Multiple Endpoints:

An Overview and New Developments

Ajit C. Tamhane and Brent R. Logan

Department of Statistics — Division of Biostatistics

Northwestern University — Medical College of Wisconsin

Division of Biostatistics

Medical College of Wisconsin

Technical Report 43

September 2003

Division of Biostatistics

Medical College of Wisconsin

8701 Watertown Plank Road

Milwaukee, WI 53226

Phone:(414) 456-8280

Multiple Endpoints: An Overview and New Developments

Ajit C. Tamhane
Department of Statistics
Northwestern University
2006 Sheridan Road, Evanston, IL 60208
E-mail: ajit@iems.northwestern.edu
and
Brent R. Logan
Division of Biostatistics
Medical College of Wisconsin
8701 Watertown Plank Rd.
Milwaukee, WI 53226
E-mail: blogan@mcw.edu

Summary

In the last two decades a large number of papers have been published on the topic of analysis of multiple endpoints in clinical trials. We provide a comprehensive review of this vast literature focusing on the statistical aspects. We make comparisons between competing procedures, present some new developments and extensions/modifications of existing procedures, make recommendations for use and note some open problems for research.

Keywords: Multiple comparisons; multiple tests; one-sided multivariate tests; Bonferroni test; chi-bar squared distribution; multivariate normal distribution; clinical decision rules; global tests; endpoint specific tests; closure method; resampling; adjusted p-values; family-wise error rate.

1. Introduction

Most clinical trials are conducted to compare a treatment group with a control group on multiple endpoints. Often, the treatment is expected to have a positive effect on all endpoints. Depending on the nature of the disease the endpoints may be grouped into primary and secondary types. We mainly focus on the case where all endpoints are primary and provide a comprehensive review of the vast literature and some new results focusing on the statistical aspects. Shorter review articles by Chi (1998), Huque and Sankoh (1997), Sankoh, Huque and Dubey (1997), Sankoh, Huque, Russell and D'Agostino (1999) and Zhang, Quan, Ng and Stepanavage (1997) also discuss some clinical aspects with examples.

Broadly speaking, there are two inferential goals when dealing with multiple endpoints. Goal 1 is to establish an overall treatment effect using a test of the global null hypothesis of no differences on any of the endpoints against a one-sided alternative. Goal 2 is to identify the individual endpoints on which the treatment is better than the control. We review procedures proposed for both these goals, make comparisons and propose some extensions. Section 2 sets the notation. Section 3 discusses test procedures for Goal 1 and Section 4 discusses test procedures for Goal 2. In Section 5 we discuss some clinical decision rules that have been proposed in practice for drug approval purposes (see, e.g., Chi 2000) which typically involve both the primary and secondary endpoints.

2. Notation and Preliminaries

Suppose that there are two independent treatment groups with n_1 and n_2 subjects on each of whom $m \geq 2$ endpoints are measured. Treatment 1 is the test treatment and treatment 2 is the control. Let x_{ijk} denote the measurement on the kth endpoint for the jth subject in the ith treatment group. For treatment group i ($i = 1, 2$), assume that $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijm})'$, $j = 1, 2, \ldots, n_i$, are independent and identically distributed (i.i.d.) random variables (r.v.'s) from an m-variate normal distribution with mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im})'$ and covariance matrix $\boldsymbol{\Sigma}_i$. In the homoscedastic case, we assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ (say). The elements of $\boldsymbol{\Sigma}$ are

$$\sigma_{kk} = \mathrm{Var}(x_{ijk}) \text{ and } \sigma_{k\ell} = \mathrm{Cov}(x_{ijk}, x_{ij\ell}) \ \ (1 \leq k \neq \ell \leq m).$$

The corresponding correlation matrix will be denoted by $\boldsymbol{R}$ with elements

$$\rho_{k\ell} = \text{Corr}(x_{ijk}, x_{ij\ell}) = \frac{\sigma_{k\ell}}{\sqrt{\sigma_{kk}\sigma_{\ell\ell}}} \ \ (1 \leq k \neq \ell \leq m).$$

In the heteroscedastic case, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are not assumed to be equal. The elements of $\boldsymbol{\Sigma}_i$ will be denoted by $\sigma_{i,k\ell}$ $(i = 1, 2; 1 \leq k \leq \ell \leq m)$. The corresponding correlation matrices will be denoted by $\boldsymbol{R}_1 = \{\rho_{1,k\ell}\}$ and $\boldsymbol{R}_2 = \{\rho_{2,k\ell}\}$, respectively.

Let $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\delta_1, \delta_2, \ldots, \delta_m)'$ denote the vector of mean differences. To establish an overall treatment effect (Goal 1), usually a *single* global null hypothesis of no difference is tested against a one-sided alternative:

$$H_0 : \boldsymbol{\delta} = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\delta} \in \mathcal{O}^+, \tag{2.1}$$

where $\mathbf{0}$ is the null vector and

$$\mathcal{O}^+ = \{\boldsymbol{\delta} | \boldsymbol{\delta} \geq \mathbf{0}, \boldsymbol{\delta} \neq \mathbf{0}\}$$

is the positive orthant. To identify the endpoints on which the treatment is better than the control (Goal 2), usually *multiple* null hypotheses are tested against one-sided alternatives:

$$H_{0k} : \delta_k = \mu_{1k} - \mu_{2k} = 0 \text{ vs. } H_{1k} : \delta_k = \mu_{1k} - \mu_{2k} > 0 \ \ (1 \leq k \leq m). \tag{2.2}$$

In this case it is required to control the familywise error rate (FWE), defined as

$$\text{FWE} = \Pr\{\text{at least one true } H_{0k} \text{ is rejected}\},$$

at a specified level α regardless of which particular H_{0k} are true. This is called the *strong* FWE control (Hochberg and Tamhane 1987, p. 3), which will be assumed throughout.

Let $\overline{\boldsymbol{x}}_{i\cdot} = (\overline{x}_{i\cdot 1}, \overline{x}_{i\cdot 2}, \ldots, \overline{x}_{i\cdot m})'$ denote the vector of sample means of the n_i subjects from the ith group and let $\widehat{\boldsymbol{\Sigma}}_i$ denote the sample covariance matrix from the ith group with $\nu_i = n_i - 1$ degrees of freedom (d.f.) $(i = 1, 2)$. In the homoscedastic case, we use the pooled estimate of $\boldsymbol{\Sigma}$ given by $\widehat{\boldsymbol{\Sigma}} = ((n_1 - 1)\widehat{\boldsymbol{\Sigma}}_1 + (n_2 - 1)\widehat{\boldsymbol{\Sigma}}_2)/(n_1 + n_2 - 2)$ with $n_1 + n_2 - 2$ d.f. Denote the elements of $\widehat{\boldsymbol{\Sigma}}$ by $\widehat{\sigma}_{k\ell}$ $(1 \leq k \leq \ell \leq m)$.

3. Global Tests

In this section we focus on the global hypothesis testing problem (2.1). We first discuss the tests proposed for the homoscedastic case and then offer their extensions for the heteroscedastic case.

3.1 Homoscedastic Case

3.1.1 Exact Likelihood Ratio (LR) Tests

It is well-known that because Hotelling's T^2 test is designed for the omnibus (two-sided) alternative $H_2 : \boldsymbol{\delta} \neq \mathbf{0}$, it lacks power for the one-sided alternative H_1 of (2.1) (Meier 1975, O'Brien 1984). Kudô (1963) derived an exact LR test when $\boldsymbol{\Sigma}$ is *known* for the one-sample problem which can be easily extended to the two-sample problem as follows. Let $\widehat{\boldsymbol{\delta}}$ be the projection of $\overline{\boldsymbol{x}}_{1\cdot} - \overline{\boldsymbol{x}}_{2\cdot}$ in the positive orthant with respect to the distance function $d(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u} - \boldsymbol{v})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{u} - \boldsymbol{v})$. Then the LR test rejects for large values of

$$\left(\frac{n_1 n_2}{n_1 + n_2}\right) \widehat{\boldsymbol{\delta}}' \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\delta}}.$$

The null distribution of this statistic is a chi-bar-squared ($\overline{\chi}^2$) distribution, which is a weighted sum of central χ^2_k $(0 \leq k \leq m)$ distributions, where $\chi^2_0 = 0$; see Robertson, Wright and Dikstra (1988). The weights are called the level probabilities that depend on $\boldsymbol{\Sigma}$. This test is not easy to implement because of the difficulty of finding $\widehat{\boldsymbol{\delta}}$ and the complicated nature of its null distribution.

Perlman (1969) derived an exact LR test for the one-sample problem when $\boldsymbol{\Sigma}$ is *unknown*. However, the null distribution of the resulting test statistic is not free of $\boldsymbol{\Sigma}$ and the test is biased. Perlman did provide sharp lower and upper bounds on the null distribution that are free of $\boldsymbol{\Sigma}$. An exact LR test has not been derived for the two-sample problem in this case. In addition to the computational and analytical difficulties mentioned above, the LR tests suffer from a more basic problem that they can sometimes reject H_0 in favor of the one-sided alternative H_1 even when all sample mean differences $\overline{x}_{1\cdot k} - \overline{x}_{2\cdot k}$ are negative; see, e.g., Follman (1995) and Silvapulle (1997). They can also be non-monotone in the sense that if the differences $\overline{x}_{1\cdot k} - \overline{x}_{2\cdot k}$ become more negative, the test statistic can get larger, thus increasing the chance of rejecting H_0. These anomalies result when the endpoints are highly positively correlated.

Tang (1994) gave an almost unbiased and uniformly more powerful test than Perlman's test. Wang and McDermott (1998) solved the problem of nuisance parameter $\boldsymbol{\Sigma}$ by deriving a LR test conditional on the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$. Sen and Tsai (1999) gave a

Stein-type two-stage test that is free of $\boldsymbol{\Sigma}$. It also has other desirable properties such as unbiasedness and monotonicity.

Perlman and Wu (1999) vigorously defended LR tests, noting that the alternative tests (e.g. those proposed by Berger 1989, Tang 1994 and Wang and McDermott 1998) that are less biased and more powerful also suffer from lack of monotonicity and nonintuitive rejection regions. Cohen and Sackrowitz (1998) suggested cone ordered monotone tests to ameliorate these difficulties. However, their rejection regions are not entirely satisfactory either since, e.g., in the bivariate case, their test can reject H_0 if one difference is highly negative and the other difference is highly positive. Thus the problem of appropriate one-sided tests in multiparameter situations remains not fully resolved.

3.1.2 Approximate Likelihood Ratio (ALR) Tests

To obviate the computational and analytical difficulties of Kudô's exact LR test, Tang, Gnecco and Geller (1989) proposed an approximate likelihood ratio (ALR) test for known $\boldsymbol{\Sigma}$. As extended to the two-sample problem, their test is as follows: First compute the transformation

$$\boldsymbol{u} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\boldsymbol{A}(\overline{\boldsymbol{x}}_{1\cdot} - \overline{\boldsymbol{x}}_{2\cdot}), \tag{3.1}$$

where $\boldsymbol{A}$ is any positive definite symmetric matrix such that $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}' = \boldsymbol{I}$. Then $\boldsymbol{u} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$, where $\boldsymbol{\theta} = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\boldsymbol{A}\boldsymbol{\delta}$ and the hypotheses (2.1) become

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\theta} \in \boldsymbol{A}(\boldsymbol{\delta}),$$

where $\boldsymbol{A}(\boldsymbol{\delta})$ is the polyhedral cone:

$$\boldsymbol{A}(\boldsymbol{\delta}) = \left\{ \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\boldsymbol{A}\boldsymbol{\delta} \mid \boldsymbol{\delta} \in \mathcal{O}^+ \right\}.$$

The matrix $\boldsymbol{A}$ used in the transformation is not unique. Tang et al. gave a method using the Cholesky decomposition for choosing $\boldsymbol{A}$ such that the center direction of $\boldsymbol{A}(\boldsymbol{\delta})$ coincides with the center direction of $\mathcal{O}^+$. An alternative method is the left-root symmetric method of Läuter, Kropf and Glimm (1998) which is both scale and order invariant. After choosing $\boldsymbol{A}$, the cone alternative $\boldsymbol{A}(\boldsymbol{\delta})$ is approximated by $\mathcal{O}^+$. Then the ALR statistic equals

$$g(\boldsymbol{u}) = \sum_{k=1}^{m} \{\max(u_k, 0)\}^2 . \tag{3.2}$$

The null distribution of $g(\boldsymbol{u})$ is the $\overline{\chi}^2$ distribution with symmetric binomial probability weights given by

$$\Pr_{H_0}\{g(\boldsymbol{u}) > c\} = \sum_{k=0}^{m} \left\{ \binom{m}{k} 2^{-m} \Pr\left(\chi_k^2 > c\right) \right\}. \tag{3.3}$$

If $\boldsymbol{\Sigma}$ is unknown and an estimate $\widehat{\boldsymbol{\Sigma}}$ is used in its place for computing $\boldsymbol{u}$ using (3.1) then the above $\overline{\chi}^2$ distribution provides too liberal an approximation to the exact null distribution of $g(\boldsymbol{u})$ (Reitmeir and Wassmer 1996). The liberalism is very high when the error d.f. ν are small in relation to m. For example, for $m = 6$, the estimated type I error rate for a nominal 0.05-level test is 0.3550 for $\nu = 10$, 0.1066 for $\nu = 30$ and 0.0830 for $\nu = 50$. To overcome this problem, Tamhane and Logan (2001) proposed the following approximation:

$$\Pr_{H_0}\{g(\boldsymbol{u}) > c\} \approx \sum_{k=0}^{m} \binom{m}{k} 2^{-m} \Pr\left\{ \left(\frac{\nu k}{\nu - m + 1} \right) F_{k,\nu-m+1} > c \right\}, \tag{3.4}$$

where $F_{0,\nu-m+1} = 0$. We refer to this as the $\overline{F}$- approximation. This approximation matches the first moment of $g(\boldsymbol{u})$ exactly and the second moment approximately. Based on simulations, it was shown to be extremely accurate even for small ν. For example, for $m = 6$, the estimated type I error rate for a nominal 0.05-level test using this approximation is 0.0464 for $\nu = 10$, 0.0476 for $\nu = 30$ and 0.0510 for $\nu = 50$.

The ALR test, although easier to apply, suffers from the same anomalies that the exact LR tests suffer. Therefore our recommendation is to use these global tests with caution. If several endpoints show moderate negative differences or if even a few show very large negative differences, then these tests should not be used because the *a priori* assumption of positive treatment effects in all endpoints is questionable.

3.1.3 Ordinary Least Squares (OLS) and Generalized Least Squares (GLS) Tests

O'Brien (1984) chose to bypass the analytical and computational difficulties of the LR tests by restricting the mean difference vector $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ to a ray: $\lambda(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{mm}})'$ where $\lambda \geq 0$. Specifically, if $\delta_k/\sqrt{\sigma_{kk}} = \lambda_k$ denotes the standardized treatment effect for the kth endpoint then he assumed that $\lambda_k = \lambda \geq 0$ for all k. In that case the hypothesis testing problem (2.1) reduces to

$$H_0 : \lambda = 0 \text{ vs. } H_1 : \lambda > 0. \tag{3.5}$$

This problem can be solved by using a univariate regression framework that models the standardized responses as

$$y_{ijk} = \frac{x_{ijk}}{\sqrt{\sigma_{kk}}} = \frac{\mu_k}{\sqrt{\sigma_{kk}}} + \frac{\lambda}{2} I_{ijk} + \epsilon_{ijk} \;\; (i = 1, 2; 1 \le j \le n_i; 1 \le k \le m), \tag{3.6}$$

where $\mu_k = (\mu_{1k} + \mu_{2k})/2, I_{ijk} = +1$ if $i = 1$ and -1 if $i = 2$, and $\epsilon_{ijk} \sim N(0, 1)$ r.v.'s with correlations

$$\text{Corr}(\epsilon_{ijk}, \epsilon_{i'j'\ell}) = \rho_{k\ell} \text{ if } i = i' \text{ and } j = j', \text{Corr}(\epsilon_{ijk}, \epsilon_{i'j'\ell}) = 0 \text{ otherwise.}$$

Note that the vectors $\boldsymbol{y}_{ij} = (y_{ij1}, y_{ij2}, \ldots, y_{ijm})'$ are independent, each with covariance (correlation) matrix $\boldsymbol{R} = \{\rho_{k\ell}\}$. Initially assume that $\boldsymbol{R}$ is known.

The OLS estimate of λ and its standard deviation (SD) equal

$$\widehat{\lambda}_{\text{OLS}} = \frac{\boldsymbol{j}'(\overline{\boldsymbol{y}}_{1\cdot} - \overline{\boldsymbol{y}}_{2\cdot})}{m} = \overline{y}_{1\cdot\cdot} - \overline{y}_{2\cdot\cdot} \text{ and } \text{SD}(\widehat{\lambda}_{\text{OLS}}) = \frac{1}{m}\sqrt{\left(\frac{n_1 + n_2}{n_1 n_2}\right)(\boldsymbol{j}'\boldsymbol{R}\boldsymbol{j})},$$

where $\boldsymbol{j}$ is a vector of all 1's of an appropriate dimension. Therefore the OLS statistic with $\boldsymbol{R}$ replaced by the sample correlation matrix $\widehat{\boldsymbol{R}}$ equals

$$t_{\text{OLS}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left[\frac{\boldsymbol{j}'(\overline{\boldsymbol{y}}_{1\cdot} - \overline{\boldsymbol{y}}_{2\cdot})}{\sqrt{\boldsymbol{j}'\widehat{\boldsymbol{R}}\boldsymbol{j}}}\right] = \frac{\boldsymbol{j}'\boldsymbol{t}}{\sqrt{\boldsymbol{j}'\widehat{\boldsymbol{R}}\boldsymbol{j}}}, \tag{3.7}$$

where $\boldsymbol{t}$ is the vector of t-statistics

$$t_k = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\overline{x}_{1\cdot k} - \overline{x}_{2\cdot k}}{\sqrt{\widehat{\sigma}_{kk}}}\right) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\overline{y}_{1\cdot k} - \overline{y}_{2\cdot k}) \;\; (1 \le k \le m) \tag{3.8}$$

for comparing the two treatment groups on the individual endpoints. Each t_k is marginally t-distributed under H_{0k} with $n_1 + n_2 - 2$ d.f.

Since the errors ϵ_{ijk} in the regression model (3.6) are not independent, the generalized least squares (GLS) estimate of λ, which is also its maximum likelihood estimate (MLE), may be preferred. The corresponding test is the Neyman-Pearson likelihood ratio test. Assuming that $\boldsymbol{\Sigma}$ is known, O'Brien (1984) showed that

$$\widehat{\lambda}_{\text{GLS}} = \frac{\boldsymbol{j}'\boldsymbol{R}^{-1}(\overline{\boldsymbol{y}}_{1\cdot} - \overline{\boldsymbol{y}}_{2\cdot})}{\boldsymbol{j}'\boldsymbol{R}^{-1}\boldsymbol{j}} \text{ and } \text{SD}(\widehat{\lambda}_{\text{GLS}}) = \sqrt{\left(\frac{n_1 + n_2}{n_1 n_2}\right)\left(\frac{1}{\boldsymbol{j}'\boldsymbol{R}^{-1}\boldsymbol{j}}\right)}.$$

The test statistic using this GLS estimate with the estimated correlation matrix $\widehat{\boldsymbol{R}}$ substituted in place of $\boldsymbol{R}$ equals

$$t_{\text{GLS}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\boldsymbol{j}'\widehat{\boldsymbol{R}}^{-1}(\overline{\boldsymbol{y}}_{1\cdot} - \overline{\boldsymbol{y}}_{2\cdot})}{\sqrt{\boldsymbol{j}'\widehat{\boldsymbol{R}}^{-1}\boldsymbol{j}}}\right) = \frac{\boldsymbol{j}'\widehat{\boldsymbol{R}}^{-1}\boldsymbol{t}}{\sqrt{\boldsymbol{j}'\widehat{\boldsymbol{R}}^{-1}\boldsymbol{j}}}. \tag{3.9}$$

We see that both the OLS and GLS statistics are standardized weighted sums of the individual t-statistics for the m endpoints. The OLS statistic uses equal weights, while the GLS statistic uses unequal weights determined by the sample correlation matrix $\widehat{\boldsymbol{R}}$. If some endpoint is highly correlated with the others then the GLS statistic gives a correspondingly lower weight to its t-statistic.

The exact small sample null distributions of t_{OLS} and t_{GLS} are not known. O'Brien (1984) proposed a t-distribution with $n_1 + n_2 - 2m$ d.f. as an approximation. For large sample sizes the standard normal (z) distribution may be used as an approximation. The t-approximation is exact for $m = 1$, but is conservative for $m > 1$; on the other hand, the z-approximation is liberal. The convergence of t_{GLS} to the standard normal distribution is slower than that of t_{OLS} because of the use of the estimated correlation matrix $\widehat{\boldsymbol{R}}$ both in the calculation of $\widehat{\lambda}_{\text{GLS}}$ and in the estimate of $\text{SD}(\widehat{\lambda}_{\text{GLS}})$. Also, the simulation study by Reitmeir and and Wassmer (1996) has shown that the powers of the OLS and GLS tests are comparable when used to test subset hypotheses in closed testing procedures (see Section 4.1). Finally, the linear combination $\boldsymbol{j}'\widehat{\boldsymbol{R}}^{-1}$ used by the GLS test can have some negative weights, which can lead to anomolous results; this problem does not occur with the OLS test. For all these reasons, the OLS test is recommended.

Finally we note that Tang, Gnecco and Pocock (1993) have generalized the GLS test statistic for an arbitrary ray alternative $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \lambda(\beta_1, \ldots, \beta_m)'$, where the vector $(\beta_1, \ldots, \beta_m)'$ with all positive elements is specified. However, if the observed mean difference $\overline{\boldsymbol{x}}_{1\cdot} - \overline{\boldsymbol{x}}_{2\cdot}$ is not close to this ray then the power of the test may be adversely affected. Since the vector $(\beta_1, \ldots, \beta_m)'$ is in general difficult to specify, Tang, Gnecco and Pocock suggest following the maxmin approach (maximize the minimum power over all ray alternatives) of Abelson and Tukey (1963).

3.1.4 Läuter's Exact Tests

Läuter (1996) proposed a class of test statistics for the hypotheses (2.1) having the property that they are exactly t-distributed with $n_1 + n_2 - 2$ d.f. under H_0. Recall that $\overline{\boldsymbol{x}}_{i\cdot} = (\overline{x}_{i\cdot 1}, \overline{x}_{i\cdot 2}, \ldots, \overline{x}_{i\cdot m})'$ denotes the vector of sample means for the ith group ($i = 1, 2$) and

let

$$\overline{\boldsymbol{x}}_{\cdot\cdot} = \frac{n_1\overline{\boldsymbol{x}}_{1\cdot} + n_2\overline{\boldsymbol{x}}_{2\cdot}}{n_1+n_2} = (\overline{x}_{\cdot\cdot 1}, \overline{x}_{\cdot\cdot 2}, \ldots, \overline{x}_{\cdot\cdot m})'$$

denote the vector of overall sample means. Define the total cross-products matrix by

$$\boldsymbol{V} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{\cdot\cdot})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{\cdot\cdot})' = (n_1+n_2-2)\widehat{\boldsymbol{\Sigma}} + \sum_{i=1}^{2} n_i(\overline{\boldsymbol{x}}_{i\cdot} - \overline{\boldsymbol{x}}_{\cdot\cdot})(\overline{\boldsymbol{x}}_{i\cdot} - \overline{\boldsymbol{x}}_{\cdot\cdot})'.$$

Let $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{V})$ be any m-dimensional vector of weights depending solely on $\boldsymbol{V}$ and $\boldsymbol{w} \neq \boldsymbol{0}$ with probability 1. Using the results from the theory of spherical distributions (Fang and Zhang 1990), Läuter (1996) showed that

$$t_{\boldsymbol{w}} = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\left(\frac{\boldsymbol{w}'\boldsymbol{t}}{\sqrt{\boldsymbol{w}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{w}}}\right)$$

is t-distributed with $n_1 + n_2 - 2$ d.f. under H_0. Various choices for $\boldsymbol{w}$ were discussed by Läuter, Kropf and Glimm (1998). We will focus on the standardized sum (SS) statistic (denoted by t_{SS}) for which $\boldsymbol{w}$ equals $(1/\sqrt{v_{11}}, 1/\sqrt{v_{22}}, \ldots, 1/\sqrt{v_{mm}})'$, where

$$v_{kk} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ijk} - \overline{x}_{\cdot\cdot k})^2$$

is the kth diagonal element of $\boldsymbol{V}$. The SS test statistic can be expressed as the t-statistic:

$$t_{\text{SS}} = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\left(\frac{\overline{y}_{1\cdot} - \overline{y}_{2\cdot}}{\widehat{\sigma}_y}\right),$$

calculated on the sum of the standardized observations for each patient:

$$y_{ij} = \sum_{k=1}^{m}\frac{x_{ijk}}{\sqrt{v_{kk}}} \quad (i = 1, 2; 1 \leq j \leq n_i),$$

where

$$\overline{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad (i = 1, 2) \text{ and } \widehat{\sigma}_y = \sqrt{\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot})^2}{n_1+n_2-2}}.$$

3.1.5 Asymptotic Power Comparison of O'Brien's OLS and Läuter's SS Tests

The OLS statistic is the sum of the t_k-statistics (3.8), which are obtained by standardizing the individual endpoints by their pooled *within* group sample standard deviations. On the other hand, the SS statistic is obtained by standardizing the data on each endpoint by its pooled *total* group sample standard deviation and then computing an overall t-statistic.

Because the total pooled standard deviation overestimates the true standard deviation since it includes the between treatment group difference, the power of the SS test would be expected to be adversely affected. In this section we compare the powers of the two tests in the asymptotic case where $n_1 = n_2 = n$ (say) and $n \to \infty$.

The limiting null and non-null distributions of t_{OLS} and t_{SS} are normal, and their powers for α-level tests can be expressed as follows (for derivations, see Logan 2001):

$$\text{Power}_{\text{OLS}} = \Phi\left(-z_\alpha + \frac{\boldsymbol{a}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}}\sqrt{\frac{n}{2}}\right)$$

and

$$\text{Power}_{\text{SS}} = \Phi\left(-z_\alpha + \frac{\boldsymbol{b}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b}}}\sqrt{\frac{n}{2}}\right),$$

where z_α is the $(1-\alpha)$th quantile of the standard normal distribution, $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)'$, $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)'$, and

$$a_k = \frac{1}{\sigma_k\sqrt{2}} \text{ and } b_k = \frac{1}{\sigma_k\sqrt{2 + \lambda_k^2/2}} \quad (1 \le k \le m).$$

Therefore

$$\text{Power}_{\text{OLS}} \ge \text{Power}_{\text{SS}} \iff \frac{\boldsymbol{a}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}} \ge \frac{\boldsymbol{b}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b}}}. \tag{3.10}$$

It is easy to show that

$$\frac{\boldsymbol{a}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}} = \frac{\sum_{k=1}^m \lambda_k}{\sqrt{\sum_{k=1}^m \sum_{\ell=1}^m \rho_{k\ell}}} \text{ and } \frac{\boldsymbol{b}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b}}} = \frac{\sum_{k=1}^m \lambda_k/\sqrt{1 + \lambda_k^2/4}}{\sqrt{\sum_{k=1}^m \sum_{\ell=1}^m \rho_{k\ell}/\sqrt{(1 + \lambda_k^2/4)(1 + \lambda_\ell^2/4)}}},$$

where $\rho_{k\ell} = 1$ if $k = \ell$. Comparison of the powers of the two tests reduces to comparison of the two expressions above.

Consider the case $\lambda_1 > 0$ and $\lambda_k = 0$ for $k > 1$. Then we have

$$\frac{\boldsymbol{a}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}} = \frac{\lambda_1}{\sqrt{\sum_{k=1}^m \sum_{\ell=1}^m \rho_{k\ell}}}$$

and

$$\frac{\boldsymbol{b}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b}}} = \frac{\lambda_1/\sqrt{1 + \lambda_1^2/4}}{\sqrt{\sum_{k=2}^m \sum_{\ell=2}^m \rho_{k\ell} + 2\sum_{k=2}^m \left(\rho_{1k}/\sqrt{1 + \lambda_1^2/4}\right) + 1/(1 + \lambda_1^2/4)}}.$$

It is simple algebra to show that the inequality (3.10) is strict in this case. Thus, if only one endpoint has a positive treatment effect then the OLS test is more powerful to detect this effect than the SS test. In fact,

$$\lim_{\lambda_1 \to \infty} \frac{\lambda_1/\sqrt{1+\lambda_1^2/4}}{\sqrt{\sum_{k=2}^m \sum_{\ell=2}^m \rho_{k\ell} + 2\sum_{k=2}^m \left(\rho_{1k}/\sqrt{1+\lambda_1^2/4}\right) + 1/(1+\lambda_1^2/4)}} = \frac{2}{\sqrt{\sum_{k=2}^m \sum_{\ell=2}^m \rho_{k\ell}}} < \infty,$$

and therefore the power of the SS test is bounded away from 1 when $\lambda_1 \to \infty$. This undesirable property of the SS test was shown by Frick (1996).

Next consider the case $\lambda_k = \lambda > 0$ for all k, which is the assumption underlying the OLS test. In this case we have

$$\frac{\boldsymbol{a}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}} = \frac{\boldsymbol{b}'\boldsymbol{\delta}}{\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b}}} = \frac{m\lambda}{\sqrt{\sum_{k=1}^m \sum_{\ell=1}^m \rho_{k\ell}}},$$

and therefore $\text{Power}_{\text{OLS}} = \text{Power}_{\text{SS}}$. It is interesting to note that in this case the OLS test has the highest power (among all comparable configurations in the positive orthant). On the other hand, in the previous case, when in a single endpoint has a treatment effect, the OLS test has the least power (again, among all comparable configurations in the positive orthant) and the SS test has even lower power. We conjecture that the OLS test is asymptotically at least as powerful as the SS test under all configurations, but we do not have a proof of this conjecture.

3.1.6 Follman's x_+^2 Test

Follman (1996) proposed an ad-hoc test, which relates to a slightly different alternative hypothesis

$$H_1 : \sum_{k=1}^m (\mu_{1k} - \mu_{2k}) > 0.$$

Unfortunately this alternative hypothesis is not very meaningful since it is dependent on the scaling used for the endpoints. His test is simple to apply: reject the global null hypothesis if Hotelling's T^2 test is significant at level 2α and $\sum_{k=1}^m (\overline{x}_{1\cdot k} - \overline{x}_{2\cdot k}) > 0$. Thus we reject the null hypothesis in favor of a one-sided alternative if we reject a two-sided test at level 2α and the average endpoint mean difference is > 0. Note that this test modifies the rejection region

of Hotelling's T^2 test by eliminating outcomes with negative differences on all endpoints, but it's rejection region is not monotone.

3.2 Heteroscedastic Case

3.2.1 Approximate Likelihood Ratio (ALR) Test

In Tamhane and Logan (2001) we proposed to extend the ALR test to the heteroscedastic case as follows. Let

$$\boldsymbol{\Omega}_i = \frac{1}{n_i}\boldsymbol{\Sigma}_i \;\; (i = 1, 2) \boldsymbol{\Omega} = \boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2 \text{ and } \boldsymbol{\Sigma} = \frac{n_1 n_2}{n_1 + n_2}\boldsymbol{\Omega}.$$

The sample estimates of these matrices are denoted by putting carets over them; thus $\widehat{\boldsymbol{\Omega}}_i = (1/n_i)\widehat{\boldsymbol{\Sigma}}_i, \widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}_1 + \widehat{\boldsymbol{\Omega}}_2$ and

$$\widehat{\boldsymbol{\Sigma}} = \frac{n_1 n_2}{n_1 + n_2}\widehat{\boldsymbol{\Omega}}.$$

The transformation matrix $\boldsymbol{A}$ in (3.1) is chosen such that $\boldsymbol{A}'\boldsymbol{A} = \widehat{\boldsymbol{\Sigma}}^{-1}$ and $\boldsymbol{A}\widehat{\boldsymbol{\Sigma}}\boldsymbol{A}' = \boldsymbol{I}$.

We suggested the same $\overline{F}$ approximation (3.4) to the null distribution of $g(\boldsymbol{u})$ in the heteroscedastic case, but with the following Welch-Satterthwaite estimated d.f. ν derived by Yao (1965) for the multivariate Behrens-Fisher problem:

$$\frac{1}{\nu} = \frac{1}{(\boldsymbol{d}'\widehat{\boldsymbol{\Omega}}^{-1}\boldsymbol{d})^2}\left[\frac{(\boldsymbol{d}'\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Omega}}_1\widehat{\boldsymbol{\Omega}}^{-1}\boldsymbol{d})^2}{n_1 - 1} + \frac{(\boldsymbol{d}'\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{\Omega}}_2\widehat{\boldsymbol{\Omega}}^{-1}\boldsymbol{d})^2}{n_2 - 1}\right],$$

where $\boldsymbol{d} = (\overline{\boldsymbol{x}}_{1\cdot} - \overline{\boldsymbol{x}}_{2\cdot})$. Note that Yao derived this formula (also using the moment matching method) to approximate the distribution of

$$\boldsymbol{u}'\boldsymbol{u} = \left(\frac{n_1 n_2}{n_1 + n_2}\right)(\overline{\boldsymbol{x}}_{1\cdot} - \overline{\boldsymbol{x}}_{2\cdot})'\widehat{\boldsymbol{\Sigma}}^{-1}(\overline{\boldsymbol{x}}_{1\cdot} - \overline{\boldsymbol{x}}_{2\cdot})$$

by Hotelling's $T^2_{m,\nu} = \left(\frac{\nu m}{\nu - m + 1}\right) F_{m,\nu-m+1}$ distribution with an estimated ν. We simply extended Yao's approximation to the $\overline{F}$ distribution. Our simulations for selected values of $m, n_1 = n_2 = n, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ showed that this approximation is quite accurate for controlling the type I error probability at the nominal level $\alpha = 0.05$ for $m = 4$ if $n \geq 20$ and for $m = 8$ if $n \geq 30$.

3.2.2 Ordinary Least Squares (OLS) and Generalized Least Squares (GLS) Tests

Pocock, Geller and Tsiatis (1987) extended O'Brien's GLS test to the heteroscedastic case as follows. Initially assume that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are known. Then the statistic for comparing the treatment with the control on the kth endpoint is

$$z_k = \frac{\overline{x}_{1\cdot k} - \overline{x}_{2\cdot k}}{\sqrt{\sigma_{1,kk}/n_1 + \sigma_{2,kk}/n_2}} \quad (1 \le k \le m). \tag{3.11}$$

Let $\boldsymbol{z} = (z_1, z_2, \ldots, z_m)'$ and $\bar{\boldsymbol{R}} = (n_1\boldsymbol{R}_1 + n_2\boldsymbol{R}_2)/(n_1 + n_2)$. In analogy with (3.9), Pocock et al. proposed the statistic

$$z_{\mathrm{GLS}} = \frac{\boldsymbol{j}'\bar{\boldsymbol{R}}^{-1}\boldsymbol{z}}{\sqrt{\boldsymbol{j}'\bar{\boldsymbol{R}}^{-1}\boldsymbol{j}}}.$$

However, this is just an ad-hoc extension. Furthermore, the covariance (correlation) matrix of $\boldsymbol{z}$ is not $\bar{\boldsymbol{R}}$, but $\boldsymbol{\Gamma} = \{\gamma_{k\ell}\}$ with elements

$$\gamma_{k\ell} = \frac{\sigma_{1,k\ell}/n_1 + \sigma_{2,k\ell}/n_2}{\sqrt{(\sigma_{1,kk}/n_1 + \sigma_{2,kk}/n_2)(\sigma_{1,\ell\ell}/n_1 + \sigma_{2,\ell\ell}/n_2)}} \quad (1 \le k < \ell \le m).$$

As a result, z_{GLS} as defined by Pocock et al. does not have the standard normal distribution under H_0. In the following we correctly derive the OLS and GLS statistics.

We use the following definition for the standardized treatment effect in the heteroscedastic case.:

$$\lambda_k = \frac{\delta_k}{\sqrt{\sigma_{1,kk} + \sigma_{2,kk}}} \quad (1 \le k \le m).$$

As in O'Brien (1984), assume that $\lambda_k = \lambda \ge 0$ for all k. To test the hypotheses (3.5), standardize the observations as

$$y_{ijk} = \frac{x_{ijk}}{\sqrt{\sigma_{1,kk} + \sigma_{2,kk}}} \quad (i = 1, 2; 1 \le j \le n_i; 1 \le k \le m).$$

Then $\boldsymbol{y}_{ij} = (y_{ij1}, y_{ij2}, \ldots, y_{ijm})'$ are independently distributed as $N(\boldsymbol{\xi}_i, \boldsymbol{\Gamma}_i)$, where $\boldsymbol{\xi}_i$ has elements

$$\xi_{ik} = \frac{\mu_{ik}}{\sqrt{\sigma_{1,kk} + \sigma_{2,kk}}} \quad (1 \le k \le m)$$

and $\boldsymbol{\Gamma}_i$ has elements

$$\gamma_{i,k\ell} = \frac{\sigma_{i,k\ell}}{\sqrt{(\sigma_{1,kk} + \sigma_{2,kk})(\sigma_{1,\ell\ell} + \sigma_{2,\ell\ell})}} \quad (1 \le k \le \ell \le m)$$

for $i = 1, 2$. Note that $\xi_{1k} - \xi_{2k} = \lambda$ for all k. Also note that $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are not correlation matrices, and $\mathbf{\Gamma} = \mathbf{\Gamma}_1 + \mathbf{\Gamma}_2$ if $n_1 = n_2$.

The hypotheses (3.5) can be tested by using a univariate regression framework as in (3.6):

$$y_{ijk} = \xi_k + \frac{\lambda}{2} I_{ijk} + \epsilon_{ijk} \quad (i = 1, 2; 1 \le j \le n_i; 1 \le k \le m), \tag{3.12}$$

where $\xi_k = (\xi_{1k} + \xi_{2k})/2$, $I_{ijk} = +1$ if $i = 1$ and -1 if $i = 2$, and $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \epsilon_{ij2}, \ldots, \epsilon_{ijm})'$ are independently distributed as $N(\mathbf{0}, \mathbf{\Gamma}_i)$. Using the same methods as those used in the homoscedastic case, the OLS and GLS statistics are as given below; for derivations, see Logan (2001).

Assuming that $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are known, it is straightforward to show that

$$\widehat{\lambda}_{\text{OLS}} = \frac{\boldsymbol{j}'(\overline{\boldsymbol{y}}_{1\cdot} - \overline{\boldsymbol{y}}_{2\cdot})}{m} = \overline{y}_{1\cdot\cdot} - \overline{y}_{2\cdot\cdot} \text{ and } \text{SD}(\widehat{\lambda}_{\text{OLS}}) = \sqrt{\boldsymbol{j}' \left(\frac{\mathbf{\Gamma}_1}{n_1} + \frac{\mathbf{\Gamma}_2}{n_2} \right) \boldsymbol{j}}.$$

Hence the OLS statistic with the $\mathbf{\Gamma}_i$ replaced by their sample estimates $\widehat{\mathbf{\Gamma}}_i$ equals

$$t_{\text{OLS}} = \frac{\boldsymbol{j}'(\overline{\boldsymbol{y}}_{1\cdot} - \overline{\boldsymbol{y}}_{2\cdot})}{\sqrt{\boldsymbol{j}'(\widehat{\mathbf{\Gamma}}_1/n_1 + \widehat{\mathbf{\Gamma}}_2/n_2)\boldsymbol{j}}}, \tag{3.13}$$

where the elements of $\widehat{\mathbf{\Gamma}}_i$ are given by

$$\widehat{\gamma}_{i,k\ell} = \frac{\widehat{\sigma}_{i,k\ell}}{\sqrt{(\widehat{\sigma}_{1,kk} + \widehat{\sigma}_{2,kk})(\widehat{\sigma}_{1,\ell\ell} + \widehat{\sigma}_{2,\ell\ell})}}.$$

For $n_1 = n_2 = n$, the above OLS statistic reduces to

$$t_{\text{OLS}} = \frac{\boldsymbol{j}'\boldsymbol{t}}{\sqrt{\boldsymbol{j}'\widehat{\mathbf{\Gamma}}\boldsymbol{j}}},$$

where $\boldsymbol{t}$ is a vector of t-statistics

$$t_k = \frac{(\overline{x}_{1\cdot k} - \overline{x}_{2\cdot k})}{\sqrt{\widehat{\sigma}_{1,kk}/n_1 + \widehat{\sigma}_{2,kk}/n_2}} \quad (1 \le k \le m) \tag{3.14}$$

for comparing the two treatment groups on the individual endpoints. These statistics are marginally approximately t-distributed under H_{0k} with d.f. estimated by the Welch-Satterthwaite formula:

$$\nu_k = \frac{(\widehat{\sigma}_{1,kk}/n_1 + \widehat{\sigma}_{2,kk}/n_2)^2}{\widehat{\sigma}^2_{1,kk}/n_1^2(n_1 - 1) + \widehat{\sigma}^2_{2,kk}/n_2^2(n_2 - 1)} \quad (1 \le k \le m).$$

Next we derive the GLS test. Assuming that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are known, it can be shown that

$$\widehat{\lambda}_{\text{GLS}} = \frac{4\boldsymbol{j}'\left(\boldsymbol{\Gamma}_1/n_1 + \boldsymbol{\Gamma}_2/n_2\right)^{-1}(\overline{\boldsymbol{y}}_{1\cdot} - \overline{\boldsymbol{y}}_{2\cdot})}{\boldsymbol{j}'[(\boldsymbol{I}-\boldsymbol{B})\boldsymbol{\Gamma}_1^{-1}/n_1 + (\boldsymbol{I}+\boldsymbol{B})\boldsymbol{\Gamma}_2^{-1}/n_2]\boldsymbol{j}}$$

and

$$\text{SD}(\widehat{\lambda}_{\text{GLS}}) = \frac{4\sqrt{\boldsymbol{j}'\left(\boldsymbol{\Gamma}_1/n_1 + \boldsymbol{\Gamma}_2/n_2\right)^{-1}\boldsymbol{j}}}{\boldsymbol{j}'[(\boldsymbol{I}-\boldsymbol{B})\boldsymbol{\Gamma}_1^{-1}/n_1 + (\boldsymbol{I}+\boldsymbol{B})\boldsymbol{\Gamma}_2^{-1}/n_2]\boldsymbol{j}},$$

where

$$\boldsymbol{B} = (n_1\boldsymbol{\Gamma}_1^{-1} - n_2\boldsymbol{\Gamma}_2^{-1})(n_1\boldsymbol{\Gamma}_1^{-1} + n_2\boldsymbol{\Gamma}_2^{-1})^{-1}.$$

Hence the GLS statistic with the $\boldsymbol{\Gamma}_i$ replaced by their sample estimates $\widehat{\boldsymbol{\Gamma}}_i$ equals

$$t_{\text{GLS}} = \frac{\boldsymbol{j}'(\widehat{\boldsymbol{\Gamma}}_1/n_1 + \widehat{\boldsymbol{\Gamma}}_2/n_2)^{-1}(\overline{\boldsymbol{y}}_{1\cdot} - \overline{\boldsymbol{y}}_{2\cdot})}{\sqrt{\boldsymbol{j}'\left(\widehat{\boldsymbol{\Gamma}}_1/n_1 + \widehat{\boldsymbol{\Gamma}}_2/n_2\right)^{-1}\boldsymbol{j}}}. \tag{3.15}$$

For $n_1 = n_2 = n$, this reduces to

$$t_{\text{GLS}} = \frac{\boldsymbol{j}'\widehat{\boldsymbol{\Gamma}}^{-1}\boldsymbol{t}}{\sqrt{\boldsymbol{j}'\widehat{\boldsymbol{\Gamma}}^{-1}\boldsymbol{j}}},$$

where the t_k in $\boldsymbol{t} = (t_1, t_2, \ldots, t_m)'$ are defined in (3.14).

3.3 p-Value Based Tests

Thus far we have assumed a multivariate normal setup. In practice, the endpoints can be quite diverse — some may be approximately normally distributed (e.g., change in tumor size), some may follow a survival distribution with possible censoring (e.g., remission time), some may be binary (e.g., death) and some may be ordinal (e.g., patient's or physician's assessment of disease condition on a five-point scale). Testing of the global null hypothesis H_0 based on such diverse metrics is facilitated by condensing the evidence of treatment efficacy on each endpoint in terms of its p-value. Denote the marginal p-value for testing the null hypothesis H_{0k} by p_k $(1 \le k \le m)$.

The simplest p-value based test is the Bonferroni test, which rejects H_0 at level α if

$$p_{\min} = \min_{1\le k\le m} p_k < \alpha/m. \tag{3.16}$$

Assuming the multivariate normal setup, this is equivalent to rejecting if the maximum t_k-statistic exceeds the upper α/m critical point of an appropriate t-distribution (exact in the homoscedastic case, approximate in the heteroscedastic case).

The problems associated with the Bonferroni test have been well-documented; see, e.g., O'Brien (1984), Pocock, Geller and Tsiatis (1987): (i) it is overly conservative especially if m is large or the endpoints are highly correlated, and (ii) it is powerful if only one endpoint has a large treatment effect, but not if most or all endpoints have moderate treatment effects.

It should be noted that the Bonferroni test is a union-intersection (UI) test when H_0 is viewed as $H_0 = \bigcap_{k=1}^{m} H_{0k}$. Therefore rejection of H_0 implies rejection of any H_{0k} with $p_k < \alpha/m$; this implied multiple test procedure for testing null hypotheses on the individual endpoints controls the FWE at level α (Hochberg and Tamhane 1987, pp. 28 -29).

An improvement on the Bonferroni test was proposed by Simes (1986). To apply the Simes test first order the p-values: $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(m)}$ and denote the corresponding hypotheses by $H_{0(1)}, H_{0(2)}, \ldots, H_{0(m)}$. Then reject H_0 if

$$p_{(k)} < \frac{(m-k+1)\alpha}{m} \text{ for some } k = 1, 2, \ldots, m. \tag{3.17}$$

Simes proved that this is an α-level test under the assumption that the p-values are independent. Sarkar and Chang (1997) showed that this result also holds for statistics having TP_2 distributions. Sarkar (1998) further extended the proof to statistics having MTP_2 and certain scale mixtures of MTP_2 distributions.

4. Endpoint-Specific Tests

4.1 Closed Tests

Kropf (1988) and Lehmacher, Wassmer and Reitmeir (1991) suggested that the closure method of Marcus, Peritz and Gabriel (1976) be used to test hypotheses (2.2) on individual endpoints. Let $M = \{1, 2, \ldots, m\}$ be the index set of all endpoints and let $K \subseteq M$ be any nonempty subset of M. Then the closure method tests and rejects an intersection hypothesis $H_{0K} = \bigcap_{k \in K}(H_{0k} : \delta_k = 0)$ at level α iff all intersection hypotheses H_{0L} for $L \supseteq K$ are tested and rejected at level α. The test procedure is applied in a step-down manner beginning with the test of the overall null hypothesis $H_0 = H_{0M}$ and testing the null hypotheses on subsets K of M of successively lower dimensions in case of rejections of the null hypotheses on the corresponding supersets. All we need to apply this procedure is an appropriate α-level

global test of each null hypothesis H_{0K} for $K \subseteq M$. Any of the global tests discussed in the previous section can be used for this purpose.

4.2 Normal Theory Based Tests

It is conceivable to test the null hypotheses H_{0k} of (2.2) on the individual endpoints using the test statistics t_k from (3.8) in the homoscedastic case and from (3.14) in the heteroscedastic case. To control the FWE at level α, we would need the upper α critical point of $\max_{1\leq k\leq m} t_k$ under the overall null hypothesis H_0 in each case. However, the joint distribution of $(t_1, t_2, \ldots, t_m)$ is not multivariate t even in the homoscedastic case because the standard deviations $\sqrt{\widehat{\sigma}_{kk}}$ used to standardize the t_k statistics are different though correlated for $k = 1, 2, \ldots, m$. Furthermore these correlations (as well as those between the numerators of the t_k statistics) are unknown being the correlations between the corresponding endpoints. Therefore the standard Dunnett-type (1955) test or its stepwise versions (Dunnett and Tamhane 1991, 1992) cannot be applied to test the hypotheses H_{0k}.

4.3 Procedures Based on Adjusted p-Values

Let p_k be the p-value for testing H_{0k} as discussed in Section 3.3 and let P_k be the corresponding r.v. This p-value is not adjusted for multiplicity of tests on all H_{0k}. A way to control the FWE at level α is to find multiplicity adjusted p-values (see Dunnett and Tamhane 1991, 1992 and Wright 1992), denoted by $\widetilde{p}_k$, and reject H_{0k} if $\widetilde{p}_k < \alpha$ $(1 \leq k \leq m)$.

The adjusted p-values corresponding to a single-step test procedure (see Hochberg and Tamhane 1987, Ch. 2) are given by

$$\widetilde{p}_k = \Pr_{H_0}\left(\min_{1\leq \ell\leq m} P_\ell \leq p_k\right) \quad (1 \leq k \leq m). \tag{4.1}$$

The joint distribution of $(P_1, P_2, \ldots, P_m)$ is unknown because of the unknown correlations among the endpoints. Therefore an approximation is often needed. The simplest such approximation is the Bonferroni adjustment (corresponding to the Bonferroni test) given by

$$\widetilde{p}_k = mp_k \quad (1 \leq k \leq m).$$

Various sharpened versions of the Bonferroni adjusted p-values are available based on the Šidák (1968) inequality and its modifications. The Šidák adjustment assumes that the P_ℓ's

are independent and is given by

$$\widetilde{p}_k = 1 - (1 - p_k)^m \ \ (1 \le k \le m).$$

If the P_ℓ's are positively dependent then this adjustment is conservative. Armitage and Parmar (1986) gave the following ad- hoc approximation to the adjusted p-values that takes into account the correlations between the endpoints:

$$\widetilde{p}_k = 1 - (1 - p_k)^{m^f} \ \ (1 \le k \le m),$$

where f is an empirically determined function of the $\rho_{k\ell}$'s. Dubey (1985) suggested using a different function $f_k = 1 - \overline{\rho}_k$ for each k, where $\overline{\rho}_k$ is the average of the correlations of the kth endpoint with the others. However, it is readily seen from the definition (4.1) of the adjusted p-value that f must be a symmetric function of all correlations. Therefore $\overline{\rho}_k$ in Dubey's formula should be replaced by $\overline{\rho}$, namely, the average of all $\rho_{k\ell}$'s. Notice that if all $\rho_{k\ell} = 0$ then we get the Šidák adjustment and if all $\rho_{k\ell} = 1$ then $\widetilde{p}_k = p_k$, i.e., there is no adjustment. Tukey, Ciminera and Heyse (1985) suggested using $f = 1/2$, i.e., $\widetilde{p}_k = 1 - (1 - p_k)^{\sqrt{m}}$, which assumes that the average correlation is $1/2$. An analytic approximation to $\widetilde{p}_k$ for jointly normally distributed endpoints was proposed by James (1991). Finally, Westfall and Young's (1989,1993) resampling method, which is distribution-free and implicitly takes the correlations between the endpoints into account can always be applied to estimate the $\widetilde{p}_k$.

For multivariate binary endpoints, a bootstrap method was given by Westfall and Young (1989) which was further extended to many other multiple testing problem in their 1993 book. Chen (1998) proposed using the generalized estimating equation (GEE) approach to estimate the unknown correlations of binary endpoints to find the adjusted p-values.

Another approach to sharpen the Bonferroni adjustment is to use a stepwise procedure for testing. The adjusted p-values for a step-down test procedure are given by

$$\begin{aligned} \widetilde{p}_{(m)} &= \Pr_{H_0}\left(\min_{1\le\ell\le m} P_\ell \le p_{(m)}\right) \text{ and} \\ \widetilde{p}_{(k)} &= \max\left[\widetilde{p}_{(k+1)}, \Pr_{H_0}\left(\min_{1\le\ell\le k} P_\ell \le p_{(k)}\right)\right] \text{ for } k = 1, \ldots, m-1. \end{aligned} \tag{4.2}$$

Conservative approximations to the above adjusted p-values can be obtained by using the Bonferroni inequality and are given by

$$\widetilde{p}_{(m)} = m p_{(m)} \text{ and } \widetilde{p}_{(k)} = \max\left[\widetilde{p}_{(k+1)}, k p_{(k)}\right] \ \ (1 \le k \le m-1).$$

These approximations correspond to Holm's (1979) step-down test procedure, which rejects $H_{0(k)}$ iff $p_{(\ell)} < \alpha/\ell$ for $\ell = k, k+1, \ldots, m$. This procedure can be derived by using the Bonferroni test (3.16) to test subset null hypotheses in the closure method.

Hommel (1988) derived a stepwise procedure by using the Simes test (3.17) to test subset null hypotheses in the closure method. Hochberg (1988) offered a slightly conservative but a much simpler procedure. It is of step-up type in that it is the exact opposite of Holm's step-down procedure in terms of sequence of testing. The adjusted p-values for the Hochberg procedure are given by

$$\widetilde{p}_{(1)} = p_{(1)} \text{ and } \widetilde{p}_{(k)} = \min\left[\widetilde{p}_{(k-1)}, kp_{(k)}\right] \quad (2 \leq k \leq m).$$

Hochberg's procedure accepts $H_{0(k)}$ iff $p_{(\ell)} \geq \alpha/\ell$ for $\ell = 1, 2, \ldots, k$. Troendle (1996) gave a bootstrap-based permutational step-up procedure.

4.4 A Hybrid Method Combining Global and Endpoint-Specific Tests

As we have seen, there are two main approaches to identify the significant endpoints: (i) adjusting the p-values of individual endpoints, and (ii) using the closure method that employs one of the global tests to test subset null hypotheses. The first approach is more powerful when only a few endpoints have positive treatment effects, while the second approach is more powerful when all or most of the endpoints have an effect. A test procedure with a more uniform power performance can be obtained by combining these two approaches along the lines of Hothorn's (1999) $T_{\max}$ testing principle.

In Logan and Tamhane (2001) we gave a closed testing procedure by combining two tests for testing each intersection hypothesis: (i) the Bonferroni $p_{\min}$ test and (ii) O'Brien's OLS test. According to this latter hybrid method, the adjusted p-value for any intersection hypothesis $H_{0K} = \bigcap_{k \in K} H_{0k}$ is defined as

$$\widetilde{p}_K = \Pr_{H_0}\left\{\min\left(\min_{k \in K} P_k, P_{K,\text{OLS}}\right) \leq \min\left(\min_{k \in K} p_k, p_{K,\text{OLS}}\right)\right\}, \tag{4.3}$$

where, as before, the lower case p's denote the unadjusted observed p-values (e.g., p_k is the p-value for H_{0k} and $p_{K,\text{OLS}}$ is the p-value for H_{0K} using the OLS test) and the upper case P's denote the corresponding r.v.'s. In Logan (2001) a third test was added, namely the ALR

test. In a closed testing procedure, a hypothesis H_{0K} is rejected at level α iff all hypotheses H_{0L} for $L \supset K$ are rejected at level α and $\widetilde{p}_K < \alpha$. In pratice, the $\widetilde{p}_K$ defined in (4.3) need to be estimated by bootstrap resampling. C language programs for this purpose for both the homoscedastic as well as the heteroscedastic case are posted on the first author's home page (`http://users.iems.northwestern.edu/~ajit`).

The simulation results given in Logan and Tamhane (2001) demonstrate that the hybrid method is quite powerful for detecting individual endpoint treatment effects and more robust to the configuration of the mean differences than both the Bonferroni $p_{\min}$ test and the OLS test. It is found to be 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 related, yet different aspects of disease recovery. In this sense very highly correlated endpoints are less informative. As a result, typical correlations range between 0.2 to 0.6, rarely exceeding 0.7 or 0.8. For such settings the hybrid method offers worthwhile power gains.

4.5 Decision Rules Based on Endpoint-Specific Tests

Inferences on individual endpoints may be mainly of scientific interest or means to arrive at a decision on the efficacy of the treatment for regulatory approval purposes. A simple decision rule is to conclude that the treatment is effective if at least r of the m endpoints show a significant improvement, where r $(1 \leq r \leq m)$ is a prespecified integer. For $r = 1$ we have a union-intersection (UI) testing problem. The Bonferroni test (3.16) offers a conservative solution to this problem; a more accurate UI test (e.g., using resampling) can be used instead. If $r = m$ then we have an intersection-union (IU) testing problem (Berger 1982):

$$H_0 : \bigcup_{k=1}^{m} (\delta_k \leq 0) \text{ vs. } H_1 : \bigcap_{k=1}^{m} (\delta_k > 0).$$

An IU test for this problem is Laska and Meisner's (1989) MIN test, which rejects H_0 if all t_k are significant at level α. Note that the null hypothesis for this test is taken as the full complement of the positive orthant. Over this null hypothesis, the least favorable (LF)

configurations at which the type I error probability is maximized ($= \alpha$) can be shown to be of the type $\delta_k = 0$ for some k and $\delta_\ell \to \infty$ for $\ell \neq k$. Cappizi and Zhang (1996) argued that the resulting MIN test is overly conservative. If the null hypothesis is restricted to $H_0 : \bigcap_{k=1}^{m}(\delta_k = 0)$ as in (2.1) then a much less conservative test is obtained. Snappin (1987) proposed a less conservative MIN-type test that uses the estimated mean differences in place of the above LF configurations. Hochberg and Mosier (2001) assumed no treatment by effect interaction. This restriction implies that H_0 is a partial complement:

$$H_0 : \bigcap_{k=1}^{m}(\delta_k \leq 0).$$

In this case the LF configuration is $\delta_1 = \cdots = \delta_m = 0$, which results in a more powerful IU test.

The above tests can be generalized for $1 \leq r \leq m$ by using the test statistic $t_{(m-r+1)}$; if $t_{(m-r+1)} > c_{m,r,\alpha}$ then the hypotheses $H_{(m-r+1)}, \ldots, H_{(m)}$ are rejected and so at least r out of the m endpoints are shown to be significant at level α. The critical constants $c_{m,r,\alpha}$ of the null distribution of $t_{(m-r+1)}$ when $t_1, t_2, \ldots, t_m$ have a multivariate t-distribution with a common known correlation ρ have been tabulated by Tamhane, Liu and Dunnett (1988) for selected values of m, r, α, ρ and error d.f. ν. However, those tables are not applicable in the present problem because the statistics t_k of (3.8) do not have a multivariate t-distribution; furthermore, the correlations between them are unknown and unequal. This testing problem remains unsolved.

An alternative rule is to declare that the treatment is effective if at least $m_1 < m$ endpoints are significant at level α_1 and the remaining $m_2 = m - m_1$ endpoints are significant at level $\alpha_2 > \alpha_1$, the idea being to show that these latter endpoints tend in the positive direction. Cappizi and Zhang (1996) suggested this rule when $m = 2, m_1 = m_2 = 1$ and $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$ or 0.20. However, as noted by Neuhäuser, Steinijans and Bretz (1999), this rule does not control the FWE at $\alpha = 0.05$.

5. Clinical Decision Rules Based on Primary and Secondary Endpoints

The global and individual endpoint tests discussed in the previous two sections are useful

for assessing the efficacy of a treatment under an intersection null hypothesis framework. The MIN test is useful for dealing with a union null hypothesis. Often, protocols for drug approval specify decision rules based on a combination of union and intersection null hypotheses. Many examples of such decision rules are given in Chi (1998, 2000). In this section we present two common types of clinical decision rules, give some examples, and discuss how formulating these decision rules as a combination of union and intersection null hypotheses can lead to FWE controlling procedures.

A typical decision rule leads to several paths for finding a significant treatment effect. For example, given three endpoints (e.g., one primary and two secondary), one might conclude effectiveness if either $\delta_1 > 0$ or ($\delta_2 > 0$ and $\delta_3 > 0$), i.e., if the primary endpoint shows an effect or both secondary endpoints show an effect. As another example, given four endpoints, two primary and two secondary, a possible decision rule might be to conclude effectiveness if at least one primary endpoint and at least one secondary endpoint is significant, i.e., if ($\delta_1 > 0$ or $\delta_2 > 0$) and ($\delta_3 > 0$ or $\delta_4 > 0$).

In each of the above cases, the decision rule corresponds to an alternative hypothesis, from which an appropriate null hypothesis can be constructed by taking the complement. Let $H_{0i} : \delta_i \leq 0$ and $H_{1i} : \delta_i > 0$ for each endpoint i. Then the alternative hypothesis for the first example is

$$H_1 : H_{11} \bigcup (H_{12} \bigcap H_{13}),$$

and the null hypothesis is

$$H_0 : H_{01} \bigcap (H_{02} \bigcup H_{03}) = (H_{01} \bigcap H_{02}) \bigcup (H_{01} \bigcap H_{03}).$$

Then applying the IU principle, we can test each intersection null hypothesis at level α and conclude that the treatment is effective if both intersection null hypotheses are rejected. Similarly for the second case, the alternative hypothesis is

$$H_1 : (H_{11} \bigcup H_{12}) \bigcap (H_{13} \bigcup H_{14}),$$

and the null hypothesis is

$$H_0 : (H_{01} \bigcap H_{02}) \bigcup (H_{03} \bigcap H_{04}).$$

Again applying the IU principle, we can test each intersection null hypothesis at level α and conclude effectiveness of the treatment if both intersection null hypotheses are rejected. Neuhäuser, Steinijans and Bretz (1999) gave an example of this method using the Simes test for each intersection null hypothesis, but any of the global tests proposed earlier in the paper would also work.

A different type of rejection rule is obtained when we require that primary endpoints not have a large negative effect. For example, with one primary and one secondary endpoint, the treatment may be regarded as effective if $\delta_1 > 0$ or if $\delta_2 > 0$ and $\delta_1 > -\delta_1^*$ where $\delta_1^* > 0$ is a specified constant representing a threshold of equivalence between the treatment and control groups on endpoint 1. Define the hypotheses

$$H_{01}^* : \delta_1 \leq -\delta_1^* \text{ and } H_{11}^* : \delta_1 > -\delta_1^*.$$

Then the rejection rule is equivalent to the alternative hypothesis

$$H_1 : H_{11} \bigcup (H_{12} \bigcap H_{11}^*),$$

and the null hypothesis is its complement, namely

$$H_0 : H_{01} \bigcap (H_{02} \bigcup H_{01}^*) = (H_{01} \bigcap H_{02}) \bigcup H_{01}^*.$$

Then both the intersection null hypothesis and the equivalency null hypothesis on primary endpoint 1 can be tested at level α. If both are rejected, then we can conclude that the treatment is effective at level α.

This idea of simultaneously testing superiority/equivalence (see, e.g., Dunnett and Gent 1996) can be easily extended to more endpoints in the same fashion. For example, consider a case with three endpoints. The treatment is regarded effective if all endpoints are equivalent ($\delta_i \geq -\delta_i^*$ for $i = 1, 2, \ldots, m$) and furthermore at least one endpoint shows superiority ($\delta_i > 0$). Then the alternative hypothesis is

$$H_1 : (H_{11}^* \bigcap H_{12}^* \bigcap H_{13}^*) \bigcap (H_{11} \bigcup H_{12} \bigcup H_{13}),$$

with corresponding null hypothesis

$$H_0 : (H_{01}^* \bigcup H_{02}^* \bigcup H_{03}^*) \bigcup (H_{01} \bigcap H_{02} \bigcap H_{03}).$$

Again using the IU principle, the resulting method is to test each equivalency hypothesis at level α and to test the intersection hypothesis at level α as well. If all hypotheses are rejected then conclude that the treatment is effective at level α.

As demonstrated above, test procedures can be constructed for desired clinical decision rules which control the error rate at a pre-specified level α and incorporate both primary and secondary endpoints in the analysis. The basic steps are to formulate the decision rule as an alternative hypothesis, take the complement to form the null hypothesis, and apply the IU principle to determine an appropriate α-level for each component hypothesis. The methods discussed earlier in the paper can be used to test those components which are actually intersection null hypotheses.

6. Concluding Remarks

In this paper we have given a comprehensive review of the statistical methods available for analyzing multiple endpoints in clinical trials. There remain many unsolved problems. Two important ones are (i) deriving one-sided multivariate tests that are unbiased, monotone and have practically acceptable rejection regions (e.g., the rejection region should not contain outcomes with large negative components), and (ii) providing a general mathematical framework for clinical decision rules based on primary and secondary endpoints, so that it is not necessary to analyze each ad-hoc rule to see if it controls the FWE.

References

1. Abelson, R. P. and Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics*, **34**, 1347–1369.

2. Armitage, P. and Parmar, M. (1986). Some approaches to the problems of multiplicity in clinical trials. *Proceedings of the XIIIth International Conference in Clinical Trials*, Seattle, WA.

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

4. Cappizi, T. and Zhang, J. (1996). Testing the hypothesis that matters for primary endpoints. *Drug Information Journal*, **30**, 949 -956.

5. Chen, J. J. (1998). *P*-value adjustment for multiple binary endpoints. *Communications in Statistics, Ser. A (Theory and Methods)*, **27**, 2791–2806.

6. Chi, G. (1998). Multiple testings: Multiple comparisons and multiple endpoints. *Drug Information Journal*, **32**, 1347S–1362S.

7. Chi, G. (2000). Clinical decision rules and multiple endpoints: A regulatory perspective. Talk presented at the Second International Conference on Multiple Comparisons, Berlin, Germany. 1347S–1362S.

8. Dubey, S. D. (1985). Adjustments of *p*-values for multiplicities of intercorrelating symptoms. *Proceedings of the VIth International Society for Clinical Biostatisticians*, Germany.

9. Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**, 1096–1121.

10. Dunnett, C. W. and Gent, M. (1996). An alternative to the use of two-sided tests in clinical trials. *Statistics in Medicine*, **15**, 1729–1738.

11. Dunnett, C. W. and Tamhane, A. C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine*, **10**, 939–947

12. Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association*, **87**, 162–170.

13. Fang, K.-T. and Zhang, Y.-T. (1990). *Generalized Multivariate Analysis*. Berlin Heidelberg: Springer.

14. Follman, D. (1995). Multivariate tests for multiple endpoints in clinical trials. *Statistics in Medicine*, **14**, 1163 -1175.

15. Follman, D. (1996). A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association*, **91**, 854-861.

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

17. Hochberg, Y. and Mosier, M. C. (2001). Intersection-union procedures for some restricted models. Unpublished manuscript.

18. Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley: New York.

19. Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75**, 383–386.

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

21. Huque, M. F. and Sankoh, A. J. (1997). A reviewer's perspective on multiple endpoint issues in clinical trials. *Journal of Biopharmaceutical Statistics*, **7**, 545–564.

22. James, S. (1991). The approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials. *Statistics in Medicine*, **10**, 1123 -1135.

23. Kieser, M., Reitmeir, P. and Wassmer, G. (1995). Test procedures for clinical trials with multiple endpoints. *Biometrie in der chemisch-pharma-zeutischen Industrie* (ed. J. Vollmar), **6**, Stuttgart: Gustav Fischer Verlag, 41 -60.

24. 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.

25. Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, **50**, 403–418.

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

27. Läuter, J. (1996). Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* **52**, 964 -970.

28. 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.

29. 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.

30. Logan, B. R. (2001). *Contributions to Multiple Endpoints and Dose Finding.* Doctoral Dissertation, Department of Statistics, Northwestern University, Evanston, IL.

31. Logan, B. R. and Tamhane, A. C.(2001). Combining global and marginal tests to compare two treatments on multiple endpoints. *Biometrical Journal*, **43**, 591–604.

32. 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.

33. Meier, P. (1975). Statistics and medical experimentation. *Biometrics* **31**, 511 -529.

34. Neuhäuser, M., Steinijans, V. W. and Bretz, F. (1999). The evaluation of multiple clinical endpoints, with application to asthma. *Drug Information Journal*, **33**, 471 -477.

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

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

37. Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, **43**, 487 -498.

38. Reitmeir, P. and Wassmer, G. (1996). One-Sided multiple endpoint testing in two-sample comparisons. *Communications in Statistics (Computation and Simulation)*, **25**, 99-117.

39. Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

40. Sankoh, A. J., Huque, M. F. and Dubey, S. D. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*, **16**, 2529–2542.

41. Sankoh, A. J., Huque, M. F., Russell, H. K. and D'Agostino, R. B., Sr. (1999). Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug Information Journal*, **33**, 119–140.

42. Sarkar, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: A proof of the Simes conjecture. *Annals of Statistics*, **26**, 494 -504.

43. Sarkar, S. K. and Chang, C-K (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, **92**, 1601 -1608.

44. Sen, P. K. and Tsai, M-T (1999). Two-stage likelihood ratio and union-intersection tests for one-sided alternatives multivariate mean with nuisance dispersion matrix. *Journal of Multivariate Analysis*, **68**, 264 -282.

45. Šidák, Z. (1968). On multivariate normal probabilities of rectangles: Their dependence on correlations. *Annals of Mathematical Statistics*, **39**, 1425 -1434.

46. Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751 - 754.

47. Silvapulle, M. J. (1997). A curious example involving the likelihood ratio test against one-sided hypotheses. *American Statistician*, **51**, 178 -180.

48. Snappin, S. M. (1987). Evaluating the efficacy of a combination therapy. *Statistics in Medicine*. **6**, 657–665.

49. Tamhane, A. C., Liu, W. and Dunnett, C. W. (1998). A generalized step-up-down multiple test procedure. *Canadian Journal of Statistics*, (1998), **26**, 353–363

50. 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.

51. Tang, D. I. (1994). Uniformly more powerful tests in a one-sided multivariate problem. *Journal of the American Statistical Association*, **93**, 380 -386.

52. Tang, D. I., Geller, N. L. and Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*, **49**, 23–30.

53. 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.

54. Troendle, J. (1996). A permutational step-up method for testing multiple outcomes. *Biometrics*, **52**, 846–859.

55. Tukey, J. W., Ciminera, J. L. and Heyse, J. P. (1985). Testing the statistical certainty of a response with increasing doses of a compound. *Biometrics*, **41**, 295–301.

56. Wang, Y. and McDermott, M. P. (1998). Conditional likelihood ratio test for a non-negative normal mean vector. *Journal of the American Statistical Association*, **89**, 380 -386.

57. Westfall, P. H. and Young, S. S. (1989). *P*-value adjustment for multiple tests in multivariate binary models. *Journal of the American Statistical Association*, **84**, 780–786.

58. Westfall, P. H. and Young, S. S. (1993) *Resampling Based Multiple Testing.* John Wiley: New York.

59. Wright, S. P. (1992). Adjusted *p*-value for simultaneous inference. *Biometrics*, 1005–1013.

60. Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem. *Biometrika*, **52**, 139 -147.

61. 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.