BIOMETRICS 55, 1188–1192
December 1999

Closed Testing Procedures for Group Sequential Clinical Trials with Multiple Endpoints

Dei-In Tang

Statistical Sciences and Epidemiology Division, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York, U.S.A.
email: tang@NKI.RFMH.ORG

and

Nancy L. Geller

Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, Maryland, U.S.A.

SUMMARY. A simple approach is given for conducting closed testing in clinical trials with multiple endpoints in which group sequential monitoring is planned. The approach allows a flexible stopping time; the earliest and latest stopping times are described. The paradigm is applicable both to clinical trials with multiple endpoints and to the one-sided multiple comparison problem of several treatments versus a control. The approach leads to enhancements of previous methods and suggestions for new methods. An example of a respiratory disease trial with four endpoints is given.

KEY WORDS: Early stopping; Several treatments versus a control; Step-down procedures; Strong type I error control.

1. Introduction

Certain clinical trials do not lend themselves to a unique primary endpoint. For example, in a recent trial on stroke treatment, four endpoints were considered without any of them having priority (National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995; Tilley et al., 1996). O'Brien (1984) first proposed tests of the global null hypothesis that two treatments are equally effective with respect to multiple endpoints against the restricted alternative that a specified treatment is more effective than the other in at least one endpoint. Tang, Gnecco, and Geller (1989a) proposed an alternative test with a more stable power function. Some of these tests were extended to the group sequential setting (Tang, Gnecco, and Geller, 1989b; Tang, Geller, and Pocock, 1993) and the closed testing setting (Lehmacher, Wassmer, and Reitmeir, 1991).

We further these developments by describing an approach for conducting closed testing in the group sequential setting. Our approach requires a group sequential boundary for each of the hypotheses to be tested in the closed testing framework but is otherwise simple. In addition, it allows a flexible stopping time. We give two specific procedures corresponding to the earliest and latest stopping times. We illustrate the procedures with an asthma trial data set. The procedures are also applicable to the problem of comparing multiple treatments to a control with a single endpoint. Comparison to existing procedures is made.

2. Notation

Let a new treatment be compared to a control treatment based on k endpoints. For simplicity, each treatment is given to m patients. Denote the response data by $\{X_{ij}, i = 1, 2$ and $j = 1, 2, \ldots, m\}$, with $i = 1$ indicating the control treatment and $i = 2$ the new treatment. Assume X_{ij}, a vector of length k, has mean vector μ_i and unknown variance–covariance matrix Σ. We write $\mu_i = (\mu_{i1}, \ldots, \mu_{ik})'$ to refer to the components of μ_i. A commonly considered null hypothesis is that there is no difference in effect between the two treatments, i.e., H_0: $\mu_1 = \mu_2$. Often it is also of interest to test the k single-endpoint null hypotheses, $H_{0,i}$: $\mu_{1i} = \mu_{2i}$, $i = 1, 2, \ldots, k$, of which the intersection is the global hypothesis H_0. We consider testing H_0 against the alternative hypothesis that $\mu_{2i} \geq \mu_{1i}$ for all $i = 1, 2, \ldots, k$ with at least one inequality strict. This alternative hypothesis reflects the clinical objective of detecting an improved new treatment. The tests of O'Brien (1984), Tang et al. (1989a), Tang et al. (1993), and Follmann (1996) are designed to be more powerful than the traditional T^2 test (Anderson, 1986) for this alternative hypothesis.

3. Background: Closed Testing for Multiple Endpoints in Fixed Sample Trials

Although rejecting the global hypothesis would be evidence for the efficacy of the new treatment, rejecting single-endpoint hypotheses would make the evidence more specific and compelling. To this end, Lehmacher et al. (1991) discussed the use of the closed testing method of Marcus, Peritz, and Gabriel

(1976) in conjunction with multiple endpoint statistics. To describe the closed testing procedure, define for any nonempty subset of the k endpoints a null hypothesis that the means of the two treatments are equal for this subset of endpoints. Refer to the number of endpoints in the subset as the dimension of the hypothesis defined. These hypotheses form a closed set in the sense that the intersection of any two member hypotheses remains a member. The closed testing procedure first tests H_0. If the test statistic based on the k endpoints is not significant at level α, then H_0 is not rejected and the procedure stops. If significant, then H_0 is rejected and the procedure steps down to test all $(k-1)$-dimensional hypotheses, each at level α. This involves deleting one endpoint, recalculating the test statistic based on the reduced data set, and, perhaps, using a new critical value. In principle, one steps down again to test $(k-2)$-dimensional hypotheses if any of the $(k-1)$-dimensional hypotheses is rejected. Further, one steps down as long as a hypothesis of dimension two or more is rejected. However, if a hypothesis is tested and not rejected, then any hypothesis that is implied by it is automatically not rejected. Thus, a hypothesis is actually tested if and only if all the other hypotheses that imply it are tested and rejected. Strong control of the type I error at level α means that the probability of rejecting one or more correct hypotheses is at most α. The closed testing method maintains strong control of the type I error at level α. Lehmacher et al. (1991) conducted simulations of various closed testing procedures and concluded power advantages of O'Brien's linear combination statistics over Hotelling's T^2 and Bonferroni-type procedures. These advantages were carried over for all lower dimensional marginal hypotheses.

4. Closed Testing for Multiple Endpoints in Group Sequential Trials

Let $K = \{1, 2, \ldots, k\}$ be the set of indices for the k endpoints. Let F denote a nonempty subset of K and $H_{0,F}$ the null hypothesis $\mu_i = 0$ for $i \in F$. Let Z_F be a test statistic for $H_{0,F}$. Consider a group sequential trial with g analyses at prespecified times. Similar results can be given for trials with unprespecified analysis times based on the use function approach (Lan and DeMets, 1983). We use $Z_{F,t}$ to indicate the dependence of Z_F on the analysis time t. Let $\{c_{F,t},\ t = 1, 2, \ldots, g\}$ be a one-sided group sequential boundary for testing $H_{0,F}$, i.e., $P_{H_{0,F}}\{Z_{F,t} > c_{F,t} \text{ for some } t\} \leq \alpha$. For a given vector μ, let $I_\mu = \{i,\ \mu_i = 0\}$.

PROPOSITION 1: *The following procedure preserves strong control of the type I error:*

Step 1. *Conduct interim analyses to test* $H_{0,K}$, *based on the group sequential boundary* $\{c_{K,t},\ t = 1, 2, \ldots, g\}$.

Step 2. *When* $H_{0,K}$ *is rejected, say at time* t^*, *stop the trial and apply the closed testing procedure to test all the other hypotheses* $H_{0,F}$ *using* Z_{F,t^*} *with* c_{F,t^*} *as the critical value.*

Step 3. *If the trial continues to the last analysis without rejection of* $H_{0,K}$, *then no hypotheses are rejected.*

Proof. A type I error occurs if, for some $F \subseteq I_\mu$, $H_{0,F}$ is rejected, where μ denotes the underlying mean vector. According to the closed testing procedure, $H_{0,F}$ can be rejected only if H_{0,I_μ} is rejected. Thus, {type I error occurs} $= \cup$\{type I error occurs at time t\} $\subseteq \cup$\{reject H_{0,I_μ} at time t\} $\subseteq \cup\{Z_{I_\mu,t} > c_{I_\mu,t}\} = \{Z_{I_\mu,t} > c_{I_\mu,t}$ for some t, $1 \leq t \leq g\}$, where each union is taken over $t = 1, 2, \ldots, g$. Hence, the probability of making at least one type I error is at most $P\{Z_{I_\mu,t} > c_{I_\mu,t}$ for some $t\} \leq \alpha$.

Proposition 1 does not allow continuation of the trial once the global test crosses its boundary. Proposition 2 below, which can be proved in essentially the same way, allows the trial to continue until all hypotheses are rejected or the last analysis is conducted.

PROPOSITION 2: *The following procedure preserves strong control of the type I error:*

Step 1. *Conduct interim analyses to test* $H_{0,K}$, *based on the group sequential boundary* $\{c_{K,t},\ t = 1, 2, \ldots, g\}$.

Step 2. *When* $H_{0,K}$ *is rejected, say at time* t^*, *apply the closed testing procedure to test all the other hypotheses* $H_{0,F}$ *using* Z_{F,t^*} *with* c_{F,t^*} *as the critical value.*

Step 3. *If any hypothesis is not rejected, continue the trial to the next stage, in which the closed testing procedure is repeated, with the previously rejected hypotheses automatically rejected without retesting.*

Step 4. *Reiterate Step 3 until all hypotheses are rejected or the last stage is reached.*

In step 3 of Proposition 2, one could choose to retest the previously rejected hypotheses, as suggested by the sufficiency principle. Since previous rejections may be revoked, this would reduce power. Propositions 1 and 2 imply that the trial may be stopped at any planned interim analysis time between the two stopping times without compromising the type I error control. The two procedures described in the propositions will be referred to as procedures 1 and 2.

5. Applications

5.1 *A Group Sequential Trial with Multiple Endpoints: A Crossover Trial in Chronic Respiratory Disease*

Pocock, Geller, and Tsiatis (1987) and Tang et al. (1993) discussed this data set. Seventeen patients with asthma or chronic obstructive airways disease were randomized in a double-blind crossover trial to an inhaled drug or placebo. Four measures of standard respiratory function (FEV_1, FVC, PEFR, and PI) were of interest. For illustration purposes, we assume that there were no period or carryover effects and that this was the first of two equally sized groups of patients to enter a group sequential trial. We also assume the data are normal and use an estimated covariance matrix as the underlying covariance matrix. For every hypothesis to be tested, we use the corresponding centered linear combination statistic (Tang et al., 1993, equation 3.1). Thus, the test for every hypothesis is a univariate one-sided normal test and we can use the same sequential boundary for all tests. Suppose we use a constant boundary of level 0.05. According to DeMets and Ware (1980), both critical values are equal to 1.88. The results for each outcome measure comparing the drug to the control are given in Table 1.

The estimated covariance matrix S is

$$\begin{bmatrix} 1.000 & 0.095 & 0.219 & -0.162 \\ & 1.000 & 0.518 & -0.059 \\ & & 1.000 & 0.513 \\ & & & 1.000 \end{bmatrix}.$$

Table 1
Results for outcome measures comparing drug to control

	FEV_1 (%)	FVC (%)	PEFR (liters/minute)	PI (%)
Mean difference	7.56	4.81	2.29	0.08
Standard deviation of difference	18.53	10.84	8.51	0.17
t-Value	1.63	1.77	1.11	1.85

Let y be the vector formed by the four t-values and d be the vector formed by the square roots of the diagonal elements of S^{-1}. The centered linear combination statistic based on all four endpoints is $Z_{\{1,2,3,4\}} = d'y/(d'Sd)^{1/2}$. For any subset of endpoints F, the test statistic Z_F follows the same formula except that y and S are redefined to include only the endpoints in F.

Suppose that this trial was designed using procedure 1. We first calculate $Z_{\{1,2,3,4\}} = 2.41$. Since this exceeds the critical value 1.88, the global hypothesis is rejected. Hence, we stop the trial and conduct closed testing. In the first stage, we obtain $Z_{\{1,2,3\}} = 2.06$, $Z_{\{1,2,4\}} = 3.17$, $Z_{\{1,3,4\}} = 2.22$, and $Z_{\{2,3,4\}} = 2.07$, all of them larger than 1.88, so the corresponding hypotheses are rejected. We can then proceed to calculate $Z_{\{1,2\}} = 2.3$, $Z_{\{1,4\}} = 2.69$, and $Z_{\{2,4\}} = 2.64$, based on which we reject the corresponding hypotheses. We also calculate $Z_{\{1,3\}} = 1.76$, $Z_{\{2,3\}} = 1.65$, and $Z_{\{3,4\}} = 1.7$, based on which we cannot reject the corresponding hypotheses. Because these last three hypotheses are not rejected, the single-endpoint hypotheses cannot be rejected either. Notice that since $Z_{\{1\}} = 1.63$, $Z_{\{2\}} = 1.77$, $Z_{\{3\}} = 1.11$, and $Z_{\{4\}} = 1.85$, we could not have rejected any of the single-endpoint hypotheses even if we had been able to test them.

If the trial were designed using procedure 2, then the trial would continue to accrue another 17 patients. In the second analysis, we would repeat the closed testing procedure subject to the condition that those hypotheses already rejected would be automatically rejected. We would retest only $H_{0,\{1,3\}}$, $H_{0,\{2,3\}}$, $H_{0,\{3,4\}}$, and all of the single-endpoint hypotheses.

The observed treatment difference in each endpoint was moderate, so a Bonferroni approach would find no treatment effect. The global test achieved rejection by taking advantage of the consistency of evidence across the endpoints. The third endpoint, PEFR, showed a relatively small difference and was highly correlated with PI. As a result, all the two-endpoint differences that involve PEFR were not significant and no single endpoint hypothesis could be actually tested. If PEFR had not been included, the step down procedure would have tested the single endpoint hypotheses, although none of them would be rejected.

If, instead, the trial had a fixed sample size and a single primary endpoint, the conclusions would depend on which endpoint had been chosen since the critical value would be 1.65. If the other endpoints were considered secondary, interpretation would be difficult. With a global test, there was no question that the trial was positive. With the increased sample size provided by the second stage, there would be increased power to find individual endpoint differences.

5.2 *Comparing Several Treatments to a Control Based on a Single Endpoint*

Suppose there are A (>2) treatments and treatment 1 is the control to which the other treatments are to be compared based on a univariate response variable. Let X_i denote the sample mean of treatment i. Consider testing the null hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_A$ against the alternative hypothesis H_1: $\mu_2 - \mu_1 \geq 0, \mu_3 - \mu_1 \geq 0, \ldots, \mu_A - \mu_1 \geq 0$, with at least one inequality strict. By considering $Y_i = X_i - X_1$ and $\theta_i = \mu_i - \mu_1$, the problem reduces to the multiple endpoint problem except that the covariance matrix now has a structure. All of the multiple endpoint statistics that have been proposed based on the nonnegative orthant alternative hypothesis are applicable here. One can, of course, use tests that are developed for the original formulation of multiple comparisons (see Tang and Lin [1997] for a survey of such tests in the light of closed testing). Propositions 1 and 2 are applicable to this problem.

Here, rejection of the global hypothesis without rejection of at least one pairwise comparison hypothesis would be unsatisfactory. One might continue the trial until at least one pairwise comparison is significant so that at least one treatment better than the control is identified, provided the maximum sample size has not been reached. This stopping time is permissible and lies between those of Propositions 1 and 2.

6. Comparison with Existing Procedures

Jennison and Turnbull (JT) (1993) proposed a formulation of the multiple endpoint problem that aims at rejection of the single-endpoint null hypotheses all at once. The null hypothesis is H_{JT}: $\mu_{2i} \leq \mu_{1i}$ for at least one i and the alternative is K_{JT}: $\mu_{2i} > \mu_{1i}$ for all i. In addition to fixed sample testing procedures, they developed group sequential procedures that have both upper and lower boundaries.

In the fixed sample setting, our approach, which reduces to that of Lehmacher et al. (1991), can accomplish the same goal of rejecting H_{JT} if rejection is obtained throughout the closed testing process. There is no loss of power if a linear combination statistic with nonnegative weights is used (see Tang and Lin [1994] for details). In the group sequential setting, similar claims can be made if the trial continues until H_{JT} is rejected or the maximum sample size is reached. An earlier stopping time is permissible if the evidence then available is deemed sufficient. Although this advantage seems unimportant if the trial's only goal is to reject H_{JT}, there is nothing to lose with our approach except for some computational effort.

Follmann, Proschan, and Geller (1994) proposed a method for sequentially comparing all pairs of three treatments. In light of our results, their procedure can be seen as a group sequential closed testing procedure. For making all pairwise comparisons among any number of treatments, Liu (1995) described a group sequential approach based on the range statistic. Since the range test can be viewed as a closed testing procedure, our approach can be applied based on it. For example, Proposition 2 permits one to use any of the sequential boundaries given by Liu to monitor the trial until all pairwise comparison hypotheses have been rejected or the last stage is reached. The group sequential approach of Liu

takes the form of repeated simultaneous confidence intervals in the spirit of Jennison and Turnbull (1984) but is equivalent to our method for testing purposes.

7. Discussion

Since the proposed procedures combine the ideas of global testing, sequential testing, and closed testing, their power performance can be assessed from available results on individual testing techniques. In the fixed sample setting, the global test approach is more powerful than the Bonferroni adjustment approach (O'Brien, 1984), the single endpoint approach (Tang et al., 1989b), and the combined endpoint approach for multiple binary outcomes (Lefkopoulou and Ryan, 1993). When a test is applied in a group sequential manner, there is in general a loss of power as compared to the corresponding fixed sample test based on the maximum sample size. This loss is justified by the possibility of early stopping. With an appropriate choice of the sequential boundary, such as the O'Brien–Fleming (O'Brien and Fleming, 1979) type, the loss can be quite small. In closed testing, the power of the global test is unaffected, although the power of the other tests is reduced because there are prerequisites for their conduct. This power characteristic is a strength of the closed testing method as compared to other methods that control type I error by lowering the levels of all of the tests to be conducted. Relevant simulation results on power performance of closed testing procedures were given by Lehmacher et al. (1991) and Tang and Lin (1997). These results indicate that the power advantage of the global test can filter down to the lower dimensional tests. The latter work also demonstrates that a global test-based closed testing procedure can have more power even for single endpoint hypotheses than a traditional simultaneous inference procedure consisting of endpointwise analyses. From the above considerations, we conclude that the proposed procedures are not power deficient, as might be suspected in view of the strong requirement of type I error control over both multiple endpoints and multiple looks. In contrast, the Bonferroni adjustment approach would obviously have a power problem.

A main criticism of the global test approach to the multiple endpoints problem has been that the rejection of the global hypothesis may not be sufficient evidence for treatment selection. So following the global test with other tests according to the closed testing method is a useful way to strengthen the evidence without sacrificing either the rigor of type I error control or the power advantage of the global test.

We have shown how to use the closed testing method to achieve a strong type I error control in conducting multiple group sequential tests. This approach can easily be extended to any other methods that achieve a strong control in a similar way, i.e., by arranging the multiple tests, each of level α, in a sequential manner such that for every possible configuration of true hypotheses there is a hypothesis whose rejection is a prerequisite for rejection of the others. Maurer, Hothorn, and Lehmacher (1995) proposed such a method for comparison of multiple treatments based on an *a priori* partial ordering of all of the hypotheses of interest. In the case of multiple endpoints, this method suggests that one could order the endpoints and test them one by one in such a way that an endpoint is tested only if all the preceeding ones have been rejected.

The applicability of our approach is limited by whether a test method has been developed for every hypothesis to be tested and, if so, whether a corresponding group sequential boundary has been developed. For example, for the all pairs comparison problem, discussed in Section 6, a frequently used test for the global hypothesis is the χ^2 test. For this test, group sequential procedures have been developed (Jennison and Turnbull, 1991). However, the test cannot be applied to those intermediate hypotheses that consist of two or more homogeneity statements for disjoint sets of means, assuming there are at least four treatments. So, for the moment, the range statistic seems the only choice.

Acknowledgement

The research of the first author was supported by grants from the National Institutes of Health.

Résumé

L'article présente une approche simple permettant de mettre en place des procédures de tests dans des essais cliniques comportant plusieurs critères d'évaluation pour lesquels il est prévu de réaliser un contrôle séquentiel. L'approche permet de définir, de manière flexible, un temps d'arrêt; le temps d'arrêt le plus précoce et le temps d'arrêt le plus tardif sont également décrits. Le modèle est applicable aussi bien à des essais cliniques avec plusieurs critères d'évaluation qu'au problème des comparaisons multiples de plusieurs traitements à un témoin. L'approche proposée a permis de mettre en valeur des méthodes antérieures et de formuler des suggestions pour de nouvelles méthodes. Un exemple d'essai sur une maladie respiratoire avec quatre critères d'évaluation est proposé.

References

Anderson, T. W. (1986). *Introduction to Multivariate Statistical Analysis,* 2nd edition. New York: Wiley.

DeMets, D. L. and Ware, J. H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika* **67,** 651–660.

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

Follmann, D. A., Proschan, M. A., and Geller, N. L. (1994). Monitoring pairwise comparisons in multiarmed clinical trials. *Biometrics* **50,** 325–336.

Jennison, C. and Turnbull, B. W. (1984). Interim analysis: The repeated confidence interval approach. *Controlled Clinical Trials* **5,** 33–45.

Jennison, C. and Turnbull, B. W. (1991). Exact calculations for sequential t, χ^2 and F tests. *Biometrika* **78,** 133–141.

Jennison, C. and Turnbull, B. W. (1993). Group sequential tests for bivariate response: Interim analysis of clinical trials with both efficacy and safety endpoints. *Biometrics* **49,** 741–752.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70,** 659–663.

Lefkopoulou, M. and Ryan, L. (1993). Global tests for multiple binary outcomes. *Biometrics* **49,** 975–988.

Lehmacher, W., Wassmer, G., and Reitmeir, P. (1991). Procedures for two sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **47,** 511–521.

Liu, W. (1995). A group sequential procedure for all-pairwise comparisons of k treatments based on the range statistic. *Biometrics* **51,** 946–955.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63,** 655–660.

Maurer, W., Hothorn, L. A., and Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and pre-clinical assays: *A priori* ordered hypotheses. In *Biometrie in der Chemisch-Pharmazeutischen Industrie,* Joachim Vollmar (ed), 3–18. Gustav Fischer Verlag: New York.

National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. (1995). Tissue plasminogen activator for acute ischemic stroke. *New England Journal of Medicine* **333,** 1581–1587.

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

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35,** 549–556.

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

Proschan, M. A., Follmann, D. A., and Geller, N. L. (1994). Monitoring multiarmed trials. *Statistics in Medicine* **13,** 1441–1452.

Tang, D.-I. and Lin, S. P. (1994). On improving some methods for multiple endpoints. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, 466–474.

Tang, D.-I. and Lin, S. P. (1997). An approximate likelihood-ratio test for comparing several treatments to a control. *Journal of the American Statistical Association* **92,** 1155–1162.

Tang, D.-I., Gnecco, C., and Geller, N. L. (1989a). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika* **76,** 577–583.

Tang, D.-I., Gnecco, C., and Geller, N. L. (1989b). Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association* **84,** 776–779.

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

Tilley, B. C., Marler, J., Geller, N. L., Lu, M., Legler, J., Brott, T., Lyden, P., and Grotta, J. (1996). Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA stroke trial. *Stroke* **27,** 2136–2142.

Received February 1998. *Revised February* 1999.
Accepted March 1999.