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 ShereAn7Mo10u

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.

The corresponding correlation matrix will be denoted by R with elements

$$\rho_k = \text{Corr}(x_{ijk}, x_{ij}) = \frac{\sigma_k}{\sqrt{\sigma_{kk}\sigma}} \quad (1 \leq k = \quad \leq m).$$

In the heteroscedastic case, Σ_1 and Σ_2 are not assumed to be equal. The elements of Σ

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 $H_2 : \delta = 0$, 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 Σ is *known* for the one-sample problem which can be easily extended to the two-sample problem as follows. Let $\hat{\delta}$ be the projection of $\bar{x}_{1\cdot} - \bar{x}_{2\cdot}$ in the positive orthant with respect to the distance function $d(u, v) = (u - v)\,\Sigma^{-1}(u - v$

Stein-type two-stage test that is free of Σ. It also has other desirable properties such as

The null distribution of $g(u)$ is the χ^2 distribution with symmetric binomial probability weights given by

$$\Pr_{H_0}\{g(u) > c\} = \sum_{k=0}^{m} \binom{m}{k} 2^{-m} \Pr\left(\chi^2_k > c\right)$$

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 $\hat{R}$. 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 $\hat{R}$ both in the calculation of $\hat{\lambda}_{GLS}$ and in the estimate of SD($\hat{\lambda}_{GLS}$). 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 $j\,\hat{R}^{-1}$ 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 $\mu_1 - \mu_2 = \lambda(\beta_1, \ldots, \beta_m)$, where the vector $(\beta_1, \ldots, \beta_m)$ with all positive elements is specified. However, if the observed mean difference $\bar{x}_{1\cdot} - \bar{x}_{2\cdot}$ 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 $\bar{x}_{i\cdot} = (\bar{x}_{i\cdot 1}, \bar{x}_{i\cdot 2}, \ldots, \bar{x}_{i\cdot m})$ denotes the vector of sample means for the ith group ($i = 1, 2$) and

let

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

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

$$V = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\cdot\cdot})(x_{ij} - \bar{x}_{\cdot\cdot}) = (n_1 + n_2 - 2)\hat{\Sigma} + \sum_{i=1}^{2} n_i(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})(\bar{x}_{i\cdot} -$$

Because the total pooled standard deviation overestimates the true standard deviation since

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

$$\Omega_i = \frac{1}{n_i}\Sigma_i \quad (i = 1, 2) \Omega = \Omega_1 + \Omega_2 \text{ and } \Sigma = \frac{n_1 n_2}{n_1 + n_2}\Omega.$$

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

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

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

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

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

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

$$u\,u = \left[\frac{n_1 n_2}{n_1 + n_2}\right](\bar{x}_{1\cdot} - \bar{x}_{2\cdot})\,\hat{\Sigma}^{-1}(\bar{x}_{1\cdot} - \bar{x}_{2\cdot})$$

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 $m, n_1 = n_2 = n, \Sigma_1$ and Σ_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 Σ_1 and Σ_2 are known. Then the statistic for comparing the treatment with the control on the kth endpoint is

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

Let $z = (z_1, z_2, \ldots, z_m)$ and $\bar{R} = (n_1 R_1 + n_2 R_2)/(n_1 + n_2)$. In analogy with (3.9), Pocock et al. proposed the statistic

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

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

$$\gamma_k = \sigma_{1,k}/n_1 + \sigma_{2,k}/n_2$$

for $i = 1, 2$. Note that $\xi_{1k} - \xi_{2k} = \lambda$ for all k. Also note that Γ_1 and Γ_2 are not correlation matrices, and $\Gamma = \Gamma_1 + \Gamma_2$ if $n_1 = n_2$.

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

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

where $\xi_k = (\xi_{1k} + \xi_{2k})/2$, $I_{ijk} = +1$ if $i = 1$ and -1 if $i = 2$, and $\varepsilon_{ij} = (\varepsilon_{ij1}, \varepsilon_{ij2}, \ldots, \varepsilon_{ijm})$ are independently distributed as $N(0, \Gamma_i)$. 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 Σ_1 and Σ_2 are known, it is straightforward to show that

$$\hat{\lambda}_{\mathrm{OLS}} = \frac{j\,(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})}{m} = \bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot} \text{ and } \mathrm{SD}(\hat{\lambda}_{\mathrm{OLS}}) = \sqrt{j\left(\frac{\Gamma_1}{n_1} + \frac{\Gamma_2}{n_2}\right)j}.$$

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

$$t_{\mathrm{OLS}} = \frac{j\,(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})}{\sqrt{j\,(\hat{\Gamma}_1/n_1 + \hat{\Gamma}_2/n_2)j}}, \tag{3.13}$$

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

$$\bar{y}_{i,k} = \frac{\bar{\sigma}_{i,k}}{(\bar{\sigma}_{1,kk} + \bar{\sigma}_{2,kk})(\bar{\sigma}_{1,\cdot\cdot} + \bar{\sigma}_{2,\cdot\cdot})}.$$

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

$$t_{\mathrm{OLS}} = \frac{j\,t}{\sqrt{j\,\hat{\Gamma} j}},$$

where t is a vector o1511.955Tf8.053r963Tf13.0353.913Td[(b)]TJrra2

Next we derive the GLS test. Assuming tha[illegible] known, it can b[illegible] sho[illegible]

$$\hat{\lambda}_{\text{GLS}} = \frac{4j\ (\Gamma_1/n_1 \ldots [illegible] \ldots - \bar{y}_{2\cdot})}{j\ [(I - B)\Gamma \ldots [illegible] \ldots)\Gamma_2^{-1}/n_2]j}$$

and

$$\text{SD}(\hat{\lambda}_{\text{GLS}}) = \frac{[illegible] \ldots {}_2/n_2)^{-1}\ j}{j\ [(I \ldots [illegible] \ldots (I + B)\Gamma_2^{-1}/n_2]j},$$

[illegible]

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 inde-

global test of each null hypothesis H_{0K} for K 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 $\max_{1 \le k \le m} t_k$ 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 $\sqrt{\bar{\sigma}_{kk}}$ used to standardize the t_k statistics are di□erent 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 $\tilde{p}_k$, and reject H_{0k} if $\tilde{p}_k < \alpha$ $(1 \le k \le m)$.

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

$$\tilde{p}_k = \Pr_{H_0}\left(\min_{1 \le \le m} P \le p_k \right) \quad (1 \le k \le m). \tag{4.1}$$

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

$$\tilde{p}_k = mp_k \quad (1 \le k \le m).$$

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 's

are independent and is given by

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

If the P'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:

$$\tilde{p}_k = 1 - (1 - p_k)^{mf} \quad (1 \le k \le m),$$

where f is an empirically determined function of the ρ_k's. Dubey (1985) suggested using a different function $f_k = 1 - \bar{\rho}_k$ for each k, where $\bar{\rho}_k$ 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 $\bar{\rho}_k$ in Dubey's formula should be replaced by $\bar{\rho}$, namely, the average of all ρ_k's. Notice that if all $\rho_k = 0$ then we get the Šidák adjustment and if all $\rho_k = 1$ then $\tilde{p}_k = p_k$, i.e., there is no adjustment. Tukey, Ciminera and Heyse (1985) suggested using $f = 1/2$, i.e., $\tilde{p}_k = 1 - (1 - p_k)^{\sqrt{m}}$, which assumes that the average correlation is 1/2. An analytic approximation to $\tilde{p}_k$ 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 $\tilde{p}_k$.

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

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

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

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

These approximations correspond to Holm's (1979) step-down test procedure, which rejects $H_{0(k)}$ if $p_{(\ell)} < \alpha/\ell$ 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

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

Hochberg's procedure accepts $H_{0(k)}$ if $p_{(\ell)} \ge \alpha/\ell$ 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 $H_{0K} = \cap_{k \in K} H_{0k}$ is defined as

$$\tilde{p}_K = \Pr_{H_0}\left[\min_{k \in K} P_k, P_{K,\text{OLS}} \le \min_{k \in K}\left(p_k, p_{K,\text{OLS}}\right)\right], \tag{4.3}$$

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 if all hypotheses H_{0L} for $L \subseteq K$ are rejected at level α and $\bar{p}_K < \alpha$. In pratice, the $\bar{p}_K$ 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`).

configurations at which the type I error probability is maximized (= α) can be shown to be of the type $\delta_k = 0$ for some k and $\delta \to \infty$ for $= k$. 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)

for assessing the e□cacy 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 e□ect. For example, given three endpoints (e.g., one primary and two secondary), one might conclude e□ectiveness if either $\delta_1 > 0$ or ($\delta_2 > 0$ and $\delta_3 > 0$), i.e., if the primary endpoint shows an e□ect or both secondary endpoints show an e□ect. As another example, given four endpoints, two primary and two secondary, a possible decision rule might be to conclude e□ectiveness 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

$$H_1 : H_{11} \cap (H_{12} \cup H_{13}),$$

and the null hypothesis is

$$H_0 : H_{01} \cup (H_{02} \cap H_{03}) = (H_{01} \cup H_{02}) \cap (H_{01} \cup H_{03}).$$

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

$$H_1 : (H_{11} \cap H_{12}) \cup (H_{13} \cap H_{14}),$$

and the null hypothesis is

$$H_0 : (H_{01} \cap H_{02}) \cup (H_{03} \cap H_{04}).$$

Again applying the IU principle, we can test each intersection null hypothesis at level α and conclude e□ectiveness 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

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 e□ective at level α.

As demonstrated above, test procedures can be constructed for desired clinical decision rules which control the error rate at a pre-specifiTJ-1agfuc11.95103l

References

1. Abelson, R. P. and Tukey, J. W. (1963). E□cient utilization of non-numerical infor-

11. Dunnett, C. W. and Tamhane, A. C. (1991). Step-down multiple tests for comparing

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,]TJ/Td[(,)-326(Stuttg)1(art955Tf222222228dures)-375(Appli22cuttg-330tionTf2326(f1.)-6

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.

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 e□cacy 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

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.