ELSEVIER

Some Statistical Methods for Multiple Endpoints in Clinical Trials

Ji Zhang, PhD, Hui Quan, PhD, Jennifer Ng, SD, and Michael E. Stepanavage, MS

Merck Research Laboratories, Clinical Biostatistics and Research Data Systems, Rahway, New Jersey

ABSTRACT: This paper summarizes, defines, and discusses multiple endpoints comparison procedures, concepts, and methodologies for applications to clinical trials. We address the more widely used methods of α-level, *p*-value, and critical value adjustments. We examine global assessment measures such as O'Brien's test and Simes' procedure and contrast them with the α-adjustment procedures of Bonferroni and Holm. We propose a global assessment procedure based on categorization of the individual endpoints to form an overall composite endpoint. Additionally, we discuss a new weighting scheme for Holm's sequentially rejective α-adjustment procedure. Investigation of the correlation between endpoints is examined in relation to adjustment of the α-level. In the context of a clinical trial, the above multiplicity procedures are applied and compared. Finally, some comments concerning ease of use and relevance are summarized for the above methods. *Controlled Clin Trials* 1997;18:204–221 © Elsevier Science Inc. 1997

KEY WORDS: *Composite endpoint, control of type I error, experimentwise error rate, p-value adjustment, α-level*

INTRODUCTION

Multiple comparison problems arise in a variety of contexts in clinical trials, including subgroup analysis, meta-analysis, sequential testing, and, particularly, testing of multiple efficacy or safety endpoints. Regulatory agencies have become increasingly interested in insuring that clinical trials, particularly phase III trials, address issues of multiplicity. Consequently, sponsors of drug trials and investigators attempt to address these issues when protocols are designed. A univariate approach with unadjusted *p*-values allows readers to reach a subjective conclusion regarding the treatment's overall efficacy, but this approach does not control the overall type I error rate.

Multiple efficacy endpoints are encountered when efficacy is evaluated using two or more biological, physiological, or social assessments or measurements. Sometimes it is possible to maintain the overall type I error rate while examining the effect of treatment on the individual endpoints and on efficacy in general by identifying a single primary endpoint and several secondary endpoints.

Address reprint requests to: Ji Zhang, Merck Research Laboratories, Clinical Biostatistics and Research Data Systems, RY 33-404, Rahway, NJ 07065-0900.

Received December 4, 1995; revised June 28, 1996; accepted July 23, 1996.

Controlled Clinical Trials 18:204–221 (1997)

655 Avenue of the Americas, New York, NY 10010

0197-2456/97/$17.00
PII S0197-2456(96)00129-8

Alternatively, one may create a composite endpoint consisting of a combination of primary endpoints. The most appropriate specific approach depends on the clinical objectives to be addressed. Composite endpoints help in establishing "nonspecific" efficacy of a test compound compared with another or with placebo, while tests and estimates of individual endpoints help in establishing specific benefits. A decision on a single primary endpoint cannot be easily made in many disease areas, for example, multifaceted diseases like asthma and benign prostatic hyperplasia.

The statistical analysis of individual endpoints using univariate procedures remains important, easier to interpret than most other methods, and is more familiar to nonstatisticians. Recently developed methods allow the assessment of overall treatment effect and the treatment effect on individual endpoints while maintaining the overall α-level.

In this paper, we summarize a variety of methods that use a single statistical measure to assess the effect of treatment on several efficacy endpoints, and examine some of the more widely used methods of α-level, p-value, and critical value adjustment that allow the assessment of the effect of treatment on individual endpoints while maintaining the overall type I error rate. We present some general concepts and principles underlying multiple testing procedures; review some global test statistics and outline their advantages and disadvantages; and present an example illustrating the multiple testing methods. We focus on some multiplicity adjustment methods for multiple endpoints that we find useful and relatively easy to implement; however, this paper is not intended to be a comprehensive survey or review of the vast field of multiple adjustment methods.

METHODS

Background

One of the most common examples of multiplicity in clinical trials arises when there are two treatments (for example, a test drug compared with a placebo), more than one endpoint, and the desire to test K individual null hypotheses against specific alternatives. Often the K null hypotheses of interest center on whether the drug and placebo are equivalent for each endpoint. Efficacy of the drug may have been evaluated separately for each endpoint. Interest therefore lies in controlling the probability of at least one erroneous rejection by testing the K individual null hypotheses simultaneously in a single multiple test procedure or by testing for an overall effect of treatment using a global test statistic.

We will use the following example to illustrate some concepts and procedures.

Example

This example comes from a randomized, multicenter, double-blinded, parallel design clinical trial conducted to assess efficacy and safety of a test drug in asthmatic patients. We consider one test drug group (n = 34) and the placebo

Table 1 Summary Statistics for the Asthma Example

Treatment	Statistics	Endpoint			
		FEV_1 (%)	Peak Expiratory Flow Rate (L/m)	Symptom Score (0–6 scale)	Additional Medication Use (Puffs/Day)
Placebo	Mean	5.7	1.6	0.34	0.15
Test drug	Mean	14.0	16.5	0.86	0.49
Pooled standard deviation		11.5	22.3	0.96	0.66
Two-sample t test		3.00	2.75	2.25	2.13
p-value		0.0037	0.0077	0.0274	0.0369

group ($n = 35$) to illustrate the statistical methods. The following four endpoints are considered to be important:

1. Forced expiratory volume in 1 second (FEV_1), in liters;
2. Peak expiratory flow rate (PEFR), in liters per minute;
3. Symptoms score (SS), 0–6 scale;
4. Additional medication use (AMU), i.e., β-agonist use, in puffs per day.

The FEV_1 and the PEFR values measure airflow obstruction. FEV_1 is measured at clinic visits, while PEFR can be measured daily at home. The symptom scores and the additional medication use reflect patient self-perceptions. These four endpoints encompass information from different dimensions of this multifaceted disease [1].

The data analysis used percent change from baseline for FEV_1 and change from baseline for the other three endpoints. For simplicity, data were multiplied by -1, if necessary, so that larger numbers corresponded to increased efficacy. Data (change or percent change values for all variables) were approximately normally distributed, and the variances for the two treatment groups were approximately equal. Patient characteristics and baseline efficacy values were comparable for the two treatment groups.

Table 1 presents a summary of the data. Table 2 presents the correlations among the four endpoints.

Table 2 Correlations (Pearson) among the Efficacy Endpoints

	FEV_1	PEFR	Symptom Score
PEFR[a]	0.25		
Symptom score	0.31	0.42	
Additional medication use	0.24	0.43	0.67

[a]PEFR: Peak Expiratory Flow Rate.

Control of the Type I Error Rate

Two approaches are available for controlling the probability of incorrectly rejecting at least one null hypothesis. One can control the "family" error rate (i.e., global α-level), or one can control the "experimentwise" error rate (i.e.,

multiple α-level). As defined by Bauer [2], Hommel [3], and Holm [4], a multiple test procedure with critical region $C_1 \ldots, C_K$ for testing the null hypotheses $H_{01}, \ldots, H_{0K}$ controls the global α-level if the probability of erroneously rejecting at least one true null hypothesis does not exceed α when all of the individual null hypotheses are true simultaneously, that is,

$$PR(\sum_{i \in I} C_i) \leq \alpha \text{ for } I \subset \{1, \ldots, K\} \text{ when } H_{01}, \ldots, H_{0K} \text{ are true.}$$

On the other hand, a multiple-test procedure controls the multiple α-level if the probability of erroneously rejecting at least one true null bypothesis is controlled by α irrespective of which and how many individual null hypotheses are true, that is,

$$PR(\sum_{i \in I} C_i) \leq \alpha \text{ for } I \subset \{1, \ldots, K\} \text{ when } H_{0i}, i \in I \text{ are true.}$$

Since control of the global α-level depends on all null hypotheses being true simultaneously, whereas the control of the multiple α-level does not depend on this assumption, a multiple-test procedure that controls the multiple α-level will also control the global α-level, but not vice versa. Therefore, in many cases, control of the multiple α-level is a more reasonable course of action. For example, in an asthma therapy project, after the phase I–II clinical trials, one might have observed that the test drug showed some beneficial effects for the endpoints of FEV_1, PEFR, symptom score, and additional medication use. These findings might contain some false positives (i.e., type I error), but investigators may believe it is unlikely that all four would be type I errors. Phase III trials are designed to confirm such preliminary findings, and thus control of the multiple α-level in these trials is more important when one wishes to establish or claim treatment effect in all four endpoints.

Closed Testing Procedures

Marcus et al [5] proposed a class of tests that control the multiple α-level. They defined testing procedures for a set of null hypotheses $\{H_0\} = \{H_{01}, H_{02}, \ldots, H_{0K}\}$ that are closed under intersection, i.e., $H_{0i}, H_{0j} \in \{H_0\}$ implies $H_{0i} \cap H_{0j} \in \{H_0\}$. (Here ∩ means "and.") For each null hypothesis, H_{0i}, i=1,2, . . ., K, it is assumed that there is a local α-level test T_i. In the closed testing procedure, a particular null hypothesis H_{0i} is tested by means of T_i if, and only if, all hypotheses that contain H_{0i} and are contained in $\{H_0\}$ have been tested and rejected. In other words, the closed testing procedure begins with the global null hypothesis and proceeds sequentially to null hypotheses involving successively fewer endpoints. The most important feature of this procedure is that each null hypothesis is tested at level α, yet the experimentwise error rate also remains at α.

For example, suppose a trial has two treatments and three endpoints. Let d_i denote the difference between treatments for endpoint i, for i=1, 2, 3. The global null hypothesis is H_0: $\{d_1=0, d_2=0, d_3=0\}$. If the global hypothesis is rejected at level α, one can proceed to the intersection null hypotheses: H_{01}: $\{d_1=0, d_2=0\}$, H_{02}: $\{d_1=0, d_3=0\}$, H_{03}: $\{d_2=0, d_3=0\}$. If any two of these hypotheses are rejected at level α, say, H_{01} and H_{02}, then one can proceed to test the elementary null hypotheses: H_{04}: $\{d_1=0\}$. If all three intersection null hypotheses,

H_{01}, H_{02}, and H_{03}, are rejected, then one can proceed to test all three of the elementary null hypotheses.

Marcus et al [5], and later Hommel [3], proved that this procedure controls the probability of making an (experimentwise) type I error to at most α.

Global Assessment Measures

Global test statistics evaluate the global null hypothesis, which corresponds to the overall assessment of treatment difference. Choice of the global test statistic depends on the alternative hypothesis of interest and the specific disease studied. Multivariate analysis of variance (MANOVA) and Hotelling's T^2 will usually provide an elliptical acceptance region, and hence would be appropriate when dealing with unrestricted alternatives, for example, both positive and negative treatment effects. When dealing with restricted alternatives (e.g., positive treatment effects in all endpoints), global test statistics such as those proposed by O'Brien [6] may be more appropriate.

In many clinical trials, a composite endpoint, i.e., a combination of several univariate endpoints, may be clinically appropriate. Often, a composite endpoint has a natural definition. For example, a clinical definition of response may be based on several endpoints. How to define and choose a composite endpoint (and hence the global null hypothesis) will depend on the specific disease or test compound under study.

We now review some general approaches for constructing a composite endpoint.

Methods of O'Brien

O'Brien [6] proposed the following three procedures: a rank-sum method, a method based on ordinary least squares (OLS), and one based on generalized least squares (GLS).

Notation

Let Y_{ijk} denote the k^{th} variable for the j^{th} subject in group i, and assume the endpoint vectors $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijK})$ are independently distributed with mean $\boldsymbol{\mu}_i$ and covariance matrix Σ $(i=1, \ldots, I, j=1, \ldots, n_i)$.

The null hypothesis is H_0: $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2= \ldots =\boldsymbol{\mu}_I$. The focus is on one-sided alternatives, $\boldsymbol{\mu}_i > \boldsymbol{\mu}_j$ elementwise for some i, j, assuming large mean values indicate beneficial effects.

Rank-Sum Procedure

The rank-sum procedure can be summarized as follows:

1. Rank observations Y_{ijk} among all values of the endpoint k in all samples: $\{Y_{ijk;}\ i=1, \ldots, I, j=1, \ldots, n_i\}$ and denote the rank by R_{ijk}.
2. Let $S_{ij} = \sum_k R_{ijk}$ be the sum of ranks for the j^{th} subject in the i^{th} group.

The original vector $\mathbf{Y}_{ij}$ for each subject is therefore reduced to the one-dimensional statistic (or a composite endpoint) S_{ij}, which is essentailly an overall

assessment of efficacy for the j^{th} subject in the i^{th} group. Therefore, appropriate statistical methods can be applied to this univariate problem.

One can, a priori, specify a set of weights when computing S_{ij}, $S_{ij} = \sum_{k=1}^{K} w_k R_{ijk}$. The weighted rank sum accounts for the relative importance of each endpoint.

Ordinary Least Squares and Generalized Least Squares Procedures

Assume Y_{ijk} is standardized $((Y_{ijk} - \bar{Y}_{.k})/S_{.k}$, where $\bar{Y}_{.k}$ is the mean for endpoint k and $S_{.k}$ is the pooled within group sample standard deviation). Furthermore, consider the following alternative, $\boldsymbol{\mu}_i = \boldsymbol{\mu} + \boldsymbol{\beta}_i \mathbf{J}$ i=1, 2, . . ., I, where $\boldsymbol{\mu}_i$ is the mean vector for group i, β_i are scalars such that $\Sigma\beta_i = 0$, and J is a column of 1's. This formulation is equivalent to saying that after standardization the treatment effects are of the same magnitude for all endpoints. The multivariate problem is then reduced to a regression problem, solvable using OLS and GLS techniques.

Assume the K endpoints have a multivariate normal distribution and a common covariance matrix for all treatment groups. Let

$$G = \sum_{i=1}^{I} \frac{n_i(\mathbf{J}'\mathbf{S}^{-1}(\bar{Y}_{i.} - \bar{Y}_{..}))^2}{(I - 1)\mathbf{J}'\mathbf{S}^{-1}\mathbf{J}}$$

where $\bar{Y}_{i.}$ is the mean vector for the i^{th} group, $\bar{Y}_{..}$ is the grand mean, and S is the pooled within-group sample covariance matrix defined by

$$\mathbf{S} = \sum_{i=1}^{I} \frac{n_i - 1}{(N - I)} \mathbf{S}_i \tag{1}$$

Here $\mathbf{S}_i$ is the sample covariance matrix for the i^{th} group, $N = \sum_{i=1}^{I} n_i$. This test procedure rejects the null hypothesis if G exceeds the $(1 - \alpha)\times 100$ percentile of the F distribution with $(I - 1)$ and $(N - IK)$ degrees of freedom.

When comparing two samples (I = 2), one may compute the two-sample *t* statistic for each endpoint as outlined in Pocock et al [7]. These test statistics can be modified to accomodate endpoints of different importance and/or different treatment magnitudes after standardization:

$$O = \frac{\mathbf{J}'\mathbf{WT}}{(\mathbf{J}'\mathbf{WSWJ})^{1/2}}, \quad G = \frac{\mathbf{J}'(\mathbf{WSW})^{-1}\mathbf{WT}}{(\mathbf{J}'(\mathbf{WSW})^{-1}\mathbf{J})^{1/2}}$$

where W is a diagonal matrix with positive weights, larger weights indicating greater importance.

O'Brien reported the results of a limited simulation study of the critical level and power of five procedures: Hotelling's T^2, Bonferroni adjustment, rank-sum, GLS, and OLS procedures. He considered alternatives based on consistent treatment differences across all endpoints. He concluded that the Hotelling's T^2 and Bonferroni methods were less powerful than the rank-sum procedure, which he recommended. O'Brien's composite endpoints approach is more powerful when all endpoints have similar treatment effect sizes (not necessarily the same), but the Bonferroni method is more powerful if one endpoint has a very small effect while another has a much larger effect. O'Brien's procedures, which are relatively simple to use, provide an overall assessment of treatment

difference. Pocock et al [7] reviewed O'Brien's procedure and introduced some extensions to the GLS procedure. Lechmacher et al [8] applied the closed multiple testing principle to O'Brien's procedures, and through simulation, showed that both the OLS and GLS procedures were more powerful than Hotelling's T^2 and Holm's procedure (see below) in detecting a difference between treatments for individual endpoints.

We find O'Brien's rank-sum procedure difficult to interpret. It provides an overall assessment of treatment effect, but offers no estimate of the magnitude of the treatment effects. Examination of individual endpoints is needed to supplement the rank-sum procedure. The following composite endpoint, D, can be another viable alternative. The interpretation of D can be an inherent feature in the definition. To construct D, one uses clinical and statistical information to transform each endpoint into an equal rating scale. These definitions should be given a priori and be validated in terms of reliability, responsiveness, and construct validity.

1. Divide each endpoint E_i into C+1 categories (C can be different for each endpoint):

   $E_i^c = j$ if $E_i \in (a_{ij}, a_{i,j+1}]$, $j = 0, \ldots, C$.

2. Define a composite endpoint as

$$D = \sum_{i=1}^{K} E_i^c \text{ or } D = \sum_{i=1}^{K} w_i E_i^c$$

where the w_i's are a set of positive weights such that $\sum w_i = 1$. Note that the composite endpoint takes values between 0 and CK, $D \in [0, CK]$.

Patients with scores of CK, 0, and nearly CK can be regarded as complete responders, nonresponders, and partial responders, respectively. The choice of a_{ij} is critical, and clinical input is essential. Many univariate statistical methods can be used to analyze this composite endpoint D. Simple applications with similar ideas were reported in the literature, for example, total scores (weighted or unweighted) for a symptom or quality of life questionnaire, where the a_{ij} are naturally defined. Choice of weights in the weighted procedures may often be difficult, especially to the medical community [9,10].

Simes' Procedure

Simes [11] developed a procedure using individual p-values to assess overall treatment effect. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$ be the ordered p-values for the individual endpoints. Then, reject the global null hypothesis, H_0, if $p_{(j)} \leq j\alpha/K$ for any $j = 1, 2, \ldots K$. The type I error $PR\{\sum_{j=1}^{K} p_{(j)} \leq j\alpha/K\}$, though not always less than α, holds for a large number of multivariate distributions, for example, many members of the multivariate Gamma distribution family [11]. Simes further showed that this procedure has type I error of α for independent tests.

Although Simes' procedure is less conservative than the Bonferroni method and Holm's procedures [4] (discussed later in this paper), it is a test for the overall treatment effect. It does not allow decisions to be made on the individual endpoint hypotheses of interest while controlling the experimentwise error rate for the individual endpoints. However, a closed multiple testing procedure can be applied to make inference on individual endpoints while controlling

the experimentwise error rate. Simes also showed that the above procedure has higher power for a given nominal α-level than the Bonferroni and Holm's procedures, especially when the endpoints are strongly correlated.

α-Level, Critical Value, and *p*-Value Adjustment Methods

When H_0 is rejected, global methods do not provide specific information about which endpoints contribute to the difference or the magnitude of the between treatment differences. Also, in some cases, biological or regulatory concerns render a single composite measure of efficacy inappropriate.

In this section we address some of the more widely used methods for adjusting the α-levels, *p*-values, and critical values. These methods differ from the global methods presented previously in that inferences concerning the overall hypothesis are made through inference on the individual endpoints while maintaining the experimentwise error rate. Therefore, these methods, unlike global assessment procedures, allow direct inference (test and estimation) regarding the individual endpoints.

α-Level Adjustment Methods

Bonferroni Method

The classic Bonferroni method is a simple α-level adjustment procedure that controls the experimentwise error rate. Application of the classic Bonferroni procedure is based upon rejecting an overall global hypothesis $H_0 = \{H_{01}, H_{02}, \ldots H_{0K}\}$ if the *p*-value for any individual test of hypotheses is less than or equal to the adjusted α-level, $\alpha_a = \alpha/K$. Each individual hypothesis will also be rejected if the *p*-value for that hypothesis is significant at the α/K level.

Sequential Rejective Bonferroni Procedure

Holm [4] developed a sequential procedure that improved upon the Bonferroni method by providing a less conservative test. Hommel [12] and Shaffer [13] developed a similar procedure. The procedure is conducted stepwise comparing successively higher *p*-values with increasingly greater significance levels:

1. Order the univariate *p*-values in increasing order such that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$ with corresponding hypotheses $H_{(01)}, H_{(02)}, \ldots, H_{(0K)}$.
2. Compare the smallest *p*-value, $P_{(1)}$, against the most conservative α-level, α/K. If $p_{(1)} \leq \alpha/K$ holds, reject $H_{(1)}$ and continue stepwise, with each successive test conducted at progressively higher significance levels, $\alpha/(K-1), \alpha/(K-2), \ldots, \alpha/1$. Otherwise, fail to reject $H_{(01)}$ and stop the procedure. Note that acceptance of $H_{(0i)}$ implies acceptance of $H_{(0j)}$ for all $j > i$.

The basic idea behind Holm's procedure is that if m hypotheses are true, an error must occur at or before step $(K - m + 1)$. Thus,

$$\begin{aligned} \text{PR (no type I errors)} &\geq \text{PR}(P_i > \alpha/m \text{ for all } i \in I) \\ &= 1 - \text{PR}(P_i \leq \alpha/m \text{ for some } i \in I) \\ &\geq 1 - \sum_{i \in I} \alpha/m = 1 - \alpha, \end{aligned}$$

where I is the set of indices of the hypotheses that are true. When most of the hypotheses are approximately true, Holm's procedure provides little increase in power over the classic Bonferroni method; however, the increase in power for the sequential procedure becomes substantial when some of the null hypotheses are rejected with high probability. In these cases, the remaining null hypotheses are tested at less stringent α-levels.

Holm's sequential procedure can be modified to accommodate endpoints that are of different importance and/or reflect different magnitudes of treatment effect. The α-levels are allocated in proportion to the weights that reflect the importance of the hypothesis or the magnitude of treatment effect:

Let $w_1, w_2, \ldots, w_K$ be positive constants indicating the "importance" of the hypotheses. Let $p_{wi} = p_i/w_i$, where $i = 1, 2 \ldots, K$, and order the p_{Wi} in increasing order, $p_{(Wi)}$. Let $w_{(i)}$ be the corresponding weights.

1. Compare the most significant result $p_{(W1)}$ with $\alpha/\sum_{i=1}^{K} w_{(i)}$, i.e., α divided by the sum of all the weights assigned to the hypotheses.
2. If the inequality $p_{(W1)} \leq \alpha/\sum_{i=1}^{K} w_{(i)}$ holds, reject the corresponding hypothesis, and continue the sequential testing procedure, checking if $p_{(W2)} \leq \alpha/\sum_{i=2}^{K} w_{(i)}$, α divided by the sum of the remaining weights, and so forth. If the inequality does not hold, fail to reject the corresponding hypothesis and stop the procedure.

When the w_i's are all equal, this procedure simplifies to Holm's sequentially rejective Bonferroni procedure. Assignment of weights to the hypotheses provides increased power for those hypotheses with larger values of w_i at the cost of reduced power for those hypotheses with smaller values of w_i.

Modified Sequentially Rejective Method

The sequentially rejective procedure assumes no a priori knowledge of the relative importance of the endpoints. The weighted sequentially rejective procedure recognizes this problem by assigning different weights to the endpoints on the basis of the importance and/or reponsiveness of the endpoint, variability, and/or its effect size. In studies with fixed sample sizes and some prior knowledge about treatment differences and variability, one may choose weights $\{w_i\}$ such that w_i is much greater than $\sum_{j=i+1}^{K} w_j$. We will then have that:

$$\frac{p_1}{w_1} \leq \frac{p_2}{w_2} \leq \ldots \leq \frac{p_K}{w_K}, \text{ i.e., } p_{(wi)} = p_{wi} = \frac{p_i}{w_i}.$$

According to the sequentially rejective weighted procedure, $P_{(wi)}$ is compared sequentially with $\alpha/\sum_{j=i}^{K} w_j$ or p_i is sequentially compared to $w_i\alpha/\sum_{j=i}^{K} w_j$. Since w_i is much greater than $\sum_{j=i+1}^{K} w_i$, the procedure is equivalent to comparing p_i to α. Thus a new sequentially rejective Bonferroni procedure rejects H_{0i} provided $p_j < \alpha$ for all $j \leq i$. This procedure is similar to a procedure developed by

Williams [14]. The advantage of the modified sequentially rejective procedure is that it utilizes prior knowledge of the first and second endpoints and it uses the full α-level for each endpoint. This procedure orders the endpoints according to their relative importance. One should avoid this procedure if one has little confidence in the prior knowledge. For example, when evaluating a bronchodilator drug for asthma, one may expect treatment effects in FEV_1 and PEFR, but might have less confidence in showing a treatment effect in the symptom score. Thus one may test, stepwise, FEV_1, PEFR, and symptom score in this "prespecified" order, each at the full α-level.

Discussion of α-Adjustment Methods

Many options are available to control the experimentwise error rate. The only requirement is that a set of test procedures yield valid p-values for each endpoint. Application of the Bonferroni method allows a decision to be made on the individual hypotheses of interest as well as a decision on the overall global hypothesis. Tarone [15] presented a modified Bonferroni procedure for discrete data. The major drawback of the Bonferroni procedure is its conservativeness and lack of power, particularly when the endpoints are correlated and none of the endpoints is highly significant.

Holm's Weighted Sequential Bonferroni procedure gives the option to weight the individual hypotheses and proportionally allocate the α-level to account for the relative "importance" of the individual endpoint. When the decision on which endpoints are primary is not clear, equal weights may be applied across endpoints.

Because the sequential methods discussed in this section are easy to apply and less conservative than the classic Bonferroni procedure, they are preferable for α-level adjustment methods.

Critical Value Adjustments

Often the results of a clinical trial are summarized by presenting the between treatment group difference for each endpoint, along with the confidence interval for the difference. Critical value adjustments can be used to calculate adjusted confidence intervals for all endpoints in order that the set of confidence intervals has the nominal coverage probability.

Let T_i be the test statistic for H_{0i} such that the T_i's have the same distribution. The T_i's, for example, could be Z- or t-statistics depending on whether the variances are known or unknown. We reject H_0 if some $|T_i|$ is larger than a critical value C, where C depends on the α-level and the method of adjustment used.

Westfall and Young Adjustment

Suppose the critical value C_{wy} is chosen such that

$$PR(|T_1| > C_{wy} \text{ or } \ldots, |T_K| > C_{wy} \mid H_0) = \alpha$$

or equivalently,

Table 3 Comparison of Critical Value Adjustment Methods

		Bonferroni; Mantel; Tukey-Ciminera-Heyse (TCH)			Westfall & Young		
K	Test	Bonferroni	Mantel	TCH	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
2	C	2.241	2.236	2.101	2.237	2.212	2.108
	α_a	0.0250	0.0253	0.0356	0.0254	0.0270	0.0350
5	C	2.576	2.569	2.279	2.568	2.511	2.274
	α_a	0.0100	0.0102	0.0227	0.0102	0.0120	0.0230
10	C	2.807	2.800	2.407	2.798	2.716	2.383
	α_a	0.0050	0.0051	0.0161	0.0052	0.0066	0.0172

C is the critical value at $\alpha = 0.05$ level, α_a is the corresponding adjusted critical level, K is the number of endpoints, and ρ is the pairwise correlation of the endpoints.

$$PR(\max) |T_i| > C_{wy} \mid H_0) = \alpha. \tag{2}$$

The above testing strategy with critical value C_{wy} produces the Westfall and Young [16,17] adjusted boundary C_{wy}. When the conditional distribution of T_i given H_0 is normal with mean 0 and variance 1, $C_{wy} \geq z_{1-\alpha/2}$. The adjusted α-level for an individual test, α_a, is equal to $PR(|T_i| > C_{wy} \mid H_0)$, and usually $\alpha_a < \alpha$. In general, it is not easy to solve equation (2) for C_{wy}, especially when K is large. The left side probability depends on the joint distribution of T_i ($i = 1, 2, \ldots, K$) and hence the correlation structure of the endpoints. To simplify the calculation, approximation methods involving only the marginal distribution could be used to approximate the probability. For example, setting $\alpha_a = \alpha/K$ yields the Bonferroni adjustment.

Mantel Adjustment

Mantel [18] suggested a simple critical value adjustment that is slightly less conservative than the Bonferroni adjustment: $\alpha_a = 1 - (1 - \alpha)^{1/K}$. The Mantel adjusted boundary C_m satisfies

$$PR(|T_i| > C_m \mid H_0) = 1 - (1 - \alpha)^{1/K}. \tag{3}$$

Tukey, Ciminera, and Heyse (TCH) Adjustment

The Mantel adjustment is appropriate when the endpoints are independent. When the endpoints are correlated but the correlation is unknown, Tukey, Ciminera, and Heyse [19] suggested replacing $1/k$ by $1/\sqrt{K}$ in equation (3): $\alpha_a = 1 - (1 - \alpha)^{1/\sqrt{K}}$. The TCH adjusted α-level is always greater, and hence less conservative, than that of Mantel. Computations of the adjusted critical values for these approximation methods are simple. They involve only the marginal distributions and are independent of the correlations among endpoints. Therefore, they do not guarantee the prespecified experimentwise error rate.

Table 3 presents the adjusted critical value and the corresponding adjusted α-level for the methods discussed above, assuming $T_1, T_2, \ldots, T_K$ have a joint multivariate normal distribution. For the Westfall and Young method, ρ is the

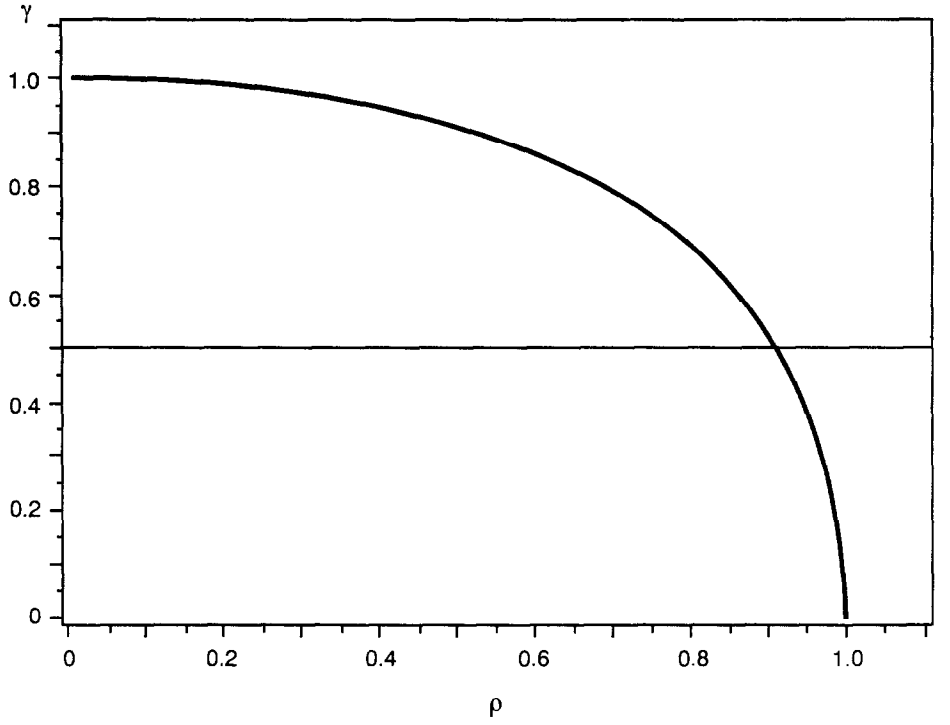

Figure 1 Relationship between the exponent γ and the correlation ρ. The adjusted α level is computed as $\alpha_a = 1 - (1 - \alpha)^{K^{-\gamma}}$, for K = 2 endpoints with correlation ρ.

pairwise correlation. The Bonferroni and Mantel adjustments are very similar. Both the Bonferroni and Mantel adjustments are not very conservative when the correlations among endpoints are fairly low (<0.5). The TCH adjustment is applicable when the correlations are very high. For moderate correlations, the exponent of K should be chosen to be somewhere between $-1/2$ and -1, i.e., $\alpha_a = 1 - (1 - \alpha)^{K^{-\gamma}}$, $1/2 \leq \gamma \leq 1$, where γ depends on the correlation structure. For example see Figure 1.

P-Value Adjustments

With the univariate procedure, the *p*-value could be adjusted in a similar fashion as the critical value adjustment. The adjusted *p*-values are always compared at the full α-level. Using the *p*-values form each endpoint, p_i, we can define the following four types of adjusted *p*-values which relate to the individual hypotheses.

Westfall and Young (W&Y) p-*value Adjustment*

The exact adjusted *p*-value for H_{0i} or the i^{th} adjusted *p*-value is defined by

$$p_{ai} = PR(\min P_j \leq p_i \mid H_0) = PR(\max |T_j| > |t_i| | H_0)$$
$$= PR(|T_1| > |t_i|, \text{ or, } \ldots, |T_K| > |t_i| | H_0)$$
$$= 1 - PR(|T_1| \leq |t_i|, \text{ and, } \ldots, |T_K| \leq |t_i| | H_0).$$

Table 4 Adjusted *p*-Values for K=2 Endpoints and a *t* statistics (T_1) of 2.2

ρ	Bonferroni	Mantel	TCH[a]	W & Y[b]
0.0	0.0556	0.0548	0.0391	0.0548
0.3	0.0556	0.0548	0.0391	0.0537
0.5	0.0556	0.0548	0.0391	0.0515
0.7	0.0556	0.0548	0.0391	0.0476
0.9	0.0556	0.0548	0.0391	0.0401
1.0	0.0556	0.0548	0.0391	0.0278

[a] Tukey, Ciminera and Heyse.

[b] Westfall and Young.

The adjusted *p*-value, a function of the i^{th} test statistic value, depends on the joint distribution of the T_i, i = 1, 2, . . ., K.

Bonferroni and Mantel p-value Adjustments

Bonferroni-adjusted *p*-value is $p_{ai} = \min (Kp_i, 1)$. Mantel-adjusted *p*-value is $p_{ai} = 1 - \{1 - PR(|T_i| > |t_i||H_0)\}^K = 1 - (1 - p_i)^K$.

Tukey, Ciminera, and Heyse (TCH) p-value Adjustment

Replacing K in the Mantel *p*-value adjustment with $K^{1/2}$ yields $p_{ai} = 1 - (1 - p_i)^{\sqrt{K}}$.

Table 4 compares the adjusted *p*-values with varying levels of correlation for the case of two endpoints (K = 2). The joint distribution of the endpoints is bivariate normal with unit variances and correlation coefficient ρ, assuming $T_1 = 2.2$. Bonferroni, Mantel, and the TCH adjusted *p*-values are independent of the correlation ρ. With significance level equal to 0.05, both the Bonferroni and Mantel approaches fail to reject H_0, while the TCH approach does reject H_0 in all cases. The exact *p*-value (W&Y) decreases from 0.0548 to 0.0278 as ρ is increased from 0.0 to 1.0, rejecting H_0 when ρ approaches 0.7.

Computational Considerations

As mentioned before, computation of the Westfall and Young adjusted *p*-values requires calculating the probability, $PR(|T_1| > |t_1|, or, . . ., |T_K| > |t_1||H_0)$, which depends on the K-dimensional joint distribution of the T_i. We can also use resampling Monte Carlo methods [17]. Suppose $X_1, . . ., X_{n_1}, Y_1, . . ., Y_{n_2}$ are the observed set of efficacy vectors from two treatments.

Bootstrap Method

1. Sample $(X_1^*, . . ., X_{n_1}^*)$, $(Y_1^*, . . ., Y_{n_2}^*)$ with replacement from the pooled sample space $X_1 - \bar{X}, . . ., X_{n_1} - \bar{X}$ and $Y_1 - \bar{Y}, . . ., Y_{n_2} - \bar{Y}$, where $\bar{X}$ and $\bar{Y}$ are the sample mean of X and Y, respectively;
2. Compute the test statistic T_i^* or the unadjusted *p*-value p_i^* (i = 1, . . ., K) with the pseudo data $(X_1^*, . . ., X_{n_1}^*)$, $(Y_1^*, . . ., Y_{n_2}^*)$;

Table 5 Results of Multiplicity Adjustments Using the Global Methods

Global methods		Overall p-value	Reject Global H_0
O'Brien[a]	Rank-sum	0.0002	Yes
	Ordinary least squares	0.0005	Yes
	Generalized least squares	0.0003	Yes
Hotelling's T^2		0.0001	Yes
Simes[b]			Yes

[a] Rank-sum (using two-sample t), T=3.88, df=67, Ordinary least squares, T=3.45, df=61, Generalized least squares, T=3.64, df=61.

[b] The individual p-values for FEV_1, PEFR, symptom score and additional medication use: 0.0037, 0.0077, 0.0274, and 0.0369 are compared to 0.0125, 0.0250, 0.0375, and 0.0500, respectively.

3. Note whether max $|T_j^*| > |t_i|$ or min $p_j^* < p_i$;
4. Repeat steps 1–3 an adequate number of times, e.g., 499 or 999 times;
5. The bootstrap estimate of p_{ai} is the proportion of bootstrap samples such that max $|T_j^*| > |t_i|$ or min $p_j^* < p_i$.

Permutation Method

The adjusted p-values could also be obtained as the proportion of permutations of the observed vector $(X_1, \ldots, X_{n1}, Y_1, \ldots, Y_{n2})$ for which the minimum of the observed univariate p-values for each endpoint in a permutation sample is smaller than the observed unadjusted p-value. This probability may be computed exactly using the multivariate hypergeometric distribution, but that is computationally complex. Alternatively, one may estimate adjusted p-values as follows:

1. Generate a set of random permutations of the pooled observed data;
2. For each of the permutations, calculate the minimum univariate p-value;
3. The estimated adjusted p-value is the proportion of the permutation samples for which the minimum p-value is less than the particular original p-value, p_i.

RESULTS

Example, Continued

In this section we apply the methods presented in the previous sections to our example. Without multiplicity adjustments, the individual null hypotheses for all four endpoints were significant at $\alpha = 0.05$ level. Table 5 presents the results of the analyses using each of the five global methods for testing the overall hypothesis of no treatment effect.

All of the global assessment procedures indicate a positive treatment effect. To determine which endpoints show a significant difference between treatments, one may apply the closed testing principle. For example, for O'Brien's method (rank-sum procedure), test of each intersection null hypothesis containing three of the four endpoints yields $p < 0.05$ for each hypothesis. Since all intersection hypotheses containing three endpoints are rejected, each of the intersection hypotheses containing two of the four endpoints can be tested,

Table 6 Results of Multiplicity Adjustments Using α-level and *p*-value Adjustment

Method		FEV_1	PEFR	Symptom Score	AMU[a]	Reject Global H_0
t statistic	*p*-values	0.0037	0.0077	0.0274	0.0369	
Bonferroni	α_a	0.0125	0.0125	0.0125	0.0125	
	Sig.?	Yes	Yes	No	No	Yes
Holm	α_a	0.0125	0.0170	0.0250	stop	
	Sig.?	Yes	Yes	No	No	Yes
Holm with	α_a	0.0100	0.0125	0.0333	0.0500	
Weights[b]	Sig.?	Yes	Yes	Yes	Yes	Yes
Mantel[c]	p_a-values	0.0151	0.0303	0.1064	0.1394	
	Sig.?	Yes	Yes	No	No	Yes
TCH[c]	p_a-values	0.0076	0.0153	0.0547	0.0723	
	Sig.?	Yes	Yes	No	No	Yes
W & Y[c,d]	p_a-values	0.0101	0.0219	0.0843	0.1121	
	Sig.?	Yes	Yes	No	No	Yes

[a] AMU: Additional medication use.

[b] Symptom score was deemed more important and given weight=2, and the remaining endpoints were given equal weights of 1.

[c] All adjusted *p*-values are compared to the 0.050 level.

[d] The correlations from Table 2 were used.

which yields $p < 0.05$ for each hypothesis. We then can proceed to test all the elementary hypotheses, which yields $p < 0.05$ for each individual endpoint. Therefore, the closed testing principle shows a significant difference between treatments for each of the four endpoints, while maintaining the experimentwise error rate at 0.05.

Simes' procedure allows an even easier application of the closed testing procedure since the test of the global null nypothesis requires only the calculation of individual endpoint *p*-values. Since the biggest *p*-value of the four endpoints was < 0.05, any hypothesis involving a subset of the endpoints will result in rejection of that hypothesis. Therefore, the closed testing principle applied to Simes' procedure leads to the same conclusion as O'Brien's procedures.

Thus, to reject an individual endpoint using any of the global methods with the closed multiple testing procedure, the *p*-value for that endpoint must be smaller than the given α-level. In other words, if all *p*-values for the prespecified set of individual endpoints are smaller than the given α-level, then all individual endpoints can be claimed statistically significant at the experimentwise error rate of α.

Table 6 presents the results using α-level adjustment methods. Each of the α-adjustment procedures also shows a significant overall treatment effect. Holm's weighted sequential method shows a significant difference between the test drug and placebo for all four endpoints, while all other α-adjustment methods find a significant difference between drug and placebo for two of the individual endpoints, FEV_1 and PEFR.

Table 6 also presents the results using *p*-value adjustment methods. As mentioned earlier, only the Westfall and Young method controls the experimentwise error rate under all correlation structures. The Westfall and Young

method indicates significant differences between treatment groups in FEV_1 and PEFR.

DISCUSSION

Many global test methods such as O'Brien's procedures introduce a univariate summary of index for each patient using multiple endpoints. Such a summary involves weighting the endpoints and using subjective biological and medical judgment. However, if it is appropriate to do so, many standard statistical methods for univariate variables can be applied. When using a global test, if the global hypothesis of no overall treatment difference is rejected, one would often like to determine which endpoints are significantly affected by treatment while maintaining the experimentwise type I error rate at α. One can proceed to the test of intersection and elementary null hypotheses as explained previously in the closed multiple test procedure. If, however, a closed testing procedure is not used, or the closed testing procedure is used but one cannot proceed to the test of elementary null hypotheses, one is left without any formal method for assessing the effects of treatment on individual endpoints. By contrast, the α-level and p-value adjustment methods provide inference on the individual endpoints, although they are usually less powerful.

The failure to identify differences between treatments for individual endpoints when an overall treatment difference has been identified may be due to insufficient statistical power. This problem underscores the value of powerful univariate tests of individual endpoints. Global test statistics are also limited in the sense that they emphasize significance testing rather than estimation.

When one chooses not to use a global testing procedure, an appropriate α-level, p-value, or critical value adjustment method should be used to control the experimentwise error rate. One should consider limiting the number of primary endpoints to three or four variables. Any remaining endpoints may be designated as secondary and exploratory, and these endpoints would not require a routine multiplicity adjustment. The secondary endpoints may then serve as supplemental evidence that provides added weight to the conclusions made from the primary variables or provide information for future research. As the α-level adjustment method, we recommend Holm's weighted or unweighted sequential procedure for it is easy to use and has higher power than the Bonferroni method. As the p-value adjustment method, we recommend Westfall & Young's p-value adjustment method with resampling techniques for its control of experimentwise error rate regardless of the correlations between endpoints. The corresponding Westfall & Young adjusted critical values can be used for constructing the proper confidence intervals.

Power of Multiple-Test Procedures

Although a multiple-test procedure can control the experimentwise error rate, assessment of the power of the study is more difficult. If the overall assessment of treatment difference is most important, the relevant power is that of the global test statistic. If the hypothesis corresponding to one endpoint is deemed most important, one may focus on the power of the test statistic for that hypothesis. Monte Carlo simulation methods can be easily used in most

cases to help evaluate the study power with a multiple test procedure; for example, see Zhang et al [20] and Capizzi et al [21].

General

We have reviewed some methods for addressing multiple endpoints in a clinical trial. Hochberg and Tamhane [22] presented well-organized statistical theory, concepts, and technical details of many other multiple-adjustment methods, and more sophisticated stepwise procedures with excellent illustrative examples, with a focus on multiple between treatment group comparisons. Tang et al [23] presented an approximate likelihood approach to the global assessment method. This overview points to the difficulty of specifying a unique or optimal statistical method or analysis strategy for studies with multiple endpoints. Rather, for each clinical trial, the data analyst must consider the study design, class of drug, disease type, the relationship between the test drug and the disease, and the relationship among endpoints. The protocol should state clear null hypotheses ad alternatives, and prespecified statistical methods and courses of action. Zhang et al [20] and Capizzi et al [21] gave several such examples in a context of asthma clinical trials.

When several endpoints are similar or measure different aspects of the same phenomenon or effect, there may be a suitable rationale for combining individual results to obtain an overall assessment of treatment effects. If so, the statistical methods suited to the study's purpose and condition should be investigated. When endpoints are not suitable for combining, adjustment of α-level, p-value, and critical value provide methods of inference for individual endpoints while maintaining the experimentwise error rate.

We thank Jim Bolognese, Tom Capizzi, Keith Soper, Joe Heyse, and especially, Janet Wittes and the referees for their helpful suggestions, which led to this improved version.

REFERENCES

1. National Asthma Education Program Expert Panel Report. *Guidelines for the Diagnosis and Management of Asthma.* Bethesda, MD: National Institutes of Health; 1991: Publication No. 91-3042.
2. Bauer P. Multiple testing in clinical trials. *Stat Med* 1991;10:871–890.
3. Hommel G. Multiple test procedures for arbitrary dependence structures. *Metrika* 1986;33:321–336.
4. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Statis* 1979;6:65–70.
5. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976;63:655–660.
6. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984;40:1079–1087.
7. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987;43:487–498.
8. Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* 1991;47:511–521.

9. Follmann D, Wittes J, Cutler JA. The use of subjective rankings in clinical trials with an application to cardiovascular disease. *Stat Med* 1992;11:427–437.
10. Tandon PK. Applications of global statistics in analyzing quality of life data. *Stat Med* 1990;9:819–827.
11. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986;73:751–754.
12. Hommel G. A stepwise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988;75:383–386.
13. Shaffer JP. Modified sequentially rejective multiple test procedures. *J Amer Stat Assn* 1986;81:826–831.
14. Williams DA. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics* 1971;27:103–117.
15. Tarone RE. A modified Bonferroni method for discrete data. *Biometrics* 1990;46:515–522.
16. Westfall PH, Young SS. p-Value adjustments for multiple tests in multivariate binomial models. *J Amer Stat Assn* 1989;84:780–786.
17. Westfall PH, Young SS. *Resampling-based Multiple Testing*. New York: John Wiley & Sons, Inc; 1993.
18. Mantel N. Assessing laboratory evidence for neoplastic activity. *Biometrics* 1980;36:381–399.
19. Tukey JW, Ciminera JL, Heyse JF. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics* 1985;41:295–301.
20. Zhang J, Shingo S, Reiss TF, Friedman BS, Capizzi TP. Statistical issues in the design and analysis of clinical trials evaluating therapy for chronic asthma. *Controlled Clin Trials* 1993;14:445.
21. Capizzi TP, Zhang J. Testing the hypothesis that matters for multiple primary endpoints. *Amer Stat Assn Biopharmaceutics Section Proceedings* 1994, pp. 460–465.
22. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. New York: John Wiley & Sons, Inc: 1987.
23. Tang DI, Gnecco C, Geller, NL. An approximate likelihood ratio test for a normal mean vector with nonnegative components with application for clinical trials. *Biometrika* 1989;76:577–583.

APPENDIX

Notation:

$I \subset \{1, \ldots, K\}$: I consists of some integer numbers between 1 and K.

$H_{0i} \in \{H_0\}$: Hypothesis H_{0i} is one of the set of hypotheses $\{H_0\}$.

$H_{0i} \cap H_{0j}$: A new hypothesis that both H_{0i} and H_{0j} holds.

(A, B]: An interval of numbers that are greater than ($>$) A, less than or equal to ($\leqslant$) B.

[A, B]: An interval of numbers that are $\geqslant$ A and $\leqslant$ B.