STATISTICS IN MEDICINE, VOL. 16, 2529–2542 (1997)

SOME COMMENTS ON FREQUENTLY USED MULTIPLE ENDPOINT ADJUSTMENT METHODS IN CLINICAL TRIALS*

A. J. SANKOH[†], M. F. HUQUE AND S. D. DUBEY

Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Rockville, Maryland 20857, U.S.A.

SUMMARY

Confirmatory clinical trials often classify clinical response variables into primary and secondary endpoints. The presence of two or more primary endpoints in a clinical trial usually means that some adjustments of the observed *p*-values for multiplicity of tests may be required for the control of the type I error rate. In this paper, we discuss statistical concerns associated with some commonly used multiple endpoint adjustment procedures. We also present limited Monte Carlo simulation results to demonstrate the performance of selected *p*-value-based methods in protecting the type I error rate. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 2529–2542 (1997)
No. of Figures: 0 No. of Tables: 5 No. of References: 22

1. INTRODUCTION

Except for a cautionary word in the 'Guidelines for the Format and Content of the Clinical and Statistical Section', there are currently no formal Agency (FDA) wide standard statistical guidelines for the assessment of efficacy evidence in the presence of multiplicity. Sponsors and reviewers often select multiplicity adjustment methods that tend to 'optimize' the chances of demonstrating clinical evidence in favour of their preconceptions regarding therapeutic benefit. This *ad hoc* selection process (at the analysis or review stage) is often done with little regard to the appropriateness of the methods to the specific objectives of the clinical trial (for example, whether to establish clinical superiority or equivalence). We believe that, to promote good clinical trial practices, there is a need to maintain uniformity in the application of these methods in the assessment of efficacy evidence in clinical trials with similar objectives and characteristics. To achieve this goal, a certain measurable level of acceptance and understanding of the problem of multiplicity and its impact on the assessment of efficacy evidence are needed.

In this overview, we try to promote such understanding by: (i) providing a general description of the problem of multiplicity in clinical trials; (ii) describing the types of error rate control approaches; (iii) outlining frequently used multiplicity adjustment procedures in the assessment of

* The views expressed here are those of the authors and not necessarily of the U.S. Food and Drug Administration.

[†] Correspondence to: A. J. Sankoh, Centre for Drug Evaluation and Research, Division of Biometrics, HFD 720, U.S. Food and Drug Administration, 5600 Fishers Lane, 9B-07, Rockville, MD 20857, U.S.A.

CCC 0277–6715/97/222529–14$17.50

Received December 1995
Revised February 1997

efficacy evidence in clinical trials, and (iv) presenting limited simulation results that clearly indicate the need for caution in using some of these procedures.

2. A GENERAL DESCRIPTION OF MULTIPLICITY IN CLINICAL TRIALS

The essence of interpretation in a clinical trial is consistency from both a clinical and a statistical perspective. The interpretation of and the conclusions drawn from the results of a clinical trial depend on a number of factors, including the disease under study, the patient population, the endpoints, the study design, the conduct of the study, the appropriateness of the statistical analysis for the given design and the sensitivity of the chosen statistical test(s) to the scientific question(s) the study seeks to investigate. One of the many factors that may make such interpretation difficult and sometimes impossible is the presence of multiplicity in a clinical trial that is unaccounted for in the design and ensuing statistical analyses. Multiplicity may enter into the design and analysis of a clinical trial from many different sources. Some of these sources are:

1. Multiple endpoints: the nature of the disease and the types of questions a clinical trial aims to investigate may necessitate multiple endpoints. For instance, repeated visits and/or measurements may be necessary for therapeutic assessments, one may need to demonstrate therapeutic benefit on several levels that correspond to the multi-dimensional appearance of a disease, there may be no consensus on the single most appropriate measure of therapeutic benefit, or it may not be practical to evaluate the therapeutic effect on a unified scale. A clinical trial to assess the therapeutic benefit in unstable angina patients, for example, could include efficacy endpoints such as mortality, myocardial infarction, urgent or emergency coronary revascularization etc., all of which are of primary clinical interest.
2. Multiple studies and/or multiple active arms: the need to demonstrate efficacy of a drug from at least two independent pivotal studies (from different centres), or to have more than one active arm (confirmatory dose ranging studies).
3. Multiple analyses and/or tests: the practice of providing evaluable subset analyses (variably dubbed as evaluable subset, per protocol subset, protocol defined subset) in addition to the all randomized (intent-to-treat) or all patients treated analyses. In many trials it is appropriate to present both analyses. What is usually lacking is a thorough discussion and interpretation of the salient differences between the two analyses. An intent-to-treat analysis is more fitting in a comparative study that seeks to demonstrate the superiority of one drug over another while a per protocol analysis may be more appropriate for a trial that seeks to show the equivalence of two drugs. Because of the heterogeneity of patient populations, subset analysis is often an important component of the report of major clinical trials. In most cases, however, one should regard such subset analysis results as hypothesis generating with findings to be tested further in other studies.

Similarly, the need to minimize the cost of obtaining data (interim analyses), to explore the many alternative statistical methods (preliminary tests), or the desire to discover new aspects of the data (subset analyses) are necessary ingredients in the conduct of a clinical trial that could also introduce multiplicity.

Thus, we see that there is an inherent multiplicity component in most clinical trials, and that it is practically impossible to conduct a clinical trial without introducing or encountering some or all of these sources of multiplicity. For instance, in a recent new drug application (NDA) submission for the symptomatic treatment of gastroesophageal reflux disease (GERD), the trial was designed as an international (from eight different participating countries spanning three continents), multi-study, multi-centre, multi-dose (four including placebo), and multiple primary endpoints (at least five per study) trial. In addition, measurements were repeatedly taken at different time points (at least 10 per study) and analyses were carried out at each of these time points. Given that this is not an isolated case in the design and analysis of clinical trials, the need to address multiplicity issues adequately in clinical trials cannot be overemphasized.

Repeated analyses of the same clinical data, if not accounted for in the ensuing analysis, increases the odds of finding significant results in favour of the alternative hypothesis when the null hypothesis is in fact true. The debate on multiplicity (or multiple endpoints, as is the focus here), therefore, is not whether to accept or not to accept the existence of multiple endpoints in a given clinical trial, or to adjust or not to adjust the resulting *p*-values for such multiplicity. It is clear that multiplicity issues pervade many clinical trials, and, therefore, such adjustment is necessary if one is to maintain good clinical trial practices. The multiplicity debate is on the appropriateness of the adjustment procedure one employs in a given situation.

3. ERROR RATES AND ERROR RATE PROTECTION

An error rate is a probabilistic measure of erroneous inferences in a given set of hypotheses from which some common conclusions are drawn. In the discussion that follows, by an error rate we mean the probability of false rejection of either the null hypothesis (referred to as type I error rate), or the alternative hypothesis (referred to as type II error rate). From a drug development perspective, an error rate of the first type occurs when an inefficacious drug gains entrance into the market; an error rate of the second type occurs when an efficacious therapy fails to reach the market.

There are two approaches to controlling the probability of false rejection of at least one of K null hypotheses in a multiple hypothesis testing problem. An experimenter who opts to control the individual type I error rates at a given nominal α-level for each of K hypotheses is controlling what is called the comparisonwise error rate. One who opts to control the overall type I error rate at a given nominal α-level for all possible hypotheses is controlling what is known as the experimentwise error rate. As an example, consider the case of $K=3$ and a prespecified nominal α-level of 0·05. A comparisonwise type I error rate method of control implies testing each of the three corresponding null hypotheses at the 0·05 nominal level, to ensure a 0·05 significance level for each test. An experimentwise type I error rate method of control, on the other hand, implies testing each of the three null hypotheses at some α-level so as to ensure an overall significance level of 0·05 or less for all three tests (for example, testing each null hypothesis at $\alpha=0{\cdot}017$, the Bonferroni level of significance).

The debate on multiple endpoint adjustments revolves around these two approaches to error rate protection. The trade-offs associated with either of these choices are well known. A well-known disadvantage associated with the comparisonwise error rate control approach is the increased chance of a higher experimentwise type I error rate. As the number of comparisons increases, the probability of at least one type I error also increases. For example, if 20 independent tests are done each at the $\alpha=0{\cdot}05$ level of significance, on average one test result will be significant even if the null hypothesis is true. The probability that at least one is significant is $1-(1-0{\cdot}05)^{20}=0{\cdot}642$. It is also known that as the type I error rate increases, the type II error rate decreases; this gives rise to more powerful tests (an experimenter's delight!). On the other hand, controlling the experimentwise error rate leads to an overall type I error rate that is less than or equal to a designated α-level. The downside of this approach (at least from an experimenter's perspective) is that the resulting experimentwise type I error rate is usually smaller than the designated α-level and this may lead to less sensitive tests (an experimenter's dismay!).

There are other known advantages associated with controlling the experimentwise (instead of the comparisonwise) error rate. Controlling the experimentwise error rate does not depend on what null hypotheses are true. Since it is clinically not practical to require realization of improvement in every component of a global test in a given clinical trial, control of the experimentwise error rate seems to be the more appropriate option. Furthermore, a multiple test procedure that controls the experimentwise error rate (in the strong sense) also controls the global error rate (that is, the probability that at least one null hypothesis is erroneously rejected when all individual null hypothese are simulatenously true). The converse is not true.[1]

D'Agostino and Hereen[2] provoked discussions of the complexity of this problem, especially as applied to the design and analysis of clinical trials for over-the-counter (OTC) drugs. They recommended that in a clinical trial where the establishment of therapeutic equality is the primary interest, power consideration should supersede type I error rate protection consideration. That is, one should prefer (the more powerful) multiple comparison procedures that control the comparisonwise error rate. For situations where the establishment of therapeutic superiority is the primary objective, protection of the type I error rate should take centre stage, and thus one should prefer adjustment procedures that control the experimentwise error rate.

4. SOME STATISTICAL APPROACHES FOR MULTIPLICITY ADJUSTMENT

Although the general consensus in the clinical and statistical community regarding multiplicity adjustment is that, for many clinical trials, some adjustment is necessary, there are a few dissonant voices.[3–5] Some of these dissenting voices argue that since no multiplicity adjustment is required in a clinical trial that compares a series of small experiments each involving a single comparison, no such adjustment should be required in a clinical trial with a single large experiment involving multiple comparisons. Others point out that the patient is the experimental unit of a clinical trial, and that to assess adequately the patients' response to treatment, it is necessary to look at different measures that adequately characterize the experimental unit. To penalize the experimenter for such a practice is tantamount to a reprimand for practising good clinical science, they argue.

Our objective in this paper is to provide a few caveats on the use of some popular multiple endpoint adjustment procedures proposed by various authors, and frequently used in the adjustment of p-values for multiplicity due to multiple endpoints in clinical trial NDA submissions. We discuss some of these procedures in the next few sections.

4.1. *p*-value-based procedures

There are two primary reasons why p-value-based procedures have enjoyed lasting popularity in the clinical trial community: (i) One can report p-value results obtained from analysing different types of data (continuous, categorical, survival) using different test statistics (t, χ^2, logrank) unambiguously, and (ii) regulatory agencies generally rely on p-values as a measure of the strength of evidence for therapeutic benefit. Multiplicity test procedures that fall in this category include the Bonferroni test and its improvements. These improvements are commonly referred to as stepwise procedures.[6–9]

4.2. The Bonferroni procedure

The Bonferroni test is the simplest multiplicity adjustment procedure and is applicable to most multiplicity adjustment problems. There are no assumptions required about the data distributions and/or about the correlation structure among the endpoints. For K endpoints, one accepts as statistically significant all those p-values $\leqslant \alpha/K$ where α is the overall type I error rate. The adjusted p-values are Kp_k where the p_k's are the observed unadjusted p-values, $k=1, 2, \ldots, K$.

This procedure ignores most of the information from the data and the correlation structure of the endpoints. It is too conservative when there are many endpoints and/or the endpoints are highly correlated. There are a number of 'less conservative' and more powerful improvements to the Bonferroni procedure also known to control the experimentwise error rate. These approaches are based on the realization that of the K null hypotheses tested, the only ones to protect against rejection (at a given step) are those not yet rejected. Among these is the Holm's procedure[6] that satisfies the Marcus *et al.* 'closure' property but is still a Bonferroni-like test.[10] There are also modifications and improvements of the Holm's procedure that are less conservative than Holm's. The Hochberg[8] (closed under independence) and Hommel[9] procedures seem to perform reasonably well when the number of endpoints is small and the endpoints are all positively correlated with small to mid-size correlations.

4.3. Bonferroni modified (stepwise) procedures

Holm[6] introduced a step-down adjustment procedure that improved the Bonferroni method by providing additional power while maintaining the experimentwise error rate. In a step-down procedure one conducts the testing in a decreasing order of significance of the (ordered) hypotheses. Significance testing continues until one accepts a null hypothesis. Then one accepts all remaining (untested) null hypotheses without further testing.

The algorithm for Holm's procedure is:

1. Let $p_1 > p_2 > \ldots > p_K$ be the ordered p-values and $H_{01}, H_{02}, \ldots, H_{0K}$ be the corresponding ordered null hypotheses.
2. Reject H_{0K} if $p_K < \alpha/K$ and go to the next step, otherwise stop and accept all null hypotheses.
3. Reject $H_{0(K-1)}$ if $p_{(K-1)} < \alpha/(K-1)$ and go to the next step, otherwise stop and accept all null hypotheses H_{0j}, $j=1, 2, \ldots, K-1$.
4. In general, reject H_{0k} if $p_j < \alpha/j$, otherwise stop and accept all null hypotheses H_{0k}, $k=1, 2, \ldots, j$. The adjusted p-values are $p_{\mathrm{ak}} = \max\{Kp_K, (K-1)p_{(K-1)}, \ldots, kp_k\}$, $k=1, 2, \ldots, K$.

Based on Simes' global method,[7] Hochberg[8] derived a step-up procedure for testing each intersection hypothesis in the closed sense of Marcus *et al.* In a step-up procedure one conducts the testing in an increasing order of significance of the (ordered) hypotheses. Significance testing continues until one rejects a null hypothesis. Then one rejects all remaining (untested) null hypotheses without further testing.

The algorithm for Hochberg's procedure is: let $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_K$ be the ordered p-values and $H_{01}, H_{02}, \ldots, H_{0K}$ be the corresponding null hypotheses. Reject H_{0k} and all H_{0j} for $j \leqslant k$ if $p_k \leqslant \alpha/k$. The adjusted p-values are $p_{\mathrm{ak}} = \min\{p_1, 2p_2, \ldots, kp_k\}$ for $k=1, 2, \ldots, K$.

Again, applying the closure principle, Hommel[9] modified Simes' procedure as follows: let $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_K$ be the ordered p-values and $H_{01}, H_{02}, \ldots, H_{0K}$ be the corresponding null hypotheses. Find the largest m for which $p_1 > \alpha$; $p_1 > \alpha$, $p_2 > \alpha/2$; $p_1 > \alpha$, $p_2 > 2\alpha/3$, $p_3 > \alpha/3$; … ; $p_1 > \alpha$, $p_2 > \alpha(m-1)/m$, $p_3 > \alpha(m-2)/m$, … , $p_m > \alpha/m$. Then reject H_{0k} for which $p_k < \alpha/m$. The adjusted p-values are $p_{ak} = mp_k$, $k = 1, 2, \ldots, K$.

For example, suppose for $K = 4$ endpoints, we observe the following (ordered) p-values: $p_1 = 0{\cdot}081 > p_2 = 0{\cdot}024 > p_3 = 0{\cdot}020 > p_4 = 0{\cdot}005$.

Under the Holm step-down procedure: (i) we reject H_{04} since $p_4 = 0{\cdot}005 < 0{\cdot}013 = 0{\cdot}05/4$ and go to the next step; (ii) but since $p_3 = 0{\cdot}020 > 0{\cdot}017 = 0{\cdot}05/3$, we stop and accept H_{03} and all remaining (untested) null hypotheses H_{0j} $(j = 1, 2)$. Here we accept all null hypotheses H_{0k} $(k = 1, 2, 3)$ since $p_3 > \alpha/3$. The adjusted p-values are $p_{ak} = \max\{4p_4, 3p_3, \ldots, kp_k\}$, $k = 1, 2, 3, 4$. That is, $p_{a1} = 0{\cdot}081$, $p_{a2} = 0{\cdot}060$, $p_{a3} = 0{\cdot}060$ and $p_{a4} = 0{\cdot}020$.

Under the Hochberg step-up procedure: (i) we accept H_{01} since $p_1 = 0{\cdot}081 > 0{\cdot}05$ and go to the next step; (ii) but since $p_{02} = 0{\cdot}024 < 0{\cdot}025 = 0{\cdot}05/2$, we stop and reject H_{02} and all remaining (untested) null hypotheses H_{0k} $(k = 3, 4)$. The adjusted p-values are $p_{ak} = \min\{p_1, 2p_2, \ldots, kp_k\}$, $k = 1, 2, 3, 4$. That is, $p_{a1} = 0{\cdot}081$, $p_{a2} = 0{\cdot}048$, $p_{a3} = 0{\cdot}048$ and $p_{a4} = 0{\cdot}020$.

Under the Hommel procedure, since $p_1 = 0{\cdot}081 > 0{\cdot}05$, $m = 1$. We reject all H_{0k} for which $p_k < 0{\cdot}05$. We therefore reject H_{02}, H_{03} and H_{04} (as in the Hochberg procedure). The adjusted p-values are $p_{ak} = mp_k = p_k$, $k = 1, 2, 3, 4$. That is, $p_{a1} = 0{\cdot}081$, $p_{a2} = 0{\cdot}024$, $p_{a3} = 0{\cdot}020$ and $p_{a4} = 0{\cdot}005$.

Note that in this example, both the Hochberg and the Hommel procedures lead to two more null hypothesis rejections (H_{02} and H_{03}) than does the Holm procedure. Hommel's procedure has been shown to be only slightly more powerful than Hochberg's, which is uniformly more powerful than Holm's, which is uniformly more powerful than the Bonferroni procedure.[11,12]

4.4. *Ad hoc* procedures

Easy-to-use *ad hoc* procedures that make use of the correlation information among the endpoints without any distributional assumptions have been developed. Although there has been little theoretical work to assess the performance of these approaches, they do lend themselves to simulation assessments.[13–15] For strongly correlated K endpoints and for a given nominal critical level α, Tukey *et al.*[13] suggested the adjustments $p_{ak} = 1 - (1 - p_k)^{\sqrt{K}}$ and $\alpha_k = 1 - (1 - \alpha)^{1/\sqrt{K}}$ where p_k and p_{ak} are, respectively, the observed and adjusted kth p-values, and α_k is the adjusted critical α-level for the kth hypothesis for $k = 1, \ldots, K$ (henceforth referred to as the TCH procedure). Dubey[14] and Armitage–Parmar[15] suggested the use of

$$p_{ak} = 1 - (1 - p_k)^{m_k} \quad \text{and} \quad \alpha_k = 1 - (1 - \alpha)^{1/m_k}$$

where

$$m_k = K^{1 - r_{.k}} \quad \text{and} \quad r_{.k} = (K - 1)^{-1} \sum_{j \neq k}^{K} r_{jk}$$

replaces $\sqrt{K}$ in the TCH formula for p_{ak}; r_{jk} is the correlation coefficient between the jth and kth endpoints (henceforth referred to as the D/AP procedure).

One of the nice properties of the D/AP procedure is that when the average of the correlation coefficients is zero, the adjustment is according to the Bonferroni test, and when it is one, the adjusted and the unadjusted p-values are the same. Note also that for equi-correlated endpoints with correlation coefficient $= 0{\cdot}5$, the D/AP procedure is equivalent to the TCH procedure.

Table I gives the results of using these procedures to adjust sponsor's reported two-sided p-values in support of a claim for the effectiveness of OTC Lactase® in the prevention/reduction

Table I. A Comparison of adjusted p-values by five (5) p-value-based methods for $K=7$ endpoints

Endpoint k	$r_{.k}$	p_k	Adjusted p-values (p_{ak})					
			D/AP	TCH	Hochb	Homm	Holm	Bonferr
1. ACs	0·4249	0·0099	0·0300	0·0260	0·0495	0·0495	0·0495	0·0693
2. Bloating	0·3652	0·0879	0·2712	0·2161	0·1758	0·4595	0·1758	0·6153
3. Belching	0·2378	0·0162	0·0694	0·0423	0·0648	0·0810	0·0648	0·1134
4. Flatulence	0·3883	0·0008	0·0026	0·0021	0·0056	0·0040	0·0056	0·0056
5. BMs	0·4709	0·0552	0·1470	0·1395	0·1656	0·2760	0·1656	0·3864
6. Vomiting	0·2097	0·2868	0·7927	0·5911	0·2868	$>0·999$	0·2868	$>0·999$
7. Diarrhoea	0·4911	0·0069	0·0185	0·0182	0·0414	0·0345	0·0414	0·0483
# $p_{ak} \leqslant 0·05$	—	4	3	4	3	3	3	2

Hochb = Hochberg, Homm = Hommel, Bonferr = Bonferroni multiple endpoint adjustment procedure, D/AP = Dubey/Armitage–Parmar, TCH = Tukey, Ciminera and Heyse multiple endpoint adjustment procedure, p_k = observed unadjusted kth p-value, and $r_{.k}$ is the k mean partial correlation coefficient ($k=1, \ldots, 7$)

of symptoms of lactose intolerance. The reviewed package for this NDA submission consisted of three pivotal studies. One of the five (composite) primary endpoints considered was symptom evaluations composed of abdominal cramps (ACs), bloating, belching, flatulence, bowel movements (BMs), vomiting and diarrhoea, each evaluated on a 5-point scale at eight different time points. Table I summarizes the estimated mean partial correlation coefficients ($r_{.k}$) for each of the $K=7$ components of the symptom primary endpoint, the sponsor's reported two-sided p-values (p_k) and the corresponding adjusted p-values (p_{ak}) for one of the three studies. For comparison bases, we also include in this table the Bonferroni, Holm, TCH, Hochberg, and Hommel adjusted two-sided p-values. Based on these adjusted p-values one could infer that the TCH procedure is the most liberal while the Bonferroni (as expected) is the most conservative of the six procedures. The D/AP, Hochberg, Hommel and Holm procedures fall in between these two. The extreme liberal nature of the TCH procedure (which is based on the assumption of highly correlated endpoints) in this example may be due partly to the fact that none of the estimated mean partial correlation coefficients is above 0·5.

5. IMPROVING SOME p-VALUE-BASED PROCEDURES

We first present Monte Carlo simulation results that compare the performances of some of the above discussed p-valued-based procedures regarding type I error rate protection. Tables II and III present the results of comparative Monte Carlo simulation experiments on the performance of the Hochberg, Hommel and D/AP methods in protecting the α-level of significance at the 0·05 nominal level. The simulation results comparing the performance of these procedures for equi-correlated endpoints appear in Table II and those for mixed-correlated endpoints in Table III. These tables also include simulation results by the TCH and by an adjustment procedure based on the R^2 adjustment (RSA) method. In the RSA method we calculate the R-square (R_k^2) value of the kth endpoint conditional on the remaining endpoints and use it in place of the average of the correlation coefficients in the D/AP

Table II. Comparisons of type I error rate protection by five (5) p-value based methods for K equi-correlated endpoints

Number of end points: K	Correlation: r	*Ad hoc* procedures			Stepwise procedures	
		D/AP	TCH	RSA	Hochberg	Hommel
2	0·1	0·054	0·070	0·050	0·050	0·050
	0·3	0·058	0·068	0·050	0·047	0·047
	0·5	0·063	0·063	0·055	0·045	0·045
	0·7	0·071	0·062	0·061	0·046	0·046
	0·9	0·064	0·051	0·060	0·040	0·040
3	0·1	0·056	0·087	0·052	0·050	0·050
	0·3	0·064	0·082	0·054	0·045	0·046
	0·5	0·081	0·081	0·068	0·047	0·048
	0·7	0·081	0·062	0·071	0·040	0·041
	0·9	0·076	0·052	0·072	0·036	0·037
5	0·1	0·057	0·104	0·053	0·045	0·046
	0·3	0·074	0·100	0·063	0·044	0·045
	0·5	0·086	0·086	0·074	0·039	0·040
	0·7	0·094	0·072	0·086	0·036	0·038
	0·9	0·091	0·052	0·087	0·028	0·031
10	0·1	0·059	0·145	0·058	0·047	0·047
	0·3	0·082	0·125	0·077	0·044	0·045
	0·5	0·107	0·107	0·102	0·038	0·038
	0·7	0·122	0·081	0·118	0·030	0·032
	0·9	0·108	0·050	0·107	0·019	0·024

Table III. Comparisons of type I error rate protection by five (5) p-value based methods for $K=3$ mixed-correlated endpoints

Correlation: $\{r_{jk}\}$	*Ad hoc* procedures			Stepwise procedures	
	D/AP	TCH	RSA	Hochberg	Hommel
0·3 0·1 0·1	0·057	0·083	0·052	0·047	0·047
0·5 0·1 0·1	0·063	0·083	0·059	0·049	0·050
0·7 0·1 0·1	0·059	0·073	0·061	0·042	0·043
0·9 0·1 0·1	0·062	0·073	0·079	0·044	0·046
0·5 0·3 0·3	0·070	0·080	0·060	0·046	0·047
0·7 0·3 0·3	0·067	0·072	0·064	0·043	0·043
0·9 0·3 0·3	0·068	0·071	0·074	0·041	0·043

$\{r_{jk}\}$ denotes the correlation matrix with jkth element r_{jk}

expression. Usually, such an R^2 is readily available from the output of most model based statistical analyses.

The entries in Tables II and III are the proportion of times the given procedure rejected the null hypothesis (at the 0·05 significance level) of no treatment effect in at least one of the endpoints based on 10,000 simulated clinical trials with two treatment arms, each treatment arm with simulated data on 100 patients. The smaller this proportion, in comparison to 0·05, the more conservative is the procedure at the 0·05 nominal level. Conversely, the larger this proportion, in comparison to 0·05, the more liberal is the procedure at the 0·05 nominal level.

We conducted these and other simulations using the SAS/IML software on a SUN Sparc 10 Station. We generated random samples on standard normal variables using 'NORMAL' function of SAS. We then transformed the random samples to random samples from K-dimensional multivariate normal distributions with a given correlation matrix using Cholesky decomposition of the given correlation matrix. We kept the normal variances at 1 and the normal means at zero, representing the null hypothesis of no difference between treatment arms.

The simulation results in Tables II and III show that both the Hochberg and Hommel procedures perform reasonably well for $K \leqslant 3$ when the correlations between the endpoints are in the low range ($\leqslant 0{\cdot}5$). For strongly correlated endpoints and as K gets larger, these two methods become conservative in protecting the nominal significance level. The results also show that the D/AP procedure is very liberal, and gets worse with increasing K. As K gets larger it seems likely that this procedure does not provide adequate protection of the overall 0·05 significance level. The RSA provides a slight improvement on the D/AP method for equi-correlated endpoints, but it is still too generous, especially for mixed-correlated endpoints. The TCH method has problems for moderate to weakly correlated endpoints ($r_{.k} \leqslant 0{\cdot}7$) but does fairly well for highly correlated endpoints. This is to be expected since the TCH method is designed specifically for highly correlated endpoints.

5.1. Improving the Hochberg and the *ad hoc* procedures

One crude way to improve the performance of the Hochberg procedure is simply to raise its nominal significance boundary by a factor c as follows:

$$\{c\alpha/K,\ c\alpha/(K-1),\ c\alpha/(K-2),\ \dots,\ c\alpha/2,\ \alpha\}$$

where c is determined via a computer optimization technique so that for given K and correlation structure among the K endpoints, the overall protection for the procedure is close to the nominal α-level. The procedure for correcting the Hochberg method for $K=3$ endpoints with a correction factor c is order the p-values from low to high as $p_1 \geqslant p_2 \geqslant p_3$. If $p_1 \leqslant \alpha$ then declare results significant for all three endpoints. However, if $p_1 > \alpha$, then compare p_2 against $c(\alpha/2)$. If $p_2 \leqslant c(\alpha/2)$, declare results significant for the remaining two endpoints corresponding to p_2 and p_3. In case, $p_2 > c(\alpha/2)$, then compare the smallest p_3 against $c(\alpha/3)$.

Despite the liberal nature of the D/AP, RSA and TCH methods in providing adequate protection for the overall significance level, these *ad hoc* procedures are appealing to some clinicians because they utilize the correlations among the endpoints, and are easy to use. With this in mind, we propose the following crude way to improve the liberal nature of these procedures. That is, use $\alpha' = c\alpha$ instead of α in $\alpha_k = 1-(1-\alpha)^{m_k}$ obtained in the D/AP, RSA, or TCH procedure. The goal is to obtain c as a function of K and the correlations between endpoints so that the overall protection is exactly α.

Tables IV and V give the simulated correction factors and the corresponding simulated α' values based on these crude methods for equi-correlated and mixed-correlated K endpoints, respectively.

Table IV. Comparisons of type I error rate protection by five (4) p-value based methods for K equi-correlated endpoints

Number of endpoints: K	Correlation: r	*Ad hoc* pocedures						Stepwise procedures	
		D/AP		TCH		RSA		Hochberg	
		c	α'	c	α'	c	α'	c	α'
2	0·1	0·94	0·050	0·72	0·050	1·00	0·050	1.00	0·050
	0·3	0·87	0·049	0·76	0·050	1·00	0·050	1·06	0·050
	0·5	0·78	0·050	0·78	0·050	0·99	0·050	1·09	0·050
	0·7	0·71	0·050	0·81	0·050	0·81	0·050	1·11	0·050
	0·9	0·76	0·050	0·99	0·050	0·81	0·050	1·35	0·050
3	0·1	0·89	0·050	0·58	0·050	0·97	0·050	1·01	0·050
	0·3	0·78	0·049	0·63	0·050	0·93	0·050	1·10	0·050
	0·5	0·63	0·051	0·63	0·051	0·74	0·051	1·11	0·051
	0·7	0·60	0·049	0·78	0·051	0·70	0·050	1·30	0·050
	0·9	0·68	0·051	0·92	0·049	0·65	0·050	1·55	0·050
5	0·1	0·85	0·049	0·45	0·049	0·93	0·050	1·09	0·049
	0·3	0·68	0·050	0·49	0·050	0·80	0·050	1·14	0·051
	0·5	0·58	0·050	0·58	0·050	0·68	0·050	1·30	0·050
	0·7	0·46	0·050	0·63	0·050	0·53	0·050	1·40	0·050
	0·9	0·51	0·051	0·96	0·050	0·52	0·050	2·20	0·050
10	0·1	0·84	0·051	0·33	0·050	0·85	0·050	1·09	0·049
	0·3	0·56	0·050	0·35	0·050	0·60	0·050	1·14	0·050
	0·5	0·45	0·051	0·45	0·051	0·46	0·050	1·40	0·050
	0·7	0·37	0·050	0·58	0·051	0·39	0·051	1·85	0·051
	0·9	0·41	0·050	1·00	0·050	0·42	0·050	3·00	0·050

α' = achieved α-level for a given correction factor c

6. NORMAL-THEORY-BASED GLOBAL PROCEDURES

Clinical evidence in favour of therapeutic benefit is strengthened if one can demonstrate treatment effectiveness consistently across multiple endpoints. Attempts to accomplish this goal have been made by resorting to global testing procedures with little or no consideration for their appropriateness regarding the nature of the endpoints and the objectives of the clinical trial. The primary aim of a global testing procedure is to demonstrate overall therapeutic benefit considering simultaneously all given endpoints. For example, given K multiple endpoints, the null hypothesis of interest is the global null hypothesis H_0: $\delta_k = 0$ ($k = 1, \ldots, K$).

Some of the commonly employed global statistical methods in this setting are the Hotelling's T^2 (HT), O'Brien's ordinary/generalized least squares (OLS/GLS), rank-sum test and Tang *et al.* asymptotic likelihood ratio (ALR) test.[16,17]

Generally speaking, these tests are appropriate in settings where the endpoints are alternative measures of the same fundamental quantity. They may not, however, be appropriate in situations where there is an interest in making separate inferences about each measure and where consistency of treatment effectiveness across endpoints cannot be guaranteed. That is, they are not designed to detect sporadic treatment effects across endpoints.

Table V. Comparisons of type I error rate protection by five (4) p-value based methods for $K=3$ mixed-correlated endpoints

Correlation: $\{r_{jk}\}$	*Ad hoc* procedures						Stepwise procedure	
	D/AP		TCH		RSA		Hochberg	
	c	α'	c	α'	c	α'	c	α'
0·3 0·1 0·1	0·88	0·050	0·60	0·049	0·96	0·050	1·06	0·049
0·5 0·1 0·1	0·71	0·051	0·59	0·051	0·83	0·050	1·02	0·050
0·7 0·1 0·1	0·85	0·050	0·69	0·050	0·82	0·050	1·19	0·051
0·9 0·1 0·1	0·80	0·051	0·90	0·051	0·65	0·051	1·14	0·050
0·5 0·3 0·3	0·74	0·050	0·63	0·050	0·85	0·050	1·09	0·050
0·7 0·3 0·3	0·74	0·050	0·69	0·051	0·70	0·050	1·16	0·050
0·9 0·3 0·3	0·75	0·050	0·73	0·050	0·67	0·050	1·22	0·050

$\{r_{jk}\}$ denotes the correlation matrix with jkth element r_{jk}; α' =achieved α-level for a given correction factor c

Moreover, given a number of multiple endpoints, clinicians desire to know specifically what endpoints are statistically significant at a given nominal significance level after proper adjustment for multiplicity. Global tests (*per se*) are therefore inadequate in this sense, because they generally lack such specificity; rejection of the global null hypothesis of no treatment difference does not imply rejection of the null hypotheses of lower order.

We say that a multiplicity adjustment procedure (for multiple endpoints or multiple comparisons) closed or satisfies the closure property if for a given nominal α-level, the rejection of the null hypothesis H_{0k} of the kth endpoint ($k=1, 2, \ldots, K$) implies the rejection of all higher dimensional hypotheses that contain this null hypothesis and *vice versa.*[10] For example, for $K=3$, we say that a multiplicity test is closed if rejection of the null hypothesis H_{01} (for $k=1$) at a given nominal α-level implies rejection of $\{H_{01}, H_{02}\}$, $\{H_{01}, H_{03}\}$ and $\{H_{01}, H_{02}, \mathrm{H}_{03}\}$ at the same nominal α-level.

Global tests of the O'Brien types that are not closed tests are, by design, power driven tests, and, in general, do not provide adequate protection of the type I error rate. They may lead to detection of treatment differences that are of no clinical significance. Such power driven tests are, therefore, more appropriate in the design and analysis of clinical trials where demonstration of therapeutic equality is the primary interest.[2] In such trials, maintenance of adequate power is usually the primary concern and not protection of the type I error rate.

A number of authors have suggested some modifications of some of these (O'Brien's OLS/GLS) global tests to address both the specificity and the power issues usually associated with these tests.[18–20] Lehmacher *et al.*[19] reported simulation results that show both the OLS and GLS tests are more powerful than the HT in detecting endpoint specific treatment group differences.

With the advent of resampling techniques, it is also now possible to incorporate the dependence structure of the data utilized in these normal-theory-based global procedures without the strong distributional assumptions associated with these methods. For highly correlated endpoints, these procedures are much less conservative than the Bonferroni procedures,[21] and, where appropriate, one should prefer these global methods.

7. CONCLUDING REMARKS

We have addressed several multiple endpoint adjustment procedures commonly used in the analyses of clinical trial data with multiple endpoints. We have chosen these procedures for discussion because of their simplicity of use and popularity among clinicians. We have highlighted the major concerns of these procedures (both from pharmaceutical and regulatory view points) by contrasting the conservative Bonferroni methods to some liberal *ad hoc* procedures regarding the protection of the type I error rate. We have also highlighted some of the primary concerns among clinicians with the use of global procedures.

We have seen extensive use of the Hochberg procedure in clinical trials that require p-value adjustments, even though it is still somewhat conservative. This is probably due to its simplicity and the fact that it is a closed test under independence. We have also seen some use of *ad hoc* methods in this regard. We have carried out limited Monte Carlo simulation studies to compare these procedures with respect to providing adequate type I error rate protection.

Our limited simulation results so far indicate that for $K=2$ or 3 both the Hochberg and Hommel procedures perform reasonably well for moderate to weakly correlated endpoints. However, for strongly correlated endpoints, and as the number of endpoints gets larger, they are somewhat conservative. Similarly, for the *ad hoc* procedures, our simulation results indicate that these methods are somewhat liberal. In line with the spirit of its design, the TCH procedure seems to perform rather well for strongly correlated endpoints but not otherwise. The D/AP procedure (the only p-value-based procedure to utilize fully the correlation structure of the data) seems to need some adjustment for all ranges of the correlations. This is the same with the RSA *ad hoc* procedure (introduced by these authors). While providing a slight improvement over the D/AP procedure, the RSA procedure does not provide much comfort to these authors, despite its interpretational appeal.

To improve on the conservative nature of the Hochberg and on the liberal nature of the *ad hoc* procedures, we have suggested the use of a (crude) correction factor c, via an optimization technique, to upsize ($c \geqslant 1$) the achieved nominal α-level (α') for the Hochberg method, and conversely, to downsize ($c \leqslant 1$) the achieved nominal α-level for the *ad hoc* methods, to the desired nominal α-level. We have provided at present these correction factor values for $\alpha = 0{\cdot}05$ and for limited values of K on assuming the data arose from K-variate normal populations. These simulation results indicate that use of the appropriate correction factor leads to reasonable nominal α-level protection in these procedures for K-variate normal samples.

Global tests such as the O'Brien GLS test and its improvements have considerable appeal because they utilize fully the correlation information among the endpoints, and test the alternative hypothesis of clinical interest. However, naive applications of these tests to multiple endpoint problems could lead to difficulty when one expects treatment effects to occur in relatively a few endpoints, or one cannot assure consistency of effects across endpoints. The issue is somewhat similar to that found in a meta-analysis of studies in combining treatment effects in the presence of inconsistent results across studies. It could be clinically difficult to interpret an overall treatment effect by a global test if a few of the endpoints showed strong treatment effect while the majority of them indicated a null treatment effect. For the above reasons, and others mentioned earlier, one should use global tests with caution.

In cases where there is a general consensus among the scientific community on the relative clinical importance or interest among endpoints for a given disease, some authors have suggested analysis of one or two at, say, 0·05 significance level and the rest at level(s) higher than 0·05 to maintain adequate power.[22] Such an approach may not be possible because, for some diseases three or more primary endpoints (measuring disparate features of patient response) may have equal clinical significance. The simple demonstration of a trend in any one of them may offer no more comfort than does the failure to establish any significant result. Furthermore, once one specifies an endpoint as primary or secondary in the protocol, one should treat it as such in the entire process of the clinical trial, including the analysis, labelling and marketing processes.

Finally, given the inherent multiplicity component of clinical trials, they should be designed with the understanding that multiplicity is sometimes unavoidable. Investigators should power trials so that they reach meaningful conclusions with regard to the overall therapeutic benefit even after making the necessary adjustments for multiplicity.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. Nancy Smith, two anonymous referees and the editor for their careful reading of the manuscript and constructive suggestions.

REFERENCES

1. Bauer, P. 'Multiple testing in clinical trials', *Statistics in Medicine*, **10**, 871–890 (1991).
2. D'Agostino, R. B. and Hereen, T. C. 'Multiple comparisons in over-the-counter (OTC) drug clinical trials with both positive and placebo controls', *Statistics in Medicine*, **10**, 1–6 (1991).
3. Perry, J. N. 'Multiple comparison procedures: a dissenting view', *Journal of Economic Entomology*, **79**, 1149–1155 (1986).
4. Rothman, K. J. 'No adjustments are needed for multiple comparisons', *Epidemiology*, **1**, 43–46 (1990).
5. Salsburg, D. S. 'Using multiple endpoints to answer clinically relevant questions', Invited presentation at the ENAR meetings, Cleveland, Ohio, 1994.
6. Holm, S. 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics*, **6**, 65–70 (1979).
7. Simes, R. J. 'An improved Bonferroni procedure for multiple tests of significance', *Biometrika*, **73**, 751–754 (1988).
8. Hochberg, Y. 'A sharper Bonferroni procedure for multiple tests of significance', *Biometrika*, **75**, 800–802 (1988).
9. Hommel, G. 'A comparison of two modified Bonferroni procedures', *Biometrika*, **76**, 624–625 (1989).
10. Marcus, R., Peritz, E. and Gabriel, K. R. 'On closed testing procedures with special reference to ordered analysis of variance', *Biometrika*, **67**, 655–660 (1976).
11. Hommel, G. 'A stagewise rejective multiple test procedure based on a modified Bonferroni test', *Biometrika*, **75**, 383–386 (1988).
12. Dunnett, C. W. and Tamhane, A. C. 'A step-up multiple test procedure', *Journal of the American Statistical Association*, **87**, 162–170 (1993).
13. Tukey, J. W., Ciminera, J. L. and Heyse, J. F. 'Testing the statistical certainty of a response to increasing doses of a drug', *Biometrics*, **41**, 295–301 (1985).
14. Dubey, S. D. 'Adjustment of *p*-values for multiplicities of intercorrelating symptoms', Proceedings of the VIth International Society for Clinical Biostatisticians, Germany, 1985.
15. Armitage, P. and Parmar, M. 'Some approaches to the problem of multiplicity in clinical trials', Proceedings of the XIIth International Biometrics Conference, Seattle, 1986.
16. O'Brien, P. C. 'Procedures for comparing samples with multiple endpoints', *Biometrics*, **40**, 1079–1089 (1984).

17. Tang, D. I., Gnecco, C. and Geller, N. L. 'An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials', *Biometrika*, **76**, 751–754 (1989).
18. Tang, D. I. and Lin, S. 'On improving some methods for multiple endpoints' Invited presentation at the ENAR meetings, Cleveland, Ohio, 1994.
19. Lehmacher, W., Wassmer, G. and Reitmeir, P. 'Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate', *Biometrics*, **47**, 511–521 (1991).
20. Pocock, S. J., Geller, N. L. and Tsiatis, A. A. 'The analysis of multiple endpoints in clinical trials', *Biometrics*, **43**, 487–493 (1987).
21. Westfall, P. H. and Young, S. S. '*p*-value adjustments for multiple tests in multivariate binomial models', *Journal of the American Statistical Association*, **84**, 780–786 (1989).
22. Capizzi, T. and Zhang, J. 'Testing the hypothesis that matters', Invited presentation at the ENAR meetings, Cleveland, Ohio, 1994.