of difference of treatment means).

Without loss of generality, we consider the case when the responses are non-negative and large responses are associated with better treatment effects. We want to test

$$H_{0\ell}: \gamma_\ell \leqslant \psi \quad \text{against} \quad H_{1\ell}: \gamma_\ell > \psi, \quad \ell = 1,2,\dots,r \tag{1}$$

where $\gamma_\ell = \mu_\ell/\mu_0$ denotes the ratio of the mean of the ℓth treatment to that of the control and $\psi<1$ is the relative non-inferiority margin. The alternative hypothesis $H_{1\ell}$ states that the ℓth treatment is non-inferior to the control. Assuming that the treatment means are non-negative, the hypotheses in (1) can equivalently be stated as

$$H_{0\ell}: \mu_\ell - \psi\mu_0 \leqslant 0 \quad \text{against} \quad H_{1\ell}: \mu_\ell - \psi\mu_0 > 0, \quad \ell = 1,2,\dots,r$$

which naturally lead us to the following test statistics. Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ and $S^2 = \sum_{i=0}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2/v$, respectively, denote the unbiased estimators of μ_i and σ^2, where $v = \sum_{i=0}^{r} (n_i - 1)$. The likelihood ratio statistics to test the hypotheses in (1) are

$$T_\ell(\psi) = \frac{\bar{Y}_\ell - \psi\bar{Y}_0}{S\sqrt{\dfrac{1}{n_\ell} + \dfrac{\psi^2}{n_0}}}, \quad \ell = 1,2,\dots,r \tag{2}$$

each of which follows a t-distribution with v degrees of freedom under $H_{0\ell}$ [4]. The vector $\mathbf{T} = (T_1,\dots,T_r)'$ of the test statistics then follows a central r-variate t-distribution with v degrees of freedom and a correlation matrix $\mathbf{R}(\psi) = [\rho_{ij}(\psi)]$, where

$$\rho_{ij}(\psi) = \frac{\psi}{\sqrt{\dfrac{n_0}{n_i} + \psi^2}} \frac{\psi}{\sqrt{\dfrac{n_0}{n_j} + \psi^2}}, \quad 1 \leqslant i \neq j \leqslant r \tag{3}$$

(see Reference [16]). This is based on the fact that $(\bar{Y}_\ell - \psi\bar{Y}_0)$, $\ell = 1,\dots,r$, jointly follows a multivariate normal distribution with mean $\mathbf{0}$ and correlation matrix $\mathbf{R}(\psi)$. We reject the null hypothesis $H_{0\ell}$ in (1) if the test statistic $T_\ell(\psi)$ is larger than a suitable multiplicity adjusted critical point c (see Reference [17]). The method in this paper aims at controlling the family-wise error rate (FWER) in the strong sense. That is, if $I_0 \subseteq I = \{1,\dots,r\}$ denotes the index set of all true null hypotheses, then

$$\text{FWER} = 1 - P\{T_\ell(\psi) \leqslant c \text{ for all } \ell \in I_0\}$$

In order to get FWER $\leqslant \alpha$ for any configuration of the true null hypotheses, we require FWER $= \alpha$ in case of the least favourable configuration $I_0 = I$. In this paper, since the interest is in tests for non-inferiority (or superiority), α is one-tailed. The critical point c is seen to be an equi-co-ordinate percentage point of $\mathbf{T}$ satisfying

$$P\{T_1(\psi) \leqslant c, \ldots, T_r(\psi) \leqslant c\} = 1 - \alpha \tag{4}$$

Note that the correlation matrix $\mathbf{R}(\psi)$ in (3) has a product correlation structure, i.e. we can factorize $\rho_{ij}(\psi)$ as $\rho_{ij}(\psi) = \lambda_i \lambda_j$, where $\lambda_i = \left(n_i\psi^2/(n_0 + n_i\psi^2)\right)^{1/2}$. This property enables us to reduce the dimension of the multivariate normal integrals involved in the computation of multivariate t equi-co-ordinate percentage points (see, e.g. Reference [18]).

Before closing this section, we give some remarks on other cases or possible generalizations of the setup above. First, in the situation above where large response values indicate better treatment effects, the choice $\psi \geqslant 1$ corresponds to a test for superiority. Secondly, for the tests in (1), one may also use unequal non-inferiority margins ψ_ℓ, $\ell = 1, \ldots, r$, if such is of interest. In all cases, the derivations in the subsequent sections are equally applicable.

3.2. Power computations

A major task in the design phase of a clinical study is that of determining sample sizes which guarantee a pre-specified power. In this section, we provide the power associated with the tests described in the previous section. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)'$ denote a vector of ratios on the basis of which we compute the power. Note that some of the θ_l's may be less than or equal to ψ (i.e. under H_{0l}). When some of the $H_{1\ell}$'s are true, the vector of test statistics $\mathbf{T}$ has a non-central r-variate t-distribution with v degrees of freedom, correlation matrix $\mathbf{R}(\psi)$, and non-centrality vector $\boldsymbol{\delta}^{\text{ratio}}$, where the elements of $\boldsymbol{\delta}^{\text{ratio}}$ are given by

$$\delta_\ell^{\text{ratio}} = \frac{\theta_\ell - \psi}{\kappa_0\sqrt{\dfrac{1}{n_\ell} + \dfrac{\psi^2}{n_0}}}, \quad \ell = 1, \ldots, r$$

and $\kappa_0 = \sigma/\mu_0$ denotes the coefficient of variation of the control group. Note that the non-centrality parameters are dimensionless as are the ratio parameters γ_ℓ. The power function depends on the non-inferiority margin ψ, the fixed ratios θ_ℓ, the coefficient of variation of the control group κ_0 and the sample sizes n_i, $i = 0, 1, \ldots, r$. For the computation of non-central multivariate t probabilities, we refer to numerical integration routines of Genz and Bretz [19] which can be directly used for the power calculations [20]. A computer program which implements the following power formulas is provided in the Appendix.

Let θ^* ($> \psi$) denote the greatest clinically irrelevant percentage of the control mean which is to be detected. Define the set of indices $I(\theta^*) = \{\ell \mid \theta_\ell > \theta^*\} = \{\ell_1, \ldots, \ell_m\}$, $1 \leqslant m \leqslant r$. All treatments with θ_ℓ values greater than θ^* are non-inferior to the control. We consider the following two power definitions.

(i) *Complete power*: Suppose that the interest is to detect all non-inferior treatments with a given power of $1 - \beta$. The power associated with this problem is called complete (or all-pairs) power and it is given by

$$\pi_{\text{Com}}(\boldsymbol{\theta}, \theta^*) = P\{T_\ell > c, \text{ for all } \ell \in I(\theta^*)\}$$
$$= \int_0^\infty \int_{-\infty}^\infty \prod_{\ell \in I(\theta^*)} \Phi\left(-\frac{c\eta - \delta_\ell^{\text{ratio}} + \lambda_\ell z}{\sqrt{1-\lambda_\ell^2}}\right) \phi(z)\varphi(\eta)\,\mathrm{d}z\,\mathrm{d}\eta \tag{5}$$

where $\Phi(.)$ and $\phi(.)$, respectively, denote the cumulative density function and the density function of the univariate standard normal distribution, and $\varphi(.)$ is the density function of $(\chi_v^2/v)^{1/2}$.

(ii) *Minimal power*: Suppose that the interest is to detect at least one non-inferior treatment with a given power of $1-\beta$. This is called minimal (or any-pair) power. The power of this test is given by

$$\pi_{\text{Min}}(\boldsymbol{\theta}, \theta^*) = P\{T_\ell > c, \text{ for some } \ell \in I(\theta^*)\}$$
$$= 1 - \int_0^\infty \int_{-\infty}^\infty \prod_{\ell \in I(\theta^*)} \Phi\left(\frac{c\eta - \delta_\ell^{\text{ratio}} + \lambda_\ell z}{\sqrt{1-\lambda_\ell^2}}\right) \phi(z)\varphi(\eta)\,\mathrm{d}z\,\mathrm{d}\eta \tag{6}$$

For other definitions of power in simultaneous testing (e.g. individual power and proportional power), we refer to Westfall *et al.* [21] and Horn and Vollandt [22].

We now return to the practical problem of determining the sample size associated with a given lower bound $1-\beta$ of the power. Note that all parameters of the distribution of $\mathbf{T}$ depend on the sample sizes for each treatment. For simplicity, we consider a balanced design with n observations per treatment. The required size n is determined iteratively by starting with a given sample size and search until the power condition is satisfied. That is, for minimal power, we look for the smallest n such that $\pi_{\text{Min}}(\boldsymbol{\theta}, \theta^*) \geqslant 1-\beta$ and similarly for complete power the smallest n for which $\pi_{\text{Com}}(\boldsymbol{\theta}, \theta^*) \geqslant 1-\beta$. In the Appendix, we also provide programs for the computation of the necessary sample size to achieve a certain pre-specified power.

A more practical allocation is to consider the case where a different number of subjects is allocated to the control group than to the other treatment groups, i.e. $n_0 \neq n_1 = \cdots = n_r$. It is well known that in case of testing for differences, the square-root allocation rule $n_0/n_\ell = \sqrt{r}$, $\ell = 1, \ldots, r$, is nearly optimal [17, 23]. If we use the same idea of minimizing $\text{Var}(\bar{Y}_\ell - \psi\bar{Y}_0)$, $\ell = 1, \ldots, r$ subject to a fixed total sample size $N = \sum_{i=0}^r n_i$, then we get the solution $n_0 = \psi\sqrt{r}n_\ell$. The power behaviour of this allocation will be discussed in later sections.

3.3. Relative efficiency of the ratio-based test

In this section, we investigate the advantage of ratio-based inference over the classical difference-based inferences in a multiple testing situation. In case of a single ratio (comparing two treatments), Laster and Johnson [3] showed that the ratio-based inference is more efficient as long as $\psi < 1$. That is, the sample size associated with the ratio-based inference is smaller than that of a comparable inference based on the difference of both means in tests for non-inferiority. In multiple testing, the results based on the ratio view can also be compared with tests based on difference (see Reference [8]). To do this, we reformulate the tests in (1) as

$$H_{0\ell} : \mu_0 - \mu_\ell \geqslant \Delta_0 \quad \text{against} \quad H_{1\ell} : \mu_0 - \mu_\ell < \Delta_0, \quad \ell = 1, 2, \ldots, r \tag{7}$$

where the *absolute non-inferiority margin* $\Delta_0 > 0$ is fixed at $\Delta_0 = (1-\psi)\mu_0$. The likelihood ratio statistics to test the hypotheses in (7) are

$$T_\ell(\Delta_0) = \frac{\bar{Y}_0 - \bar{Y}_\ell - \Delta_0}{S\sqrt{\dfrac{1}{n_0} + \dfrac{1}{n_\ell}}}, \quad \ell = 1, 2, \ldots, r$$

Under the null hypotheses in (7), $T_\ell(\Delta_0)$ follows a central t-distribution with v degrees of freedom. Jointly, $\mathbf{T}(\Delta_0) = (T_1(\Delta_0), \ldots, T_r(\Delta_0))'$ is distributed as a central r-variate t-distribution with v degrees of freedom and a correlation matrix $\mathbf{R}(\Delta_0) = [\rho_{ij}(\Delta_0)]$. The correlation $\rho_{ij}(\Delta_0)$ has also a product correlation structure and can be written as $\rho_{ij}(\Delta_0) = \lambda_i \lambda_j$, where $\lambda_i = (n_i/(n_0 + n_i))^{1/2}$. Unlike in the case of the ratio-based inference, note that the correlation matrix for the inference based on difference does not depend on the non-inferiority margin. If the design is balanced, we have $\rho_{ij}(\Delta_0) = 0.5,\ 1 \leqslant i \neq j \leqslant r$. The $H_{0\ell}$ hypotheses in (7) will be rejected if $T(\Delta_0) < -c$, where c is an equi-co-ordinate percentage point of $\mathbf{T}(\Delta_0)$. Now, to obtain a comparable power with that of the ratio-based inference, we set the vector of mean differences $(\mu_0 - \mu_1, \ldots, \mu_0 - \mu_r)'$ to $\Delta = (\Delta_1, \ldots, \Delta_r)'$, where $\Delta_\ell = (1-\theta_\ell)\mu_0,\ \ell = 1, \ldots, r$. Under the alternative hypotheses, $\mathbf{T}(\Delta_0)$ has a non-central r-variate t-distribution with v degrees of freedom, a correlation matrix $\mathbf{R}(\Delta_0)$ and non-centrality vector $\boldsymbol{\delta}^{\text{diff}}$, where

$$\delta_\ell^{\text{diff}} = \frac{\Delta_\ell - \Delta_0}{\sigma\sqrt{\dfrac{1}{n_0} + \dfrac{1}{n_\ell}}} = \frac{\psi - \theta_\ell}{\kappa_0\sqrt{\dfrac{1}{n_0} + \dfrac{1}{n_\ell}}}, \quad \ell = 1, \ldots, r$$

Consider the minimal power $\pi_{\text{Min}}(\Delta, \Delta^*)$, where $\Delta^* = (1-\theta^*)\mu_0$. The sample size n required per treatment (in a balanced design) is the smallest n such that $\pi_{\text{Min}}(\Delta, \Delta^*) \geqslant 1 - \beta$. It is observed that the power function increases as both the elements of the correlation matrix and the non-centrality parameters increase, and *vice versa*. For a given $\psi < 1$, κ_0 and $\boldsymbol{\theta}$, note that $\rho_{ij}(\Delta_0) > \rho_{ij}(\psi)$ but $0 < -\delta_\ell^{\text{diff}} < \delta_\ell^{\text{ratio}}$. Therefore, there is no easy analytical way of comparing the power functions of the ratio-based and difference-based inferences as in the single-ratio case. However, from the plot of the power against the elements of the correlation matrix and the non-centrality parameters (not shown here), it is observed that the elements of the correlation matrix have little impact on the power compared with the impact of the non-centrality parameters. Thus, the difference in powers of the two approaches is mainly due to the differences in the non-centrality parameters. Let n_{ratio} and n_{diff} denote the number of observations required per treatment by the ratio and difference approaches, respectively. In Section 4, we show by various numerical examples that $n_{\text{ratio}} \leqslant n_{\text{diff}}$ if we have increasing effect (i.e. for the hypotheses in (1)) in tests for non-inferiority with $\psi < 1$.

3.4. Least favourable configuration

As noted in Section 3.2, power computation in multiple testing relies on the knowledge about the configuration of the m true alternative hypotheses with $\theta_\ell > \theta^*$. Typically, the number m

of true alternatives is not known in advance. One possibility is to evaluate the power at the least favourable configuration (LFC), i.e. at the parameter configuration under the alternative hypotheses at which the smallest power is obtained. For inferences based on the difference of location parameters, Horn and Vollandt [22] derived LFCs for various power definitions. Along a similar line, we obtain LFCs associated with the ratio-based inference for complete and minimal power in one-sided tests for non-inferiority (superiority). Suppose that *a priori* one knows upper and lower bounds on m, i.e. $g \leqslant m \leqslant h$, where g and h are integers such that $1 \leqslant g \leqslant h \leqslant r$. We consider the case $n_0 \neq n_1 = \cdots = n_r = n$, that is $\lambda_\ell = \lambda$. From the expression for complete power in (5), we see that

$$\begin{aligned}\pi_{\mathrm{Com}}(\boldsymbol{\theta}, \theta^*) &= \int_0^\infty \int_{-\infty}^\infty \prod_{\ell \in I(\theta^*)} \Phi\left(\frac{\theta_\ell - \psi}{\kappa_0}\sqrt{n} - \frac{c\eta + \lambda_\ell z}{\sqrt{1-\lambda_\ell^2}}\right) \phi(z)\varphi(\eta)\,\mathrm{d}z\,\mathrm{d}\eta \\ &> \int_0^\infty \int_{-\infty}^\infty \Phi^m\left(\frac{\theta^* - \psi}{\kappa_0}\sqrt{n} - \frac{c\eta + \lambda z}{\sqrt{1-\lambda^2}}\right) \phi(z)\varphi(\eta)\,\mathrm{d}z\,\mathrm{d}\eta \\ &\geqslant \int_0^\infty \int_{-\infty}^\infty \Phi^h\left(\frac{\theta^* - \psi}{\kappa_0}\sqrt{n} - \frac{c\eta + \lambda z}{\sqrt{1-\lambda^2}}\right) \phi(z)\varphi(\eta)\,\mathrm{d}z\,\mathrm{d}\eta \\ &= P\{T_1 > c, \ldots, T_h > c \mid \theta_1 = \cdots = \theta_h = \theta^*\}\end{aligned}$$

Thus, a LFC for the complete power is $\theta_1 = \cdots = \theta_h = \theta^*$, $\theta_\ell < \theta^*$ for $\ell > h$, if $g \leqslant m \leqslant h$. That is, if we compute the power at a LFC, for any other configuration of $\boldsymbol{\theta}$, the resultant power is larger than $1-\beta$. Note that $\theta_{r-h+1} = \cdots = \theta_{r-1} = \theta_r = \theta^*$, $\theta_\ell < \theta^*$ for $\ell \leqslant r-h$ is also a LFC. Therefore, the LFCs are permutation invariant. When there is no prior information about m (i.e. $g=1$ and $h=r$), $\theta_1 = \cdots = \theta_r = \theta^*$ is a LFC. In a similar manner, from (6) it can be shown that

$$\pi_{\mathrm{Min}}(\boldsymbol{\theta}, \theta^*) \geqslant 1 - P\{T_1 < c, \ldots, T_g < c \mid \theta_1 = \cdots = \theta_g = \theta^*\}$$

Therefore, $\theta_1 = \cdots = \theta_g = \theta^*$, $\theta_\ell < \theta^*$ for $\ell > g$, constitutes a LFC for the minimal power if *a priori* $g \leqslant m \leqslant h$, and $\theta_1 = \theta^*$, $\theta_\ell < \theta^*$ for $\ell > 1$ is a LFC if there is no prior information about m.

4. RESULTS

4.1. Numerical study

In this section, we investigate the power and the associated sample sizes for both the ratio-based and difference-based inferences. Various scenarios of the coefficient of variation for the control group ($\kappa_0 100$ per cent) and specific ratio parameter configurations (θ) under the alternative hypotheses are considered. Suppose that large response values indicate better treatment effect. Figure 1 shows the minimal power function at LFC for the ratio-based test with three comparisons ($r=3$). The figure compares the power functions for various sample sizes n in a balanced design. As one would expect, the power is an increasing function of the clinically irrelevant percentage θ^* and larger sample size lead to larger power values.

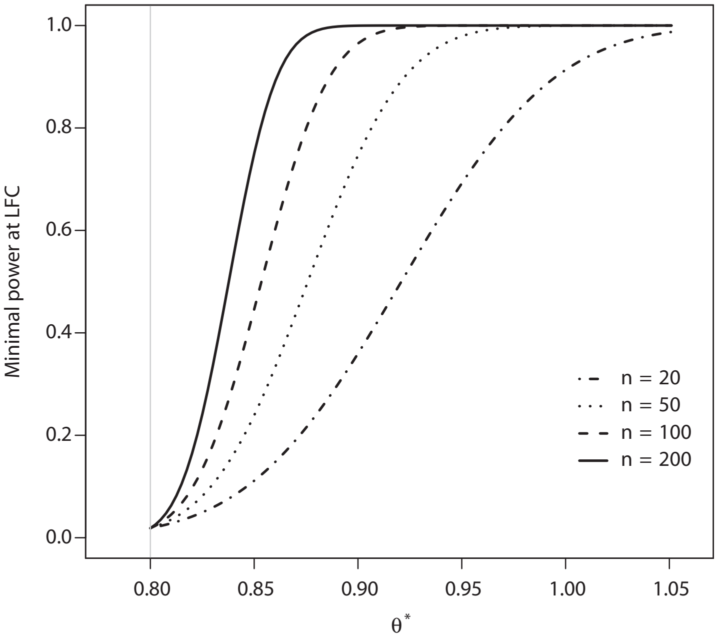

Figure 1. Comparisons of minimal power curves at LFC for various balanced sample sizes n when $r=3$, $g=1$, $\psi=0.8$, $\kappa_0=0.2$, and $\alpha=0.05$.

Figure 2 shows the minimal power differences between the ratio and difference approach $\pi_{\text{Min}}(\boldsymbol{\theta},\theta^*)-\pi_{\text{Min}}(\Delta,\Delta^*)$ when there is no prior information about the correct configuration of the θ_ℓs for different number of comparisons r. From this figure, we see that the ratio-based is more efficient (in terms of power) than the difference-based in tests for non-inferiority with $\psi<1$ (assuming that large response values correspond to better treatment effects). This result is in line with the results obtained by Laster and Johnson [3] for a single ratio ($r=1$). Figure 2, thus, indicates that this result also holds true for multiple ratios ($r>1$). If the θ values fall far to the right of ψ, the power functions for both ratio and difference-based tests are close to one and the two approaches practically do not differ, i.e. the power difference is close to zero.

When interest lies in controlling the complete power, we have a slightly different situation in the power differences $\pi_{\text{Com}}(\boldsymbol{\theta},\theta^*)-\pi_{\text{Com}}(\Delta,\Delta^*)$ at LFCs. As shown in Figure 3, the power of the difference-based testing is slightly greater than that of the ratio for θ^* values near the non-inferiority margin ψ. In other words, for a fixed very small complete power, n_{diff} can be smaller than n_{ratio}, even if $\psi<1$.

We now consider the impact of different parameter constellations on the resulting sample sizes. Sample sizes are determined based on LFC when no prior information concerning m is available (i.e. $g=1$ when controlling the minimal power and $h=r$ for complete power). Tables I and II consist of the sample sizes required for a given minimal power and complete power, respectively. From the tables, it can be seen that the sample size required for the

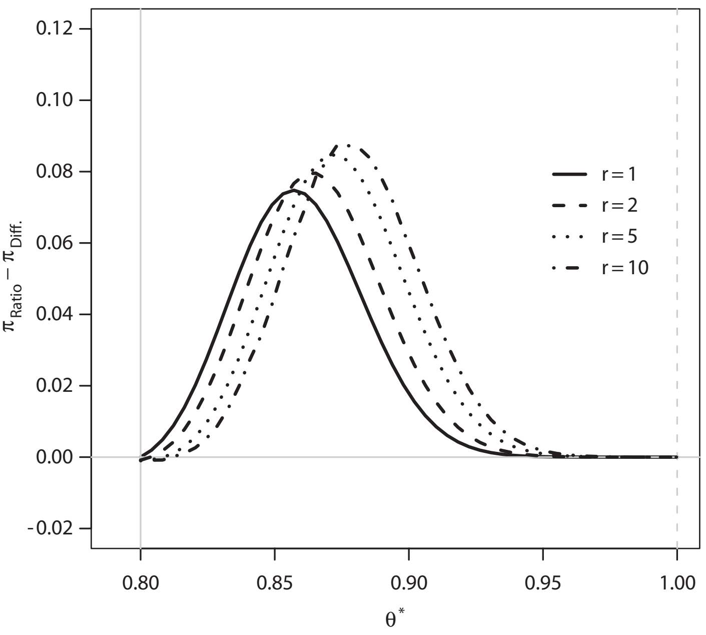

Figure 2. Comparisons of the differences between the minimal power at LFC for the ratio-based and the difference-based tests when $g=1$, $\psi=0.80$, $\kappa_0=0.2$, and $n=100$ (balanced design).

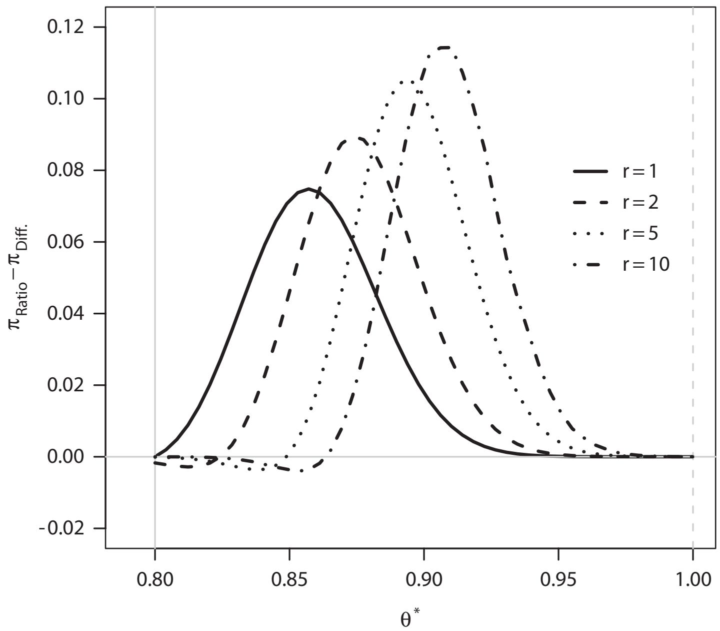

Figure 3. Comparisons of the differences between the complete power at LFC for the ratio-based and the difference-based tests when $h=r$, $\psi=0.80$, $\kappa_0=0.2$, and $n=100$ (balanced design).

Table I. Comparisons of the required sample sizes $n_{\text{ratio}}(n_{\text{diff}})$ in tests for non-inferiority given a minimal power of $1-\beta$ when large response values indicate better treatment effects ($r=3$, $g=1$, $\psi=0.80$, $\alpha=0.05$).

		θ^*				
$\kappa_0(\%)$	$1-\beta$	0.85	0.90	0.95	1	1.10
10	0.75	51 (61)	14 (16)	7 (8)	4 (5)	3 (3)
	0.80	57 (68)	15 (18)	7 (9)	5 (5)	3 (3)
	0.90	75 (90)	20 (23)	9 (11)	6 (7)	3 (4)
	0.95	92 (111)	24 (28)	11 (13)	7 (8)	4 (4)
20	0.75	201 (241)	51 (61)	23 (28)	14 (16)	7 (8)
	0.80	226 (271)	57 (68)	26 (31)	15 (18)	7 (9)
	0.90	298 (359)	75 (90)	34 (41)	20 (23)	9 (11)
	0.95	366 (441)	92 (111)	42 (50)	24 (28)	11 (13)
50	0.75	1249 (1499)	313 (375)	140 (167)	79 (95)	36 (43)
	0.80	1404 (1687)	352 (423)	157 (188)	89 (106)	40 (48)
	0.90	1858 (2237)	465 (560)	207 (249)	117 (141)	53 (63)
	0.95	2281 (2749)	571 (688)	254 (306)	144 (173)	64 (77)

Table II. Comparisons of the required sample sizes $n_{\text{ratio}}(n_{\text{diff}})$ in tests for non-inferiority given a complete power of $1-\beta$ when large response values indicate better treatment effects ($r=3$, $h=3$, $\psi=0.80$, $\alpha=0.05$).

		θ^*				
$\kappa_0(\%)$	$1-\beta$	0.85	0.90	0.95	1	1.10
10	0.75	73 (85)	19 (22)	9 (10)	6 (6)	3 (3)
	0.80	79 (93)	21 (24)	10 (11)	6 (7)	3 (4)
	0.90	98 (116)	25 (30)	12 (14)	7 (8)	4 (5)
	0.95	115 (137)	29 (35)	14 (16)	8 (10)	4 (5)
20	0.75	289 (339)	73 (85)	33 (38)	19 (22)	9 (10)
	0.80	315 (371)	79 (93)	36 (42)	21 (24)	10 (11)
	0.90	388 (461)	98 (116)	44 (52)	25 (30)	12 (14)
	0.95	456 (544)	115 (137)	52 (61)	29 (35)	14 (16)
50	0.75	1801 (2114)	451 (529)	201 (236)	113 (133)	51 (60)
	0.80	1962 (2312)	491 (579)	219 (258)	124 (145)	55 (65)
	0.90	2423 (2879)	606 (720)	270 (321)	152 (181)	68 (81)
	0.95	2843 (3395)	711 (849)	317 (378)	179 (213)	80 (95)

complete power (Table II) is substantially larger than that of the minimal power (Table I). In the tables, we compare the sample sizes required by the ratio-based inference (n_{ratio}) with that of the inferences based on difference of means (n_{diff}). To this end, the mean of the control group is assumed unchanged and is fixed at an arbitrary value of μ_0 and the common standard deviation is fixed at $\sigma=\kappa_0\mu_0$. Tables I and II show that smaller sample sizes are associated

Table III. Comparisons of the required sample sizes $n_{\text{ratio}}(n_{\text{diff}})$ in tests for superiority given a minimal power of $1-\beta$ when large response values indicate better treatment effects ($r=3$, $g=1$, $\psi=1.20$, $\alpha=0.05$).

		θ^*				
κ_0(%)	$1-\beta$	1.25	1.30	1.35	1.40	1.50
10	0.75	73 (61)	19 (16)	9 (8)	6 (5)	3 (3)
	0.80	82 (68)	21 (18)	10 (9)	6 (5)	3 (3)
	0.90	109 (90)	28 (23)	13 (11)	8 (7)	4 (4)
	0.95	133 (111)	34 (28)	16 (13)	9 (8)	5 (4)
20	0.75	288 (241)	73 (61)	33 (28)	19 (16)	9 (8)
	0.80	325 (271)	82 (68)	37 (31)	21 (18)	10 (9)
	0.90	431 (359)	109 (90)	49 (41)	28 (23)	13 (11)
	0.95	531 (441)	133 (111)	60 (50)	34 (28)	16 (13)
50	0.75	1797 (1499)	450 (375)	201 (167)	113 (95)	51 (43)
	0.80	2025 (1687)	507 (423)	226 (188)	127 (106)	57 (48)
	0.90	2691 (2237)	673 (560)	300 (249)	169 (141)	76 (63)
	0.95	3312 (2749)	829 (688)	369 (306)	208 (173)	93 (77)

with the ratio-based inference for the cases under investigation. As remarked in Section 3.1, for large response values indicating better treatment effects and when the margin ψ is greater than 1, we have test for superiority problem. In this case, it is found that $\pi_{\text{Min}}(\boldsymbol{\theta},\theta^*)<\pi_{\text{Min}}(\Delta,\Delta^*)$ for $\psi<\theta$, and hence ratio-based inferences lead to lower power. Table III consists of the sample sizes required for this setting. The superiority margin is chosen to be $\psi=1.20$. As can be seen from the table, the sample size required for the ratio-based test is larger than that of the difference. Comparing Tables I and III, one also discerns the symmetry in the sample size required by the inference based on difference (n_{diff} is the same in Tables I and III). This is not only a numerical finding but it is also theoretically expected (since the correlation matrix and the non-centrality parameters for the two tables are identical). For the ratio approach, this kind of symmetry does not hold true.

In summary, for the ratio-based inference, a smaller sample size is required in tests for non-inferiority with $\psi<1$ compared with the common difference-based inference.

4.2. The case of small response values indicating better treatment effects

For small response values indicating better treatment effects, the choice of $\psi<1$ leads to test for superiority while $\psi>1$ is test for non-inferiority. Therefore, in simultaneous tests for non-inferiority with small response values indicating better treatment benefits, the hypotheses to be tested are

$$H_{0\ell} : \gamma_\ell \geqslant \psi \quad \text{against} \quad H_{1\ell} : \gamma_\ell<\psi, \quad \ell=1,2,\ldots,r \tag{8}$$

where now the non-inferiority margin $\psi>1$. In this case, we decide the ℓth treatment to be non-inferior to the control if $T_\ell(\psi)<-c$. The critical point c is computed as in equation (4) by using the value of ψ fixed in (8). From the symmetry of the distribution of $T_\ell(\psi)$ under the null, it can be shown that the power behaviours of the problem in (8) are exactly the same as the power behaviours of simultaneous tests for superiority when large response values indicate better treatment effect and $\psi > 1$. The sample sizes associated with the latter scenario is given in Table III. Thus, there is no sample size advantage in simultaneous tests for non-inferiority with $\psi > 1$. In order to maintain the sample size advantage of the ratio view, if it gives sense, one may make inference about the ratio of control mean to that of the test treatments (as suggested for a two-sample problem [3]). Thus, if we invert the ratios in (8), the hypotheses to be tested are

$$H_{0\ell} : \mu_0/\mu_\ell \leqslant \psi_1 \quad \text{against} \quad H_{1\ell} : \mu_0/\mu_\ell > \psi_1, \quad \ell = 1, 2, \ldots, r \tag{9}$$

where $\psi_1 < 1$. The associated joint distribution of the test statistics to test the hypotheses in (9) has a multivariate t-distribution with off-diagonal elements of the correlation matrix given by

$$\rho_{ij}(\psi_1) = \frac{1}{\sqrt{1 + \dfrac{n_0}{n_i}\psi_1^2}} \frac{1}{\sqrt{1 + \dfrac{n_0}{n_j}\psi_1^2}}, \quad 1 \leqslant i \neq j \leqslant r$$

For power computations, the corresponding non-centrality parameters are given by

$$\delta_\ell^{\text{ratio}}(\psi_1) = \frac{1 - \psi_1(\mu_0/\mu_\ell)^{-1}}{\kappa_0\sqrt{\dfrac{1}{n_0} + \dfrac{1}{n_\ell}\psi_1^2}}$$

Now, for comparing the power associated with the tests in (9) with the difference-based test, we consider the hypotheses

$$H_{0\ell} : \mu_\ell - \mu_0 \geqslant \Delta_\ell \quad \text{against} \quad H_{1\ell} : \mu_\ell - \mu_0 < \Delta_\ell, \quad \ell = 1, 2, \ldots, r$$

where $\Delta_\ell = (1 - \psi_1)\mu_\ell$. Here, we encounter *varying absolute non-inferiority margins* which depend on the mean of the new treatments. For these choices of the delta-margins, the corresponding ratio-formatted tests are more powerful. It might be more desirable to have identical absolute delta-margins across the comparisons. In this case, from the relationship between μ_0 and μ_ℓ on the boundaries of the $H_{0\ell}$ hypotheses in (9), we can write $\mu_\ell = \mu_0/\psi_1$. Substituting this in $\Delta_\ell = (1 - \psi_1)\mu_\ell$, we get another delta-margin of $(1/\psi_1 - 1)\mu_0$. Often, there exists more prior information about the standard treatment (the control) than the new treatments. Thus, the latter approach seems to be a more practical way of choosing the delta-margins. However, for this second choice of the delta-margins, the difference-based test is more efficient.

4.3. *Re-visiting the examples*

Now, let us determine the sample sizes required for the two motivating examples in Section 2. All computations are carried out for the least favourable configuration.

4.3.1. Example 1. For the osteoporosis study described in Section 2.1, we compute the sample size associated with a non-inferiority margin of $\psi = 0.70$, a minimal power of 80 per cent, $\alpha = 0.05$, and $\kappa_0 = 0.50$. We further assume that the greatest clinically irrelevant percentage of the control mean which is to be detected is $\theta^* = 0.95$. Since there is no prior information

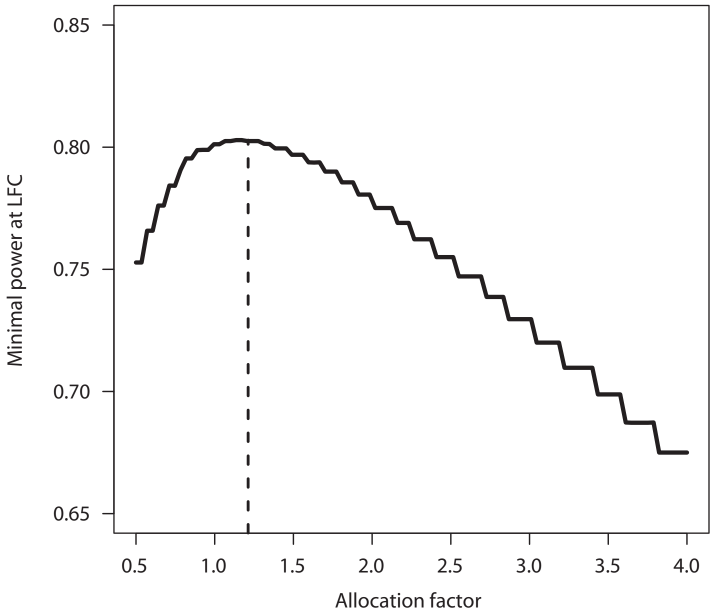

Figure 4. Minimal power at LFC *versus* allocation factor (n_0/n_ℓ) for a fixed total sample size $N=208$. The vertical dotted line is at the allocation factor $\psi\sqrt{r}=1.212$ (Example 1).

about m, a LFC is $\theta_1=0.95$, $\theta_2,\theta_3<0.95$. Under these conditions, for the one-sided simultaneous test of three treatment schedules *versus* an active control, the required number of observations per treatment is $n_{\text{ratio}}=52$ ($n_{\text{diff}}=68$) in a balanced design. For the ratio-based testing, if we use the allocation rule $n_0=\psi\sqrt{r}n_\ell$ with a fixed total sample size of $N=4\times 52=208$, the number of observations required per treatment are $n_0=60$ and $n_1=n_2=n_3=50$. The power of this allocation is 0.807. The graph of power *versus* other allocation factors (n_0/n_ℓ) is shown in Figure 4. The power attains its maximum at the allocation factor $0.7\times\sqrt{3}$ (the vertical dotted line in Figure 4). Note that this holds true when $g=1$. If $g>1$ or if the interest is to control the complete power (with $h=3$), the maximum power over all possible allocations is slightly greater than the power at the allocation $0.7\times\sqrt{3}$. For the difference-based inference, using the allocation rule $n_0=\sqrt{r}n_\ell$ with a fixed total sample size of $N=4\times 68=272$, the number of observations required per treatment are $n_0=100$ and $n_1=n_2=n_3=58$.

4.3.2. Example 2. The clinical data example described in Section 2.2 is the case when small response values indicate better treatment effect. The superiority margin is $\psi=0.90$, the given minimal power is 0.80, $\alpha=0.025$, and $\kappa_0=0.17$. Let the largest clinically relevant percentage to be detected be $\theta^*=0.85$. Since there is no prior information about m, a LFC is $\theta_1=0.85$, $\theta_2,\theta_3>0.85$. Thus, the number of observations required per treatment is $n_{\text{ratio}}=215$ ($n_{\text{diff}}=237$) in a balanced design. Therefore, in terms of sample size, the ratio approach is more efficient in designing a new confirmatory clinical trial. For the same trial, if one wishes to control the complete power, the LFC is $\theta_1=0.85$, $\theta_2=0.85$, $\theta_3=0.85$, and the required sample size for each treatment is $n_{\text{ratio}}=290$ ($n_{\text{diff}}=315$).

We remark that when controlling minimal power with $g>1$, the power at the allocation factor $\psi\sqrt{r}$ is slightly smaller than the maximum power over all possible allocations. In this case, it is again possible to determine the optimum sample sizes (associated with the maximum power) iteratively by first finding the optimum allocation factor.

5. CONCLUDING REMARKS

In this paper, we considered the problem of sample size and power computations in simultaneous tests for non-inferiority based on the ratio view. The efficiency of this approach is also compared with that of tests based on the difference of location parameters. From the various numerical studies, the following results are observed. The ratio approach has advantage (i) in tests for non-inferiority with relative margins less than one and in situations where large response values indicate better treatment effects and (ii) in tests for superiority with superiority margins less than one and in situations where small response values indicate better treatment effects. The latter case is illustrated using a clinical data example (Example 2). It is not directly investigated in this paper but can be shown analogously to tests for non-inferiority in the case of large response values indicating better treatment effects. The reduction in sample size by applying the ratio approach can be clinically relevant. For instance, in Example 1, we have about 25 per cent reduction in the number of observations per treatment by applying the ratio approach. In tests for non-inferiority with small response values indicating better treatment effects (or tests for superiority with large response values indicating better treatment benefits), one may make inference for the ratios of the control to that of the test treatments. Therefore, this generalizes the ratio view for the two-sample case to that of multiple testing situation.

In conclusion, from the perspective of higher power and problem-adequate interpretation, ratio-based multiple testing can be recommended for selected non-inferiority (or superiority) trials when the interest is to control the minimal power. The related R code for the design is provided in the Appendix.

APPENDIX

Power and sample size computations are done in R, an open source statistical software package available at www.r-project.org. For the computation of multivariate t probabilities and the equi-co-ordinate percentage points c, the *pmvt* function from the *mvtnorm* package (a package which computes multivariate normal probabilities and critical points) is used, which is also available under the URL above. See Reference [24] for further details. The code below is for the ratio-based inference when large response values indicate better treatment effects, but it can also be easily modified for inferences based on differences by redefining the correlation matrix and the non-centrality parameters. In the program, *n*.ratio is a function that computes the smallest sample size (balanced design) given the number of comparisons r, the value of g (often there is no prior information about m, therefore, $g=1$), the relative non-inferiority margin (psi), the minimal power (min.power), the coefficient of variation of the control group ($k0$), the smallest clinically irrelevant percentage of the control to be detected (theta.star), the familywise type I error rate (alpha), and a starting value for the sample size (*n*.start).

```
library (mvtnorm)
n.ratio <- function (r, g, psi, min.power,k0,theta.star,alpha,n.start)  {
rho <- (psi^2)/sqrt((psi^2+1)*(psi^2+1))
RHO <- matrix(rep(rho,r*r), nr=r)
  diag(RHO) <- rep(1,r) # correlation matrix (balanced design)

  n <- n.start
  power <- 0
  eps <- 0.00001

  while(power < min.power) {
    nu <- (r+1)*(n-1)
    probq <- function(q) {
    pmvt(rep(-Inf,r),rep(q,r),nu,corr=RHO,delta=rep(0,r),abseps=eps)-(1-alpha)}
    C0 <- uniroot(probq, lower=0, upper=4) $root #computes the critical point c

    theta.vec <- rep(theta.star,g )
    deltaR <- (theta.vec - psi)/(k0*sqrt(1/n + (psi^2)/n)) #non-centrality para.
    RHO.LFC <- matrix(rep(rho,g*g), nr = g)
    diag(RHO.LFC) <- rep(1,g)

   power <- 1-pmvt(rep(-Inf,g),rep(C0,g),nu,corr=RHO.LFC,delta=deltaR,abseps=eps)
    n <- n + 1
    }
  cbind(c(sample.size=round(n-1,0),power=round(power,4)))
  }
```

For example, for the pharmacological study in Example 1, we have $r = 3$, $g = 1$, $\psi = 0.70$, $1 - \beta = 0.80$, $\kappa_0 = 0.5$, $\theta^* = 0.95$ and $\alpha = 0.05$. Set the starting value for the sample size to 2 and run the following command:

```
n.ratio(r=3, g=1, psi=0.7, min.power=0.8, k0=0.5, theta.star=0.95,
   alpha=0.05, n.start=2)
```

For sample size computations based on the complete power, we change only the fifth line from the end of the *n*.ratio function to the following command, and replace *g* by *h*.

```
power <- pmvt(rep(C0,h),rep(Inf,h),nu,corr=RHO,delta=deltaR,abseps=eps)
```

ACKNOWLEDGEMENTS

We would like to thank the anonymous referee for providing us a number of constructive comments on an earlier version of this paper. The Work of Gemechis Dilba, Frank Bretz, and Ludwig A. Hothorn is supported by the Deutsche Forschungsgemeinschaft; contract/grant number: BR 2202/1.

REFERENCES

1. D'Agostino RBSr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Statistics in Medicine* 2003; **22**:169–186.

2. Rashid MM. Rank-based tests for non-inferiority and equivalence hypotheses in multi-centre clinical trials using mixed models. *Statistics in Medicine* 2003; **22**:291–311.
3. Laster LL, Johnson MF. Non-inferiority trials: the ‘at least as good as’ criterion. *Statistics in Medicine* 2003; **22**:187–200.
4. Röhmel J. Therapeutic equivalence investigations: statistical considerations. *Statistics in Medicine* 1998; **17**: 1703–1714.
5. Hwang IK, Morikawa T. Design issues in noninferiority/equivalence trials. *Drug Information Journal* 1999; **33**:1205–1218.
6. Hauschke D, Kieser M, Diletti E, Burke M. Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Statistics in Medicine* 1999; **18**:93–105.
7. Pigeot I, Schäfer J, Röhmel J, Hauschke D. Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine* 2003; **22**:883–899.
8. Blackwelder WC. ‘Proving the null hypothesis’ in clinical trials. *Controlled Clinical Trials* 1982; **3**:345–353.
9. Hauschke D, Kieser M. Multiple testing to establish non-inferiority of k treatments with a reference based on the ratio of two means. *Drug Information Journal* 2001; **35**:1247–1251.
10. Dilba G, Bretz F, Guiard V. Simultaneous confidence sets and confidence intervals for multiple ratios. *Journal of Statistical Planning and Inference*, in press.
11. Bretz F, Hothorn LA, Hsu J. Identifying effective and/or safe doses by stepwise confidence intervals for ratios. *Statistics in Medicine* 2003; **22**:847–858.
12. Tamhane AC, Shi K, Strassburger K. Power and sample size determination for a stepwise test procedure for finding the maximum safe dose. *Journal of Statistical Planning and Inference*, in press.
13. Hothorn LA, Bauss F. Biostatistical design and analyses of long-term animal studies simulating human postmenopausal osteoporosis. *Drug Information Journal* 2004; **38**:47–56.
14. Kanis JA, Oden A, Johnell O, Caulin F, Bone H, Alexandre J-M, Abadie E, Lekkerkerker F. Uncertain future of trials in osteoporosis. *Osteoporosis International* 2002; **13**:443–449.
15. Knapp HH, Schrott H, Ma P, Knopp R, Chin B, Gaziano JM, Donovan JM, Burke SK, Davidson MH. Efficacy and safety of combination simvastatin and colesevelam in patients with primary hypercholesterolemia. *American Journal of Medicine* 2001; **110**:352–360.
16. Kotz S, Nadarajah S. *Multivariate t Distributions and Their Applications*. Cambridge University Press: Cambridge, MA, 2004.
17. Hochberg J, Tamhane A. *Multiple Comparison Procedures*. Wiley: New York, 1987.
18. Tong YL. *The Multivariate Normal Distribution*. Springer: New York, 1990; 191–193.
19. Genz A, Bretz F. Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation* 1999; **63**:361–378.
20. Bretz F, Genz A, Hothorn LA. On the numerical availability of multiple comparison procedures. *Biometrical Journal* 2001; **43**:645–656.
21. Westfall PH, Tobias RD, Rom D, Wolfinger RD, Hockberg Y. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Institute Inc.: Cary, NC, 1999.
22. Horn M, Vollandt R. Sample sizes for comparisons of k treatments with a control based on different definitions of power. *Biometrical Journal* 1998; **40**:589–612.
23. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**:1096–1121.
24. Hothorn T, Bretz F, Genz A. On multivariate t and Gauβ probabilities. *R News* 2001; **1**(2):27–29.