Biometrika (2004), **91,** 3, *pp.* 715–727

Printed in Great Britain

A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials

BY AJIT C. TAMHANE

Department of Statistics, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, U.S.A.

ajit@iems.northwestern.edu

AND BRENT R. LOGAN

Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin 53226, U.S.A.

blogan@mcw.edu

SUMMARY

This paper considers the problem of comparing a new treatment with a control based on multiple endpoints. The hypotheses are formulated with the goal of showing that the treatment is equivalent, i.e. not inferior, on all endpoints and superior on at least one endpoint compared to the control, where thresholds for equivalence and superiority are specified for each endpoint. Roy's (1953) union-intersection and Berger's (1982) intersection-union principles are employed to derive the basic test. It is shown that the critical constants required for the union-intersection test of superiority can be sharpened by a careful analysis of its type I error rate. The composite UI-IU test is illustrated by an example and compared in a simulation study to alternative tests proposed by Bloch et al. (2001) and Perlman & Wu (2004). The Bloch et al. test does not control the type I error rate because of its nonmonotone nature, and is hence not recommended. The UI-IU and the Perlman & Wu tests both control the type I error rate, but the latter test generally has a slightly higher power.

Some key words: Bootstrap; Hotelling's T^2 test; Intersection-union principle; Multivariate one-sided likelihood ratio test; Union-intersection principle

1. INTRODUCTION

Many clinical trials compare a treatment with a control on multiple endpoints. Often, the treatment is expected to have a positive effect on most, but not necessarily on all, endpoints. However, in order for the treatment to be preferred to the control, it may be sufficient to show that the treatment is not inferior, i.e. not much worse, on any of the endpoints and is strictly superior on at least one endpoint, or some specified number of endpoints. We formulate this as a combination of a union-intersection and an intersection-union testing problem, and propose a test based on the corresponding testing principles.

Bloch et al. (2001) considered a similar formulation to ours, but used Hotelling's T^2 statistic to test for superiority. Perlman & Wu (2004) suggested replacing the T^2 statistic

in the Bloch et al. test with a one-sided likelihood ratio statistic. We compare both these tests with our proposed test via simulation.

2. PRELIMINARIES AND NOTATION

Consider a treatment group, group 1, and a control group, group 2, with n_1 and n_2 patients. Suppose that $m \geqslant 2$ endpoints are measured on each patient. Denote the random data vectors from group i by $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijm})$, for $i = 1, 2$ and $j = 1, 2, \ldots, n_i$. We assume that the X_{ij} are independent and identically distributed random vectors from an m-variate normal distribution with mean vector $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im})$ and a common unknown covariance matrix $\Sigma = \{\sigma_{k\ell}\}$ with $\sigma_{kk} = \sigma_k^2 = \mathrm{var}(X_{ijk})$ and $\sigma_{k\ell} = \mathrm{cov}(X_{ijk}, X_{ij\ell})$ for $k \neq \ell$. Denote the correlation matrix by R with off-diagonal entries $\rho_{k\ell} = \mathrm{corr}(X_{ijk}, X_{ij\ell}) = \sigma_{k\ell}/\sigma_k\sigma_\ell$. Let $\theta_k = \mu_{1k} - \mu_{2k}$ and let $\theta = (\theta_1, \ldots, \theta_k) = \mu_1 - \mu_2$ be the vector of mean differences between the treatment and control.

The treatment is regarded as superior to the control on the kth endpoint if $\theta_k > \delta_k$ and equivalent, i.e. non-inferior, to the control if $\theta_k > -\varepsilon_k$, where the constants $\delta_k, \varepsilon_k \geqslant 0$ are specified. Note that often $\delta_k = 0$ is used because most experimental treatments are expected to show only small improvements over the control which are nonetheless regarded as beneficial. The hypotheses for showing the superiority and equivalence of the treatment on the kth endpoint are as follows:

$$H_{0k}^{(S)}: \theta_k \leqslant \delta_k \quad \text{versus} \quad H_{1k}^{(S)}: \theta_k > \delta_k, \quad H_{0k}^{(E)}: \theta_k \leqslant -\varepsilon_k \quad \text{versus} \quad H_{1k}^{(E)}: \theta_k > -\varepsilon_k.$$

Let

$$H_0^{(S)} = \bigcap_{k=1}^{m} H_{0k}^{(S)}, \quad H_1^{(S)} = \bigcup_{k=1}^{m} H_{1k}^{(S)}, \quad H_0^{(E)} = \bigcup_{k=1}^{m} H_{0k}^{(E)}, \quad H_1^{(E)} = \bigcap_{k=1}^{m} H_{1k}^{(E)}.$$

It is desired to test

$$H_0 = H_0^{(S)} \cup H_0^{(E)} \quad \text{versus} \quad H_1 = H_1^{(S)} \cap H_1^{(E)} \tag{2·1}$$

at a preassigned level α. For $m = 2$, the regions of the parameter space corresponding to H_0 and H_1 are shown in Fig. 1. Note that (2·1) is a combination of union-intersection

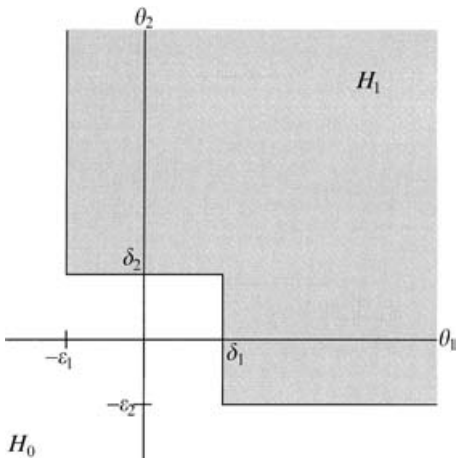

Fig. 1. Regions of parameter space corresponding to hypotheses H_0 and H_1.

and intersection-union testing problems. If $\delta_k = \varepsilon_k = 0$ for all k then $H_{0k}^{(S)} = H_{0k}^{(E)} = H_{0k}$, say, and $H_{1k}^{(S)} = H_{1k}^{(E)} = H_{1k}$, say. In that case, (2·1) reduces to $H_0 = \cup_{k=1}^m H_{0k}$ versus $H_1 = \cap_{k=1}^m H_{1k}$, which can be tested using the intersection-union approach of Berger (1982) resulting in the MIN test of Laska & Meisner (1989).

3. The simultaneous confidence intervals approach

Let $\bar{X}_{1\cdot k}$ and $\bar{X}_{2\cdot k}$ be the sample means for the kth endpoint for groups 1 and 2, respectively. Furthermore, let $S_1^2, S_2^2, \ldots, S_m^2$ be the pooled sample variances based on $v = n_1 + n_2 - 2$ degrees of freedom. We follow the usual convention of upper-case letters for random variables and the corresponding lower case letters for their observed values.

The pivotal random variable for θ_k is

$$T_k = \frac{(\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k}) - \theta_k}{S_k\sqrt{(1/n_1 + 1/n_2)}} \quad (1 \leqslant k \leqslant m). \tag{3·1}$$

Each T_k is marginally distributed as t_v. The joint distribution of $(T_1, T_2, \ldots, T_m)$ is the multivariate generalisation of a bivariate t-distribution considered by Siddiqui (1967).

Since the joint distribution of $(T_1, T_2, \ldots, T_m)$ depends on the unknown correlation matrix R, the exact critical constants needed to compute simultaneous $100(1-\alpha)\%$ confidence intervals for the θ_k are not available. Based on the Bonferroni method, conservative lower one-sided confidence intervals are given by

$$\theta_k \geqslant L_k = \bar{x}_{1\cdot k} - \bar{x}_{2\cdot k} - t_{v,\alpha/m} s_k \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad (1 \leqslant k \leqslant m), \tag{3·2}$$

where $t_{v,\alpha/m}$ is the upper α/m critical point of the t_v distribution. We reject H_0 if all $L_k > -\varepsilon_k$ and at least one $L_k > \delta_k$. Defining the t-statistics for testing the superiority and equivalence of the treatment on the kth endpoint by

$$t_k^{(S)} = \frac{\bar{x}_{1\cdot k} - \bar{x}_{2\cdot k} - \delta_k}{s_k\sqrt{(1/n_1 + 1/n_2)}}, \quad t_k^{(E)} = \frac{\bar{x}_{1\cdot k} - \bar{x}_{2\cdot k} + \varepsilon_k}{s_k\sqrt{(1/n_1 + 1/n_2)}} \quad (1 \leqslant k \leqslant m), \tag{3·3}$$

we see that the above test reduces to

$$\min_{1\leqslant k\leqslant m} t_k^{(E)} > t_{v,\alpha/m}, \quad \max_{1\leqslant k\leqslant m} t_k^{(S)} > t_{v,\alpha/m}. \tag{3·4}$$

In fact, since all inferences follow from a single set of simultaneous confidence bounds (3·2), all endpoints can be classified with $1-\alpha$ confidence as follows: on the kth endpoint the treatment is not equivalent, i.e. inferior, if $L_k \leqslant -\varepsilon_k$, is equivalent but not superior if $-\varepsilon_k < L_k \leqslant \delta_k$, and is superior if $L_k > \delta_k$. In the next section we show how the test (3·4) can be sharpened by applying the union-intersection and intersection-union principles of test construction.

4. A test based on union-intersection and intersection-union principles

4·1. *The union-intersection and intersection-union*, UI-IU, *test*

An α-level test of (2·1) derived by applying the intersection-union principle is as follows: test $H_0^{(S)} = \cap_{k=1}^m H_{0k}^{(S)}$ and $H_0^{(E)} = \cup_{k=1}^m H_{0k}^{(E)}$ separately at level α, and reject H_0 if both are rejected. The union-intersection test (Roy, 1953) of $H_0^{(S)}$ rejects at level α using the

Bonferroni approximation if $\max_{1\leqslant k\leqslant m} t_k^{(S)} > t_{\nu,\alpha/m}$. The intersection-union test (Berger, 1982) of $H_0^{(E)}$ rejects at level α if $\min_{1\leqslant k\leqslant m} t_k^{(E)} > t_{\nu,\alpha}$. We will refer to this procedure as the UI-IU test. Note the smaller critical constant, $t_{\nu,\alpha}$, for the intesection-union test of $H_0^{(E)}$ compared to that used by the simultaneous confidence interval test (3·4).

This UI-IU test is conservative because it requires that the type I error probability be separately controlled for $H_0^{(E)}$ and $H_0^{(S)}$, which assumes the least favourable configuration that one of the two hypotheses is true and the other is infinitely false. It is possible to have $H_0^{(E)}$ true and $H_0^{(S)}$ infinitely false; for example, we can have $\theta_k = -\varepsilon_k$ and $\theta_\ell \to \infty$ for $\ell \neq k$. In fact, this is the least favourable configuration for the intersection-union test. However, we cannot have $H_0^{(S)}$ true and $H_0^{(E)}$ infinitely false because if $\theta_k \leqslant \delta_k$ for all k then it cannot be simultaneously true that $\theta_\ell \to \infty$ for some ℓ. This suggests that, although the critical constant $t_{\nu,\alpha}$ for the intersection-union test of $H_0^{(E)}$ cannot be reduced, it may be possible to reduce the critical constant $t_{\nu,\alpha/m}$ for the union-intersection test of $H_0^{(S)}$. From now on, we will use a general notation, c and d, for the critical constants of $H_0^{(E)}$ and $H_0^{(S)}$, respectively. In the next section we investigate how to find the smallest possible values of c and d.

4·2. *Sharpened critical constants for the* UI-IU *test*

By using the relationship

$$t_k^{(S)} = t_k^{(E)} - \frac{\delta_k + \varepsilon_k}{s_k\sqrt{(1/n_1 + 1/n_2)}} \quad (1 \leqslant k \leqslant m),$$

we can write the UI-IU test as

$$\min_{1\leqslant k\leqslant m} \left\{ t_k^{(S)} + \frac{\delta_k + \varepsilon_k}{s_k\sqrt{(1/n_1 + 1/n_2)}} \right\} > c, \quad \max_{1\leqslant k\leqslant m} t_k^{(S)} > d. \tag{4·1}$$

For $m = 2$, $\sigma_k = 1$, $\delta_k = 0$, $\varepsilon_k = 0{\cdot}5$ and $n_k = 50$, this rejection region is shown in Fig. 2; the rejection regions for the Bloch et al. (2001) and the Perlman & Wu (2004) tests, which are discussed in § 6, are also shown in Fig. 2 for an easy comparison. Note that, if $H_0^{(E)}$ is

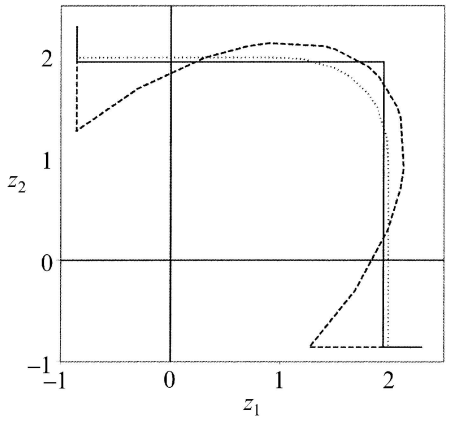

Fig. 2. Rejection regions of the UI-IU test, solid line, Bloch et al. (2001) test, dashed line, and Perlman & Wu (2004) test, dotted lines, for $m = 2$.

rejected and if

$$d < \max_{1 \leqslant k \leqslant m} \left\{ c - \frac{\delta_k + \varepsilon_k}{s_k \sqrt{(1/n_1 + 1/n_2)}} \right\},$$

then $H_0^{(S)}$ is automatically rejected; thus superiority need not be tested separately.

We obtain an expression for the type I error probability of the UI-IU test (4·1) in Lemma 1, and then find its least favourable configuration in Lemma 2.

LEMMA 1. *Define*

$$Z_k = \frac{\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k} - \theta_k}{\sigma_k \sqrt{(1/n_1 + 1/n_2)}} \sim N(0, 1), \quad U_k = \frac{S_k}{\sigma_k} \sim \sqrt{\frac{\chi_\nu^2}{\nu}} \quad (1 \leqslant k \leqslant m),$$

so that $Z = (Z_1, \ldots, Z_m)$ *has an* m*-variate standard normal distribution with correlation matrix* R *independently of* $U = (U_1, \ldots, U_m)$. *Denote the density functions of* Z *and* U *by* $\phi_m(z|R)$ *and* $h_{m,\nu}(u|R)$, *respectively. Let*

$$\delta_k^* = \frac{\delta_k}{\sigma_k \sqrt{(1/n_1 + 1/n_2)}}, \quad \varepsilon_k^* = \frac{\varepsilon_k}{\sigma_k \sqrt{(1/n_1 + 1/n_2)}}, \quad \theta_k^* = \frac{\theta_k}{\sigma_k \sqrt{(1/n_1 + 1/n_2)}}. \tag{4·2}$$

Furthermore, let

$$a_k = \theta_k^* + \varepsilon_k^*, \quad b_k = \theta_k^* - \delta_k^* \quad (1 \leqslant k \leqslant m). \tag{4·3}$$

Then the probability that the UI-IU *test* (4·1) *rejects* H_0 *of* (2·1) *can be written as*

$$Q(\theta) = \int_0^\infty \cdots \int_0^\infty \Psi(\theta|u) h_{m,\nu}(u|R) du, \tag{4·4}$$

where

$$\Psi(\theta|u) = \int_{cu_1 - a_1}^\infty \cdots \int_{cu_m - a_m}^\infty \phi_m(z|R) dz - \int_{cu_1 - a_1}^{du_1 - b_1} \cdots \int_{cu_m - a_m}^{du_m - b_m} \phi_m(z|R) dz. \tag{4·5}$$

In the above, if $d < c - (a_k + b_k)/u_k$ *for any* $u_k \geqslant 0$ $(1 \leqslant k \leqslant m)$ *then the second integral is taken to be zero.*

Proof. Write the desired probability as the difference between

$$\text{pr}\{T_k^{(S)} > c - (\delta_k^* + \varepsilon_k^*)/U_k \text{ for all } k\}, \quad \text{pr}\{c - (\delta_k^* + \varepsilon_k^*)/U_k \leqslant T_k^{(S)} \leqslant d \text{ for all } k\}.$$

Then, by conditioning on the U_k, we can write the two probabilities as multivariate normal integrals shown in (4·5). The final expression (4·4) is obtained by removing the conditioning on U. □

To simplify the notation, from now on, we will assume that Σ and R are known or equivalently that $\nu \to \infty$. Therefore $U_k \to 1$ for all k and $Q(\theta) \to \Psi(\theta|u = 1_m)$ almost surely, where 1_m is an m-vector of ones.

LEMMA 2. *The type* I *error probability of the* UI-IU *test is maximised at one or more of the following configurations*:

$$\text{LFC}_0 = \{\theta_1 = \delta_1, \ldots, \theta_m = \delta_m\} \quad \text{or} \quad \text{LFC}_k = \{\theta_k = -\varepsilon_k, \theta_\ell \to \infty, \ell \neq k\} \quad (1 \leqslant k \leqslant m). \tag{4·6}$$

Proof. To find the maximum of $Q(\theta)$ with respect to θ_k over $H_0 = H_0^{(S)} \cup \{\cup_{k=1}^m H_{0k}^{(E)}\}$, take the derivatives of $Q(\theta)$ with respect to $\theta_k^* \propto \theta_k$. In particular, for $k=1$, using $\partial a_1/\partial\theta_1^* = \partial b_1/\partial\theta_1^* = 1$, we obtain

$$\begin{aligned}
\frac{\partial Q(\theta)}{\partial \theta_1^*} &= \int_{c-a_2}^{\infty} \cdots \int_{c-a_m}^{\infty} \phi_m(c-a_1, z_2, \ldots, z_m|R)dz_2 \ldots dz_m \\
&\quad - \left\{ - \int_{c-a_2}^{d-b_2} \cdots \int_{c-a_m}^{d-b_m} \phi_m(d-b_1, z_2, \ldots, z_m|R)dz_2 \ldots dz_m \right. \\
&\quad \left. + \int_{c-a_2}^{d-b_2} \cdots \int_{c-a_m}^{d-b_m} \phi_m(c-a_1, z_2, \ldots, z_m|R)dz_2 \ldots dz_m \right\} \\
&= \left\{ \int_{c-a_2}^{\infty} \cdots \int_{c-a_m}^{\infty} \phi_m(c-a_1, z_2, \ldots, z_m|R)dz_2 \ldots dz_m \right. \\
&\quad \left. - \int_{c-a_2}^{d-b_2} \cdots \int_{c-a_m}^{d-b_m} \phi_m(c-a_1, z_2, \ldots, z_m|R)dz_2 \ldots dz_m \right\} \\
&\quad + \int_{c-a_2}^{d-b_2} \cdots \int_{c-a_m}^{d-b_m} \phi_m(d-b_1, z_2, \ldots, z_m|R)dz_2 \ldots dz_m \\
&> 0.
\end{aligned}$$

Thus, $Q(\theta)$ is increasing in each θ_k and hence is maximised over $H_0^{(S)}$ at LFC_0 and over $H_{0k}^{(E)}$ at LFC_k for $1 \leqslant k \leqslant m$. The global maximum is found by evaluating $Q(\theta)$ at each of these $m+1$ least favourable configurations, and taking the overall maximum. □

Let

$$e_k = \delta_k^* + \varepsilon_k^* = \frac{\delta_k + \varepsilon_k}{\sigma_k \sqrt{(1/n_1 + 1/n_2)}} \quad (1 \leqslant k \leqslant m). \tag{4·7}$$

Then for LFC_0 we obtain $a_k = e_k$ and $b_k = 0$, so that

$$\begin{aligned}
Q_{\max,0} &= \sup_{\theta \in H_0^{(S)}} Q(\theta) \\
&= \int_{c-e_1}^{\infty} \cdots \int_{c-e_m}^{\infty} \phi_m(z|R)dz - \int_{c-e_1}^{d} \cdots \int_{c-e_m}^{d} \phi_m(z|R)dz \\
&= \text{pr}\left\{ \min_{1 \leqslant k \leqslant m} (Z_k + e_k) > c \text{ and } \max_{1 \leqslant k \leqslant m} Z_k > d \right\}.
\end{aligned} \tag{4·8}$$

For LFC_k $(1 \leqslant k \leqslant m)$ we obtain $a_k = 0$, $b_k = -e_k$ and a_ℓ, $b_\ell \to \infty$ for $\ell \neq k$, so that

$$\begin{aligned}
Q_{\max,k} &= \sup_{\theta \in H_{0k}^{(E)}} Q(\theta) \\
&= \int_{c}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi_m(z|R)dz - \int_{c}^{d-e_k} \int_{-\infty}^{-\infty} \cdots \int_{-\infty}^{-\infty} \phi_m(z|R)dz \\
&= 1 - \Phi(c).
\end{aligned} \tag{4·9}$$

Equating this to α, we obtain $c = z_\alpha$. For $v < \infty$, the above equation can be shown to be $Q_{\max,k} = 1 - F_v(c)$, where $F_v(.)$ is the distribution function of the t_v distribution. Setting $Q_{\max,k} = \alpha$, we see that the smallest value of c is $t_{v,\alpha}$, as conjectured earlier. In the following lemma we obtain the limiting values of d.

LEMMA 3. *If $e_k = \delta_k^* + \varepsilon_k^* \to 0$ for all k then $d = c = z_\alpha$ controls the type* I *error rate conservatively at level α. If $e_k \to \infty$ for all k then $d = z_{m,R,\alpha}$, the upper α critical point of $\max_{1 \leqslant k \leqslant m} Z_k$, controls the type* I *error rate exactly at level α.*

Proof. If $e_k \to 0$ for all k then by substituting $d = c$ in (4·8) we obtain

$$Q_{\max,0} = \int_c^\infty \ldots \int_c^\infty \phi_m(z|R)dz \leqslant 1 - \Phi(c) = \alpha.$$

Therefore, the type I error probability is controlled below α. If $e_k \to \infty$ for all k then we obtain

$$Q_{\max,0} = 1 - \int_{-\infty}^d \ldots \int_{-\infty}^d \phi_m(z|R)dz = 1 - \text{pr}\{\max(Z_1, \ldots, Z_m) \leqslant d\} = \alpha$$

by substituting $d = z_{m,R,\alpha}$. □

From the first result in Lemma 3 we see that it is possible to have $d \leqslant c$ for small e_k. Note also that $z_{\alpha/m}$ is the Bonferroni upper bound on $z_{m,R,\alpha}$.

4·3. *Bootstrap implementation of the* UI-IU *test*

The previous results show that, to apply the UI-IU test at level α, we must set $c = t_{v,\alpha}$ and then solve for d by setting the finite degrees of freedom version of (4·8) equal to α; that is

$$\text{pr}\left\{\min_{1 \leqslant k \leqslant m}(T_k + e_k) > c \text{ and } \max_{1 \leqslant k \leqslant m} T_k > d\right\} = \alpha, \tag{4·10}$$

where $T_1, T_2, \ldots, T_m$ have the generalised multivariate t distribution referred to earlier. Evaluation of this probability requires the knowledge of Σ. To obviate this difficulty, we propose the following bootstrap algorithm. This algorithm does not directly compute d, but determines if $\max_{1 \leqslant k \leqslant m} t_k^{(S)} > d$ or not, and hence whether $H_0^{(S)}$ can be rejected or not. In conjunction with the test of $H_0^{(E)}$, this enables us to conduct the UI-IU test in (4·1).

The algorithm is as follows.

Step 0. If $\min_{1 \leqslant k \leqslant m} t_k^{(E)} \leqslant c = t_{v,\alpha}$ then accept $H_0^{(E)}$ and hence H_0 and stop. Otherwise go to Step 1.

Step 1. Centre the observed data vectors x_{ij} by subtracting the sample mean vector $\bar{x}_{i\cdot}$. Denote the centred data vectors by x_{ij}^* $(i = 1, 2; 1 \leqslant j \leqslant n_i)$.

Step 2. Draw B bootstrap samples with replacement from the pooled sample of centred data vectors. Denote the bth bootstrap sample by x_{ijb}^* $(i = 1, 2; 1 \leqslant j \leqslant n_i; 1 \leqslant b \leqslant B)$.

Step 3. Calculate the statistics $t_{kb}^{*(E)}$ and $t_{kb}^{*(S)}$ for the bootstrap data using (3·3) for $k = 1, 2, \ldots, m$ and $b = 1, 2, \ldots, B$.

Step 4. For the bth bootstrap sample, if $\min_{1\leqslant k\leqslant m} t_{kb}^{*(E)} > c = t_{\nu,\alpha}$ and $\max_{1\leqslant k\leqslant m} t_{kb}^{*(S)} > \max_{1\leqslant k\leqslant m} t_k^{(S)}$ then reject H_0; otherwise accept H_0. Repeat this for all B bootstrap samples.

Step 5. Let A be the number of bootstrap samples in which H_0 is rejected and let $\hat{p} = A/B$ be the corresponding proportion. If $\hat{p} < \alpha$ then reject $H_0^{(S)}$ and hence H_0 at level α.

The following points should be noted regarding the bootstrap implementation of the UI-IU test:

(i) the algorithm essentially sets $d = \max_{1\leqslant k\leqslant m} t_k^{(S)}$ and estimates the p-value for rejecting $H_0^{(S)}$ conditional on having rejected $H_0^{(E)}$ at level α;
(ii) it does not explicitly make use of normality, other than using $c = t_{\nu,\alpha}$, and in this respect the algorithm is similar to the Bloch et al. (2001) algorithm;
(iii) it can be readily modified to allow for heteroscedastic covariance matrices, in which case the t-statistics in (3·3) must be also modified to use separate variance estimates for the treatment and control groups.

5. EXAMPLE

We use an example from Tang et al. (1993) about the efficacy of an inhaled drug for asthma compared to placebo. Seventeen patients were randomised in a double-blind cross-over trial. There were four standard respiratory function measures, i.e. endpoints, namely forced expiratory volume, FEV, forced vital capacity, FVC, peak expiratory flow rate, PEFR, and penetration index, PI. There was no period or crossover effect, so the comparisons

Table 1. *Data for the asthma example*

	FEV	FVC	PEFR	PI
Mean difference	7·56	4·81	2·29	0·081
Std dev. of difference	18·53	10·84	8·51	0·17
t-statistic	1·682	1·830	1·110	1·965
p-value	0·0560	0·0430	0·1417	0·0335

FEV, forced expiratory volume; FVC, forced vital capacity; PEFR, peak expiratory flow rate; PI, penetration index.

for the individual endpoints could be performed using paired t-statistics. The summary statistics are shown in Table 1 and the estimated correlation matrix is

$$\begin{bmatrix} 1{\cdot}000 & 0{\cdot}095 & 0{\cdot}219 & -0{\cdot}162 \\ & 1{\cdot}000 & 0{\cdot}518 & -0{\cdot}059 \\ & & 1{\cdot}000 & 0{\cdot}513 \\ & & & 1{\cdot}000 \end{bmatrix}.$$

For these data, the ordinary least squares and generalised least squares statistics of O'Brien (1984) are highly significant indicating a global improvement. However, none of the individual endpoints can be shown to have significant improvement at $\alpha = 0{\cdot}05$ using the Bonferroni method or one of its sharpened versions.

Suppose $\delta_k = 0$ and $\varepsilon_k = \lambda\sigma_k$ with $\lambda = 0{\cdot}50$ for $1 \leqslant k \leqslant 4$. Then we have

$$e_k = \frac{\delta_k + \varepsilon_k}{\sigma_k\sqrt{(1/n)}} \doteq 0{\cdot}50\sqrt{17} = 2{\cdot}062;$$

here $\sqrt{(1/n_1 + 1/n_2)}$ is replaced by $\sqrt{(1/n)}$ since this is essentially a paired-sample study with n patients. The t-statistics given Table 1 are the superiority t-statistics, $t_k^{(S)}$, since we assumed $\delta_k = 0$ for all k.

For $\alpha = 0{\cdot}05$, we have $c = t_{16,0{\cdot}05} = 1{\cdot}746$. Applying (4·1) and taking s_k as approximately equal to σ_k, we find that

$$\min_{1 \leqslant k \leqslant 4} \{t_k^{(S)} + 2{\cdot}062\} = \min \{3{\cdot}744, 3{\cdot}892, 3{\cdot}172, 4{\cdot}027\} > c = 1{\cdot}746,$$

so that $H_0^{(E)}$ is rejected.

We next apply the bootstrap algorithm of § 3 to test $H_0^{(S)}$, but, since only summary statistics are available for these data, we cannot directly apply the bootstrap algorithm since it is given for raw data. Instead, we applied a parametric version in which we drew samples from a four-variate normal distribution with a null mean vector and the estimated correlation matrix along with the sample standard deviations given in Table 1. A total of 100 000 bootstrap samples were drawn and the estimated proportion of rejections of $H_0^{(S)}$ was observed to be 0·04488. Therefore H_0 is rejected at the 5% level and the inhaled drug is shown to be preferred to the placebo.

6. Two previously proposed tests

6·1. *The Bloch, Lai & Tubert-Bitter* (2001) *test*

Bloch et al. (2001) considered the superiority-equivalence formulation, for the special case of all $\delta_k = 0$, in a general nonparametric setting using a bootstrap approach similar to ours. To test $H_0^{(S)}$ they employed a one-sided version of Hotelling's T^2 statistic, modified to allow for unequal covariance matrices, which equals T^2 if $H_0^{(E)}$ is rejected and is zero otherwise. To test $H_0^{(E)}$ in the normal setting, they used the same intersection-union test that we used, with rejection region $\min_{1 \leqslant k \leqslant m} t_k^{(E)} > t_{\nu,\alpha}$. If we denote the indicator function of an event A by $I(A)$, their test rejects H_0 if

$$T^2 \times I\left(\min_{1 \leqslant k \leqslant m} t_k^{(E)} > t_{\nu,\alpha} \right) > d, \tag{6·1}$$

where $d > 0$ is a critical constant that is determined via bootstrap to make the type I error rate equal to the specified level α under the null configuration. For the normal, homoscedastic setting of the present paper, we have

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

where $\bar{x}_{i\cdot} = (\bar{x}_{i\cdot 1}, \bar{x}_{i\cdot 2}, \ldots, \bar{x}_{i\cdot m})'$, for $i = 1, 2$, and

$$\hat{\Sigma} = \frac{\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_{1\cdot})(x_{1j} - \bar{x}_{1\cdot})' + \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_{2\cdot})(x_{2j} - \bar{x}_{2\cdot})'}{n_1 + n_2 - 2}$$

is the pooled sample covariance matrix.

It should be noted that the Bloch et al. test is non-monotone (Cohen & Sackrowitz, 1998) for certain choices of ε_k; see Fig. 2. Thus the type I error of the test in the interior of $H_0^{(S)} : \cap_{k=1}^m \theta_k \leqslant 0$ may exceed that at the null configuration, $\theta_k = 0$ for all k, which is the configuration used to determine the critical constant d, so that the test may become anti-conservative. Perlman & Wu (2004) noted this drawback of the Bloch et al. test in

their Footnote 2 and in their §§ 3 and 4, where they stated that the Bloch et al. test does not have the desired monotonicity property and may reject H_0 for sample points actually inside H_0. We will demonstrate this phenomenon in our simulations.

6·2. *The Perlman & Wu* (2004) *test*

Perlman & Wu (2004) replaced the T^2 statistic in the Bloch et al. (2001) test by the multivariate one-sided likelihood ratio statistic derived by Perlman (1969). Furthermore, they tested $H_0^{(S)}$ and $H_0^{(E)}$ separately at level α, the latter using the same intersection-union test as ours. Their test statistic for testing $H_0^{(S)}$ is the difference between the observed vector $\bar{x}_{1\cdot} - \bar{x}_{2\cdot}$ and its projection on to the nonpositive orthant $\mathcal{O}^- = \{\theta | \theta_k \leqslant 0 \text{ for all } k\}$ with respect to the norm $\|x\|^2 = x'\hat{\Sigma}^{-1}x$. Denote this statistic by U^2. Then the Perlman & Wu test rejects if

$$U^2 > d, \quad \min_{1 \leqslant k \leqslant m} t_k^{(E)} > t_{\nu,\alpha}, \tag{6·2}$$

where d is the upper α critical constant for testing $H_0^{(S)}$ and is the solution to the equation (Perlman, 1969)

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

In their simulation study, Perlman & Wu (2004) used a modified form of the Bloch et al. (2001) test analogous to their own as follows:

$$T^2 > T^2_{m,n_1+n_2-m-1,\alpha}, \quad \min_{1 \leqslant k \leqslant m} t_k^{(E)} > t_{\nu,\alpha},$$

where $T^2_{m,n_1+n_2-m-1,\alpha}$ is the upper α critical constant of T^2. Note that this form of the Bloch et al. test as well as the Perlman & Wu test given in (6·2), which test $H_0^{(S)}$ and $H_0^{(E)}$ separately each at level α, are conservative. In our simulations, we used the original form (6·1) of the Bloch et al. test and we modified the Perlman & Wu test (6·2) to conform to the same form by evaluating its critical constant d via bootstrap so that the overall type I error probability is controlled at α.

7. SIMULATION STUDY

The simulation study was aimed at investigating the control of the type I error rate over the entire null space by the UI-IU, Perlman & Wu (2004) and Bloch et al. (2001) tests, as well as comparing their powers. Throughout, we used $\sigma_k^2 = 1$, $\delta_k = 0$, $\varepsilon_k = \lambda\sigma_k = \lambda$, for all k, and $\alpha = 0{\cdot}05$. Furthermore, we restricted attention to the equicorrelated case with $\rho_{k\ell} = \rho$. A straightforward modification of the bootstrap algorithm given in § 3 was used to determine the critical value d for the Bloch et al. test and the Perlman & Wu test, using T^2 for the Bloch et al. test and U^2 for the Perlman & Wu test.

First, we investigated the type I error rates for two types of null configuration, one where $\theta_k = 0$ for all k, that is all endpoints have a zero treatment effect, and the other where $m/2$ of the endpoints have a zero treatment effect and the remaining $m/2$ have a treatment effect equal to $-\lambda/2$. We considered $m = 2, 4, 8$, $\lambda = 0{\cdot}2, 0{\cdot}5, 0{\cdot}8$, $n_1 = n_2 = 50$ and an equicorrelated matrix with $\rho = 0{\cdot}0, 0{\cdot}5$. The estimated type I error rates are given in Table 2. Note that the Bloch et al. test has excessive type I error rates, as high as 0·244, when $\varepsilon_k = \lambda = 0{\cdot}8$ and θ is an m-vector of $m/2$ zeros followed by $m/2$ elements equal to

Table 2. *Simulation estimates of type* I *error rates of* UI-IU, *Perlman & Wu and Bloch et al. tests for* $\alpha = 0{\cdot}05$

			$\theta = (0, 0, \ldots, 0)$			$\theta = (0, \ldots, 0, -\lambda/2, \ldots, -\lambda/2)$		
m	λ	ρ	UI-IU	PW	BLT	UI-IU	PW	BLT
2	0·2	0·0	0·052	0·052	0·052	0·024	0·024	0·024
		0·5	0·050	0·050	0·049	0·026	0·027	0·027
	0·5	0·0	0·052	0·052	0·052	0·014	0·011	0·013
		0·5	0·050	0·049	0·050	0·022	0·020	0·035
	0·8	0·0	0·048	0·048	0·050	0·014	0·011	0·109
		0·5	0·047	0·047	0·049	0·025	0·023	0·211
4	0·2	0·0	0·004	0·004	0·004	0·001	0·001	0·001
		0·5	0·047	0·047	0·047	0·023	0·023	0·023
	0·5	0·0	0·053	0·049	0·049	0·007	0·004	0·005
		0·5	0·049	0·049	0·048	0·018	0·018	0·032
	0·8	0·0	0·053	0·051	0·050	0·009	0·005	0·101
		0·5	0·047	0·048	0·052	0·026	0·024	0·228
8	0·2	0·0	0·000	0·000	0·000	0·000	0·000	0·000
		0·5	0·020	0·020	0·020	0·008	0·008	0·008
	0·5	0·0	0·047	0·045	0·046	0·001	0·001	0·001
		0·5	0·051	0·050	0·053	0·018	0·017	0·032
	0·8	0·0	0·048	0·046	0·046	0·004	0·001	0·060
		0·5	0·049	0·047	0·050	0·025	0·024	0·244

For the right-hand columns, θ is an m-vector of $m/2$ zeros and $m/2$ elements of $-\lambda/2$.
UI-IU, proposed test; PW, Perlman & Wu test (2004); BLT, Bloch et al. (2001) test.

$-\lambda/2$, which is a point on the boundary of the null space. On the other hand, the UI-IU and the Perlman & Wu tests control the type I error rates for all configurations. This is because the Bloch et al. superiority test is only a test of the point null hypothesis $\theta_k = 0$ for all k and does not have a monotone rejection region with respect to the positive orthant

$$\mathcal{O}^+ = \{\theta | \theta_k \geqslant 0 \text{ for all } k, \text{ with at least one } \theta_k > 0\}.$$

As a result, the least favourable configuration of the test over the entire null space does not occur at the point null hypothesis. This problem of the Bloch et al. test worsens as ρ increases because the rejection region of the test becomes more elliptical and less monotone. The excessive type I error rate of the Bloch et al. test was not observed in the simulations reported in Bloch et al. (2001) and Perlman & Wu (2004) because these papers did not examine null configurations where this error rate is near maximum. The UI-IU and Perlman & Wu tests of superiority are both appropriate tests of $H_0 : \theta_k \leqslant 0$ for all k, and have rejection regions that are cone-order monotone (Logan, 2003).

Next we investigated the powers of the three procedures to detect treatment differences. The global powers to identify the treatment as non-inferior on all endpoints and superior on at least one endpoint were compared for the same scenarios as above, but for different treatment-effect configurations. The power results are given in Table 3 for $\lambda = 0{\cdot}5$. Results for other values of λ were similar with the differences in power among the tests increasing

Table 3. *Simulation estimates of powers of* UI-IU, *Perlman & Wu and Bloch et al. tests for* $\delta_k = 0$, $\varepsilon_k = \lambda = 0{\cdot}5$, $\alpha = 0{\cdot}05$

m	θ	ρ	UI-IU	PW	BLT
2	(0·4, 0)	0	0·455	0·455	0·458
		0·5	0·479	0·470	0·507
	(0·4, 0·4)	0	0·800	0·855	0·849
		0·5	0·686	0·699	0·656
4	(0·4, 0, 0, 0)	0	0·280	0·275	0·276
		0·5	0·357	0·340	0·386
	(0·4, 0·4, 0, 0)	0	0·486	0·520	0·517
		0·5	0·498	0·499	0·547
	(0·4, 0·4, 0·4, 0·0)	0	0·704	0·745	0·743
		0·5	0·615	0·626	0·644
	(0·4, 0·4, 0·4, 0·4)	0	0·938	0·977	0·974
		0·5	0·745	0·771	0·663
8	(0·3, 0·3, 0, 0, 0, 0, 0, 0)	0	0·171	0·182	0·179
		0·5	0·235	0·230	0·272
	(0·3, 0·3, 0·3, 0·3, 0, 0, 0, 0)	0	0·338	0·359	0·356
		0·5	0·357	0·361	0·417
	(0·3, 0·3, 0·3, 0·3, 0·3, 0·3, 0, 0)	0	0·573	0·596	0·592
		0·5	0·444	0·464	0·484
	(0·3, 0·3, 0·3, 0·3, 0·3, 0·3, 0·3, 0·3)	0	0·901	0·919	0·918
		0·5	0·554	0·585	0·422

UI-IU, proposed test; PW, Perlman & Wu (2004) test; BLT, Bloch et al. (2001) test.

with λ. For $\rho = 0$, the Perlman & Wu and Bloch et al. tests have similar powers. For $\rho = 0{\cdot}5$, the Bloch et al. test has higher power when the number of endpoints with a positive treatment effect is low, while the Perlman & Wu test has higher power when most of the endpoints have a positive treatment effect. The higher power of the Bloch et al. test is mainly due to its inflated type I error rate. When comparing the UI-IU test with the Perlman & Wu test, we see that the former has slightly higher power in general when fewer than half of the endpoints have positive treatment effects, while the Perlman & Wu test has significantly higher power when half or more of the endpoints have positive treatment effects. This is consistent with the previous findings on superiority tests (Logan, 2003). It is interesting to note that the power of the procedures is sensitive to the correlation in a specific pattern. When half or fewer of the endpoints have a nonzero treatment effect, the power is higher for correlated endpoints, but, when more than half have a nonzero treatment effect, the power is higher for uncorrelated endpoints.

Based on the simulation results, we conclude that the Perlman & Wu test has the best performance of the three tests. The UI-IU test is a close second. The Bloch et al. test has power comparable to the Perlman & Wu test, but it does not control the type I error rate and hence is not recommended.

All of the above tests address a single global hypothesis in (2·1). It might be useful to derive a stepwise multiple test procedure that can determine which of the endpoints show a superior treatment effect with familywise error rate control (Hochberg & Tamhane,

1987, p. 3). The UI-IU test is most convenient for this purpose. Finally, the approach given here can be generalised to deal with the goal of showing that the treatment is equivalent on all endpoints and superior on at least r endpoints, where r is specified ($1 \leqslant r < m$).

Acknowledgement

We would like to thank Michael Perlman, two referees and the editor for their many useful comments and suggestions which helped to improve the paper.

References

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**, 295–300.
Bloch, D. A., Lai, T. L. & Tubert-Bitter, P. (2001). One-sided tests in clinical trials with multiple endpoints. *Biometrics* **57**, 1039–47.
Cohen, A. & Sackrowitz, H. B. (1998). Directional tests for one-sided alternatives in multivariate models. *Ann. Statist.* **26**, 2321–38.
Hochberg, Y. & Tamhane, A. C. (1987). *Multiple Comparison Procedures.* New York: Wiley.
Laska, E. M. & Meisner, M. J. (1989). Testing whether an identified treatment is best. *Biometrics* **45**, 1139–51.
Logan, B. R. (2003). A cone order monotone test for the one-sided multivariate testing problem. *Statist. Prob. Lett.*, **63**, 315–23.
O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–87.
Perlman, M. D. (1969). One-sided testing problems in multivariate analysis. *Ann. Math. Statist.* **40**, 549–67.
Perlman, M. D. & Wu. L. (2004). A note on one-sided tests with multiple endpoints. *Biometrics* **60**, 276–9.
Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.* **24**, 220–38.
Tang, D.-I., Geller, N. L. & Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49**, 23–30.

[*Received October* 2002. *Revised December* 2003]