Biometrical Journal **43** (2001) 5, 591–604

Combining Global and Marginal Tests to Compare Two Treatments on Multiple Endpoints

BRENT R. LOGAN and AJIT C. TAMHANE

Department of Statistics
Northwestern University
Evanston, USA

Summary

We consider the problem of comparing two treatments on multiple endpoints where the goal is to identify the endpoints that have treatment effects, while controlling the familywise error rate. Two current approaches for this are (i) applying a global test within a closed testing procedure, and (ii) adjusting individual endpoint *p*-values for multiplicity. We propose combining the two current methods. We compare the combined method with several competing methods in a simulation study. It is concluded that the combined approach maintains higher power under a variety of treatment effect configurations than the other methods and is thus more power-robust.

Key words: Multiple endpoints; One-sided alternative; Closed test procedure; O'BRIEN's test; Bootstrap; Approximate likelihood ratio test.

1. Introduction

Many clinical trials are conducted to compare two treatment groups (e.g., a new treatment and a control or placebo) with respect to several endpoints. Generally, the treatment is expected to have a positive effect on each of the endpoints. There are two possible inferential goals in these clinical trials. One goal is to establish a significant overall difference between the groups using a global test of the overall null hypothesis of no differences on any of the endpoints. The other goal is to determine on which of the endpoints the treatment differs from the control. One cannot simply test each endpoint separately at level α since this will inflate the probability of at least one false rejection, called the familywise error rate (FWE). For the second goal, an appropriate method must account for the multiplicity of the tests involved, strongly controlling the FWE (HOCHBERG and TAMHANE, 1987) at a specified level α.

There are two standard approaches to controlling the FWE in the literature: applying a global test of no treatment effect on any of the endpoints within a closed testing procedure (MARCUS, PERITZ, and GABRIEL, 1976) and adjusting the single endpoint *p*-values for multiplicity. O'BRIEN (1984) proposed a simple glo-

© WILEY-VCH Verlag Berlin GmbH, 13086 Berlin, 2001 0323-3847/01/0509-0591 $ 17.50+.50/0

bal test, which is optimal when the treatment has the same positive standardized effect on all of the endpoints. Other global tests have also been proposed (see, e.g., TANG, GNECCO, and GELLER, 1989; FOLLMAN, 1996; LÄUTER, 1996). Any of these global tests can be applied in a closed testing procedure to determine which of the endpoints are significant while controlling the FWE (KROPF, 1988; LEHMACHER, WASSMER, and REITMEIR, 1991; WANG, 1998). Adjusting the single endpoint p-values for multiplicity can be done through bootstrap resampling (WESTFALL and YOUNG, 1993). This individual endpoint adjustment is generally more effective when only a few of the endpoints are significant, as it is based on the minimum individual p-value (or maximum individual t-statistic).

In this paper we propose a method that combines the closed testing procedure based on O'BRIEN's global test and the resampling based adjustment to individual endpoint p-values. In LOGAN (2001) we have extended this procedure to include the approximate likelihood ratio test of TANG et al. (1989) as well. This extension is not discussed here because of space constraints, and will be reported elsewhere. Such combined procedures, since they share many of the power characteristics of the separate procedures, are more robust to the configuration of mean differences than individual methods separately.

The organization of the paper is as follows. In Section 2 we outline the problem and set the notation. In Section 3 we review the existing approaches mentioned above. We describe the combined approach in Section 4. In Section 5 we present the results of simulations to compare the proposed procedure with existing procedures. Finally, in Section 6 we discuss the results and draw conclusions.

2. Notation and Problem Formulation

Let X_{ijk} denote the measurement on the kth endpoint for the jth subject in the ith treatment group. Assume that there are two independent treatment groups with n_1 and n_2 subjects. For treatment group i, assume that $\boldsymbol{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijK})'$ $(i = 1, 2, j = 1, 2, \ldots, n_i)$, are independent and identically distributed (i.i.d.) random vectors from a K-variate normal distribution with mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iK})'$ and a common covariance matrix $\boldsymbol{\Sigma}$. Treatment 1 refers to the test treatment and treatment 2 refers to the control.

We are interested primarily in identifying those endpoints that demonstrate a significant improvement over the control. Therefore the family of hypotheses of interest is the set of individual endpoint hypotheses,

$$H_{0k} : \mu_{1k} - \mu_{2k} = \delta_k = 0 \text{ vs. } H_{1k} : \mu_{1k} - \mu_{2k} = \delta_k > 0 \quad (k = 1, 2, \ldots, K). \tag{2.1}$$

Here it is assumed that a higher mean response represents a positive treatment effect for every endpoint.

3. Review of Existing Procedures

3.1 *Closure Method*

The closure method was proposed by MARCUS et al. (1976). From the family of individual endpoint hypotheses (2.1), we first form a closure family by including in it all intersections $H_{0I} = \bigcap_{k \in I} H_{0k}$ for $I \subseteq \{1, 2, \ldots, K\}$. Then we reject any null hypothesis H_{0I} (including any individual hypothesis H_{0k}) at level α iff all null hypotheses $H_{0I'}$ implying it (i.e. $I \subseteq I'$) are rejected by their corresponding α-level tests.

KROPF (1988) and LEHMACHER et al. (1991) have given illustrations of the steps of the closed test procedure in a multiple endpoint situation. All we need to apply the closed testing procedure are appropriate α-level tests of all null hypotheses H_{0I}.

3.2 *Ordinary Least Squares (OLS) Test*

The global hypothesis testing problem for endpoints $k \in I$ can be stated as

$$H_{0I} : \boldsymbol{\mu}_{1I} = \boldsymbol{\mu}_{2I} \text{ vs. } H_{1I} : \boldsymbol{\delta}_I = \boldsymbol{\mu}_{1I} - \boldsymbol{\mu}_{2I} \geq \mathbf{0}$$
$$\text{with} \quad \delta_k > 0 \quad \text{for at least one} \quad k \in I, \tag{3.1}$$

where $\boldsymbol{\mu}_{1I}$ and $\boldsymbol{\mu}_{2I}$ are the mean vectors of those endpoints that are in I for treatments 1 and 2, and $\mathbf{0}$ denotes the null vector of an appropriate dimension. KUDÔ (1963) and PERLMAN (1969) derived the likelihood ratio tests for the above one-sided alternative; however, these tests are very complicated and the null distributions of the test statistics are difficult to evaluate. O'BRIEN (1984) bypassed this difficulty by using a restricted alternative hypothesis. He assumed that the standardized treatment effect δ_k/σ_k is the same (equal to λ) for each endpoint. Therefore the hypothesis testing problem (3.1) reduces to the simple problem $H_0 : \lambda = 0$ vs. $H_1 : \lambda > 0$. The test statistic, using the ordinary least squares (OLS) estimate of λ, depends on the unknown covariance matrix. In practice, the estimated covariance matrix is substituted, resulting in the test statistic

$$T_{\text{OLS}} = \frac{\hat{\lambda}_{\text{OLS}}}{SE(\hat{\lambda}_{\text{OLS}})} = \frac{\boldsymbol{J}'\boldsymbol{t}}{(\boldsymbol{J}'\boldsymbol{R}\boldsymbol{J})^{1/2}}, \tag{3.2}$$

where $\boldsymbol{J}$ is a vector of 1's, $\boldsymbol{R}$ is the estimated correlation matrix and $\boldsymbol{t}$ is the vector of t-statistics for the endpoints in I. The t-statistics are given by

$$t_k = \frac{\bar{x}_{1\cdot k} - \bar{x}_{2\cdot k}}{s_k} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad \text{for} \quad k \in I, \tag{3.3}$$

where s_k is the usual pooled standard deviation for endpoint k and $\bar{x}_{i\cdot k}$ is the sample mean of the ith group for endpoint k. We do not use O'BRIEN's GLS test statistic because (i) it has slower convergence to the standard normal distribution compared to T_{OLS}, and (ii) REITMEIR and WASSMER (1996) have shown that, in terms of the power to reject null hypotheses on individual endpoints within a closed testing procedure, the OLS and GLS tests have similar power performances.

O'Brien proposed a t-distribution with $n_1 + n_2 - 2|I|$ degrees of freedom as an approximation to the null distribution of T_{OLS} for small sample sizes. This gives rather conservative results, especially if $|I|$ is large relative to n_1 and/or n_2, but a better approximation with guaranteed control of the type I error probability is not yet available. LÄUTER's (1996) test is exact; however, its power performance is not satisfactory in a certain asymptotic case as noted by FRICK (1996).

3.3 *Approximate Likelihood Ratio (ALR) Test*

TANG et al. (1989) derived an approximation to the likelihood ratio test of (3.1) that is easy to compute and has a tabulable null distribution. The first step is to compute the transformation

$$\boldsymbol{z} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, \boldsymbol{A}(\bar{\boldsymbol{x}}_{1\cdot} - \bar{\boldsymbol{x}}_{2\cdot})\,, \tag{3.4}$$

where $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{A}'\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{I}$, so that $\boldsymbol{z}$ is a normal vector with the identity covariance matrix $\boldsymbol{I}$. The alternative hypothesis in terms of the transformed vector $\boldsymbol{z}$ is now the polyhedral cone $\boldsymbol{A}(\boldsymbol{\delta}) = \{\boldsymbol{A}\boldsymbol{\delta} \mid \boldsymbol{\delta} \geq \boldsymbol{0}\}$. This cone alternative is approximated by the positive orthant. The exact likelihood ratio statistic for this approximate alternative can be easily derived because the components of $\boldsymbol{z}$ are i.i.d. $N(0, 1)$ under H_0. The resulting statistic is called the approximate likelihood ratio (ALR) test statistic and it equals

$$g(\boldsymbol{z}) = \sum_{k=1}^{K} \{\max(z_k, 0)\}^2\,.$$

The null distribution of $g(\boldsymbol{z})$ is a special case of the chi-bar-squared distribution with binomial weights, given by

$$P(g(\boldsymbol{z}) \geq c) = \sum_{k=0}^{K} \left[\binom{K}{k} \frac{1}{2^K}\, P(\chi_k^2 \geq c) \right], \tag{3.5}$$

where $\chi_0^2 = 0$.

The matrix $\boldsymbol{A}$ used in the transformation is not unique. TANG et al. (1989) suggested using the Cholesky decomposition with a further restriction that $\boldsymbol{A}$ be chosen so that the center direction of $\boldsymbol{A}(\boldsymbol{\delta})$ and the positive orthant coincide. TANG

Table 1
Critical Values for ALR test based on unknown $\mathbf{\Sigma}$

n	K						
	2	3	4	5	6	7	8
10	5.14	7.17	9.38	11.83	14.61	18.00	22.20
50	4.38	5.69	6.91	8.05	9.11	10.17	11.30
∞	4.23	5.44	6.50	7.48	8.41	9.29	10.16

et al. note that a further requirement of order invariance could be imposed, but the statistic $g(\boldsymbol{z})$ does not change much. The left-root symmetric method of LÄUTER, KROPF, and GLIMM (1998) for finding the decomposition $\boldsymbol{A}'\boldsymbol{A} = \mathbf{\Sigma}^{-1}$ is both scale and order invariant.

The result (3.5) holds if and only if the covariance matrix is known. In practice, the covariance matrix is estimated, but the sample sizes must be large for (3.5) to provide a good approximation (REITMEIR and WASSMER, 1996). For small to moderate sample sizes, the distribution of the ALR test using the estimated covariance matrix can be obtained using simulation, since it does not depend on the true covariance matrix $\mathbf{\Sigma}$. This result is proved in TAMHANE and LOGAN (2001), where an accurate approximation to the critical values of the ALR test statistic in case of an estimated covariance matrix is also given. The 5% critical values obtained via simulation and the chi-bar-squared approximation are given in Table 1 for equal sample sizes of $n = 10$ and $n = 50$, and for $K = 2, \ldots, 8$. These estimated correct critical values are used in all simulations.

3.4 *Multiplicity Adjustments to Tests for Individual Endpoints*

Another way to control the FWE for this family is to adjust the p-values of the individual hypotheses (2.1) for multiplicity. Let $p_{(1)} \geq \ldots \geq p_{(K)}$ denote the ordered p-values corresponding to the one-sided t-statistics (given in equation (3.3)) for the hypotheses $H_{0(1)}, \ldots, H_{0(K)}$. Adjusted p-values for the step-down test procedure are given by

$$\tilde{p}_{(K)} = P\left(\min_{1\leq\ell\leq K} P_\ell \leq p_{(K)}\right) \quad \text{and}$$

$$\tilde{p}_{(k)} = \max\left[\tilde{p}_{(k+1)}, P\left(\min_{1\leq\ell\leq k} P_\ell \leq p_{(k)}\right)\right] \quad \text{for} \quad k = 1, \ldots, K-1, \quad (3.6)$$

where P_ℓ refers to the random variable associated with the p-value for the ℓth endpoint.

The bootstrap method of WESTFALL and YOUNG (1993) allows one to estimate the above adjusted p-values by resampling. Because a common covariance matrix

is assumed, the data from the two treatments can be pooled before drawing bootstrap samples. The bootstrap method implicitly takes into account the correlations between endpoints. Note, however, that at each stage of the closed testing procedure, the test is based solely on the minimum p-value. Therefore using individual endpoint adjustments can have less power compared to the OLS closed testing procedure when all of the endpoints have treatment effects that are similar in magnitude.

4. Proposed Combined Closed Testing Procedure

The individual endpoints method performs better when only a few endpoints have treatment effects, while the OLS method performs better when all of the endpoints have treatment effects. To exploit their complementary areas of strength, we propose to combine the two in a method analogous to HOTHORN's (1999) $T_{\max}$ testing principle, which uses the maximum of several t-statistics optimal under different alternative configurations to test a given hypothesis. It should be noted that the choice of the OLS test in the combined procedure is somewhat arbitrary; any other global test could be used. However, this specific choice does not detract from the principal advantage of the combined approach.

To test H_{0I}, we propose to use the test statistic,

$$\min\left(\left(\min_{k\in I} p_k\right), p_{I,\mathrm{OLS}}\right), \tag{4.1}$$

i.e., the minimum p-value between the individual endpoints in I and the OLS global test on I. Because the minimum individual p-value and the OLS p-value are positively correlated, taking the minimum of the two will not change the α-level critical point substantially from that of each statistic separately. Therefore this combined approach will mimic the power characteristics of the better method for each situation, and will result in a method more power-robust to the configuration of treatment effects.

The distribution of the test statistic to be used in the closed testing procedure is complicated, but it can be easily estimated using bootstrap resampling. To estimate the p-value for the test of H_{0I}, we use the following algorithm. The C language code for the algorithm is available from the web site `http://users.iems.northwestern.edu/~ajit`.

1. Mean center the data:

   $$y_{ijk} = x_{ijk} - \bar{x}_{i\cdot k}\,.$$

   This mean centering allows one to pool the data from the two samples together, assuming a common covariance matrix $\mathbf{\Sigma}$, and to mimic the null hypothesis of no difference between the two treatment mean vectors.

2. Generate M pairs of bootstrap samples of sizes n_1 and n_2 with replacement from the pooled mean-centered data, y_{ijk}. The mth bootstrap sample ($m = 1, \ldots, M$) is denoted by

$$\left\{ y_{ijk}^{*(m)} \text{ for } i = 1, 2,\ j = 1, \ldots, n_i,\ k = 1, \ldots, K \right\}.$$

3. For each bootstrap sample, calculate the bootstrap t-statistics corresponding to the individual hypotheses H_{0k},

$$t_k^{*(m)} = \frac{\bar{y}_{1\cdot k}^{*(m)} - \bar{y}_{2\cdot k}^{*(m)}}{s_k^{*(m)} \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \quad \text{for} \quad k \in I, \tag{4.2}$$

where $s_k^{*(m)}$ is the pooled standard deviation for the kth endpoint of the mth bootstrap sample. Then calculate the unadjusted bootstrap p-values

$$p_k^{*(m)} = P\left(T_{n_1+n_2-2} > t_k^{*(m)} \right) \quad \text{for} \quad k \in I, \tag{4.3}$$

where $T_{n_1+n_2-2}$ is a Student's t random variable with $n_1 + n_2 - 2$ degrees of freedom.

4. For each bootstrap sample, calculate the bootstrap OLS t-statistic corresponding to the hypothesis H_{0I},

$$t_{I,\mathrm{OLS}}^{*(m)} = \frac{\boldsymbol{J}'\boldsymbol{t}^{*(m)}}{\left(\boldsymbol{J}'\boldsymbol{R}^{*(m)}\boldsymbol{J}\right)^{1/2}}, \tag{4.4}$$

where $\boldsymbol{R}^{*(m)}$ is the estimated correlation matrix and $\boldsymbol{t}^{*(m)}$ is the vector of t-statistics for endpoints in I, both computed from the mth bootstrap sample. Then the bootstrap OLS p-value is given by

$$p_{I,\mathrm{OLS}}^{*(m)} = P\left(T_{n_1+n_2-2|I|} > t_{I,\mathrm{OLS}}^{*(m)} \right). \tag{4.5}$$

5. The p-value to test H_{0I} is given by

$$p_I = P\left\{ \min\left(\min_{k \in I} P_k,\ P_{I,\mathrm{OLS}} \right) \leq \min\left(\min_{k \in I} p_k,\ p_{I,\mathrm{OLS}} \right) \right\}, \tag{4.6}$$

where P_k is the random variable associated with the individual p-value for the kth endpoint, $P_{I,\mathrm{OLS}}$ is the random variable associated with the OLS p-value for the endpoints in I, p_k is the observed individual p-value for the kth endpoint, and $p_{I,\mathrm{OLS}}$ is the observed OLS p-value for the endpoints in I. Note that if I contains only one endpoint k then p_I equals p_k. These p-values can be estimated using

$$\hat{p}_I = \frac{1}{M}\left\{ \#m \,\middle|\, \min\left(\min_{k \in I} p_k^{*(m)},\ p_{I,\mathrm{OLS}}^{*(m)} \right) \leq \min\left(\min_{k \in I} p_k,\ p_{I,\mathrm{OLS}} \right) \right\}. \tag{4.7}$$

This bootstrap p-value can be compared to α to test H_{0I} in the closed testing procedure.

One can also obtain the adjusted p-values for each subset hypothesis in the closed testing procedure. The adjusted p-value for any H_{0I} is simply the maximum of the unadjusted p-values for all hypotheses $H_{0I'}$ implying H_{0I}, i.e., $I \subseteq I'$. However, this calculation requires evaluation of bootstrap p-values for all $2^K - 1$ subset hypotheses. It is computationally quicker to apply α-level tests rather than compute adjusted p-values, since if a certain hypothesis is accepted, the hypotheses implied by it are accepted by implication, and their bootstrap p-values need not be computed.

The bootstrap $\max t$ approach uses a shortcut based on the union-intersection nature of the test, allowing it to bypass the sub-hypotheses of the closed testing procedure and directly test the individual hypotheses. A similar shortcut can be applied when using the combined test: if rejection of the overall null hypothesis is caused by a particular individual hypothesis H_{0k}, which produced the minimum individual p-value, then that hypothesis can be rejected and removed from subsequent consideration. This shortcut is conjectured to be valid; however, we do not have an analytical proof of this conjecture. While it remains to be proven, simulation studies support the shortcut method's control of the FWE. All of the simulation studies in this paper apply the shortcut method.

5. Simulations

5.1 *Design of Simulation Studies*

Simulations were designed to compare the proposed combined method with the bootstrap $\max t$ and the closed testing procedures based on the OLS and the ALR tests. First, we simulated the FWE under the overall null configuration, where all null hypotheses $H_{0k} : \delta_k = 0$ are true. Sample sizes of $n = 10$ and $n = 50$ were investigated for $K = 4$ and $K = 8$ endpoints. Four correlation matrices were considered: equal correlation with $\rho = 0.0, 0.5$ and 0.7, and block correlation with two blocks, $(\{1,2\},\{3,4\})$ for $K = 4$ and $(\{1,2,3,4\},\{5,6,7,8\})$ for $K = 8$ with values of $\rho = 0.5$ within blocks and $\rho = 0.1$ between blocks. In each simulation run $n = 10$ or 50 deviates were generated for each treatment group from an MVN distribution with mean vector $\mathbf{0}$ and the given correlation matrix. For each configuration, a total of 10,000 runs were made. The bootstrap methods were based on 5,000 bootstrap samples.

The power functions of the four methods were also compared through simulation. Five different mean difference configurations ($\boldsymbol{\delta}$) were studied for $K = 4$ and $K = 8$ endpoints with either $1/4$, $1/2$, $3/4$ or all of the endpoints with common $\delta_k > 0$. In addition, a configuration was studied where all endpoints had positive treatment differences, but half had large δ_k values and half had small δ_k values. The same four correlation structures as in the FWE simulations

were studied. In each simulation run, $n = 50$ MVN deviates were generated from both treatments, but with mean vector $\boldsymbol{\delta}$ for treatment 1 and $\mathbf{0}$ for treatment 2. The criterion reported is the average power, which is the average of the marginal powers for detecting each positive treatment difference. The powers for four endpoints are reported in Table 2 and the powers for eight endpoints are reported in Table 3.

Although the average of the marginal powers to detect treatment differences is of primary interest when trying to identify which of the endpoints has a treatment effect, sometimes the researcher is interested only in testing whether there is a global treatment effect. To compare the performances of the given methods in this situation, a final simulation study was designed. The same correlation structures were used as above and the same treatment mean configurations, but with different $\boldsymbol{\delta}$ values (to avoid powers close to 1). The criterion given is the power to reject the global null hypothesis. The powers for four endpoints are reported in Table 4 and the powers for eight endpoints are reported in Table 5.

Table 2
Average Power ($K = 4$, $n = 50$)

Corr. Structure	$\boldsymbol{\delta}$	Average Power			
		Bootstrap max t	Bootstrap max t/OLS	Closed OLS	Closed ALR
Equal (0.0)	(0.5, 0, 0, 0)	0.590	0.575	0.192	0.524
	(0.5, 0.5, 0, 0)	0.623	0.615	0.321	0.600
	(0.5, 0.5, 0.5, 0)	0.655	0.664	0.511	0.674
	(0.5, 0.5, 0.5, 0.5)	0.711	0.771	0.784	0.776
	(0.25, 0.25, 0.75, 0.75)	0.601	0.623	0.614	0.627
Equal (0.5)	(0.5, 0, 0, 0)	0.623	0.619	0.165	0.619
	(0.5, 0.5, 0, 0)	0.645	0.640	0.253	0.659
	(0.5, 0.5, 0.5, 0)	0.671	0.665	0.407	0.691
	(0.5, 0.5, 0.5, 0.5)	0.717	0.737	0.759	0.693
	(0.25, 0.25, 0.75, 0.75)	0.611	0.616	0.544	0.607
Equal (0.7)	(0.5, 0, 0, 0)	0.658	0.657	0.174	0.737
	(0.5, 0.5, 0, 0)	0.674	0.673	0.236	0.719
	(0.5, 0.5, 0.5, 0)	0.698	0.695	0.383	0.689
	(0.5, 0.5, 0.5, 0.5)	0.729	0.738	0.759	0.652
	(0.25, 0.25, 0.75, 0.75)	0.625	0.627	0.526	0.616
Block (0.5, 0.1)	(0.5, 0, 0, 0)	0.597	0.586	0.178	0.540
	(0.5, 0.5, 0, 0)	0.622	0.616	0.288	0.555
	(0.5, 0.5, 0.5, 0)	0.663	0.658	0.459	0.679
	(0.5, 0.5, 0.5, 0.5)	0.704	0.742	0.761	0.735
	(0.25, 0.25, 0.75, 0.75)	0.612	0.620	0.594	0.609

Table 3
Average Power ($K = 8$, $n = 50$)

Corr. Structure	$\boldsymbol{\delta}$	Average Power			
		Bootstrap max t	Bootstrap max t/OLS	Closed OLS	Closed ALR
Equal (0.0)	(0.5, 0, 0, 0)	0.503	0.492	0.038	0.362
	(0.5, 0.5, 0, 0)	0.524	0.519	0.011	0.461
	(0.5, 0.5, 0.5, 0)	0.554	0.574	0.316	0.581
	(0.5, 0.5, 0.5, 0.5)	0.609	0.733	0.762	0.746
	(0.25, 0.25, 0.75, 0.75)	0.545	0.576	0.547	0.579
Equal (0.5)	(0.5, 0, 0, 0)	0.543	0.542	0.079	0.416
	(0.5, 0.5, 0, 0)	0.564	0.562	0.125	0.514
	(0.5, 0.5, 0.5, 0)	0.587	0.584	0.246	0.573
	(0.5, 0.5, 0.5, 0.5)	0.636	0.670	0.714	0.578
	(0.25, 0.25, 0.75, 0.75)	0.569	0.575	0.456	0.541
Equal (0.7)	(0.5, 0, 0, 0)	0.597	0.597	0.086	0.567
	(0.5, 0.5, 0, 0)	0.610	0.610	0.130	0.596
	(0.5, 0.5, 0.5, 0)	0.626	0.625	0.232	0.559
	(0.5, 0.5, 0.5, 0.5)	0.664	0.675	0.716	0.515
	(0.25, 0.25, 0.75, 0.75)	0.588	0.591	0.434	0.555
Block (0.5, 0.1)	(0.5, 0, 0, 0)	0.526	0.521	0.055	0.322
	(0.5, 0.5, 0, 0)	0.549	0.546	0.130	0.338
	(0.5, 0.5, 0.5, 0)	0.574	0.573	0.267	0.551
	(0.5, 0.5, 0.5, 0.5)	0.632	0.697	0.734	0.649
	(0.25, 0.25, 0.75, 0.75)	0.564	0.576	0.500	0.542

5.2 *Results*

The FWE's are not reported since all procedures maintained FWE's reasonably well, with none exceeding the .05-level critical value of $0.05 + 1.96\sqrt{\dfrac{.05 \times .95}{10,000}} = 0.054$. It should be remarked that the $t_{n_1+n_2-2|I|}$ approximation for the OLS test is overly conservative for $n_1 = n_2 = 10$ and $K = 8$ with FWE's < 0.03. In this situation, the degrees of freedom are too low because the sample sizes are too small relative to K.

In terms of the average power to detect treatment differences, several trends are worth noting. The two methods involving individual endpoints, the bootstrap $\max t$ and the combined max t/OLS, have stable powers under more configurations than does the OLS test by itself in a closed testing procedure. When only one-fourth of the endpoints have treatment effects, the power of the OLS test drops dramatically compared to the other two methods. When all of the endpoints have equal treatment differences, the OLS test is most powerful, as expected. While the

Table 4
Power to Reject Global Null Hypothesis ($K = 4$, $n = 50$)

Corr. Structure	$\boldsymbol{\delta}$	Power			
		Bootstrap $\max t$	Bootstrap $\max t$/OLS	OLS	ALR
Equal (0.0)	(0.4, 0, 0, 0)	0.419	0.418	0.267	0.406
	(0.4, 0.4, 0, 0)	0.640	0.668	0.637	0.739
	(0.4, 0.4, 0.4, 0)	0.782	0.852	0.912	0.896
	(0.4, 0.4, 0.4, 0.4)	0.868	0.966	0.991	0.967
	(0.2, 0.2, 0.4, 0.4)	0.717	0.837	0.913	0.857
Equal (0.5)	(0.4, 0, 0, 0)	0.426	0.422	0.155	0.475
	(0.4, 0.4, 0, 0)	0.596	0.592	0.346	0.703
	(0.4, 0.4, 0.4, 0)	0.685	0.682	0.599	0.757
	(0.4, 0.4, 0.4, 0.4)	0.750	0.762	0.823	0.696
	(0.2, 0.2, 0.4, 0.4)	0.617	0.619	0.594	0.601
Equal (0.7)	(0.4, 0, 0, 0)	0.466	0.465	0.142	0.625
	(0.4, 0.4, 0, 0)	0.586	0.584	0.297	0.833
	(0.4, 0.4, 0.4, 0)	0.655	0.655	0.525	0.834
	(0.4, 0.4, 0.4, 0.4)	0.688	0.690	0.723	0.601
	(0.2, 0.2, 0.4, 0.4)	0.597	0.596	0.522	0.609
Block (0.5, 0.1)	(0.4, 0, 0, 0)	0.427	0.419	0.191	0.447
	(0.4, 0.4, 0, 0)	0.590	0.587	0.453	0.557
	(0.4, 0.4, 0.4, 0)	0.728	0.734	0.732	0.819
	(0.4, 0.4, 0.4, 0.4)	0.802	0.854	0.920	0.834
	(0.2, 0.2, 0.4, 0.4)	0.631	0.667	0.733	0.655

OLS test is optimal for equal mean differences, it cannot be recommended due to its loss in power when only a few of the endpoints have treatment effects.

The ALR test has more stable power than does the OLS test when only a few of the endpoints have treatment effects. For $K = 4$ endpoints the ALR test is never far from the best performing method, usually performing better than the others when $1/2$ or $3/4$th of the endpoints have treatment effects, and slightly worse when either all or $1/4$th of the endpoints have effects. The differences become more pronounced when $K = 8$. The ALR test loses 14–20% power relative to the individual endpoint methods when $1/4$th of the endpoints have treatment effects. For the equal correlation of 0.5 and block correlation matrices, the ALR test is dominated by the combined test. Under independence, its power increases by no more than 1.5% over the combined test when at least half of the endpoints have effects. In terms of average power, the combined test appears to do better overall than the ALR test.

In comparing the combined test with the bootstrap $\max t$ test, note first that the worst the combined test does is a 1.5% loss in power when one of the four independent endpoints has a treatment effect. On the other hand, the combined test

Table 5
Power to Reject Global Null Hypothesis ($K = 8$, $n = 50$)

Corr. Structure	$\boldsymbol{\delta}$	Power			
		Bootstrap max t	Bootstrap max t/OLS	OLS	ALR
Equal (0.0)	(0.4, 0, 0, 0)	0.529	0.540	0.410	0.605
	(0.4, 0.4, 0, 0)	0.774	0.837	0.873	0.919
	(0.3, 0.3, 0.3, 0)	0.652	0.826	0.932	0.893
	(0.3, 0.3, 0.3, 0.3)	0.751	0.962	0.995	0.970
	(0.2, 0.2, 0.4, 0.4)	0.818	0.966	0.994	0.980
Equal (0.5)	(0.4, 0, 0, 0)	0.498	0.496	0.166	0.642
	(0.4, 0.4, 0, 0)	0.650	0.649	0.378	0.848
	(0.3, 0.3, 0.3, 0)	0.507	0.507	0.448	0.578
	(0.3, 0.3, 0.3, 0.3)	0.553	0.557	0.635	0.469
	(0.2, 0.2, 0.4, 0.4)	0.661	0.661	0.633	0.642
Equal (0.7)	(0.4, 0, 0, 0)	0.512	0.512	0.147	0.826
	(0.4, 0.4, 0, 0)	0.618	0.618	0.315	0.948
	(0.3, 0.3, 0.3, 0)	0.457	0.457	0.360	0.669
	(0.3, 0.3, 0.3, 0.3)	0.501	0.501	0.534	0.389
	(0.2, 0.2, 0.4, 0.4)	0.614	0.613	0.530	0.669
Block (0.5,0.1)	(0.4, 0, 0, 0)	0.476	0.472	0.210	0.575
	(0.4, 0.4, 0, 0)	0.634	0.635	0.510	0.555
	(0.3, 0.3, 0.3, 0)	0.552	0.559	0.593	0.646
	(0.3, 0.3, 0.3, 0.3)	0.613	0.655	0.796	0.629
	(0.2, 0.2, 0.4, 0.4)	0.688	0.718	0.800	0.673

generally has higher power when more than half the endpoints have treatment effects. In addition, in the case where all endpoints have identical treatment effects, the combined test enjoys a 6–12% gain in power for four and eight endpoints under independence, a 2–4% gain in power under the equal correlation of $\rho = 0.5$, and a 4–6% gain in power under the block correlation structure. However, when $\rho = 0.7$, the combined test and the bootstrap max t test have essentially identical powers. Thus the combined test is more robust to the treatment difference configuration than the bootstrap $\max t$ test.

The effect of increasing the correlation is generally, though not uniformly, to reduce the relative effectiveness of the OLS test and to increase the relative effectiveness of the bootstrap max t method. This ends up lessening the differences between the bootstrap max t and the combined method, making the extra effort of including the OLS test less worthwhile. However, there is still some small improvement when all of the endpoints have treatment effects. As mentioned above, this improvement gets less pronounced as the correlations increase.

Next consider the case where the researcher's primary interest is detection of an overall treatment effect. First we examine the simulations under independence.

The ALR test dominates the max t and the combined approach under independence. If fewer than half of the endpoints have treatment effects then the ALR test performs better than the OLS test, while if more than half of the endpoints have treatment effects then the OLS test performs best. In all configurations the combined approach does better than just the bootstrap max t.

The situation changes if we look at endpoints with equal correlation of 0.5 and 0.7. In these cases, the bootstrap $\max t$ and combined methods perform almost identically. The OLS test is less powerful than the other methods in all cases except when all endpoints have equal treatment effects. If all endpoints have treatment effects then the individual endpoint methods perform better than the ALR test. Otherwise, the ALR test has higher power. The results for block correlation fall in between those for independence and equal correlation of 0.5 or 0.7.

6. Discussion and Conclusions

Based on the simulation results we conclude that the combined test is very powerful for detecting individual endpoint treatment effects and is more robust to the configuration of the mean differences than the other methods. It is especially advantageous when the correlations between the endpoints are low and the treatment effects are similar for all endpoints. There is very little loss of power in other situations. The only drawback of the combined test is the additional computation.

Clinical researchers often choose the endpoints that measure different aspects of disease recovery. As a result, typical correlations range between 0.2 to 0.6, rarely exceeding 0.7 or 0.8. In this sense very highly correlated endpoints are less informative. For such settings the combined test offers definite power gains.

In terms of the power to reject the global null hypothesis, the combined test is more robust to the treatment effect configuration and correlation matrix than either the individual endpoint test or the global OLS test alone. In this case the ALR test performs better than the combined test in most situations and requires less computations.

The proposed combined test could easily be extended to MVN data with unequal covariance matrices, and even to non-normal data. All that is required are appropriate individual endpoint tests and a global test. Then the bootstrap method allows us to easily estimate the null distribution of the complicated combined test statistic, while guaranteeing asymptotic control of the FWE. The work for unequal covariance matrices will be reported in a future paper.

Acknowledgments

The authors are grateful to two referees for helpful comments and pointing out some relevant references. The authors also wish to thank Professor Ludwig Ho-

thorn, the editor of this special issue of the journal and the organizer of the Second International Conference on Multiple Comparisons, whose seminar talk at Northwestern University motivated this research.

References

FOLLMAN, D., 1996: A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association* **91**, 854–861.

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

HOCHBERG, Y. and TAMHANE, A. C., 1987: *Multiple Comparison Procedures*. John Wiley: New York.

HOTHORN, L., 1999: The $T_{\max}$ testing principle. Seminar given at the Department of Statistics, Northwestern University, Evanston, IL.

KROPF, S., 1988: Application of multivariate test procedures to the combination of multivariate and univariate tests with varying variable sets. *Biometrical Journal* **4**, 461–470.

KUDÔ, A., 1963: A multivariate analogue of the one-sided test. *Biometrika* **50**, 403–418.

LÄUTER, J. 1996: Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* **52**, 964–970.

LÄUTER, J., KROPF, S., and GLIMM, E., 1998: Exact stable multivariate tests for applications in clinical research. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, 46–55.

LEHMACHER, W., WASSMER, G., and REITMEIR, P., 1991: Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **47**, 511–521.

LOGAN, B. R., 2001: Contributions to Multiple Endpoints and Dose Finding. Doctoral Dissertation, Department of Statistics, Northwestern University, Evanston, IL.

MARCUS, R., PERITZ, E., and GABRIEL, K. R., 1976: On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

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.

REITMEIR, P. and WASSMER, G., 1996: One-sided multiple endpoint testing in two-sample comparisons. *Communications in Statistics-Simulation* **25**, 99–117.

TAMHANE, A. C. and LOGAN, B. R., 2001: Accurate critical constants for the one-sided approximate likelihood ratio test of a normal mean vector when the covariance matrix is estimated. Submitted for publication.

TANG, D. I., GNECCO, C., and GELLER, N. L., 1989: An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika* **76**, 577–583.

WANG, S.-J., 1998: A closed procedure based on Follman's test for the analysis of multiple endpoints, *Communications in Statistics – Theory and Methodology* **27**, 2461–2480.

WESTFALL, P. H. and YOUNG, S. S., 1993: *Resampling Based Multiple Testing*. John Wiley: New York.

BRENT R. LOGAN and AJIT C. TAMHANE
Dept. of Statistics
Northwestern University
2006 Sheridan Rd.
Evanston, IL 60208-4070
E-mail: ajit@iems.northwestern.edu

Received, September 2000
Revised, April 2001
Accepted, May 2001