Drug Information Journal, Vol. 30, pp. 523–533, 1996
Printed in the USA.
0092-8615/96

STATISTICAL CONSIDERATIONS FOR MULTIPLICITY IN CONFIRMATORY PROTOCOLS

GARY G. KOCH, PHD, AND STUART A. GANSKY, MS

Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina

Statistical issues concerning multiple response criteria, multiple treatment groups, and multiple subgroups in clinical trials require careful attention in order to avoid an inappropriately high prevalence of chance findings, as well as to avoid unsatisfactorily low power to detect real treatment differences. An underlying goal is using a 0.050 significance level as often as possible for separate assessments while maintaining a 0.050 level for all assessments taken together, so statistical power is not compromised. For multiple response criteria, a useful assessment strategy is composite ranking as a single criterion first and then its individual components. Multiple treatment comparisons can often be effectively addressed with closed testing procedures with hierarchical evaluation. This hierarchy must be well specified since significance at its first stage is required before testing is allowed at the next stage.

In most clinical trials, subgroups are of supportive interest after statistical significance for all patients is shown. A subgroup hierarchy, however, permits primary evaluation in conjunction with all patients through significance level spending function methods as in interim analyses, for example, the O'Brien-Fleming method. The rationale is the analogy between a subgroup hierarchy and the patient hierarchy at successive interim analyses. With this method, the significance level for all patients' evaluation typically ranges between 0.040 and 0.045 and that for subgroups ranges from 0.005–0.020.

The methods outlined here for multiple response criteria, multiple treatment groups, and subgroups, or related counterparts, should be prespecified in the protocol for a clinical trial. If not in the protocol, they should be incorporated in the analysis plan prior to study unmasking.

Key Words: Multiple endpoints; Multiple treatments; Subgroups; Significance level spending function; Closed testing procedure

INTRODUCTION

THERE ARE SEVERAL well-known statistical principles in study design. One is randomization to provide statistically comparable groups and to avoid selection bias in treatment assignment. Another involves stratifying patients into subgroups for separate randomization. This further strengthens the comparability or balance of groups; it also sometimes provides better power through smaller variability for estimated treatment differences. Finally, stratification can assure that a study has appropriate generalizability by including the relevant segments of a potential study population.

Presented at the DIA Workshop "Clinical Trials in Biotechnology: Planning to Prevent the Pitfalls," January 30–February 1, 1994, Newport Beach, California.

Reprint address: Gary G. Koch, PhD, Department of Biostatistics, School of Public Health, University of North Carolina, CB# 7400, Chapel Hill, NC 27599–7400.

Another important statistical principle is the masking of patients, investigators, and sponsors to treatment assignments. The purpose here is to support treatment group comparability by minimizing the likelihood of potential bias from different sources of variability that may come from aspects of patient management or measurement occurring after randomization.

Beyond these generally traditional study design principles, issues from the area known as "multiplicity" must be taken into account in study planning. The main theme concerning multiplicity is that multiple assessments lead to multiple opportunities for findings to be due to chance and so need control. For example, with the usual 0.050 criterion for statistical significance, five tests each at the 0.050 level have a probability as high as 0.226 for at least one showing statistical significance by chance alone, depending upon the extent of correlation among the tests. Thus, performing multiple tests makes the traditional 0.050 level (or whatever level might be the basic criterion) no longer necessarily applicable.

To have statistical analyses operate rigorously at the designated significance level, one must address multiplicity. Multiplicity comes in many different forms. One form comes from multiple interim analyses, that is, examining the data part-way through the study, as well as at the end of the study. Multiplicity comes from evaluating multiple primary endpoints, that is, different criteria for judging therapy efficacy. Multiplicity comes from evaluating one or more subgroups, in addition to all patients. In studies that involve more than two treatment groups (eg, three or four treatment groups), multiplicity comes from making multiple comparisons among those treatments.

The main issue for multiple assessments is to avoid loss of validity due to inflating the Type I error (significance level) from insufficient control, while simultaneously avoiding power loss by excessive Type II error from over-control. For example, performing five tests at the 0.010 significance level so as not to exceed an overall 0.050 significance level often over-controls for Type I error for a study with a sample size planned at the 0.050 significance level through the implied loss of power. Alternatively, planning at the 0.010 level potentially gives a much larger sample size than might be realistically needed. Thus, one must balance the previously stated considerations so as to maintain validity by assuring that the overall Type I error or significance level is controlled at the 0.050 level, while simultaneously maintaining adequate power and preventing excessive sample size resulting from over-control for multiplicity through an inefficient strategy for analysis.

MULTIPLE ENDPOINTS

One area for multiplicity is the presence of multiple primary response variables. This matter can be addressed in several ways. The first, of course, is to specify the primary response variable(s) and note several of its (their) features. Sometimes, several variables are related; in hypertension treatment studies, for example, sitting and standing diastolic and systolic blood pressures at the end of the study are the primary variables. Usually, those four variables are sufficiently highly correlated to yield similar results; hence, multiplicity in this case is not usually a major issue (although to be more conservative, it is often appropriate to identify one dominant primary variable—typically resting diastolic blood pressure). Other studies may have several families of variables. In ulcer disease studies (1,2), primary evaluation may focus on endoscopic verification of the ulcer's presence or absence after treatment, as well as on relief of symptoms. Often, one first tests the primary family (ie, healing by endoscopy), and then one proceeds to the symptom variable(s). In some studies for respiratory disorders, two families of variables are of interest: one for alleviating symptoms and the other for various aspects of physiologic response for lung function. In some cases, one can separately assess such different families, but often it is important to order them in a logical hierarchy to ac-

count for the possibility that one yields statistical significance while another may not.

An important consideration for multiple endpoints is to recognize that if the objective is to show statistical significance for all endpoints, then all can be tested at the 0.050 level without any correction; the reason is that the probability for all tests having $p \leq 0.050$ when any one of them would yield this result by chance alone is less than the probability that any one of them corresponding to a result by chance alone would yield $p \leq 0.050$. Thus, if one requires significance for all tests, then all tests can be assessed at the 0.050 level. Sample size, however, usually needs to be increased for this situation because each of these tests has a Type II error probability for failing to yield $p \leq 0.050$; and so the overall probability that at least one of the tests fails to yield $p \leq 0.050$ can be as large as the sum of these Type II errors. Thus, such studies often have to be designed with 0.90 or 0.95 power for each endpoint so that across the multiple endpoints the overall global power is in the neighborhood of 0.80.

A multicenter ulcer disease study (Table 1, Example 1) further illustrates the previously discussed issues (1). In this case, there is interest in showing that not only is a particular treatment good for healing ulcers, but it also enables prevention of recurrence after healing. A main consideration for the design of this study is that the goal of showing two significant outcomes requires greater power for each one separately than would be necessary if statistical significance for only one were the objective. As indicated in Koch et al. (1), 100 patients in each group provide appropriate power for both comparisons for the two outcomes of interest (as well as for comparisons concerning multiple treatments).

A very flexible and helpful procedure to deal with multiple endpoints is the method of composite ranking. This method, sometimes called an "O'Brien" method (3), involves ranking the patients on each response criterion, then averaging those ranks across the criteria for each patient, and applying a nonparametric test to the average of the rankings. Typically, in practice, not only must this composite ranking exhibit a significant result to show overall benefit, but also at least one of the individual measures must show significance in order for conclusions to be reasonably interpretable. The important point, however, is that the statistical tests for both the composite ranking and each of the components are applied at the 0.050 significance level; a refinement of this method which more formally supports statistical inference for the individual components is discussed in Lehmacher (4).

An example of this method of composite ranking is a multicenter, multivisit clinical trial (Table 1, Example 2) to compare two treatments for a respiratory disorder (5). There were two centers, an ordered response variable with five categories for global evaluation, and four visits. The p-values from a direct visit-by-visit analysis of treatment differences with a nonparametric method were 0.052 at Visit 1, less than 0.001 at Visit 2, 0.002 at Visit 3, and 0.020 at Visit 4. In this particular study, everything was strongly positive for more favorable response with test treatment than placebo. Furthermore, this particular study's protocol specified Visit 3 as the primary visit because this trial was for a respiratory condition for which certain concomitant therapies were withdrawn over the course of the visits; Visit 3 was at the time of essentially maximal withdrawal, because at Visit 4 concomitant therapies could be reintroduced. Visit 3, the primary endpoint, had a p-value of less than 0.050.

At one time during the planning of the study, however, there was discussion as to whether either Visit 2 or Visit 3 could be primary. In this case, the smaller of the p-values for Visit 2 and Visit 3 tests would utilize a 0.025 level; both Visits 2 and 3 met that criterion. It is important to have all of these criteria prespecified, because Visit 1 does not quite show statistical significance, at least by the usual method. Using a method invoking covariance adjustment, a more powerful method (that accounts for correlation with baseline), provides p-values that are all clearly less than 0.050, in fact, even

TABLE 1
Multiplicity Examples

1. Ulcer Disease	Multicenter 2 endpoints (healing and no recurrence) 3 treatments (test, active control, placebo)
2. Respiratory Disorder	2 centers Ordinal response (5 levels), also additional variables for symptoms and lung function 4 endpoints (visits) 2 treatments (test, placebo)
3. Gastrointestinal Disorder	2 centers Ordinal response Multivisit 3+ endpoints (pain symptoms, well-being, physiologic status) 3 treatments (A, B, placebo) 2 subgroups (baseline status: moderate, severe)
4. Dental Pain	2 centers 3 endpoints (SPID, TOTPAR, TOTGONE) 5 treatments (test-high, test-low, active control-high, active control-low, placebo)
5. Hypertension	Response surface study 1 primary endpoint (supine diastolic blood pressure) 12 treatments (3 HCTZ doses $\times$ 4 ACE doses, including placebo)
6. Chronic Pain	Multicenter 1 endpoint (pain recurrence) 2 treatments (test, placebo) 2 subgroup factors (prior treatment type, prior treatment experience)

less than 0.010. With this method, there would be no multiplicity problem, but the method for covariance analysis and the covariables to be used would need specification in the study protocol or in a formal analysis plan before the unmasking of treatments for the study. With a composite ranking, a p-value of less than 0.001 is obtained regardless of covariance adjustment.

Example 2 utilized not only the global evaluation variable, but also physiologic lung function variables and symptom variables; composite rankings were used across all of them to address statistical significance ultimately for a spectrum of variables that actually included as many as 30 endpoints (across visits and family of response variable). In this way, one was first able to demonstrate an overall pattern of significant differences between treatments, then significant patterns of differences within each of the respective families for endpoints, and finally significant differences for the majority of individual variables.

In another example with multiple endpoints (Table 1, Example 3), a multicenter study with multiple visits is used to compare treatments for a gastrointestinal disorder. There are two centers and three kinds of variables: ordered outcomes for pain symptoms, measures of well-being, and measures of physiologic status. This study's analysis plan addresses multiple endpoints by focusing on an average of rankings over the measures; if there is a significant result, then each component for this composite ranking would be examined separately. In this way, multiple endpoints can all be addressed at the 0.050 level as long as there is an overall effect at the 0.050 level.

The most useful multiple endpoint procedure in many situations is the O'Brien method. As stated previously, the application of this method involves ranking each of the

endpoints, then averaging the rankings, and then testing that average ranking; statistical significance implies an overall pattern of differences between treatment groups. Additionally, one should test each endpoint individually; if at least one of these is significant, then the result from the composite ranking method is interpretable. Thus, in this case, only the results from two tests are really required to be significant—at least one for the separate endpoints, together with that for the composite ranking. One must recognize, however, that the effective use of this method requires that the components being averaged for the rankings be sound endpoints which have reasonable potential for the detection of real treatment differences. Thus, the method does not bypass the difficulty of having insensitive or unreliable endpoints. Good endpoints, for which there is reasonable information for the detection of treatment differences, are necessary for the average ranking method to work well for addressing multiplicity. Moreover, it will typically have more power than the methods for the individual endpoints themselves.

MULTIPLE TREATMENTS

Some studies compare multiple treatment groups. For instance, a study may compare three treatments: a test treatment, an active control, and a placebo. Another example involves two doses of a test treatment and placebo. A third example has three doses of a test treatment. An example with four treatments includes a combination treatment, each of its monotherapy components and placebo; and another has two doses of a test treatment, an active control treatment, and placebo. Additionally, there are a variety of study designs for dose-response studies, which may have any number of doses, with or without placebo, with or without an active control treatment (6). More generally, there are response surface studies that involve combinations of two or more doses of monotherapy components. With such studies, one has to plan the analysis so as many comparisons as possible can be tested at more practical significance levels, such as 0.050 or 0.025, in order to avoid excessively conservative adjustments for multiplicity and correspondingly unsatisfactory statistical power or excessive sample size.

One way to proceed is to specify primary comparisons. If all such comparisons must be significant at the 0.050 level, they can each be tested at the 0.050 level, as stated previously. Often, in analgesic studies, one expects the test treatment to be better than placebo, the active control treatment to be better than placebo, and the test treatment to be better than the active control treatment minus a tolerance that corresponds to support for clinical equivalence. So three significant results are desired, and all testing can be done at the 0.050 level. Increased sample size, however, may be necessary to avoid excessive Type II error, and a 2 : 2 : 1 allocation may be useful for sample size being no larger than necessary; here, the two-fold sample sizes apply to the pair of treatments with the smallest difference relative to the applicable standard deviation.

In another situation, however, where only one or two comparisons need to be significant, such as in a confirmatory study with multiple doses of a test treatment versus placebo, a method must adjust the significance level of the separate tests to address the multiplicity from comparisons of placebo to multiple doses. One way to make the testing more effective is to first apply certain overall comparisons; one useful overall method is testing the average of several doses versus placebo to establish that at least one dose is better than placebo (ie, dose-response in some sense), and then applying pairwise testing versus placebo for the individual doses. This can often be a very useful procedure. The averaging can involve different subsets of treatments. One could also look at more than one average, given that some multiplicity adjustment is made for the number of such averages. Finally, one will often examine relationships between the treatment groups and the response to document further the overall pattern of treatment differences.

One interesting situation to consider fur-

ther has three treatment groups: a test treatment, an active control treatment, and placebo. The first test in the hierarchy is performed at the 0.050 level to compare the test treatment to placebo because this is the most important comparison for assessment of whether the test treatment provides any benefit at all. Given statistical significance for that test, calculate a 0.95 confidence interval to judge the equivalence of the test treatment and the active control treatment through the ratio of their differences from placebo. If the lower bound of that confidence interval exceeds some quantity such as 0.67, clinical equivalence is supported. If the lower bound exceeds 1.00, potential superiority is suggested; and if it exceeds some value such as 1.50, a clearer version of superiority is supported. Finally, in order to verify that this kind of assessment of test treatment versus active control is clinically meaningful, it is important to show that the active control is better than placebo. These tests are done in a hierarchical manner, that is, they are applied sequentially where testing does not proceed to the next step unless the previous one is significant. In that way, the overall significance level is 0.050. For the previous example, the sequence is applied to show test treatment is better than placebo, test treatment is at least as good as active control, active control is better than placebo, and test treatment is better than active control, in that order.

In a situation with low-dose, high-dose, and placebo, one might assume both doses are effective. In that case, the average of the low and high doses can be tested against placebo with relatively high power. If that result is significant, all the pairwise comparisons can be performed at the 0.050 level, because an overall criterion has indicated a significant difference (and so at most, one of the three pairwise comparisons can be null). In some confirmatory studies with two doses versus placebo, both doses may not be effective; in this case, an adjustment must be made for two comparisons between test treatment and placebo. One approach is to evaluate the comparison with the smaller p-value at 0.025 using the Bonferroni-Holm method (7). If that result is significant, the other comparison to placebo can be tested at the 0.050 level. If at least one of the two doses is found to be significantly better than placebo, the two doses can be compared with each other at the 0.050 level to address dose-response. Thus, *a priori* rules to address multiplicity will allow testing in a relatively efficient way.

The gastrointestinal study mentioned earlier (Table 1, Example 3) has three treatment groups: treatment A, treatment B, and placebo. First, all patients are evaluated for the comparisons of treatment A versus placebo and treatment B versus placebo. These comparisons are tested with the Bonferroni-Holm method. The comparison with the smaller p-value is tested at 0.025; if that is significant, the other one is tested at 0.050. Subsequently, subgroups for this example will be evaluated for these treatment comparisons, given their statistical significance for all patients. This further analysis is an additional stage of the hierarchical testing procedure.

Example 4 (Table 1), a two-center trial concerning dental pain relief (8), has an active control treatment at two doses, a test treatment at two doses, and placebo. (There was interest in the dose-response structure for the two active treatments, as well as in comparisons against placebo.) Rules were specified whereby one could compare each of the two doses against placebo by the method described earlier, for each of the two treatments in a sequential order; so all comparisons versus placebo could be applied at either the 0.025 or 0.050 levels, provided there was statistical significance for at least one test at each stage of the hierarchy.

The next example (Table 1, Example 5) is a response surface study (9) for blood pressure reduction. There were two active levels of hydrochlorothiazide (HCTZ) and a null level, as well as three active levels of an angiotensin converting enzyme (ACE) and a null level; thus, there were 12 treatment combinations altogether, with placebo corresponding to the combination of both monotherapies' null levels. The primary response variable was supine diastolic blood pressure change. In order to address multiple compari-

son issues, two regions of the response surface were identified. One region involved four combinations from both positive HCTZ doses and the lower two active ACE doses, while the other region involved four combinations from both active HCTZ doses and the higher two active ACE doses. Treatment averages in those regions were then compared against the averages for their monotherapy counterparts with two overall tests, each at the 0.025 level (although the 0.050 level could have been used for the high p-value if the smaller one was less than 0.025). Both results were significant, indicating there was efficacious response for the combination treatments in those response surface regions; thus, it was possible to examine each individual combination treatment against its monotherapy counterpart at the 0.050 level as well, because an overall effect was already established (although for more rigorous assessment, testing for averages of pairs of combinations within the regions could be done as a connecting intermediate step for the hierarchical procedure).

Overall, the two-stage procedure identified three combination treatments (the lower active HCTZ dose with each of the highest two active ACE doses and the highest HCTZ dose with the middle active ACE dose) as being efficacious in the sense of being significantly better than their monotherapy components. Thus, although this particular study design involved 11 different test treatments and one placebo, it was possible to specify multiple comparison procedures that allowed all testing at the 0.025 or 0.050 level. In addition to identifying efficacious combination treatments for this type of study, one should also apply analyses that describe dose-response relationships; these analyses indicated that all four combinations involving the two active HCTZ doses and the higher two ACE doses had relatively similar effects on blood pressure reduction.

MULTIPLE SUBGROUPS

With subgroups, there are a number of issues which can be important. Of course, one is to specify the primary subgroups in a protocol so those subgroups can be evaluated in an inferential way. Otherwise, subgroups are purely *ad hoc* and only useful for exploratory purposes. Another reason for examining subgroups is to address the generalizability of findings so as to confirm that treatment differences are as applicable, for example, for men as they are for women, or for younger people as they are for older people. Often, this type of analysis is accomplished by examining statistical interactions; in many studies, subgroups have a supportive role since the first objective is to show statistical significance for all patients, and then attention is given to evaluating homogeneity of treatment differences across the subgroups.

For instance, in Example 3 (Table 1), the gastrointestinal study considered previously, there were two primary subgroups based on the severity level of the disorder, but the primary analysis was for all patients. Once statistical significance for all patients was shown, one would examine the subgroups. A subgroup of particular interest is the moderately severe patients, who are two thirds of the total; one can further test treatment differences within the moderately severe patients and within the severe patients, given statistically significant results for all patients. If within a subgroup, one of the two test treatments is found to be significantly better than placebo by a method which suitably addresses multiplicity in treatment comparisons, one can then compare the two treatments within that subgroup. An interesting feature of this study, which is currently being planned, is that treatment A is expected to be better than treatment B for moderately severe patients, but treatment B is expected to be better than treatment A for severe patients. Addressing such hypotheses without having a hierarchical testing procedure would require a variety of multiplicity adjustments. With a suitably specified hierarchical procedure for sequential testing, however, all tests can be done at the 0.025 or 0.050 level.

There are some situations where subgroups are primary, and these will require carefully selected strategies. For Example 6

(Table 1), a multicenter study for recurrence of a chronic pain condition, there were two primary subgroups: one was based on type of prior treatment for patients and the other was based on an aspect of patient experience with that prior treatment. Thus, the subgroups are hierarchical with one containing about one-third of the patients and the other containing about two-thirds of the patients (including those in the smaller subgroup). Investigators were interested in classifying the study a success if statistical significance applied to all patients or to either subgroup. One strategy here is to order the patients according to the subgroups in a sequential hierarchy with independent increments such as in interim analyses; one basically views the smaller subgroup as a first interim analysis at one-third through the study, the larger subgroup as a second interim analysis at two-thirds through the study, and all patients as the basis for final analysis of the completed study (10). With an alpha-spending function method such as the O'Brien-Fleming method (11), the analysis for all patients can apply statistical testing at about a 0.047 level, the subgroup with about two-thirds of the patients can do so at the $p \leq 0.015$ level, and the smaller subgroup can do so at about the 0.0005 level. Of course, there would have to be very dramatic results in the smaller subgroup to achieve statistical significance. There are ways in which the 0.047 could be reduced and the 0.0005 could be increased (12), but the applicable concept is to maintain the overall significance level at $p \leq 0.050$ with some spending function.

Alternatively, a closed hierarchical method such as that described for multiple endpoints could be implemented for subgroups. For example, with three subgroups, the primary subgroup would be assessed first; if the primary test was significant at 0.050, the next group could be examined, also at 0.050; and given significance of the second test, the last subgroup could be tested at 0.050. Note that with three subgroups there are three factorial ways to order this hierarchy. Thus, the order of tests used in analysis must be prespecified in order to allow testing at the 0.050 level for each assessment.

Without reviewing details of statistical methods for interim analysis, one should simply remember that their nature, purposes, and methods should be specified in the protocol for a study, particularly the spending function method for controlling the multiplicity. Also, from the viewpoint of not adversely influencing the study's further conduct, one should strictly limit who knows the interim analyses' results.

STATISTICAL APPROACHES

As noted throughout this discussion, a variety of statistical issues involve multiplicity, and they affect the study design, the sample size, and data analysis. When planning sample size, one should account for the design, the multiplicity of treatments, the multiplicity of subgroups, the multiplicity of response criteria, the multiplicity of interim assessments, and a variety of other issues that have to do with the study's data structure and conduct. For taking all of these matters into account, there are general ways to determine sample size through the following principles. First one creates a replicate reflecting the study design and data structure (ie, number of treatments, allocation ratios for treatments, and parallel/crossover structure). Then one determines the number of replicates (r) needed and uses this quantity to determine the total sample size n from $n = rn_r$, where n_r is the sample size within each replicate. A method for the determination of r is $r = (z_\alpha + z_\beta)^2\ Var/\Delta^2$, where z_θ is the 100 $(1 - \theta)$th percentile of the standard normal distribution, Δ is the effect size, and Var is the applicable variance for a replicate of n_r patients with the design structure and possible covariates taken into account. As the variance increases, power decreases, unless sample size is increased to offset the increasing variability of the data.

After a study has been completed, its data need to be analyzed; there are two different

analysis postures that one can apply (13). One is the use of nonparametric methods, such as Mantel-Haenszel tests (14–17), on the basis of randomization and its implied permutation distributions for data. These methods have the advantage of requiring minimal assumptions about homogeneity of treatment differences across centers or other factors or about sample sizes for centers. For $h = 1, \ldots, H$ levels of cross-classified stratification factors, $Q_R = (\Sigma w_h d_h)^2/\Sigma w_h^2 v_h$ can be used to compare two treatments (or with s treatments, to assess trend), where d_h is a general measure for the difference between treatments in the hth stratum. Typically, d_h is a difference in proportions or means of actual values or ranks, but it can include covariance adjustment; w_h is a weight for the hth stratum that is usually based on sample size $((n_{h1}n_{h2})/(n_{h1} + n_{h2}))$, but other choices are possible (eg, $w_h = 1$, for all h, weights strata equally or $w_h = ((n_{h1}n_{h2})/(n_{h1} + n_{h2}))^c$ with $0 \leq c \leq 1$ weights strata according to sample size with lessening degree as c decreases from 1 to 0); v_h is the applicable variance for d_h under the null hypothesis (no treatment difference). For large samples, Q_R is approximately chi-square in distribution with one degree of freedom. Another virtue of these methods is that their application has the same basic form, regardless of whether the outcome is dichotomous, ordered, continuous, or time-to-event. Although the advantages of these nonparametric methods are appealing, any plan to apply them to a study should be specified in the protocol or in a statistical amendment that is prepared before any unmasking of the treatments for the study.

The other analysis posture involves statistical models. These methods are advantageous in explaining the role of treatment differences in the variation of response variables. These methods, however, usually require additional nonstatistical arguments to justify assumptions that the data under study are like a statistically random sample; since centers and patients in most studies are selected for inclusion by convenience (18), the fundamental assumptions for modeling methods are debatable. The scope of modeling methods includes: logistic regression for dichotomous response, polychotomous logistic regression for nominal response, ordinal logistic (proportional odds model) regression for ordinal response, multiple linear regression for continuous response, and proportional hazards regression for time-to-event response. Table 2 (adapted from 18) simply gives references for dichotomous, ordinal, continuous, and time-to-event data for various analyses (13,19). These methods can be specified in some form in protocols and statistical amendments for analysis plans.

In summary, when attempting to address multiplicity, one should decide how to allocate the significance level to the various multiplicity components. In this regard, one should first address the interim analyses to determine the overall significance level for each interim analysis. Then, one should partition the significance level per interim analysis according to the subgroups, if there are multiple subgroups. After that, one should focus on treatment comparisons and define ways to avoid reducing the significance level any further because of multiplicity of treatment comparisons.

One way to do this is to use closed or hierarchical procedures; another is to use the Bonferroni-Holm (20,21) method. A third option, called the Hailperin-Rüger method (7,22,23), simply applies an adjusted significance level ($\alpha^* = \alpha j/k$, where j/k is the proportion of k total comparisons required to be significant to declare overall significance). For example, for four tests with the goal of having any three of four as significant, at the 0.050 level, one would test each at three-fourths of 0.050 (ie, a 0.0375 level). Another procedure is the Westfall-Young bootstrap multiple comparison method (24); it is advantageous because it accounts for correlation between tests (to the extent of allowing for redundancies); it can be used with small sample sizes or for assessment of outcomes with small probabilities of occurrence; and it can incorporate a hierarchical procedure for testing. Finally, one should address the

TABLE 2
Summary for Some Statistical Methods and Related References

	Nonparametric Methods Based on Randomization		Model-based Methods for Random Samples
Measurement Scale	One Stratum	Combined Strata	Combined Strata
Continuous determinants	Wilcoxon rank sum test, Kruskal-Wallis test, Rank analysis of covariance (13,19)	van Elteren test, Stratified rank analysis of covariance (13)	Multiple linear regression, analysis of covariance (19)
Dichotomous categories	Fisher's exact test, chi-square tests (13)	Mantel-Haenszel test (13)	Logistic regression (13,19)
Ordinal categories or discrete counts	Wilcoxon rank sum test, Kruskal-Wallis test with midranks for ties (13)	Extended Mantel-Haenszel test (13)	Proportional odds model, Equal adjacent odds model (13)
Censored survival times	Logrank test (13)	Extended Mantel-Haenszel (logrank) test (13)	Proportional hazards regression (13)

Adapted from (18).

multiple response criteria. Here, one could use a closed hierarchical procedure, the O'Brien procedure (ie, averaging the ranks), the Bonferroni-Holm method, the Hailperin-Rüger method, the Westfall-Young bootstrap method, or Lachin's multivariate nonparametric method (25).

It is important to recognize that strategies to address multiplicity require increased sample size to achieve lower significance levels and to avoid reduced power from multiple opportunities for not achieving statistical significance. Sample size may also need to be increased to avoid reduced power from multiple analyses being required to address such issues as intent-to-treat versus protocol compatible analyses, or issues raised by regulatory reviewers, with all such analyses being required to yield statistical significance (26). Because requiring all multiple analyses to yield statistical significance increases the Type II error and reduces power in somewhat unexpected ways, one may need to increase the sample size 25–30% beyond that which the scientific principles of study design might indicate for a perfectly conducted study which did not have any difficulties in its data structure (eg, no missing data, no patients withdrawing from the study prior to its completion, and no required multiple analyses).

CONCLUSION

Statistical issues concerning multiplicity require careful attention in the design, analysis, and interpretation of confirmatory clinical studies in order to avoid bias in estimates of treatment effects, to avoid excessive likelihood for chance findings, and to avoid low power to detect real treatment differences. Strategies are available to address these statistical issues in ways which can provide satisfactory power for primary objectives. These strategies should be specified in the protocol (or prior to unmasking a study) in order to be applicable.

Acknowledgments—The authors wish to thank Dr. Susan Ellenberg for her helpful comments and suggestions in reviewing the manuscript. Further, they would like to thank Cara Davis, Kim Kusy, and Brian Mabe for their word processing assistance. Stuart Gansky's work was partially supported by the Department of Veterans Affairs, Office of Academic Affairs' Predoctoral Fellowship in Health Services Research.

REFERENCES

1. Koch GG, Amara IA, Forster J, McSorley D, Peace KE. Statistical issues in the design and analysis of ulcer healing and recurrence studies. *Drug Inf J*. 1993;17:805–824.
2. Elashoff JD, Koch GG. Statistical methods in trials of anti-ulcer drugs. In: Swabb EA, Szabo S, eds. *Ulcer Disease Investigation and Basis for Therapy*. New York: Marcel Dekker, Inc.; 1991:375–406.
3. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1984;40:1079–1087.
4. Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics*. 1991;47:511–521.
5. Koch GG, Carr GJ, Amara IA, Stokes ME, Uryniak TJ. Categorical data analysis. In: Berry DA, ed. *Statistical Methodology in the Pharmaceutical Sciences*. New York: Marcel Dekker, Inc.; 1990;389–473.
6. Koch GG. Comment. *Stat Med*. 1991;10:13–16.
7. Bauer P. Multiple testing in clinical trials. *Stat Med*. 1991;10:871–890.
8. Gansky SA, Koch GG, Wilson J. Statistical evaluation of relationships between analgesic dose and ordered ratings of pain relief over an eight-hour period. *J Biopharm Stat*. 1994;4(2):233–265.
9. Phillips JA, Cairns V, Koch GG. The analysis of a multiple-dose, combination-drug clinical trial using response surface methodology. *J Biopharm Stat*. 1992;2(1):49–67.
10. Koch GG. Discussion. *Biopharm Report*. 1993;2(1):7–8.
11. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35:549–556.
12. DeMets DL, Lan KKG. Interim analysis: The alpha spending function approach. *Stat Med*. 1994;13(13/14):1341–1352.
13. Koch GG, Edward S. Clinical efficacy trials with categorical data. In: Peace KE, ed. *Biopharmaceutical Statistics for Drug Development*. New York: Marcel Dekker; 1988:403–457.
14. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959;22:719–748.
15. Mantel N. Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *J Am Stat Assoc*. 1963;58:690–700.
16. Koch GG, Amara IA, Davis GW, Gillings DB. A review of some statistical methods for covariance analysis of categorical data. *Biometrics*. 1982;38:563–595.
17. Kurtiz SJ, Landis JR, Koch GG. A general overview of Mantel-Haenszel methods: Applications and recent developments. *Ann Rev Pub Health*. 1988;9:123–160.
18. Koch GG, Sollecito WA. Statistical considerations in the design, analysis, and interpretation of comparative clinical studies: An academic perspective. *Drug Inf J*. 1984;18:131–151.
19. Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York: Wiley; 1986.
20. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian J Stat*. 1979;6:65–70.
21. Shaffer J. Modified sequentially rejective multiple test problems. *J Am Stat Assoc*. 1986;81:826–831.
22. Hailperin R. Best possible inequalities for the probability of a logical function of events. *Am Math Monthly*. 1965;72:343–359.
23. Rüger B. Das maximale signifikanzniveau des tests: "Lehne H_0 ab, wenn *k* unter *n* gegebenen tests zur ablehnung führen." *Metrika*. 1978;25:171–178.
24. Westfall PH, Young SS. *Resampling-based Multiple Comparison Testing: Examples and Methods for p-Value Adjustment*. New York: Wiley; 1993.
25. Lachin J. Some large-sample distribution-free estimators and tests for multivariate, partially incomplete data from two populations. *Stat Med*. 1992;11:1151–1170.
26. Gillings D, Koch GG. The application of intention-to-treat to the analysis of clinical trials. *Drug Inf J*. 1991;25:411–424.