Biometrical Journal **41** (1999) 3, 261–277

Inference on Multiple Endpoints in Clinical Trials with Adaptive Interim Analyses

MEINHARD KIESER

Dr. Willmar Schwabe Pharmaceuticals
Department of Biometry
Karlsruhe
Germany

PETER BAUER

Department of Medical Statistics
University of Vienna
Austria

WALTER LEHMACHER

Department of Medical Statistics, Informatics and Epidemiology
University of Cologne
Germany

Summary

Planned interim analyses which permit early stopping or sample size adaption of a trial are desirable for ethical and scientific reasons. Multiple test procedures allow inference about several hypotheses within a single clinical trial. In this paper, a method which combines multiple testing with adaptive interim analyses whilst controlling the experimentwise error rate is proposed. The general closed testing principle, the situation of *a priori* ordered hypotheses, and application of the Bonferroni-Holm method are considered. The practical application of the method is demonstrated by an example.

Key words: Adaptive interim analyses; Clinical trials; Early stopping; Experimentwise error rate; Multiple endpoints; Multiple inference.

1. Introduction

For ethical, scientific and economic reasons alike, there is an increasing demand in many clinical trials to allow inference on several clinically relevant parameters, with the smallest possible number of patients, and within a minimum of time. Multiplicity issues arise quite often in clinical trials. For example, in some disease areas the judgement of the efficacy of a treatment cannot be based on a single outcome measure, but multiple endpoints have to be evaluated. An overview of examples and methods for multiple testing in clinical trials is given by BAUER (1991), and specific methods for the multiple endpoint situation when comparing two treatments in classical fixed sample size designs have been derived (see, e.g., O'BRIEN, 1984; LEHMACHER, WASSMER, and REITMEIR, 1991; LÄUTER, 1996). Planning interim analyses permits an early detection of treatment superiority as well as an early termination of a trial due to lack of efficacy or to safety concerns. Group sequential designs (POCOCK, 1977; O'BRIEN and FLEMING, 1979; LAN and DEMETS, 1983) are widely used to accomodate these demands by periodically monitoring safety and efficacy during the course of the study.

A major drawback of group sequential designs lies in the fact that they require an *a priori* specification of the maximum number of patients to be included in the trial. The concept of adaptive interim analyses proposed by BAUER and KÖHNE (1994) overcomes this difficulty. As an advantage to the group sequential approach it does not only allow for early stopping but additionally provides the opportunity to re-evaluate assumptions that have been made when the protocol was written, and to correct possible misjudgements in the course of the trial, for example with respect to the variability of the outcome parameter.

A combination of multiple endpoint testing and planned interim analyses seems quite appealing. LEE (1994) presented an overview about the determination of local significance levels for various multivariate tests to be applied in group sequential designs. However, only methods for the assessment of the global hypothesis in interim analyses were considered up to now, whereas the important problem of follow-up inference about the individual endpoints has not been examined.

It is the purpose of this article to propose methods for multiple inference on multiple endpoints in studies with planned interim analyses which control the experimentwise error rate. We restrict the basic arguments to the standard multiple endpoint scenario, where the endpoints are measured throughout the trial (leaving aside the situation that omitting endpoint measurements would result in a substantial reduction of the burden to the patients or of costs). For adaptive two-stage designs BAUER and KIESER (1999) presented a general multiple test procedure allowing that the set of hypotheses to be tested is reduced after the interim analysis. The situation of no restriction in the choice of a single null hypothesis for the second stage has been investigated by HOMMEL (1997). Although we are focussing on adaptive interim analyses, the methods presented in this paper can also be applied in group sequential trials.

In the following section, the concept of adaptive interim analyses and the multiple test problem to be assessed in this setting are described. Section 3 presents a closed test procedure for clinical trials with an adaptive design. As special cases of this general procedure the corresponding strategies for *a priori* ordered hypotheses and the Bonferroni-Holm method are described in Section 4. Furthermore, some hints concerning the choice of the critical boundaries for interim and final analysis are given. Rules for the calculation of the sample size for the second stage are presented in Section 5, and practical application of the procedures is demonstrated in Section 6 by a clinical trial example. Section 7 contains concluding remarks and a discussion of the proposed methods.

2. Formulation of the Problem

We consider clinical trials with planned adaptive interim analyses (Bauer and Köhne, 1994). One purpose of an adaptive interim analysis is to stop the trial as soon as the decision about the interesting null hypotheses can be made. Although the method in principle allows a substantial re-design of the trial, we will primarily discuss the aspect of re-estimation of sample size after the interim analysis.

In the simplest case of one interim analysis and a single one-sided null hypothesis H_0 to be tested in a confirmatory analysis the method works as follows. Let p_1 and p_2 be the one-sided p-values corresponding to H_0 for the samples of the first and the second stage of the trial, respectively. Note that a one-sided formulation of null hypotheses is adopted in order to prevent possible conflicting directional decisions between stages. (This will be particularly important for treatment comparisons along multiple endpoints, where the endpoints within and between the stages should deviate in the same "direction".) The trial stops after the interim analysis with rejection of H_0 if $p_1 \leq \alpha_1$, and with acceptance if $p_1 \geq \alpha_0$. If $\alpha_1 < p_1 < \alpha_0$, the trial is continued. Using Fisher's combination test, H_0 can be rejected after the second stage if $p_1 \cdot p_2 \leq c_{\alpha_2}$ while H_0 is accepted otherwise. Figure 1 illustrates the adaptive design for the situation of a single test problem H_0 versus H_1 and one interim analysis. The approach can also be generalized to designs with more than one interim analysis (Bauer and Köhne, 1994; Bauer and Röhmel, 1995).

In the original paper of Bauer and Köhne (1994) only the situation $\alpha_2 = \alpha$ was considered. Other choices of α_2 correspond to other α-spending (Lan and DeMets, 1983) between interim analysis (local level α_1) and final combination test (local level α_2) (see Bauer and Röhmel, 1995; Bauer, Bauer, and Budde, 1998). The critical boundaries α_1 and α_2 are to be determined such that the Type I error rate is controlled by α. Using the fact that under H_0 the p-values p_1 and p_2 for the tests performed in stochastically independent samples are independently and uniformly distributed (or stochastically not smaller than the uniform distribution), the critical boundary c_{α_2} for Fisher's combination test at a local level $\alpha_2 \leq \alpha$ is given by

$$c_{\alpha_2} = \exp\left[\left(-\tfrac{1}{2}\,\chi_4^2(1-\alpha_2)\right]\right., \tag{1}$$

where $\chi_4^2(1-\alpha_2)$ denotes the $(1-\alpha_2)$-quantile of the χ^2-distribution with 4 degrees of freedom.

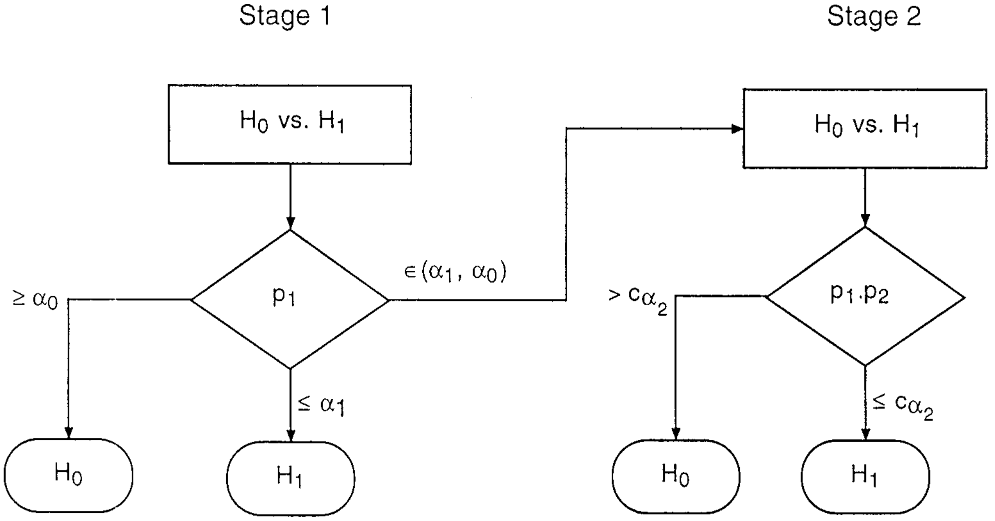

Fig. 1. The adaptive design with one interim analysis and the same null hypothesis H_0 tested in both stages

The Type I error rate of the two-stage design is given by

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_{0}^{c_{\alpha_2}/p_1} \mathrm{d}p_2\,\mathrm{d}p_1 = \alpha_1 + c_{\alpha_2} \cdot (\ln \alpha_0 - \ln \alpha_1)\,. \tag{2}$$

Therefore, in order to control the experimentwise error rate by α, the critical boundaries α_0, α_1 and α_2 have to be determined by equating the expression (2) to α. Note that $\alpha_1 \geq c_{\alpha_2}$ can be assumed: If $p_1 \leq c_{\alpha_2}$ holds after the first stage it follows that H_0 would be rejected in the final analysis due to $p_1 \cdot p_2 \leq c_{\alpha_2}$. For every choice of α, α_2 and α_0, $\alpha_2 \leq \alpha \leq \alpha_0$, there is a unique value $\alpha_1 \in [c_{\alpha_2}, \alpha]$ for which (2) is equal to α. Additionally, for fixed α the level α_1 to be applied in the interim analysis increases for decreasing α_0 and for increasing α_2. A further discussion of the characteristics of the critical boundaries will be given in Section 4.2.2. Some special cases are worth noting. Assuming continuous test statistics, for $\alpha_0 = 1$ ($\alpha_1 = 0$) no early stopping with acceptance (rejection) of H_0 is possible. If no early stopping is intended at all ($\alpha_0 = 1$ and $\alpha_1 = 0$), the first part of the trial serves as an ‘internal pilot study’ for the second stage.

In the following, we suppose that the one-sided null hypotheses $H_0^1, \ldots, H_0^K$ with corresponding alternatives $H_1^1, \ldots, H_1^K$, $K \geq 1$, are to be tested within this adaptive setting. The multiple test procedures given in the following sections control the experimentwise error α in the strong sense, i.e., the probability of the rejection of at least one of the true null hypotheses is controlled by α, irrespective of which and how many of the null hypotheses $H_0^1, \ldots, H_0^K$ are in fact true (HOCHBERG and TAMHANE, 1987).

To keep the presentation simple, we describe the methods for the situation of one interim analysis. However, the procedures can be extended to designs with more than two stages if the corresponding critical levels and decision rules are taken into account.

3. A Closed Test Procedure

The closed testing principle is a general method for constructing multiple test procedures controlling the experimentwise error rate α. In the multiple endpoint situation, the family of null hypotheses $\mathcal{H} = \bigcup_{J \subseteq \{1,\ldots,K\}} \mathrm{H}_0^J$ is considered, where $\mathrm{H}_0^J = \bigcap_{i \in J} \mathrm{H}_0^i$, $J \subseteq \{1, \ldots, K\}$. $\mathcal{H}$ is closed under intersection, i.e., the intersection of any two elements of $\mathcal{H}$ is also contained in $\mathcal{H}$. Furthermore, application of the closed testing principle requires that there is a level-α test for every null hypothesis $\mathrm{H}_0^J \in \mathcal{H}$. A number of test procedures exist for the intersections of individual endpoint hypotheses (see, e.g., WASSMER et al., 1999, for a review). The performance of some of these methods was examined by LEHMACHER et al. (1991), KIESER, REITMEIR, and WASSMER (1995), and REITMEIR and WASSMER (1996). The following closed test procedure controls the experimentwise level α in the strong sense (MARCUS, PERITZ, and GABRIEL, 1976). A null hypothesis $\mathrm{H}_0^J \in \mathcal{H}$ is rejected, if it is rejected at level α and all null hypotheses of $\mathcal{H}$ implying H_0^J, i.e., all H_0^I, $I \supseteq J$, are rejected at level α, too. One difficult problem is to prove under which conditions a closed test procedure for a set of two-sided null hypotheses controls directional (Type III) errors, too (MARCUS et al., 1976; SHAFFER, 1980). Formulating the multiple endpoint problem in the natural one-sided way helps to avoid this complication.

Taking into account the rules of the two-stage level-α test for early rejection or acceptance of the null hypotheses in the interim analysis and for the combination test in the final analysis, applying the closed testing principle in trials with one adaptive interim analysis leads to the following decisions for the null hypotheses $\mathrm{H}_0^J \in \mathcal{H}$. Let p_{Ij} denote the p-value of a level-α test associated to hypothesis H_0^I, $I \subseteq \{1, \ldots, K\}$ and analysis j, $j = 1, 2$ ($j = 1$ denotes the interim analysis and $j = 2$ the final analysis).

Case I: $p_{I1} \leq \alpha_1$ or ($\alpha_1 < p_{I1} < \alpha_0$ and $p_{I1} \cdot p_{I2} \leq c_{\alpha_2}$) for all $I \supseteq J$
- H_0^J can be rejected.

Case II: Otherwise
- H_0^J is accepted.

This strategy controls the experimentwise level α in the strong sense: Case I means that all null hypotheses implying H_0^J are rejected either in the interim analysis (first condition) or after the second stage with Fisher's combination test (second condition). All null hypotheses are, therefore, tested at level α because the critical levels of the adaptive design are applied. Furthermore, a rejection of null hypotheses occurs along the rules of the closed test procedure. This completes the proof.

Figure 2 illustrates the procedure for the closed set of null hypotheses $\mathcal{H} = \{H_0^1, H_0^2, H_0^{\{1,2\}}\}$ with $H_0^{\{1,2\}} = H_0^1 \cap H_0^2$. For example, the null hypothesis H_0^1 can be rejected after the first stage if $H_0^{\{1,2\}}$ and H_0^1 can be rejected in the interim analysis, i.e., the corresponding p-values fall below the critical boundary α_1. If the intersection hypothesis $H_0^{\{1,2\}}$ can be rejected after the first stage but for the p-value p_{11} the inequality $\alpha_1 < p_{11} < \alpha_0$ holds, H_0^1 can be rejected after the second stage if additionally $p_{11} \cdot p_{12} \leq c_{\alpha_2}$. Note that there are a number of tests available to derive the necessary p-values $p_{\{1,2\}j}$, $j = 1, 2$, for $H_0^{\{1,2\}}$. In the next section, two special cases are considered more closely: *a priori* ordered endpoints, where $p_{\{1,2\}j} = p_{1j}$, and a Bonferroni-type procedure, where $p_{\{1,2\}j} = 2 \min (p_{1j}, p_{2j})$.

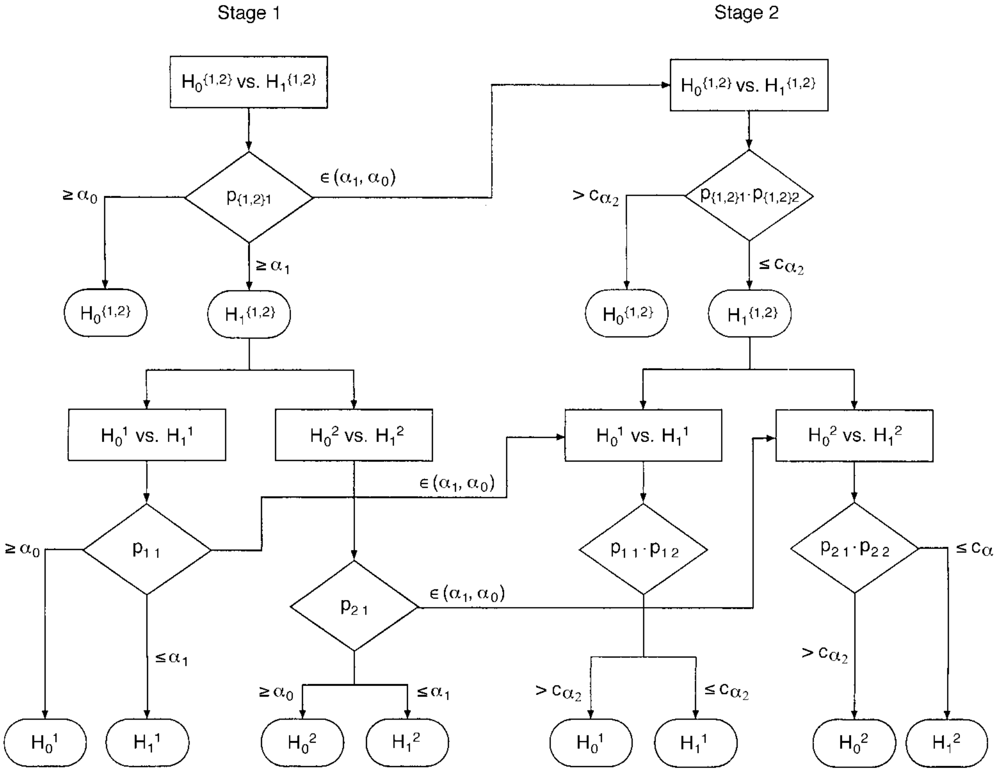

Fig. 2. Closed test procedure for the adaptive design with one interim analysis and two null hypotheses H_0^1, H_0^2 and intersection $H_0^{\{1,2\}} = H_0^1 \cap H_0^2$

4. Important Special Cases

4.1 *A priori ordered hypotheses*

In many clinical trial settings, the hypotheses to be tested have different priorities with regard to the objectives of the study. This enables a hierarchical *a priori* ordering of the null hypotheses in the planning phase of the experiment. Let the sequence of priority of the hypotheses be given by the ordering $H_0^1, \ldots, H_0^K$. In this situation, the rejection of H_0^i is only of concern if all $H_0^1, \ldots, H_0^{i-1}$ can also be rejected. In terms of comparing multiple endpoints between two treatments there may be a "natural" hierarchy among the outcome measures. Superiority in a lower level may be irrelevant if no benefit in a higher level can be established. Also the higher level endpoint may have a higher chance for showing a difference, and the sample size chosen is focussed on the test for this higher level endpoint (see Section 5). The following procedure described by MAURER, HOTHORN, and LEHMACHER (1995) for fixed sample size designs tests these hypotheses at a local error level α while also controlling an experimentwise error level α. In step $i, i = 1, \ldots, K$, of the multiple test procedure, a level-α test for the test problem H_0^i vs. H_1^i is performed; let p_i denote the corresponding p-value. If $p_i \leq \alpha$ reject H_0^i; if $i < K$ continue with step $i + 1$, if $i = K$, stop the procedure. If $p_i > \alpha$ then stop. If the procedure stops at step i, $1 \leq i \leq K$, without rejecting the null hypothesis H_0^i, all null hypotheses $H_0^i, \ldots, H_0^K$ are accepted, while all $H_0^1, \ldots, H_0^{i-1}$ are rejected at experimentwise level α ($H_0^0 = \emptyset$). This procedure is a special application of the closed testing principle (MAURER et al., 1995).

The concepts of testing *a priori* ordered hypotheses and performing adaptive interim analyses can be combined into a strategy which controls the experimentwise error rate α in the strong sense. Let the set of null hypotheses $H_0^1, \ldots, H_0^K$ be *a priori* ordered in this sequence and let denote p_{ij}, $i = 1, \ldots, K$, the p-values associated to hypothesis H_0^i and analysis $j, j = 1, 2$.

Interim analysis:

In the interim analysis, the local Type I error level to be applied is set to α_1. In step i of the test procedure, $i = 1, \ldots, K$, the null hypotheses $H_0^1, \ldots, H_0^{i-1}$ have already been rejected, and the following decisions can be made:

Case I: $p_{i1} \geq \alpha_0$

- All null hypotheses $H_0^i, \ldots, H_0^K$ are accepted.
- Stop the test procedure and end of the study.

Case II: $p_{i1} \leq \alpha_1$

- H_0^i is rejected.
- If $i < K$: Continuation of the test procedure with step $i + 1$.
  If $i = K$: Stop the test procedure and end of the study.

Case III: $\alpha_1 < p_{i1} < \alpha_0$

- No definitive decision about H_0^i in the interim analysis.
- Stop the test procedure in the interim analysis. Planning of the second part of the study.
- In the final analysis the test procedure continues with step i, where H_0^i is tested with Fisher's combination test.

Final analysis:

In order to reject a null hypothesis H_0^i in the final analysis, the p-value of the interim analysis has to fulfill the condition $p_{i1} < \alpha_0$. If the index s is defined as $s = \max\limits_{i=1,\dots,K} \{\text{i: } p_{j1} < \alpha_0 \text{ for all } j \leq i\}$, only null hypotheses H_0^i, $i \leq s$, can be rejected in the final analysis. (If the maximum does not exist $s := 0$.)

In step i of the final analysis, the critical limit c_{α_2} is applied to the product $p_{i1} \cdot p_{i2}$:

Case I: $p_{i1} > \alpha_1$ and $p_{i1} \cdot p_{i2} > c_{\alpha_2}$

- All null hypotheses $H_0^i, \dots, H_0^K$ are accepted.
- Stop the test procedure.

Case II: $p_{i1} \leq \alpha_1$ or $p_{i1} \cdot p_{i2} \leq c_{\alpha_2}$

- H_0^i is rejected.
- If $i < s$: Continuation of the test procedure with step $i + 1$.
  If $i = s$: Stop the test procedure.

In each step the local level is controlled by α, because the critical levels of the adaptive procedure are applied. Furthermore, the decisions are made along the rules of the multiple test procedure for *a priori* ordered hypotheses. Therefore, the test procedure controls the experimentwise error rate α.

Note that for $\alpha_0 = 1$ (no early acceptance) the whole procedure becomes much simpler. In the interim analysis only Case II and Case III (applying the condition $\alpha_1 < p_{i1}$) have to be considered. In the final analysis the product test is applied to all null hypotheses not already rejected at the first stage ($s \equiv K$).

4.2 *The Bonferroni-Holm procedure*

4.2.1 Description of the test procedure

The Bonferroni-Holm method (Holm, 1979) is a simple test procedure which controls the experimentwise error rate α for the general situation that there is no hierarchy between the hypotheses. In trials without interim analysis let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(K)}$ denote the ordered p-values and $H_0^{(1)}, H_0^{(2)}, \dots, H_0^{(K)}$ the corresponding null hypotheses. Starting with the lowest p-value $p_{(1)}$ and proceeding stepwise, $p_{(i)}$ is compared with the critical level $\alpha/(K - i + 1)$; rejection occurs as long as

$p_{(i)} \leq \alpha/(K - i + 1)$, $i = 1, \ldots, K$. The Bonferroni-Holm procedure can be derived from the closed testing principle by using Bonferroni global tests with local level α for all hypotheses $H_0^J \in \mathcal{H}$: an intersection H_0^J is rejected if at least one of the null hypotheses H_0^j, $j \in J$, can be rejected at level $\alpha/|J|$, where $|J|$ denotes the number of elements in the set J.

The results of Section 3 can therefore be used to extend the Bonferroni-Holm procedure to trials with adaptive interim analysis. For this purpose, appropriate critical levels for the Bonferroni global test for interim and final analysis have to be defined. Let for a two-stage level-α test with a fixed value of α_0 denote $\alpha_1(\alpha)$ and $\alpha_2(\alpha)$ the critical levels to be spent at the interim and the final analysis, respectively, and $c_{\alpha_2(\alpha)}$ the corresponding critical value for Fisher's combination test. The following Bonferroni global test controls the level α for adaptive designs with one interim analysis: reject an intersection $H_0^J \in \mathcal{H}$, if at least one of the null hypotheses H_0^j, $j \in J$, can be rejected in the interim analysis at level $\alpha_1(\alpha/|J|)$, or if for at least one of the null hypotheses H_0^j, $j \in J$, the conditions $\alpha_1(\alpha/|J|) < p_{j1} < \alpha_0$ and $p_{j1} \cdot p_{j2} \leq c_{\alpha_2(\alpha/J)}$ hold. The corresponding Bonferroni-Holm procedure now follows directly from the general closed test procedure described in Section 3 by applying the two-stage Bonferroni global tests to all null hypotheses in the closed system $\mathcal{H}$.

Interim analysis:

Let $p_{(1)1} \leq p_{(2)1} \leq \ldots \leq p_{(K)1}$ denote the ordered p-values after the first stage for the corresponding null hypotheses $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(K)}$.

Case I: $p_{(1)1} \geq \alpha_0$

- All null hypotheses $H_0^{(i)}$, $i = 1, \ldots, K$, are accepted.
- End of the study.

Case II: $p_{(1)1} < \alpha_0$

- $H_0^{(i)}$ is rejected for $i = 1, \ldots, r$, where $r = \max\limits_{i=1,\ldots,K}\{\text{i: } p_{(j)1} \leq (\alpha_1(\alpha/(K - j + 1))$ for all $j = 1, \ldots, i\}$. (If the maximum does not exist $r := 0$ and $H_0^{(0)} = \emptyset$.)
- $H_0^{(i)}$ is accepted for those i with $p_{(i)1} \geq \alpha_0$.
- No definitive decision about $H_0^{(i)}$ in the interim analysis for $i = r + 1, \ldots, s$, where $s = \max\limits_{i=1,\ldots,K}\{i\text{: } p_{(i)1} < \alpha_0\}$. (If the maximum does not exist $s := 0$ and $H_0^{(0)} = \emptyset$.) In the final analysis the null hypotheses $H_0^{(i)}$, $i = r + 1, \ldots, s$, can be tested again.

Final analysis:

For simplicity of notation we assume that $H_0^1, \ldots, H_0^r$ have already been rejected, and that $H_0^{r+1}, \ldots, H_0^s$, $r + 1 \leq s \leq K$, have neither been accepted nor rejected in the interim analysis and can, therefore, be tested after the second stage. Let $(p_1 \cdot p_2)_{(i)}$ denote the ordered products of p-values for the null hypotheses H_0^i, $i = r + 1, \ldots, s$, and assume the null hypotheses to be ordered such that $(p_1 \cdot p_2)_{(r+1)}$ corresponds to H_0^{r+1}. Note that this ordering based on the product of the p-values will in general differ from the ordering $H_0^{(1)}, \ldots, H_0^{(K)}$ based on the p-values from the first stage.

Case I: $(p_1 \cdot p_2)_{(r+1)} > c_{\alpha_2(\alpha/(K-r))}$
- All null hypotheses H_0^i, $i = r + 1, \ldots, s$, are accepted.

Case II: $(p_1 \cdot p_2)_{(r+1)} \leq c_{\alpha_2(\alpha/(K-r))}$
- H_0^{r+1} is rejected.
- If $r + 1 < s$: Continuation of the test procedure with step $r + 2$.
- If $r + 1 = s$: Stop the test procedure.

Step $r + 2$ of the procedure:
$p_{(r+2)1} > \alpha_1(\alpha/(K - r - 1))$ and $(p_1 \cdot p_2)_{(r+2)} > c_{\alpha_2(\alpha/(K-r-1))}$
- All null hypotheses $H_0^i i$, $i = r + 2, \ldots, s$, are accepted. Stop the test procedure.

$p_{(r+2)1} \leq \alpha_1(\alpha/(K - r - 1))$ or $(p_1 \cdot p_2)_{(r+2)} \leq c_{\alpha_2(\alpha/(K-r-1))}$
- H_0^i is rejected, where i is the index with $p_{i1} = p_{(r+2)1}$ or $(p_{i1} \cdot p_{i2}) = (p_1 \cdot p_2)_{(r+2)}$. (If i is not defined uniquely, this does not matter, because any H_0^i fulfilling at least one of the conditions will be rejected in this or one of the following steps.)
- If $r + 2 < s$: Continuation of the test procedure with step $r + 3$.
- If $r + 2 = s$: Stop the test procedure.

Again, the simplification for the situation of no early acceptance ($\alpha_0 = 1$) is straightforward ($s \equiv K$).

It is worth noting that a null hypothesis that could not be rejected in the interim analysis may be rejected in the final analysis solely on the basis of the p-value of the first stage. This is due to the fact that the critical boundaries for both interim and final analysis increase when rejections occur. For example, suppose that $K = 3$ endpoints are investigated and that for the p-values of the first stage $\alpha_1(\alpha/3) < p_{11} = p_{(1)1} \leq \alpha_1(\alpha/2)$, $p_{21} < \alpha_0$, and $p_{31} < \alpha_0$ holds. None of the null hypotheses is therefore rejected or accepted in the interim analysis. If in the final analysis $p_{21} \cdot p_{22} \leq c_{\alpha_2(\alpha/3)}$, then H_0^2 can be rejected, and H_0^1 can be rejected too, regardless of the result in the second stage (endpoint 1 may not be investigated after the interim analysis at all). H_0^3 can be rejected if $p_{31} \leq \alpha_1(\alpha)$ or $p_{31} \cdot p_{32} \leq c_{\alpha_2(\alpha)}$.

4.2.2 Considerations on the choice of the critical limits

Clearly the critical limits $\alpha_1(\alpha)$ and $\alpha_2(\alpha)$ for the test decision in the Bonferroni-Holm procedure should be chosen to be increasing with the corresponding significance level α. This is guaranteed if the following monotonicity conditions are fulfilled:

$$\alpha_1(\alpha') \leq \alpha_1(\alpha'') \quad \text{for} \quad \alpha' \leq \alpha'' , \tag{3a}$$

$$\alpha_2(\alpha') \leq \alpha_2(\alpha'') \quad \text{for} \quad \alpha' \leq \alpha'' . \tag{3b}$$

For $\alpha_2 = \alpha$ as proposed by BAUER and KÖHNE (1994) condition (3b) is obviously fulfilled, and it can be shown that (3a) holds too (see Appendix). For $\alpha_2 < \alpha$ the situation is more complex because there is an infinite number of possible choices for α_1 and α_2. Additionally, the spending of the Type I error rate can be varied for the different levels $\alpha/|J|$ to be applied. One sensible choice is to hold the ratio between the levels to be spent at the first and the second stage constant for all significance levels $\alpha/|J|$, $|J| = 1, \dots, K$:

$$\frac{\alpha_1(\alpha/|J|)}{c_{\alpha_2(\alpha/J)} \cdot (\ln \alpha_0 - \ln \alpha_1(\alpha/|J|))} = \kappa \quad \text{for} \quad |J| = 1, \dots, K\,, \quad \kappa > 0\,. \qquad (4)$$

It can be shown (see Appendix) that for every $\kappa > 0$ the monotonicity conditions (3a) and (3b) are fulfilled for this choice of α_1 and α_2.

Table 1 gives values for $\alpha_1(\alpha/K)$ and $c_{\alpha/K}$ for $\alpha = 0.05$ and different choices of $K = 1, \dots, 5$ and $\alpha_0 = 0.3$, 0.4, 0.5, 0.6, 0.7 and 1.0; in the final analysis Fisher's combination test is applied at the significance level α/K. In Table 2 the critical levels $\alpha_1(\alpha/K)$ and $\alpha_2(\alpha/K)$ to be applied in the interim and final analysis, respectively, and the critical limits $c_{\alpha_2(\alpha/K)}$ for the combination test are given for $\alpha = 0.05$ and the same values for K and α_0 as in Table 1; here it is assumed that the significance level α/K is equally spent for the interim and final analysis (i.e., $\kappa = 1$). As can be seen from the tables the critical levels for interim and final analysis decrease considerably for increasing K. For $\alpha = 0.05$, $\alpha_0 = 0.5$, $K = 5$ and α_2 $(\alpha/K) = \alpha/K$, and application of the Bonferroni-Holm procedure, for example, at least one p-value below 0.0035 is necessary to reject any of the null hypotheses already in the interim analysis; if the significance level is equally spent between interim and final analysis the respective critical level for the lowest p-value $p_{(1)1}$ is 0.005. Note that for $\kappa = 1$ in case of $\alpha_0 = 0.3$, $K = 1$, 2, and

Table 1

The local significance level $\alpha_1(\alpha/K)$ to be applied in the interim analysis and critical limits $c_{\alpha/K}$ for the final Fisher combination test, respectively, for the Type I error probabilities α/K. $\alpha = 0.05$, $K = 1, \dots, 5$, and α_0 denotes the critical limit for early acceptance. This corresponds to the situation that the local significance level $\alpha_2(\alpha/K) = \alpha/K$ is applied in the final analysis

		$\alpha_1(\alpha/K)$				
	K	1	2	3	4	5
	α/K	0.05	0.025	0.0167	0.0125	0.01
α_0	$c_{\alpha/K}$	0.00870	0.00380	0.00237	0.00169	0.00131
0.3		0.0299	0.0131	0.0081	0.0058	0.0045
0.4		0.0263	0.0115	0.0071	0.0051	0.0040
0.5		0.0233	0.0102	0.0063	0.0045	0.0035
0.6		0.0207	0.0090	0.0056	0.0040	0.0031
0.7		0.0183	0.0080	0.0050	0.0036	0.0027
1.0		0.00870	0.00380	0.00273	0.00169	0.00131

Table 2

The local significance levels $\alpha_1(\alpha/K) = \alpha/K$ and $\alpha_2(\alpha/K)$ to be applied in the interim analysis and critical limit $c_{\alpha_2(\alpha/K)}$ for the final Fisher combination test, respectively, for the Type I error probabilities α/K. $\alpha = 0.05$, $K = 1, \ldots, 5$, and α_0 denotes the critical limit for early acceptance. This corresponds to the situation of equal spending of the Type I error probability α/K between the two stages of the adaptive procedure

		$\alpha_2(\alpha/K)$ $c_{\alpha_2(\alpha/K)}$				
	K	1	2	3	4	5
	α/K	0.05	0.025	0.0167	0.0125	0.01
α_0	$\alpha_1(\alpha/K)$	0.025	0.0125	0.00833	0.00625	0.005
0.3		0.0563 0.01006	0.0257 0.00393	0.0164 0.00233	0.0120 0.00161	0.00941 0.00122
0.4		0.0515 0.00902	0.0239 0.00361	0.0154 0.00215	0.0113 0.00150	0.00887 0.00114
0.5		0.0483 0.00835	0.0227 0.00339	0.0146 0.00204	0.0108 0.00143	0.00850 0.00109
0.6		0.0460 0.00787	0.0217 0.00323	0.0141 0.00195	0.0104 0.00137	0.00821 0.00104
0.7		0.0442 0.00750	0.0210 0.00311	0.0137 0.00188	0.00101 0.00132	0.00799 0.00101
1.0		0.0406 0.00678	0.0196 0.00285	0.0128 0.00174	0.00948 0.00123	0.00752 0.00094

$\alpha_0 = 0.4$, $K = 1$, the local level α_2 to be applied in the final combination test exceeds α. If the condition $\alpha_2 \leq \alpha$ is to be maintained and the level α should be exhausted, less than half of the overall Type I error probability has to be spent in the interim analysis in these cases.

Up to now we have chosen α_0 to be constant at any stage of the stepwise procedure. In principle, we could choose α_0 depending on the number t of null hypotheses rejected so far. Since quite obviously α_1 is decreasing with α_0 under all previous scenarios, an increasing $\alpha_0(t)$ has to be chosen carefully in order to guarantee that α_1 and α_2 are still increasing in t. One reason to choose an increasing $\alpha_0(t)$ would be to facilitate early stopping with the acceptance of the global null hypothesis $\bigcap_{i=1}^{K} H_0^i$ as compared to early acceptance of individual null hypotheses otherwise.

5. Sample Size Re-Assessment

It is worth noting that the procedures described above allow the application of some simple rules for the choice of the sample size at the second stage. As an example, the situation of comparing two treatments with respect to $K = 3$ endpoints is considered. Suppose that there is an *a priori* ordering among the three endpoints. No rejection has occurred in the interim analysis, but the results were in the intended direction and $p_{i1} < \alpha_0$, $i = 1, 2, 3$. In order to get a rejection of all three endpoints' null hypotheses, the p-values for the second stage will have to meet the conditions $p_{12} \leq c_{\alpha_2}/p_{11}$ and $p_{22} \leq c_{\alpha_2}/p_{21}$ and $p_{32} \leq c_{\alpha_2}/p_{31}$. The sample size for the second stage would then have to be chosen such that under a particular alternative $H_1 = H_1^1 \cap H_1^2 \cap H_1^3$ all three p-values for the three endpoint comparisons fall below their respective critical limits c_{α_2}/p_{i1}. Let $\beta_{i2}(n_2)$ be the corresponding probability of a Type II error for the test of endpoint i at stage 2 when a sample size n_2 is applied. Assuming positive correlations between the p-values, the probability $P_{123} = \prod_{i=1}^{3} (1 - \beta_{i2}(n_2))$ under independence could serve as a conservative estimate for the probability of the joint final rejection of all three endpoints. Common power calculations for fixed sample size tests applying the somewhat unusual (data-driven) significance levels c_{α_2}/p_{i1} can be used to find a sample size n_2 which, for a given alternative, assures that P_{123} is equal to a chosen value.

A simpler approach would concentrate on the null hypothesis H_0^1 highest up in the hierarchy. Then the sample size for the second stage based on a given (conditional) power $1 - \beta_{i2}$ can be calculated from the fixed sample size test of the first endpoint with an adjusted significance level c_{α_2}/p_{11} in a straightforward way. If no order relation holds among the endpoints, n_2 could be chosen so that $\max_i (1 - \beta_{i2}(n_2))$ is equal to a given number. However, the adjusted significance levels $c_{\alpha_2}/(3p_{i1})$, $i = 1, 2, 3$, have to be applied now. This choice guarantees that the probability of getting at least one correct rejection (if the corresponding alternative applies) can be controlled.

6. Example

A randomized, double-blind, placebo-controlled clinical trial was conducted to investigate efficacy and safety of the kava-kava special extract WS 1490 in outpatients suffering from anxiety disorders (Malsch and Kieser, 1999). An adaptive interim analysis was planned after 40 completed patients, the experimentwise error rate, the level to be applied in the final analysis, and the boundary for early stopping were fixed to $\alpha = 0.05$, $\alpha_2 = \alpha$, and $\alpha_0 = 0.6$, respectively, resulting in a local level $\alpha_1 = 0.0207$ for the interim analysis. Endpoints for confirmatory analysis were the Hamilton Anxiety Scale (HAMA), the Adjective Mood Scale (Bf-S), and the rate of patients with withdrawal symptoms; the corresponding hypotheses were *a priori* ordered in this sequence. The p-values in the interim analysis were $p_{11} = 0.0103$, $p_{21} = 0.0032$ (both U-test, one-sided) and $p_{31} = 0.215$ (χ^2-test, one-sided). Hence,

H_0^1 and H_0^2 could be rejected in the interim analysis. The study was stopped with this result and acceptance of H_0^3. If rejection of H_0^3 had also been aspired, the study would have to be continued, and the respective *p*-value in the final analysis would have to fulfill $p_{32} \leq c_\alpha / p_{31} = 0.040$. Sample size calculation for the second part of the trial would therefore have to be based on this Type I error rate.

We now assume that the null hypotheses were not *a priori* ordered but should be analyzed with the Bonferroni-Holm procedure. In this case, the critical levels for the interim analysis were $\alpha_1(\alpha/3) = 0.0056$, $\alpha_1(\alpha/2) = 0.0090$, $\alpha_1(\alpha) = 0.0207$, respectively. Null hypothesis H_0^2 could have been rejected in the interim analysis because of $p_{(1)1} = p_{21} = 0.0032 < \alpha_1(\alpha/3)$, but no early decision could be made about H_0^1 and H_0^3 due to $\alpha_1(\alpha/2) < p_{(2)1} = p_{11} < \alpha_0$ and $p_{(3)1} = p_{31} < \alpha_0$. In order to reject both H_0^1 and H_0^3 at the second stage, one of the corresponding products of *p*-values has to fall below $c_{\alpha_2(\alpha/2)} = 0.0038$ and the other below $c_{\alpha_2(\alpha)} = 0.0087$. Sample size re-assessment has to be done according to Section 5. Suppose that after the second stage the product of the *p*-values for H_0^3 fulfills $p_{31} \cdot p_{32} \leq c_{\alpha_2(\alpha/2)}$ (i.e., $p_{32} \leq 0.017$). Clearly H_0^3 can be rejected, but additionally H_0^1 can be rejected, too, irrespective of the result for this hypothesis in the second stage. The reason lies in the fact that after rejection of H_0^3 the critical boundary for the *p*-value p_{11} is $\alpha_1(\alpha) = 0.0207$.

7. Discussion

By the application of multiple test procedures in studies with planned interim analyses several objectives can be investigated within one clinical trial. At the same time a high degree of flexibility is achieved with regard to early stopping or sample size adjustment based on the results of the interim analysis. It is known that, e.g., for normal means, only a small loss of power is connected with the application of the combination test to an artificially partitioned sample in a classical nonadaptive experiment as compared to the optimal test in the total sample (BAUER and KÖHNE, 1994; BANIK, KÖHNE, and BAUER, 1996). However, the possible gain in power arising from the adaptive nature of the design may be substantial (BAUER and RÖHMEL, 1995; BAUER and KIESER, 1999).

The adaptation can go far beyond a re-assessment of sample size allowing, for example, the reduction of the set of null hypotheses to be tested after the interim analysis (BAUER and KIESER, 1999). In the context of multiple endpoints, selection of a subset of the initially investigated hypotheses for the second stage of the trial may occur if for one or more of the target parameters there is no distinct treatment effect. It could be an attractive option to drop the corresponding null hypothesis after the interim stage with early acceptance if, for example, the values of an endpoint are determined by an invasive or costly examination.

Early rejection of the global null hypothesis may lead to a substantial reduction of the sample size. If, however, only some but not all endpoints show significant results after the first stage there will generally be no saving in sample size. Moreover, the following complication may arise in such a situation. If also those endpoints are measured at the second stage of the study that have already been rejected in the interim analysis, one would have to point at a possible interaction between treatment and trial stage if conflicting results evolve in the second part of the study.

Planning studies in which multiple test procedures are to be applied is usually much harder than for the case of only one question of interest. For example, uncertainty about distribution, effects and variability to be expected for several parameters and the complexity of the decision process makes a precise sample size determination nearly impossible; for the same reason an optimal choice of the test statistics to be used in the analysis is difficult. From this point of view, planning studies with multiple hypotheses in an adaptive design seems to be quite appealing, because the knowledge gained from the results of the interim analysis can be used for the subsequent part of the study. This may go as far as choosing the tests to be used at the second stage depending on the results of the interim analysis, e.g., by estimating suitable scores from the first stage. Under the null hypothesis the p-value of the respective test performed in a stochastically independent sample will be independently and uniformly distributed (or stochastically not smaller than the uniform distribution). Therefore, such a data-driven procedure does not affect the Type I error.

These impressive gains in flexibility are not quite free of charge. Typical problems such as a higher expenditure of logistics or the effect of over-running occurring when recruitment continues while an interim analysis is performed not only hold for group sequential trials but also for studies with adaptive interim analyses. Furthermore, if multiple test procedures are incorporated in such trials the trade-off for the gain in information lies in a higher complexity of the study design and decision alternatives. Considering the numerous options available within this approach, it is of particular importance to establish that the decision rules to be employed are specified in advance.

Appendix

Proof of the monotonicity conditions (3a) and (3b) for $\alpha_2 = \alpha$

The critical limit $\alpha_1(\alpha)$ can be obtained by solving the equation

$$F(\alpha, \alpha_0, \alpha_1) = \alpha_1 + c_\alpha \cdot (\ln \alpha_0 - \ln \alpha_1) - \alpha = 0\,. \tag{5}$$

From the implicit function theorem it follows

$$\begin{aligned}\left(\frac{\partial \alpha_1(\alpha)}{\partial \alpha}\right) &= -\left(\frac{\partial F}{\partial \alpha_1}\right)^{-1} \cdot \left(\frac{\partial F}{\partial \alpha}\right)\\ &= \frac{\alpha_1(\alpha)}{c_\alpha - \alpha_1(\alpha)} \cdot \left(\frac{\partial c_\alpha}{\partial \alpha} \cdot (\ln \alpha_0 - \ln \alpha_1(\alpha)) - 1\right).\end{aligned} \tag{6}$$

Using the definition of c_α and the relation between the quantiles and the density function of the χ^2-distribution with 4 degrees of freedom it can be shown that

$$\frac{\partial c_\alpha}{\partial \alpha} = \frac{c_\alpha}{\alpha - c_\alpha} \, . \tag{7}$$

Furthermore, from (2) it follows that

$$c_\alpha \cdot (\ln \alpha_0 - \ln \alpha_1(\alpha)) = \alpha - \alpha_1(\alpha) \, . \tag{8}$$

Imputing (7) and (8) in equation (6) leads to

$$\left(\frac{\partial \alpha_1(\alpha)}{\partial \alpha}\right) = \frac{\alpha_1(\alpha)}{\alpha_1(\alpha) - c_\alpha} \cdot \left(1 - \frac{\alpha - \alpha_1(\alpha)}{\alpha - c_\alpha}\right) > 0 \, . \tag{9}$$

This proves condition (3a). Condition (3b) follows directly from $\alpha_2(\alpha) = \alpha$.

Proof of the monotonicity conditions (3a) and (3b) for constant ratio between the Type I error rates in the two stages of the adaptive design

Constant ratio for spending of the Type I error rate in both stages for all values of α, $0 \leq \alpha \leq 1$, means that the critical limits $\alpha_1(\alpha)$ and $\alpha_2(\alpha)$ are chosen such that

$$\frac{\alpha_1(\alpha)}{c_{\alpha_2(\alpha)} \cdot (\ln \alpha_0 - \ln \alpha_1(\alpha))} = \kappa \, , \qquad \kappa > 0 \, . \tag{10}$$

It follows that

$$\alpha_1(\alpha) = \frac{\kappa}{(1+\kappa)} \cdot \alpha \tag{11}$$

$$c_{\alpha_2(\alpha)} = \frac{\dfrac{\alpha}{(1+\kappa)}}{\ln \dfrac{\alpha_0}{\kappa} - \ln \dfrac{\alpha}{(1+\kappa)}} \, . \tag{12}$$

Therefore, both $\alpha_1(\alpha)$ and $c_{\alpha_2(\alpha)}$ are increasing with α. Taking into account that $\alpha_2(\alpha)$ is strictly increasing with $c_{\alpha_2(\alpha)}$ completes the proof.

References

Banik, N., Köhne, K., and Bauer, P., 1996: On the power of Fisher's combination test for two stage sampling in the presence of nuisance parameters. *Biometrical Journal* **38,** 25–37.

Bauer, P., 1991: Multiple testing in clinical trials. *Statistics in Medicine* **10,** 871–890.

Bauer, M., Bauer, P., and Budde, M., 1998: A simulation program for adaptive two stage designs. *Computational Statistics & Data Analysis* **26,** 351–371.

Bauer, P. and Kieser, M., 1999: Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* **18,** to appear.

Bauer, P. and Köhne, K., 1994: Evaluation of experiments with adaptive interim analyses. *Biometrics* **50,** 1029–1041. Correction in *Biometrics* **52,** 380.

BAUER, P. and RÖHMEL, J., 1995: An adaptive method for establishing a dose-response relationship. *Statistics in Medicine* **14,** 1595–1607.

HOCHBERG, Y. and TAMHANE, A., 1987: *Multiple Comparison Procedures*. Wiley, New York.

HOLM, S., 1979: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6,** 65–70.

HOMMEL, G., 1997: Tests of individual hypotheses for experiments with interim analyses and adaptive choice of hypotheses. Paper given at the Biometric Colloquium of the German Region of the International Biometric Society, Munich.

KIESER, M., REITMEIR, P., and WASSMER, G., 1995: Test procedures for clinical trials with multiple endpoints. In: J. Vollmar (ed.): *Biometrie in der chemisch-pharmazeutischen Industrie*. Fischer, Stuttgart, Volume **6**, 41–60.

LAN, K. K. G. and DEMETS, D. L., 1983: Discrete sequential boundaries for clinical trials. *Biometrika* **70,** 659–663.

LÄUTER, J., 1996: Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* **52,** 964–970.

LEE, J. B., 1994: Group sequential testing in clinical trials with multivariate observations: a review. *Statistics in Medicine* **13,** 101–111.

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.

MALSCH, U. and KIESER, M., 1999: Efficacy of the special extract of kava rhizomes WS 1490 in patients with anxiety disorders of non-psychotic origin. Submitted.

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 preclinical assays: a-priori ordered hypotheses. In: *Biometrie in der chemisch-pharmazeutischen Industrie.* Fischer, Stuttgart, Volume **6**, 3–18.

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., 1977: Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191–199.

REITMEIR, P. and WASSMER, G., 1996: One-sided multiple endpoint testing in two-sample comparisons. *Communications in Statistics: Simulation and Computation* **25***,* 99–117.

SHAFFER, J., 1980: Control of directional errors with stagewise multiple test procedures. *Annals of Statistics* **8,** 1342–1348.

WASSMER, G., REITMEIR, P., KIESER, M., and LEHMACHER, W., 1999: Procedures for testing multiple endpoints in clinical trials: an overview. *Journal of Statistical Planning and Inference* **47,** to appear.

Dr. MEINHARD KIESER
Dr. Willmar Schwabe GmbH & Co.
Fachbereich Biometrie
Willmar-Schwabe-Strasse 4
D-76227 Karlsruhe
Germany

Received, May 1998
Revised, February 1999
Accepted, February 1999