Superiority Inferences on Individual Endpoints Following Noninferiority Testing in Clinical Trials

Brent R. Logan[*,1] and **Ajit C. Tamhane**[2]

[1] Division of Biostatistics Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226-0509, USA

[2] Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208-3119, USA

Received 17 September 2007, revised 28 January 2008, accepted 1 May 2008

Summary

We consider the problem of drawing superiority inferences on individual endpoints following non-inferiority testing. Röhmel et al. (2006) pointed out this as an important problem which had not been addressed by the previous procedures that only tested for global superiority. Röhmel et al. objected to incorporating the non-inferiority tests in the assessment of the global superiority test by exploiting the relationship between the two, since the results of the latter test then depend on the non-inferiority margins specified for the former test. We argue that this is justified, besides the fact that it enhances the power of the global superiority test. We provide a closed testing formulation which generalizes the three-step procedure proposed by Röhmel et al. for two endpoints. For the global superiority test, Röhmel et al. suggest using the Läuter (1996) test which is modified to make it monotone. The resulting test not only is complicated to use, but the modification does not readily extend to more than two endpoints, and it is less powerful in general than several of its competitors. This is verified in a simulation study. Instead, we suggest applying the one-sided likelihood ratio test used by Perlman and Wu (2004) or the union-intersection $t_{\max}$ test used by Tamhane and Logan (2004).

Key words: Closed test procedure; Familywise error rate; Intersection-union test; Likelihood ratio test; Multiple comparisons; Multiple endpoints; Non-inferiority; Union-intersection test.

1 Introduction

Recently there has been much interest in the so-called "superiority-non-inferiority approach" to the multiple endpoints problem in which the goal is to demonstrate that the treatment is not inferior by more than a specified margin to the control on all endpoints, and superior on at least one endpoint. This is formulated as a hypothesis testing problem with the type I error controlled at a designated level α. The multivariate normal model (stated in Section 2) is commonly assumed. All procedures use separate α-level t-tests on individual endpoints for non-inferiority testing since it is an intersection-union (IU) testing problem (Berger, 1982; Laska and Meisner, 1989). The procedures differ in the global tests that they use for superiority. The Bloch, Lai and Tubert-Bitter (2001) procedure uses a one-sided version of Hotelling's T^2-test. The Perlman and Wu (2004) procedure uses the one-sided likelihood ratio (LR) test of Perlman (1969). The Tamhane and Logan (2004) procedure uses the union-intersection (UI) test (Roy 1953) based on the $t_{\max}$ statistic. The Röhmel et al. (2006) procedure uses a modified (to achieve monotonicity) Läuter's (1996) exact test (or the Holm 1979 test). We refer to these procedures as the BLT, PW, TL and RGBL procedures, respectively. Recently, Bloch et al.

[*] Corresponding author: e-mail: blogan@mcw.edu, Phone: +001 414 456 8849, Fax: +001 414456 6513

(2007) have generalized the BLT procedure to arbitrary functionals of the treatment and control distributions (instead of just the mean differences) and a nonparametric setting.

Since the overall goal of demonstrating both non-inferiority *and* superiority is itself an IU testing problem, a simple way to combine the tests for superiority and non-inferiority is to perform them separately, each at the α-level. However, this can be overly conservative since it assumes that the least favorable configuration (LFC) for the non-inferiority problem, which is that the treatment is inferior on one endpoint and is infinitely superior on all the others, holds simultaneously with the LFC for the superiority problem, which is that the treatment approaches borderline superiority on all endpoints. It is clear that the two LFCs are incompatible. In fact, the two tests are closely related and a more powerful test can be derived by utilizing the fact that demonstration of non-inferiority on all endpoints adds credence to superiority on at least one endpoint (more on this later). Bloch et al. (2001) and Tamhane and Logan (2004) used this idea to sharpen the critical constant of the superiority test. Since the exact evaluation of the critical constant (or equivalently the p-value) for the resulting test is difficult, these authors employed the bootstrap method. Perlman and Wu (2004) did not utilize this method to sharpen their superiority test, and applied the latter independently at the α-level.

Röhmel et al. (2006) dismissed these procedures arguing that they result in the dependence of the p-value for the superiority test on the inferiority margins which they were "unable to find a good reason" for and hence found them "difficult to understand." We will provide a counter-argument to this based in part on the discussion in the previous paragraph. They emphasized the importance of drawing superiority inferences on individual endpoints. Restricting to two endpoints, they proposed a three-step procedure in which the first step is the non-inferiority tests, the second step is a modified (for monotonicity) global test of Läuter (1996) or the Holm (1979) test for superiority, and the third step is the separate α-level t-tests for superiority on individual endpoints.

We agree with Röhmel et al. that a global test of superiority is often insufficient and inferences on individual endpoints are necessary. To derive FWER controlling procedures, we show how the superiority-non-inferiority problem with inferences on individual endpoints can be formulated as a closed testing problem and how to define the necessary families of hypotheses. We show that the stepwise procedure proposed by Röhmel et al. (2006) for the two endpoint problem is a particular application of this closed procedure and how it can be extended to more than two endpoints. Using simulations we demonstrate that if the $t_{\max}$ or the PW global test for superiority is sharpened as indicated above, the resulting three-step procedure is generally more powerful.

The paper is organized as follows. Section 2 introduces the notation, assumptions and the problem formulation. Section 3 reviews the procedures mentioned above. The closed testing formulation is given in Section 4. Two examples to illustrate the competing procedures are given in Section 5. Simulation results for FWER and power are given in Section 6. Conclusions and extensions are stated in Section 7.

2 Preliminaries and Notation

Consider a treatment group labelled 1 and a control group labelled 2 with n_1 and n_2 patients. Suppose that $m \geq 2$ endpoints are measured on each patient. Denote the random data vectors from group i by $\boldsymbol{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijm})'$ and their observed values by $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijm})'$ $(i = 1, 2, j = 1, 2, \ldots, n_i)$. We assume that the $\boldsymbol{X}_{ij}$ are independent and identically distributed (i.i.d.) random vectors from an m-variate normal distribution with mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im})'$ and a common covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{k\ell}\}$ with $\sigma_{kk} = \sigma_k^2 = \text{var}\,(X_{ijk})$ and $\sigma_{k\ell} = \text{cov}\,(X_{ijk}, X_{ij\ell})$ for $k \neq \ell$. Let $\delta_k = \mu_{1k} - \mu_{2k}$ and let $\delta = (\delta_1, \delta_2, \ldots, \delta_m)' = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ be the vector of mean differences between the treatment and the control. The treatment is regarded as superior to the control on the k-th endpoint if $\delta_k > \xi_k$ and non-inferior to the control if $\delta_k > -\varepsilon_k$, where the constants $\xi_k \geq 0$ and $\varepsilon_k > 0$ are prespecified. It is common to specify all $\xi_k = 0$, but we continue with the more general case in the remainder of the paper. The hypotheses for showing superiority are

$$H_{0k}^{(S)} : \delta_k \le \xi_k \quad \text{vs.} \quad H_{1k}^{(S)} : \delta_k > \xi_k \ \ (1 \le k \le m) \tag{1}$$

and those for showing non-inferiority are

$$H_{0k}^{(N)} : \delta_k \le -\varepsilon_k \quad \text{vs.} \quad H_{1k}^{(N)} : \delta_k > -\varepsilon_k \ \ (1 \le k \le m)\,. \tag{2}$$

Note that the superiority hypothesis is simply a shift of the non-inferiority hypothesis that requires a higher threshold to be cleared for its proof. To show superiority of the treatment on at least one endpoint and non-inferiority on all endpoints leads to the following UI and IU testing problems:

$$H_0^{(S)} = \bigcap_{k=1}^{m} H_{0k}^{(S)} \quad \text{vs.} \quad H_1^{(S)} = \bigcup_{k=1}^{m} H_{1k}^{(S)} \quad \text{and} \quad H_0^{(N)} = \bigcup_{k=1}^{m} H_{0k}^{(N)} \quad \text{vs.} \quad H_1^{(N)} = \bigcap_{k=1}^{m} H_{1k}^{(N)}\,. \tag{3}$$

The overall global superiority-non-inferiority hypothesis testing problem is

$$H_0 = H_0^{(S)} \cup H_0^{(N)} \quad \text{vs.} \quad H_1 = H_1^{(S)} \cap H_1^{(N)}\,. \tag{4}$$

Rejection of this global null hypothesis means that all endpoints have cleared the non-inferiority threshold while at least one endpoint has cleared the higher superiority threshold.

Let $\bar{x}_{i\cdot k}$ be the sample mean of the k-th endpoint for the i-th group ($i = 1, 2, k = 1, 2, \ldots, m$), $\bar{x}_{i\cdot} = (\bar{x}_{i\cdot 1}, \bar{x}_{i\cdot 2}, \ldots, \bar{x}_{i\cdot m})'$ and let $\boldsymbol{S}$ be the pooled sample covariance matrix based on $\nu = n_1 + n_2 - 2$ degrees of freedom (d.f.) with diagonal entries s_k^2 and off-diagonal entries $s_{k\ell}$. Then the t-statistics for testing superiority and non-inferiority of the treatment on the k-th endpoint are given by

$$t_k^{(S)} = \frac{\bar{x}_{1\cdot k} - \bar{x}_{2\cdot k} - \xi_k}{s_k\sqrt{1/n_1 + 1/n_2}} \quad \text{and} \quad t_k^{(N)} = \frac{\bar{x}_{1\cdot k} - \bar{x}_{2\cdot k} + \varepsilon_k}{s_k\sqrt{1/n_1 + 1/n_2}} \quad (1 \le k \le m)\,. \tag{5}$$

3 Review of Procedures

As mentioned before, to demonstrate non-inferiority on all endpoints, all foregoing procedures use the IU test which rejects $H_0^{(N)}$ at level α if

$$\min_{1 \le k \le m} t_k^{(N)} > t_{\nu,\alpha}\,, \tag{6}$$

where $t_{\nu,\alpha}$ is the upper α critical point of the t-distribution with ν d.f. Since the overall hypothesis testing problem (4) is also an IU testing problem, a simple α-level test of H_0 is to test both $H_0^{(S)}$ and $H_0^{(N)}$ at the α-level. A more powerful test is obtained by considering the type I error of the combined test of $H_0^{(S)}$ and $H_0^{(N)}$ as in (7) below. Different procedures use different global tests for $H_0^{(S)}$.

The BLT procedure uses Hotelling's T^2-statistic for testing $H_0^{(S)}$:

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right) (\bar{\boldsymbol{x}}_{1\cdot} - \bar{\boldsymbol{x}}_{2\cdot} - \boldsymbol{\xi})' \ \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}}_{1\cdot} - \bar{\boldsymbol{x}}_{2\cdot} - \boldsymbol{\xi})\,,$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)'$. The critical constant d for testing H_0 is determined from

$$P\left\{\left(\min_{1 \le k \le m} t_k^{(N)} > t_{\nu,\alpha}\right) \cap (T^2 > d)\right\} = \alpha \tag{7}$$

using bootstrap.

The PW procedure uses the LR statistic to test $H_0^{(S)}$. Define

$$z_k = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,(\bar{x}_{1\cdot k} - \bar{x}_{2\cdot k} - \xi_k)\,,$$

and

$$W = (n_1 + n_2 - 2)\, S\,.$$

Then the LR statistic is

$$U^2 = (z - \tilde{\boldsymbol{\delta}})'\; W^{-1}(z - \tilde{\boldsymbol{\delta}})\,,$$

where $\tilde{\boldsymbol{\delta}}$ is the projection of the vector z on to the nonpositive orthant $\mathcal{O}^- = \{\boldsymbol{\delta} \mid \delta_k \leq 0 \text{ for all } k\}$ with respect to the norm $\|x\|^2 = x'\, W^{-1}\, x$.

Perlman and Wu (2004) applied the U^2-test at the α-level independently of the IU test (6) for non-inferiority. In that case, the following equation from Perlman (1969) gives the critical constant d for U^2:

$$\frac{1}{2}\, P\left(\frac{\chi^2_{m-1}}{\chi^2_{n_1+n_2-m}} > d\right) + \frac{1}{2}\, P\left(\frac{\chi^2_m}{\chi^2_{n_1+n_2-m-1}} > d\right) = \alpha\,. \tag{8}$$

Tamhane and Logan (2004) sharpened this test by accounting for non-inferiority tests in the same way as for the BLT procedure with U^2 replacing T^2 in (7).

The TL procedure uses the UI test statistic $\max_{1\leq k\leq m} t_k^{(S)}$ for testing superiority. If this test is conducted independently of the IU test (6) for non-inferiority then the Bonferroni critical constant $t_{\nu,\alpha/m}$ can be used as a conservative upper bound on the exact α critical point, which depends on the unknown correlations among the endpoints. A sharper critical constant d can be obtained by accounting for non-inferiority tests in the same way as for the BLT procedure with $\max_{1\leq k\leq m} t_k^{(S)}$ replacing T^2 in (7).

Röhmel et al. (2006) criticized the incorporation of the non-inferiority test in determining the critical constant of the superiority test arguing that different choices of non-inferiority margins, ε_k, lead to different results for the superiority test which is unacceptable. For example, if the ε_k are made smaller then the p-value of the superiority test becomes smaller, making it easier to reject. This phenomenon is readily explained by the fact that if one can establish non-inferiority at more stringent thresholds, then that lends more credence to the superiority hypothesis. Refer back to the remark following (2) that the superiority hypothesis is simply a shift of the non-inferiority hypothesis. In particular, if all $\varepsilon_k \to 0$ then establishing non-inferiority is the same as establishing superiority, and a separate test for superiority is not needed. It should be noted that all the aforementioned tests have the option of not incorporating the non-inferiority test, but the resulting procedures are less powerful.

Röhmel et al. (2006) devoted much discussion to the lack of monotonicity of some one-sided LR tests which was not particularly relevant to the superiority-non-inferiority problem. Lack of monotonicity exists if as the sample mean differences, $\bar{x}_{1\cdot k} - \bar{x}_{2\cdot k}$ get larger, the test for superiority becomes less significant instead of becoming more significant. This curious phenomenon has been well-known and was first pointed out by Silvapulle (1997). Note that the UI test for superiority used by the TL procedure is monotone, and hence does not encounter this problem. Perlman and Wu (2003) have shown that nonmonotonicity of the LR tests is not due to the LR principle itself, but rather due to the misspecification of the global superiority hypothesis as a point null hypothesis, $\boldsymbol{\delta} = \mathbf{0}$, instead of $H_0^{(S)}$ from (3), which is a full complement of $H_1^{(S)}$.

Röhmel et al. (2006) modified the Läuter (1996) test to make it monotone as follows. They restricted to the $m = 2$ case, and considered the problem of testing $H_{0\boldsymbol{\eta}} : \delta_1 = \eta_1, \delta_2 = \eta_2$ where $\eta_1, \eta_2 \geq 0$. Let

$$\bar{x}_{\cdot\cdot k} = \frac{n_1\bar{x}_{1\cdot k} + n_2\bar{x}_{2\cdot k}}{n_1 + n_2}\,,$$

denote the overall sample mean for endpoint k and let

$$SS_k(\xi) = \sum_{j=1}^{n_1}\left(x_{1jk} - \bar{x}_{\cdot\cdot k} - \frac{n_2}{n_1+n_2}\,\eta_k\right)^2 + \sum_{j=1}^{n_2}\left(x_{2jk} - \bar{x}_{\cdot\cdot k} + \frac{n_1}{n_1+n_2}\eta_k\right)^2$$

note a modified total sum of squares for endpoint k. Define the weight vector $\boldsymbol{w}(\boldsymbol{\eta}) = (w_1(\boldsymbol{\eta}), w_2(\boldsymbol{\eta}))'$ where

$$w_k(\boldsymbol{\eta}) = \frac{1}{\sqrt{SS_k(\boldsymbol{\eta})}}\ .$$

Then the Läuter's SS statistic to test $H_{0\boldsymbol{\eta}}$ is given by

$$t_{SS}(\boldsymbol{\eta}) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left[\frac{\sum_{k=1}^{2} w_k(\boldsymbol{\eta})\,(\bar{x}_{1\cdot k} - \bar{x}_{2\cdot k} - \eta_k)}{\sqrt{\boldsymbol{w}(\boldsymbol{\eta})'\,\boldsymbol{S}\boldsymbol{w}(\boldsymbol{\eta})}} \right],$$

which follows a t_ν distribution when $\delta_1 = \eta_1, \delta_2 = \eta_2$. Therefore an α-level test of $H_0^{(S)}$ rejects if $t_{SS}(\xi_1, \xi_2) > t_{\nu,\alpha}$.

To ensure monotonicity of the test over the non-inferiority region, $-\varepsilon_k \le \delta_k \le \xi_k$ for $k = 1, 2$, this basic test is supplemented with additional requirements resulting in the following decision rule: Reject $H_0^{(S)}$ if $t_{SS}(\xi_1, xi_2) > t_{\nu,\alpha}$ and $S_{12} > 0$ or if $\min\,(t_{SS}(\xi_1, \xi_2), t_{SS}(-\varepsilon_1, \xi_2),\ t_{SS}(\xi_1, -\varepsilon_2)) > t_{\nu,\alpha}$.

Röhmel et al. (2006) proposed the following three-step procedure for $m = 2$ endpoints in which the first two steps are similar to those of the other procedures, and the third step tests superiority of the treatment on individual endpoints.

Step 1 Perform the non-inferiority tests (6). If $H_0^{(N)}$ is rejected, go to Step 2; otherwise accept H_0 and stop testing.

Step 2 Use the Läuter (1996) test as modified above to guarantee monotonicity to test $H_0^{(S)}$. If $H_0^{(S)}$ is rejected, go to Step 3; otherwise accept H_0 and stop testing. Alternatively, use the Holm (1979) test as follows: Let $t_{(1)}^{(S)} \le t_{(2)}^{(S)}$ be the ordered t-statistics for superiority. If $t_{(2)}^{(S)} > t_{\nu,\alpha/2}$ then reject $H_0^{(S)}$ and go to Step 3; otherwise accept H_0 and stop testing.

Step 3 Use separate one-sided α-level t-tests on each endpoint. If the Holm test is used in Step 2, then it only remains to check if $t_{(1)}^{(S)} > t_{\nu,\alpha}$.

The insistence on using the modified Läuter test makes the RGBL procedure unnecessarily unwieldy. It is difficult to see why the condition $S_{12} > 0$ is relevant to testing the superiority hypothesis. Also, there is no easy way to extend this modified test to more than two endpoints. Finally, the Läuter test has the undesirable property that its power is bounded strictly below 1 if one of the $\delta_k \to \infty$ as has been shown by Frick (1996) and Logan and Tamhane (2004). Röhmel et al. also found in their simulations that if "similar beneficial effects in both variables can be assumed, Läuter's SS procedure is superior to Holm's procedure" but not in other cases unless "the correlation between both variables is low or negative" (which is often not the case for related endpoints). In general, Läuter's test is not as powerful as some of the alternative procedures such as O'Brien's (1985) OLS or GLS test because it uses the total sum of squares instead of the within sum of squares in the definition of its test statistic. This is also confirmed in the Röhmel et al. simulation study where they found that the inferiority in terms of power of Läuter's SS compared to O'Brien's (1985) GLS is "negligible (usually not more than 1–2%)." In the power simulations reported in Section 6, the RGBL procedure generally performed worse than its competitors.

4 Closed Testing Formulation

For $m = 2$, the goal of showing that the treatment is superior on a particular endpoint and at least non-inferior on the other one can be met by testing the following hypotheses:

$$H_{01} : H_{01}^{(S)} \cup H_{02}^{(N)} \quad \text{and} \quad H_{02} : H_{01}^{(N)} \cup H_{02}^{(S)}\ . \tag{9}$$

If H_{01} is rejected then we conclude that the treatment is superior on endpoint 1 and non-inferior on endpoint 2. Similarly, if H_{02} is rejected then we conclude that the treatment is superior on endpoint 2 and non-inferior on endpoint 1. If both H_{01} and H_{02} are rejected then we conclude that the treatment is superior on both the endpoints. Denote this as family $F = \{H_{01}, H_{02}\}$. To control the FWER for this family we consider its closure: $\overline{F} = \{H_{01}, H_{02}, H_{01} \cap H_{02}\}$. Note that

$$\begin{aligned} H_{01} \cap H_{02} &= (H_{01}^{(S)} \cap H_{01}^{(N)}) \cup (H_{01}^{(S)} \cap H_{02}^{(S)}) \cup (H_{02}^{(N)} \cap H_{01}^{(N)}) \cup (H_{02}^{(N)} \cap H_{02}^{(S)}) \\ &= H_0^{(S)} \cup H_0^{(N)}\,, \end{aligned}$$

which is the global null hypothesis H_0 in (4). The closed test procedure is as follows. First test H_0 at the α level. If it is rejected, then proceed to test H_{01} and H_{02}, each at the α level. To test H_{01} and H_{02}, we only need to test their superiority components, i.e., $H_{01}^{(S)}$ and $H_{02}^{(S)}$, respectively, since the non-inferiority components, $H_{01}^{(N)}$ and $H_{02}^{(N)}$, are already tested and rejected in the test of H_0. This results in the three-step procedure stated by RGBL except that in the second step any other global test of $H_0^{(S)}$ can be used. We will use the $t_{\max}$ or the U^2 test. Furthermore, this global test can be sharpened as explained before. In general, if there are $m \geq 2$ endpoints then $m+1$ steps of testing are required by the closed procedure. For example, if there are three endpoints then, analogous to (9), we have three hypotheses:

$$H_{01} = H_{01}^{(S)} \cup H_{02}^{(N)} \cup H_{03}^{(N)}, \qquad H_{02} = H_{01}^{(N)} \cup H_{02}^{(S)} \cup H_{03}^{(N)}, \qquad H_{03} = H_{01}^{(N)} \cup H_{02}^{(N)} \cup H_{03}^{(S)},$$

and the closure includes three pairwise intersections and one overall intersection. A closure procedure involves hierarchical testing of these hypotheses. One can show that $H_0 = H_0^{(S)} \cup H_0^{(N)} = H_{01} \cap H_{02} \cap H_{03}$, which is rejected as a result of the first two steps of testing. Step 3 tests the pairwise intersections, $H_{01} \cap H_{02}$, $H_{01} \cap H_{03}$, and $H_{02} \cap H_{03}$, each at the α level. Since the non-inferiority hypotheses $H_{0k}^{(N)}$ $(1 \leq k \leq 3)$ have already been rejected and because the pairwise intersection can be written as $H_{0i} \cap H_{0j} = (H_{0i}^{(S)} \cap H_{0j}^{(S)}) \cup H_0^{(N)}$, this can be done by simply testing the superiority hypotheses alone, $H_{01}^{(S)} \cap H_{02}^{(S)}$, $H_{01}^{(S)} \cap H_{03}^{(S)}$ and $H_{02}^{(S)} \cap H_{03}^{(S)}$, each at the α level. Alternatively, one can recognize that $H_{0i} \cap H_{0j} \supseteq (H_{0i}^{(S)} \cap H_{0j}^{(S)}) \cup (H_{0i}^{(N)} \cup H_{0j}^{(N)})$, which can be tested using the global superiority-non-inferiority test of Tamhane and Logan (2004). Finally, depending on which of these pairwise intersections are rejected, the singletons $H_{01}^{(S)}$, $H_{02}^{(S)}$ and $H_{03}^{(S)}$ are tested individually at the α level at Step 4.

We now give the algorithm for the closed test procedure in the general $m > 2$ case. The hypotheses of interest are $\{H_{0k} = H_{0k}^{(S)} \cup H_0^{(N)}\}$. Let $I \subseteq \{1, \ldots, m\}$, so that $H_{0I} = \cap_{k \in I} H_{0k}$. The steps are as follows:

Step 1 Test for non-inferiority by rejecting $H_0^{(N)}$ at level α and continuing to step 2 if $\min_{1 \leq k \leq m} t_k^{(N)} > t_{\nu,\alpha}$.

Step 2 Test all intersection hypotheses $H_{0I}, I \subseteq \{1, \ldots, m\}$.

This can be done either by ignoring the non-inferiority hypotheses and simply applying a standard superiority test, or by incorporating the non-inferiority hypotheses and applying a global superiority-non-inferiority test. For example, the Bonferroni adjustment ignoring non-inferiority would reject H_{0I} if $\max_{k \in I} t_k^{(S)} > t_{\nu,\alpha/m'}$, where $m' = \mid I \mid$. To use a global superiority-non-inferiority test, recognize that

$$\begin{aligned} H_{0I} &= \left\{ \bigcap_{k \in I} H_{0k}^{(S)} \right\} \cup H_0^{(N)} \\ &\supseteq \left\{ \bigcap_{k \in I} H_{0k}^{(S)} \right\} \cup \left\{ \bigcup_{k \in I} H_{0k}^{(N)} \right\} \end{aligned}$$

Rejection of the latter hypothesis implies rejection of H_{0I}. This latter hypothesis can be directly tested by computing the p-value

$$p_I = P\left\{ \left(\min_{k \in I} T_k^{(N)} > t_{\nu,\alpha} \right) \cap \left(\max_{k \in I} T_k^{(S)} > \max_{k \in I} t_k^{(S)} \right) \right\}, \tag{10}$$

analogous to expression (7), and rejecting H_{0I} at level α if $p_I < \alpha$. This p-value can be estimated using bootstrap.

Step 3 Test the individual hypotheses H_{0k} by rejecting H_{0k} $if t_k^{(S)} > t_{\nu,\alpha}$ and all $H_{0I}, I \supseteq \{k\}$ are also rejected at level α. Equivalently, compute the adjusted p-value $\tilde{p}_k = \max_{I \supseteq \{k\}} p_I$ and reject H_{0k} if $\tilde{p}_k < \alpha$.

5 Examples

We demonstrate the described procedures on two examples. In the first example, we use the data from Röhmel et al. (2006) to contrast their procedure with ours in the two-endpoint setting. In the second example, we illustrate the procedure on the data from a clinical trial for asthma with four endpoints.

5.1 Example 1

Röhmel et al. (2006) give an example of a clinical trial comparing an active treatment A and a placebo P based on two approximately normally distributed primary endpoints. Patients were randomized in a 2:1 ratio to treatment versus placebo. Since low values of the endpoints were considered desirable, all test statistics were based on the control minus treatment differences, $P - A$. Röhmel et al. used non-inferiority margins of $\varepsilon_1 = 1$ and $\varepsilon_2 = 2$ for the two endpoints, superiority margins of $\xi_1 = \xi_2 = 0$, and an overall α level of 0.025. The summary data are given in Table 1 for each group.

The t-statistics for non-inferiority are computed to be $t_1^{(N)} = 3.945$ and $t_2^{(N)} = 2.990$. Since both statistics are above the critical value of 1.96, we reject $H_0^{(N)}$ and proceed to Step 2.

In Step 2, we apply a global test of $H_0^{(S)}$. The two test statistics for superiority are $t_1^{(S)} = 2.653$ and $t_2^{(S)} = 0.788$, so that $t_{\max} = 2.653$. The critical value for the $t_{\max}$ test is either 2.114 (obtained using bootstrap by accounting for rejection of the non-inferiority hypotheses) or 2.220 (the upper 0.0125 critical point of the standard normal distribution which does not account for rejection of the non-inferiority hypotheses). In either case the superiority hypothesis $H_0^{(S)}$ is rejected, and we proceed to Step 3. If the PW test is used, the test statistic is $U^2 = 0.0108$, while the critical value is either 0.007335 (obtained using bootstrap by accounting for rejection of the non-inferiority hypotheses) or 0.007935 (obtained using expression (8)). In either case, the global superiority hypothesis is rejected. Röhmel et al. (2006) computed the modified Läuter test statistic to be 2.0416 with a critical value of 1.964, which also leads to rejection of $H_0^{(S)}$.

In the final step, individual test statistics for superiority are compared to the critical value of 1.96. Only the hypothesis for endpoint 1 is rejected, leading us to conclude that there is a significant difference between treatment and placebo on endpoint 1, while the treatment group is non-inferior on endpoint 2.

Table 1 Summary data for Example 1.

		Sample means		Covariance matrix	
				$\boldsymbol{S}$	
Group	n	$\bar{x}_{i\cdot 1}$	$\bar{x}_{i\cdot 2}$		
A	442	13.269	22.796	78.60082	36.12524
				36.12524	111.65005
P	211	15.322	23.512	100.13374	53.62950
				53.62950	130.84153

Table 2 Summary data for Example 2.

Group		Endpoint			
		FEV$_1$	PEFR	SS	AMU
Treatment	Mean	14.0	16.5	0.86	0.49
Placebo	Mean	5.7	1.6	0.34	0.15
Pooled SD		11.5	22.3	0.96	0.66
$t^{(S)}$		3.00	2.75	2.25	2.13
p (one-sided)		0.002	0.004	0.014	0.018
$t^{(N)}$		3.83	3.58	3.08	2.96
Correlation	FEV$_1$	1			
Matrix	PEFR	0.25	1		
	SS	0.31	0.42	1	
	AMU	0.24	0.43	0.67	1

5.2 Example 2

Zhang et al. (1997) consider a randomized, multicenter, double-blind, parallel arm clinical trial to assess the efficacy and safety of a new drug in asthma patients. Four endpoints are considered: Forced expiratory volume in 1 second (FEV$_1$); Peak expiratory flow rate (PEFR) in litres per minute; Symptoms score (SS) on a scale from 0–6; and Additional medication use (AMU) in puffs per day. These four endpoints are meant to encompass different aspects of the disease which may respond to the new therapy. The outcomes were measured as percent change from baseline for FEV$_1$ and change from baseline for all others. There were 34 patients in the treatment group and 35 patients in the placebo group. The summary results are given in Table 2.

We are interested in determining which endpoints are superior while requiring that all endpoints be demonstrated to be non-inferior. We use non-inferiority thresholds of $\varepsilon_k = 0.2\sigma_k$, and superiority thresholds of $\xi_k = 0$. Using a one-sided type I error rate of 0.025, the non-inferiority critical value is $c = 2.00$, so that all four endpoints are considered non-inferior. Several strategies are possible for assessing superiority; we illustrate two here. First, one could separate the superiority and non-inferiority testing using the intersection-union principle. Then simple application of the Holm procedure yields adjusted p-values of $(0.008, 0.011, 0.028, 0.028)$, so that the treatment is concluded to be superior with respect to the first two endpoints (FEV$_1$and PEFR) at the 0.025 significance level.

Alternatively, one could apply a full closed test procedure incorporating the non-inferiority hypotheses into each intersection null hypothesis. Here the intersection null hypothesis for the set of hypoth-

Table 3 Adjusted p-values for example 2, using the $t_{\max}$ superiority test, adjusting for the non-inferiority testing.

Hypothesis H_{0I}	$\tilde{p}$	Hypothesis H_{0I}	$\tilde{p}$
$\{1,2,3,4\}$	0.0011	$\{2,3\}$	0.0044
$\{1,2,3\}$	0.0013	$\{2,4\}$	0.0044
$\{1,2,4\}$	0.0011	$\{3,4\}$	0.0175
$\{1,3,4\}$	0.0016	$\{1\}$	0.0017
$\{2,3,4\}$	0.0044	$\{2\}$	0.0044
$\{1,2\}$	0.0013	$\{3\}$	0.0175
$\{1,3$	0.0017	$\{4\}$	0.0184
$\{1,4\}$	0.0016		

Table 4 Familywise error rates.

ϱ	ε	δ	RGBL	TL*	TL	PW*	PW	Non-inferiority power
0	0.2	(0,0)	0.011	0.012	0.008	0.012	0.010	0.089
	0.33	(0,0)	0.016	0.024	0.016	0.021	0.017	0.421
	0.5	(0,0)	0.016	0.025	0.022	0.021	0.021	0.886
0.5	0.2	(0,0)	0.020	0.025	0.017	0.024	0.018	0.153
	0.33	(0,0)	0.022	0.025	0.027	0.025	0.025	0.499
	0.5	(0,0)	0.022	0.025	0.025	0.025	0.025	0.900

eses $I \subseteq \{1, 2, 3, 4\}$ is interpreted as at least one of the endpoints in I is inferior, or none of the endpoints are superior. Applying the full closed test procedure with the $t_{\max}$ superiority test and accounting for non-inferiority requirements yields the adjusted p-values in Table 3. The adjusted p-values for the individual endpoint hypotheses are $(0.002, 0.004, 0.018, 0.018)$, so that we can conclude that the treatment is superior to the placebo on all four endpoints.

6 Simulation Study

We performed a simulation study to investigate the FWERs as well as the powers of the three competing procedures: TL, PW and RGBL. Two versions of TL and PW were studied: the basic versions that did not employ the adjustment due to non-inferiority testing (denoted by TL and PW) and the sharpened versions that did (denoted by TL* and PW*). Sample sizes of $n_1 = n_2 = 100$ were used, with covariance matrices consisting of 1's on the diagonal and a common correlation $\varrho = 0.0$ or 0.5. A superiority threshold of 0 was used in all cases. Three different non-inferiority margins were considered, $\varepsilon = 0.2, 0.33, 0.5$. Values of the true mean difference vector δ include scenarios where both endpoints have a positive effect, as well as those where only one endpoint has a positive effect. A one-sided significance level of 0.025 was used for all simulations. A total of 10,000 replications were simulated for each estimate. The FWER estimates are given in Table 4, and the power estimates to reject at least one of the false hypotheses are given in Table 5. We also include in each table the proportion of times that the non-inferiority hypotheses were rejected.

In Table 4, notice that while all procedures control the FWER at the 0.025 level, the RGBL procedure tends to be more conservative than the others. This is possibly because the global superiority test in Step 2 does not account for the non-inferiority tests. All of the procedures are conservative when $\varepsilon = 0.2$; this is possibly because of the low likelihood of satisfying the non-inferiority requirement.

The power results in Table 5 indicate that the RGBL procedure performs worse than the others in a majority of cases, especially when there is a positive effect in only one of the endpoints. This is possibly because of the poor performance of the Läuter test in this configuration, which has been noted by Logan and Tamhane (2004) and Frick (1996). The RGBL procedure has higher power than TL and PW procedures only when the effects in both endpoints are equal. Overall, the PW procedure performs slightly better than the TL procedure because it has better power when both endpoints have a positive effect, although the differences in most cases are nonsignificant; this is the same finding as in Tamhane and Logan (2004). However, the U^2 statistic is hard to compute, its critical values are not commonly available and it is difficult to interpret to practitioners. The $t_{\max}$ statistic does not have these drawbacks.

Also, we point out that the impact of the bootstrap to account for the non-inferiority on the TL and PW procedures is somewhat variable. Generally, there are higher power gains in the $\varrho = 0$ case than in the $\varrho = 0.5$ case. However, in a number of cases, the power gains are not significant. This is because we are looking at the power to reject individual endpoint hypotheses, and the power advantage of the global test is not carrying through, especially when there is a large effect in only one of the endpoints, e.g., $(\delta_1, \delta_2) = (0.66, 0)$, in which case there are slight power losses. This indicates that

Table 5 Power to reject at least one individual endpoint hypothesis.

ϱ	ε	δ	RGBL	TL*	TL	PW*	PW	Non-inferiority power
0	0.2	(0.4,0)	0.239	0.243	0.214	0.243	0.228	0.292
		(0.66,0)	0.287	0.287	0.303	0.287	0.304	0.305
		(0.4,0.2)	0.698	0.702	0.631	0.702	0.684	0.796
		(0.33,0.33)	0.822	0.826	0.739	0.826	0.808	0.926
	0.33	(0.4,0)	0.431	0.510	0.467	0.496	0.468	0.638
		(0.66,0)	0.642	0.650	0.644	0.649	0.643	0.649
		(0.4,0.2)	0.789	0.798	0.746	0.808	0.799	0.960
		(0.33,0.33)	0.848	0.837	0.777	0.855	0.845	0.993
	0.5	(0.4,0)	0.491	0.687	0.681	0.676	0.670	0.941
		(0.66,0)	0.860	0.931	0.932	0.930	0.930	0.941
		(0.4,0.2)	0.798	0.780	0.769	0.818	0.816	0.998
		(0.33,0.33)	0.848	0.788	0.778	0.848	0.847	1.000
0.5	0.2	(0.4,0)	0.251	0.275	0.265	0.273	0.263	0.294
		(0.66,0)	0.295	0.296	0.301	0.296	0.301	0.301
		(0.4,0.2)	0.642	0.672	0.647	0.672	0.651	0.803
		(0.33,0.33)	0.739	0.734	0.701	0.744	0.724	0.932
	0.33	(0.4,0)	0.355	0.538	0.536	0.530	0.528	0.646
		(0.66,0)	0.612	0.643	0.648	0.643	0.648	0.650
		(0.4,0.2)	0.682	0.737	0.737	0.740	0.738	0.965
		(0.33,0.33)	0.750	0.707	0.703	0.731	0.724	0.993
	0.5	(0.4,0)	0.364	0.710	0.703	0.697	0.690	0.936
		(0.66,0)	0.745	0.937	0.940	0.936	0.939	0.939
		(0.4,0.2)	0.681	0.745	0.742	0.747	0.744	0.999
		(0.33,0.33)	0.745	0.713	0.715	0.733	0.736	1.000

when one is interested in the power to reject individual endpoint hypotheses, one can use the simpler test of superiority based on testing $H_0^{(S)}$ without accounting for the non-inferiority if it is expected that only one of the endpoints will show a large positive effect.

The column containing the power of the non-inferiority tests can be useful in interpreting the overall power results. The non-inferiority test power is dictated by the separation between the non-inferiority threshold and the smallest δ_k, i.e., for a given non-inferiority threshold the power is smaller when one of the endpoints has no effect ($\delta_k = 0$) and higher when all the endpoints have an effect. The power of the non-inferiority tests can also be used to determine the relative impact of the non-inferiority versus superiority tests. The difference between the overall power and the power of the non-inferiority test represents the probability that the study will pass the non-inferiority requirement but not pass the superiority test. For example, in the first row of Table 5, the power for the RGBL procedure is 0.239 and the non-inferiority power is 0.292. This indicates that 5.3% of the simulated trials were able to show non-inferiority but not superiority.

When the non-inferiority hypotheses are unlikely to be rejected ($\varepsilon = 0.2$ and at least one $\delta_k = 0$), little differences in the superiority tests can be seen. However, when $\varepsilon = 0.33$ or 0.5 so that the non-inferiority tests are less of a hurdle, the differences in the superiority tests are more pronounced. For example, when $\varepsilon = 0.5$ so that there is a greater than 90% chance of demonstrating non-inferiority, the RGBL procedure performs poorly when only one of the endpoints has a positive effect and performs well when both have an effect, matching the expected performance based on superiority testing alone.

7 Concluding Remarks

We have conclusively argued that accounting for non-inferiority testing in the global test for superiority is justified and enhances its power. However, simulation results indicate that this power advantage does not always carry through by the same amount to the superiority tests on individual hypotheses. Thus, the TL and PW global tests can be employed in their original simplified forms if only one of the endpoints is expected to have a large effect. Even in their simplified forms, use of these tests instead of the Läuter test results in higher power except when the endpoints have equal effects. Although the PW test is slightly more powerful, it is also more difficult to compute and apply in practice. The closed testing formulation shows that the three-step procedure controls the FWER for two endpoints; for more than two endpoints additional steps of testing are required to test all hierarchical intersections.

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

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

Bloch, D. A., Lai, T. L., Su, Z., and Tubert-Bitter, P. (2007). A combined superiority and non-inferiority approach to multiple endpoints in clinical trials. Statistics in Medicine **26**, 1193–1207.

Bloch, D. A., Lai, T. L., and Tubert-Bitter, P. (2001). One-sided tests in clinical trials with multiple endpoints. *Biometrics* **57**, 1039–1047.

Frick, H. (1996). On the power behaviour of Läuter's exact multivariate one-sided tests. *Biometrical Journal* **38**, 405–414.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Laska, E. M. and Meisner, M. J. (1989). Testing whether an identified treatment is best. *Biometrics* **45**, 1139–1151.

Läuter, J. (1996). Exact t and F tests for analyzing clinical trials with multiple endpoints. *Biometrics* **52**, 964–970.

Logan, B. R. and Tamhane, A. C. (2004). On O'Brien's OLS and GLS tests for multiple endpoints. *Recent Developments in Multiple Comparison Procedures*, IMS Lecture Notes-Monograph Series **47**, 76–88.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

Perlman, M. D. (1969). One-sided testing problems in multivariate analysis. *Annals of Mathematical Statistics* **40**, 549–567.

Perlman, M. D. and Wu, L. (2003). On the validity of the likelihood ratio and maximum likelihood methods. *Journal of Statistical Planning and Inference* **117**, 59–81.

Perlman, M. D. and Wu, L. (2004). A note on one-sided tests with multiple endpoints. *Biometrics* **60**, 276–279; Correction in (2007), *Biometrics* **63**, 622.

Röhmel, J., Gerlinger, C., Benda, N., and Läuter, J. (2006). On testing simultaneously non-inferiority in two multiple primary endpoints and superiority in at least one of them. *Biometrical Journal* **48**, 916–933.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics* **24**, 220–238.

Silvapulle, M. J. (1997). A curious example involving the likelihood ratio test against one-sided alternatives. *The American Statistician* **51**, 178–181.

Tamhane, A. C. and Logan, B. R. (2004). A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. *Biometrika* **91**, 715–727.

Zhang, J., Quan, H., Ng, J., and Stepanavage, M. E. (1997). Some Statistical Methods for Multiple Endpoints in Clinical Trials. *Controlled Clinical Trials* **18**, 204–221.