Recent Developments in Multiple Comparison Procedures
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 47 (2004) 33–47

Multiple testing in adaptive designs – A review

Martin Hellmich[1] **and Gerhard Hommel**[2]

University of Cologne and University of Mainz

Abstract: During the course of a study it would be desirable to take advantage of new internal or external information in order to modify design features laid down in the study protocol. However, for conventional controlled experiments this does not seem feasible without casting doubt on the statistical validity of the whole experiment. In contrast, modifications pertaining to treatment arms, endpoints, hypotheses, statistical methods etc. are possible within adaptive designs, thus enabling the conduct of complex controlled experiments which may still be tailored after onset to meet ethical and scientific as well as economic requirements.

This article briefly reviews recent statistical methods for adaptive study design, particularly those built from p-value combination rules or, equivalently, conditional error functions. Its main focus, however, is on the application of adaptive testing methods to clinical experiments with multiple objectives, e.g., multiple treatment arms or endpoints. The authors demonstrate in this overview that as a consequence of using an adaptive interim analysis, null hypotheses may be dropped or added, and test statistics may be exchanged, whilst the studywise type I error rate remains under (strong) control. Moreover, adaptive designs may be applied to experiments aiming to establish dose-response relationships, or to demonstrate non-inferiority, superiority or equivalence of multiple treatment arms. An example from the literature, which has not previously been discussed from an adaptive viewpoint, is provided as a worked illustration.

1. Introduction

Multiplicity due to multiple hypotheses, endpoints, treatment arms or subgroups is present in virtually all clinical trials. In order to guard against the opportunity of selecting the most favorable result from a pool of analyses, adequate control and pre-specification of statistical procedures is required (CPMP/EWP, 2003). A conventional approach to limit the type I error α of the statistical procedure (also called *consumer's risk* in case of an efficacy endpoint) whilst granting reasonable power ($1-\beta$, where β denotes the type II error of the statistical procedure or the *producer's risk* in case of an efficacy endpoint) is to control the studywise rate of false positive conclusions, i.e., the probability of one or more erroneous conclusions drawn from the results of the same trial needs to be kept below α. The more questions a clinical trial is set out to answer, the more complex are the statistical procedures employed to safeguard adequate type I error control. In the planning phase of a clinical trial, any lack of information on important design quantities, such

[1]Institute of Medical Statistics, Informatics and Epidemiology, University of Cologne, D-50924 Cologne, Germany. e-mail: martin.hellmich@medizin.uni-koeln.de

[2]Institute of Medical Biometrics, Epidemiology and Informatics, University of Mainz, D-55101 Mainz, Germany.

Keywords and phrases: multiple testing, adaptive designs, multiple hypotheses, multiple endpoints, multiple treatments.

AMS 2000 subject classifications: primary 62-02; secondary 62L99.

as the expected treatment effect or sample variance, may lead to a serious deficiency in sample size, and thus in power (see Westfall et al. (1999) for a description of relevant power concepts pertaining to multiple testing).

Sequential analysis was pioneered by Wald (1947), but first applied to clinical trials by Armitage (1957; 1975), with the aim of reducing the average sample size. Some limitations of the classical sequential approach, i.e., the observed data must be paired and continuously monitored, were overcome by the development of group-sequential methods (Pocock, 1977; O'Brien and Fleming 1979) with interim analyses scheduled according to a prefixed series of equal-sized groups. The more general α-spending approach by Lan and DeMets (1983) neither requires the number, nor the time, of interim analyses to be specified in advance, whilst the sample size still needs to be predetermined. In contrast, re-calculation of the final sample size can be performed using an internal pilot study where the variance is re-estimated in a blinded or unblinded interim analysis (Wittes and Brittain, 1990; Gould and Shih, 1992; Birkett and Day, 1994; Friede and Kieser 2002). A comprehensive survey of group-sequential methods was given by Jennison and Turnbull (2000). In recent years, adaptive designs as initiated by Bauer (1989) have been advocated to be superior to classical group-sequential clinical trials since only the former bear the potential for substantial data-driven re-design such as re-calculation of sample size or modification of treatment arms, endpoints, hypotheses, statistical methods etc.

The aim of this article is to demonstrate the merits of applying adaptive testing methods in multiple testing situations as entailed by multiple endpoints, treatment arms or subgroups. An overview of methods for adaptive study design is given in Section 2. Some basics of multiple hypotheses testing, particularly properties of the closure procedure, are presented in Section 3. The advantages of using adaptive designs in multiple testing situations, e.g., midtrial abandonment or inclusion of hypotheses, are described in detail in Section 4. In Section 5, an example from the literature is analyzed from a new (adaptive) viewpoint. The article concludes with a critical discussion of the possible malpractice of the extensive flexibility introduced by adaptive designs to clinical trials.

2. Adaptive designs

Detailed reviews of adaptive design methodology have recently been presented by Bauer et al. (2001a) and Wassmer et al. (2001); thus only a brief overview will be given below.

A one-sided null hypothesis H_0, say, on the difference θ in mean efficacy of two treatments, i.e., $H_0 : \theta \leq -\delta$ vs. $H_a : \theta > -\delta$ for some $\delta \geq 0$, is tested at level α using a two-stage adaptive group-sequential design as follows: In the planning phase, the investigator needs to fix (i) the design of the first stage, including the test statistic to calculate the (one-sided) p-value p_1 from the sample of the first stage, (ii) the function $C(p_1, p_2)$ to be used if a second stage (yielding a p-value p_2) is performed for combination of the p-values from both stages and (iii) the early decision boundaries α_1, α_0 with $0 \leq \alpha_1 < \alpha < \alpha_0 \leq 1$. These three ingredients are used as following: After completion of the first stage, H_0 is rejected if $p_1 \leq \alpha_1$ and accepted if $p_1 > \alpha_0$. In either case, the trial is stopped. If $\alpha_1 < p_1 \leq \alpha_0$, the second stage can be planned using all information collected so far, both from inside and outside of the trial. After completion of the second stage, H_0 is rejected if $C(p_1, p_2) \leq c$ where c is calculated from α, α_1, α_0 and $C(p_1, p_2)$ to control the level α. Otherwise, H_0 is accepted.

The combination function $C(p_1, p_2)$ is assumed to be increasing in both arguments, strictly increasing in at least one, and left continuous in p_2, for all $p_1 \in]\alpha_1, \alpha_0]$, $p_2 \in [0, 1]$. Moreover, the p-values p_1 and p_2 are required to be p-clud (Brannath et al., 2002), meaning that under H_0 the distributions of p_1 and p_2 conditional on p_1 are larger or equal to the uniform distribution on $[0, 1]$, in formulae

$$P_{H_0}(p_1 \leq \alpha) \leq \alpha \quad \text{and} \quad P_{H_0}(p_2 \leq \alpha | \, p_1) \leq \alpha \quad \text{for all} \quad 0 \leq \alpha \leq 1.$$

This is fulfilled, for example, if independent samples are recruited at the different stages of the trial and the applied hypotheses tests are level α tests for any pre-chosen significance level α. Essentially, any midtrial design adaptation which preserves this distributional property of the p-values will not compromise the type I error control.

Throughout this article, null hypotheses are usually assumed to be one-sided in order to simplify the presentation. This is not an essential restriction, because any two-sided null hypothesis H_0 can be tested at level 2α by combination of both one-sided tests at level α (exact only for $\alpha_0 < 0.5$; Wassmer, 1999) as follows: Let p_1 and p_2 denote, as above, the one-sided p-values from the two stages corresponding to a specific test direction. Then, after completion of the first stage, H_0 is rejected if $p_1 \leq \alpha_1$ or $1 - p_1 \leq \alpha_1$ and accepted if $\alpha_0 \leq p_1 \leq 1 - \alpha_0$. In either case, the trial is stopped. If $\alpha_1 < p_1 < \alpha_0$ or $\alpha_1 < 1 - p_1 < \alpha_0$, the second stage can be planned using all information collected so far, both from inside and outside of the trial. After completion of the second stage, H_0 is rejected if $C(p_1, p_2) \leq c$ or $C(1 - p_1, 1 - p_2) \leq c$ where c is calculated from α, α_1, α_0 and $C(p_1, p_2)$ to control the level α. Otherwise, H_0 is accepted.

Adaptive two-stage designs were described above in terms of a combination function for p-values and corresponding decision boundaries. Such designs can equivalently be formulated by means of a conditional error function $\alpha(p_1) : [0, 1] \to [0, 1]$ which is nondecreasing in p_1 and fulfills

$$\int_0^1 \alpha(p_1) \, dp_1 \leq \alpha$$

(Proschan and Hunsberger, 1995; Müller and Schäfer, 2001). This function $\alpha(p_1)$ determines the conditional type I error to be controlled by any design and test procedure for H_0 chosen for a contingent second stage. The one-to-one mapping between decision procedures based on p-value combination and conditional error functions was worked out in general by Brannath et al. (2002). Important examples were given by Posch and Bauer (1999) and Wassmer (1999).

If the stage-wise order of the sample space is assumed (Armitage, 1957; Tsiatis et al. 1984), a (global) p-value function for the combination test can be defined as

$$q(p_1, p_2) = \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_1 + \displaystyle\int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{[C(x,y) \leq C(p_1, p_2)]} \, dx \, dy & \text{otherwise} \end{cases}.$$

To give a simple example, Brannath et al. (2002) derived the p-value function for Fisher's product test, i.e.,

$$q(p_1, p_2) = \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_1 + p_1 \cdot p_2 \left(\ln \alpha_0 - \ln \alpha_1\right) & \text{if } p_1 \in (\alpha_1, \alpha_0] \text{ and } p_1 \cdot p_2 \leq \alpha_1 \\ p_1 \cdot p_2 \left[1 + \ln \alpha_0 - \ln(p_1 \cdot p_2)\right] & \text{if } p_1 \in (\alpha_1, \alpha_0] \text{ and } p_1 \cdot p_2 > \alpha_1 \end{cases} .$$

Upon conclusion of an adaptive group-sequential study, median unbiased point estimates and (monotone) confidence intervals may be constructed using the stopping rule specified at the outset (Brannath et al., 2002). Alternatively, one may construct a sequence of repeated confidence intervals (conservative, simultaneous coverage probability) that are independent of any particular stopping rule (Jennison and Turnbull, 1989; Lehmacher and Wassmer, 1999).

The generalization to designs with an arbitrary, not necessarily pre-planned number of stages is straightforward. For instance, assume p' in the formula for the global p-value $p = q(p_1, p')$ stems from another two-stage design specified just after completion of the first stage such that $p' = q'(p_2, p_3)$ and, thus, $p = q(p_1, q'(p_2, p_3))$. Thus, a level α recursive combination test can be obtained for any finite number of stages (Brannath et al., 2002). Alternatively, at any stage of the trial any design and test procedure that does not exceed the respective conditional type I error may be chosen for the continuation of the trial (Müller and Schäfer, 2001). Since any classical group-sequential design corresponds to a specific sequence of combination functions, it is a special case of the presented general class of adaptive designs.

3. Multiplicity control

Let $\mathcal{H}$ be a family of null hypotheses of interest. For any test of a null hypothesis $H_0 \in \mathcal{H}$ the local significance level or comparisonwise error rate (CER) is defined as $\text{CER} = P(\text{Reject } H_0 \,|\, H_0 \text{ is true})$. In contrast, the familywise error rate (FWE) for a subfamily $\mathcal{H}' \subset \mathcal{H}$ is defined as $\text{FWE} = P(\text{Reject at least one } H_0 \in \mathcal{H}' \,|\, \text{All } H_0 \in \mathcal{H}' \text{ are true})$. A multiple testing or comparison procedure (MCP) for the family of null hypotheses $\mathcal{H}$ is said to control the FWE in the weak sense if it protects the FWE for $\mathcal{H}' = \mathcal{H}$ but not necessarily for all subfamilies $\mathcal{H}' \subset \mathcal{H}$. If $\max_{\mathcal{H}' \subset \mathcal{H}} \text{FWE}$ is protected, the MCP is said to control the FWE in the strong sense. It is widely acknowledged that control of the CER without reference to the corresponding family $\mathcal{H}$ is insufficient. Instead, strong control of the FWE is required.

A general method for devising MCPs that strongly controls the FWE was formally introduced by Peritz (1970) and Marcus et al. (1976) and is known as the closure test. Let $\mathcal{H}$ be a finite family of null hypotheses closed under intersection, i.e., $H', H'' \in \mathcal{H}$ implies $H' \cap H'' \in \mathcal{H}$. For every $H \in \mathcal{H}$ fix a local level α test ϕ_H. Any null hypothesis $H \in \mathcal{H}$ is tested by means of ϕ_H if and only if all hypotheses $H' \subset H$, $H' \in \mathcal{H}$ have been tested and rejected using $\phi_{H'}$. Thus, strong control of the FWE is guaranteed.

Specifically, the presence of interim analyses does not affect the error properties of the closure test as long as each $H \in \mathcal{H}$ is decided upon by its prefixed ϕ_H (e.g., a group-sequential or adaptive test) and local levels are kept at α. However, if the design of the experiment is modified in consequence of the interim results (e.g., discontinuation of arms or assessment of endpoints), the subsequent application of the prefixed local level α tests may be problematic (Hellmich, 2001). For example, consider a conventional group-sequential trial with 4 arms and 1 interim analysis (i.e., 2 stages), all pairwise comparisons of means being of interest, see Figure 1.

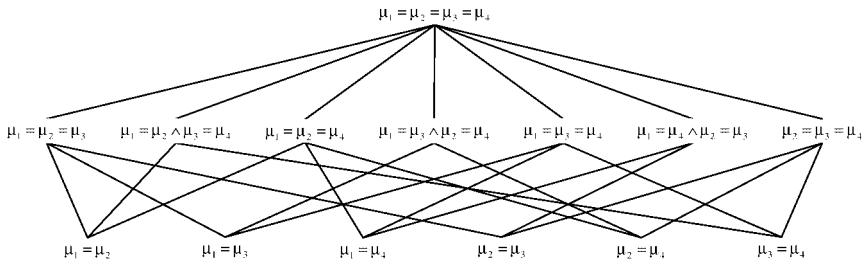

Figure 1: Closed system of null hypotheses for all pairwise comparisons of four groups.

Assume, after completion of stage 1 the null hypotheses (of equal means) H_0^{1234}, H_0^{123}, H_0^{124}, $H_0^{12\wedge 34}$, and H_0^{12} are rejected at local level α and arm 1 is dropped because of demonstrated inferiority. Suppose no further null hypotheses can be rejected. Continue the trial until completion of stage 2. Then, for comparison of arms 3 and 4, the null hypotheses H_0^{34}, H_0^{234}, and H_0^{134} have to be rejected at local level α whilst H_0^{1234} and $H_0^{12\wedge 34}$ need not be retested. Testing H_0^{134} might be problematic since the inferior arm 1 has been dropped at stage 1. Unless some adaptive testing method is used, any change of the prefixed test statistic (e.g., from F to t) is not covered by the closed testing principle and, therefore, has to be avoided. (Note, conventional group-sequential methods require a fixed allocation, either balanced or unbalanced, of subjects to treatment arms. Allocations cannot be adapted to design changes, particularly to the abandonment of any arms.)

Since closure procedures become cumbersome even for a moderate number of null hypotheses, shortcuts have been investigated. For example, if the family $\mathcal{H} = \{H_1, H_2, \ldots, H_n\}$ satisfies the free combination condition (Holm, 1979; Westfall and Young, 1993), i.e., for every subfamily $\mathcal{H}_J \subset \mathcal{H}$, $J \subset \{1, 2, \ldots, n\}$ the simultaneous truth of $\mathcal{H}_J$ and falsehood of $\mathcal{H}_{\{1,2,\ldots,n\}\setminus J}$ is a plausible event, then all intersection hypotheses differ from each other. Hence, the general closure test can be simplified without loss of power in the following way (Hommel, 1986): For each index set $J \subset \{1, 2, \ldots, n\}$ choose a level α test for $H_J = \bigcap \{H_j \,|\, j \in J\}$, e.g., all tests of the same type. Thus, a hypothesis H_j is rejected if and only if all H_J with $j \in J$ can be rejected. If the free combination condition is not satisfied, strong control of the FWE is guaranteed while power may be lost compared to the general closure test because a hypothesis may be tested differently with the simplified version. Important applications (see Hommel and Kropf, 2001) include (i) Holm's procedure where H_J is locally rejected if $\min\{p_j \,|\, j \in J\} \leq \alpha/|J|$ (Bonferroni tests) and (ii) fixed sequences of hypotheses (a priori ordered) $H_1 \succ H_2 \succ \cdots \succ H_n$ where H_J is locally rejected if the front hypothesis H_f with front index $f = \min J$ is rejected (the symbol $\succ$ denotes 'is more important than').

While being more powerful than single step procedures, closure procedures do not in general yield corresponding confidence sets (Hsu, 1996, p. 45) and that may be viewed as an important drawback (CPMP/EWP, 2003). Two additional problems may arise, (i) for hypotheses $H', H'' \in \mathcal{H}$ whose local tests $\phi_{H'}, \phi_{H''}$ give (unadjusted) p-values $p' \leq p''$, it may occur that H'' is rejected and H' is not, and (ii) directional errors are not always controlled by closure procedures, i.e., for

two-sided tests false inferences on the sign of the effect may occur. Hence directional inference in closure procedures should be considered with care (Westfall et al., 1999, p. 160).

4. Multiple testing and adaptive designs

An application of the closure test for non-adaptive group-sequential designs was described by Tang and Geller (1999). Let $\{H_1, H_2, \ldots, H_n\}$ denote the set of null hypotheses under investigation. For each intersection null hypothesis $H_J = \bigcap \{H_j \,|\, j \in J\}$, $J \subset \{1, 2, \ldots, n\}$, let $Z_J(t)$ be a group-sequential test statistic with one-sided boundary $c_J(t)$, i.e., $P_{H_J}\{Z_J(t) > c_J(t) \text{ for some } t\} \leq \alpha$. Tang and Geller proposed the following procedure for the multiple endpoint setting: Conduct interim analyses of the global null hypothesis $\bigcap \{H_1, H_2, \ldots, H_n\}$. As soon as this is rejected at some time t^*, apply the closure procedure to the subhypotheses H_J using $Z_J(t^*)$ and $c_J(t^*)$. Otherwise, no hypothesis is rejected. If any subhypothesis is not rejected, continue the trial and repeat the closure procedure, not retesting previously rejected hypotheses, until all hypotheses are rejected or the final stage of the trial has been reached. In fact, this proposition holds for any intersection-closed family of null hypotheses. Retesting of previously rejected hypotheses may result in reduced power and possibly contradictory findings, therefore it is not recommended. For adaptive designs, several applications of the closure test have been described by Bauer and Röhmel (1995), Kieser et al. (1999), Bauer and Kieser (1999), Lehmacher et al. (2000) and Kropf et al. (2000). A general theory was described by Hommel (2001).

In order to apply the closure principle to an adaptive two-stage design, for each intersection null hypothesis H_J fix an adaptive test with local tests ϕ_{J1}, ϕ_{J2} and conditional error function α_J. After completion of the first stage, ϕ_{J1} yields the p-value p_{J1}. If $p_{J1} \leq \alpha_1$, H_J is rejected and need not to be tested again (after completion of the second stage). If $p_{J1} > \alpha_0$, H_J can never be rejected. If $\alpha_1 < p_{J1} \leq \alpha_0$, (possibly) adapt the local test ϕ_{J2} yielding the p-value p_{J2} after completion of the second stage. If $p_{J2} \leq \alpha_J(p_{J1})$ H_J is rejected. Otherwise, it is accepted. The merits of using adaptive designs in multiple testing situations will be evaluated below. For ease of presentation, only two-stage designs are considered and singletons $\{j\} \subset J$ are written without brackets in subscripts, thus $H_{ji} = H_{\{j\}i}$, $i = 1, 2$ etc. (The following ideas may be applied without modification to designs with more than two stages.)

First, let $\{H_1, H_2, \ldots, H_n\}$ denote the initial set of null hypotheses tested within an adaptive two-stage design, where the hypotheses with indices in $E \subset \{1, 2, \ldots, n\}$ are excluded after the completion of the first stage because they have been rejected, retained or became irrelevant. Hence, for the combination test, the local test ϕ_{J2} of H_J, $J \subset \{1, 2, \ldots, n\}$ has to be based on the possibly restricted set of hypotheses with indices in $J \setminus E$ to be investigated at the second stage (e.g., a Bonferroni test with $p_{J2} = |J \setminus E| \cdot \min\{p_{j\,2} \,|\, j \in J \setminus E\}$), see Figure 2. If $J \setminus E$ is empty, the corresponding combination test cannot be carried out.

This strategy was applied by Kieser et al. (1999) for inference on multiple endpoints and by Bauer and Kieser (1999) to the (related) problem of multiple comparisons with a common control. Hellmich (2001) focussed on all pairwise comparisons between multiple treatment arms and showed that the appealing sequentially rejective strategy proposed by Follmann et al. (1994) does not in general guarantee strong control of the FWE. When using an adaptive design in contrast, treatment arms may be terminated as a consequence of interim analysis, or as a result of safety

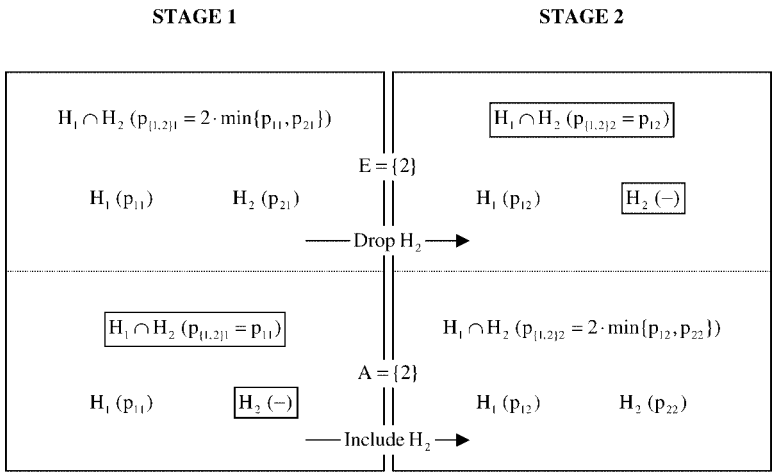

Figure 2: Inclusion or exclusion of a null hypothesis at interim analysis; Bonferroni-adjusted p-values are shown in brackets.

assessment, without corrupting strong error control. Various authors investigated the properties of closure procedures for dose-response analysis (incl. Marcus et al., 1976; Bauer and Budde, 1994; Rom et al., 1994; Tamhane et al., 1996). Usually, a monotone dose-response relationship is assumed, i.e., the expected responses in k groups given increasing dose-levels fulfill the order constraint $\mu_0 \leq \mu_1 \leq \cdots \leq \mu_k$, where μ_0 corresponds to placebo. In order to establish a global trend and determine the minimum effective dose, the nested (and therefore intersection-closed) set of null hypotheses to be considered is given by $H_{0i} : \mu_0 = \mu_1 = \cdots = \mu_i$ vs. $H_{ai} : \mu_0 \leq \mu_1 \leq \cdots \leq \mu_i$ $(i = 1, 2, \ldots, k)$ with $\mu_{j-1} < \mu_j$ for at least one $j \in \{1, 2, \ldots, i\}$. The larger family of null hypotheses generated by hypotheses corresponding to pairwise comparisons of adjacent doses $H_{0i} : \mu_{i-1} = \mu_i$ vs. $H_{ai} : \mu_{i-1} < \mu_i$ $(i = 1, 2, \ldots, k)$, additionally allows to test for efficient dose steps, e.g., the highest dose level which still provides a clinically relevant step in the response as compared to the adjacent lower dose, thus providing further details on the dose-response relationship. The combination of adaptive designs and closure procedures for dose-response analysis was recommended by Lehmacher et al. (2000) since for such studies sufficient sample size estimation is usually not feasible. Bauer and Röhmel (1995) emphasized the importance of the choice of doses to be included in the study. If the experiment was performed in a late or early plateau of the dose-response relationship the (indirect) demonstration of efficacy might fail (CPMP/ICH, 1994). Therefore, they advocated the use of an adaptive method starting with a few, say two, doses from the conjectured therapeutic dose range. After completion of the first stage, if an insufficient trend is visible, the doses to be investigated in the second stage may be changed, for example by lowering the low dose and/or increasing the high dose. The efficacy decision then relies on the combination test based on the p-values from the separate stages. Moreover, controlled multiple inference on the null hypotheses tested at the separate stages can be achieved by a prefixed closure procedure.

Secondly, the aforementioned method for dose-response analysis suggests that within adaptive designs even new hypotheses may be included at interim analyses. Let $\{H_1, H_2, \ldots, H_n\}$ denote the final set of null hypotheses tested within an adap-

tive two-stage design where the hypotheses with indices in $A \subset \{1, 2, \ldots, n\}$ may be included after completion of the first stage because they have become relevant. For interim analysis, the (prefixed) local test ϕ_{J1} of H_J, $J \subset \{1, 2, \ldots, n\}$ is based on the possibly restricted set of hypotheses with indices in $J \setminus A$ investigated at the first stage (e.g., a Bonferroni test with $p_{J1} = |J \setminus A| \cdot \min\{p_{j\,1} \,|\, j \in J \setminus A\}$), see Figure 2. If $J \setminus A$ is empty, the corresponding combination test is degenerate and only uses the p-value p_{J2} from the second stage (fixed sample situation). Thus, seen from a methodological viewpoint, the inclusion and exclusion of null hypotheses at interim analyses are just reverse strategies.

Thirdly, as a consequence of an adaptive interim analysis the order of a fixed sequence of hypotheses may be altered, reflecting a corresponding shift in interest or importance. Suppose that the fixed sequence of null hypotheses $H_1 \succ H_2 \succ \cdots \succ H_n$ is to be tested in a two-stage adaptive design. Hence, at the interim analysis, H_J, $J \subset \{1, 2, \ldots, n\}$ is locally tested by $\phi_{J1} = \phi_{f1}$ with $f = \min J$ (front index), yielding the p-value $p_{J1} = p_{f1}$. For the second stage, a new sequence (permutation) π of the null hypotheses may be fixed, say $H_{\pi(1)} \succ H_{\pi(2)} \succ \cdots \succ H_{\pi(n)}$. Thus, after completion of the second stage, H_J, $J \subset \{1, 2, \ldots, n\}$ is locally tested by $\phi_{J2} = \phi_{g2}$ with front index $g = \pi(\min\{j \,|\, \pi(j) \in J\})$, yielding the p-value $p_{J2} = p_{g2}$, see Figure 3. In fact, the rearrangement of a fixed sequence of null hypotheses is a special case of an adaptive choice of test statistics for specific hypotheses in order to gain power. Another example of this strategy is the adaptive choice of weights for multiple endpoints (Westfall et al., 1998) or hypotheses (Hommel, 2001).

The three strategies presented above may be applied simultaneously within the same experiment, meaning that at an adaptive interim analysis, null hypotheses may be dropped, new null hypotheses may be included and the sequencing of null hypotheses may be altered. In the extreme, consider the case in which all null hypotheses tested in the interim analysis are dropped from the study, and only newly introduced null hypotheses are tested in the final analysis. While, from a clinical point of view, this seems to be a procedure of questionable value, there is nothing wrong with the method.

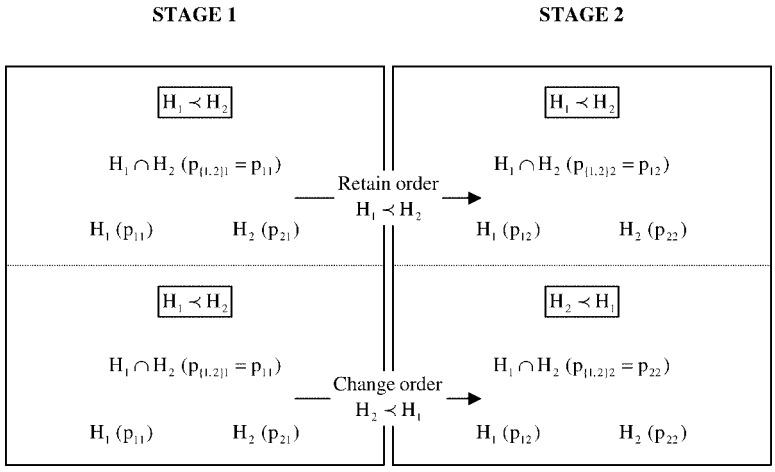

Figure 3: Change of the order of a fixed sequence of null hypotheses; p-values are shown in brackets.

5. An example from the literature

Bauer et al. (1998) presented an analysis of a fixed sample clinical trial with five treatment arms to investigate the dose-response relationship of a new calcium channel blocker (CCB) in three doses (50, 100 and 150 mg) versus placebo (zero dose) and an active control (10 mg amlodipine) in patients with chronic stable angina pectoris. Summary results for the primary endpoint 'post/pre-treatment difference of exercise test duration in seconds' are shown in Table 1. Based on these results, imagine an adaptive two-stage clinical trial to test the following fixed sequence of null hypotheses, $H_1 : \mu_{\text{ac}} \leq \mu_0$, i.e., active control is not superior to placebo, $H_2 : \mu_{150} \leq \mu_0$, i.e., CCB_{150} is not superior to placebo and $H_3 : \mu_{150} \leq \mu_{\text{ac}} - 15$ seconds, i.e., CCB_{150} is relevantly inferior to active control (margin $\delta = 15$ seconds), thus $H_1 \succ H_2 \succ H_3$. For each intersection null hypothesis an one-sided adaptive test according to Bauer and Köhne (1994) is prefixed with $\alpha = 0.025$, $\alpha_1 = 0.0102$, $\alpha_0 = 0.5$ and $c = 0.0038$. The maximum sample size allocated in the fictive trial protocol is 300 with the interim analysis to take place after 15 patients have been randomized to each group. The null hypotheses H_1 and H_3 are decided upon by t-tests and H_2 by the linear contrast test with coefficients $(-3, -1, 1, 3)$. All tests use the pooled standard deviation for all five treatment arms (82.9 seconds). Of course, tests for relevant inferiority or superiority of the other CCB dose groups (50 and 100 mg) could be incorporated in the fixed sequence but this would make the presentation of this example unduly complicated.

Assume the means and standard deviations of Table 1 are those observed at the interim analysis. Since all p-values are greater than $\alpha_1 = 0.0102$, none of the null hypotheses can yet be rejected (see Table 2). Unexpectedly, the active control does not show a clear effect. Consequently, one of the following four strategies could be pursued. Strategy 1: The trial is completed as planned, i.e., no adaptation takes place at the interim analysis. Strategy 2: The trial is stopped and a completely new trial is planned and conducted with a (hopefully) better active control. Strategy 3: The old active control treatment is dropped and a better active control is chosen for the second stage. Thus, in contrast to conducting a completely new trial, most of the data from the first stage are still available for the final analysis using the combination test. Strategy 4: The new CCB is reckoned a promising treatment possibly superior to the present standard and safe even at high dose. Pursuing the last strategy a bit further, a new null hypothesis is included, $H_4 : \mu_{150} \leq \mu_{\text{ac}}$, i.e., CCB_{150} is not superior to active control, and a new sequence of null hypotheses is fixed for the second stage, say $H_2 \succ H_4 \succ H_1 \succ H_3$ (see Table 2). Moreover, the contrast statistic for H_2 is tailored to the leveling response in the highest dose group,

Table 1: Summary results of post/pre-treatment differences of exercise test duration (in seconds) in patients with chronic stable angina pectoris on one of three doses of a new calcium channel blocker (CCB), placebo or active control (see Bauer et al., 1998).

	Dose of new CCB				Active control
	0	50	100	150	
Mean	57.5	76.8	109.5	105.3	67.3
SD	75.0	75.5	87.1	85.7	90.1
N	62	60	60	62	59

Table 2: Interim results of a fictive adaptive two-stage clinical trial based on the summary data given in Table 1 with a fixed sequence of null hypotheses. According to Strategy 4 (see text) a new null hypothesis is included and a new sequence is fixed for Stage 2.

STAGE 1 $H_1 \succ H_2 \succ H_3$			STAGE 2 $H_2 \succ H_4 \succ H_1 \succ H_3$		
Index set J	Front index	p_{J1}	$\alpha_J(p_{J1})$	Index set J	Front index
$\{1,2,3\}$	1	0.373	0.010	$\{1,2,3,4\}$	2
..				$\{1,2,3\}$	2
$\{1,2\}$	1	0.373	0.010	$\{1,2,4\}$	2
..				$\{1,2\}$	2
$\{1,3\}$	1	0.373	0.010	$\{1,3,4\}$	4
..				$\{1,3\}$	1
$\{2,3\}$	2	0.033	0.114	$\{2,3,4\}$	2
..				$\{2,3\}$	2
$\{1\}$	1	0.373	0.010	$\{1,4\}$	4
..				$\{1\}$	1
$\{2\}$	2	0.033	0.114	$\{2,4\}$	2
..				$\{2\}$	2
$\{3\}$	3	0.040	0.094	$\{3,4\}$	4
..				$\{3\}$	3
—	—	—	0.025	$\{4\}$	4

i.e., the coefficients $(-3,-1,2,2)$ are chosen, and the initially planned sample size of 45 per group for the second stage is reduced to 40 because of high conditional power. Applying the closure procedure, an intersection null hypothesis H_J, $J \subset \{1,2,3,4\}$ could then be rejected after completion of the second stage, if both (i) $p_{J\setminus\{4\}1} \leq \alpha_1$ or $[p_{J\setminus\{4\}1} \leq \alpha_0$ and $p_{J2} \leq \alpha_J(p_{J\setminus\{4\}1}) = c/p_{J\setminus\{4\}1}]$ and (ii) any subset intersection null hypothesis $H_{J'}$, $J' \subset J$ were rejected, beforehand. A similar example was presented by Kropf et al. (2000) and Hommel and Kropf (2001).

Since $H_3 \subset H_4$, or (relevant) inferiority implies non-superiority, the stronger null hypothesis H_3 may be dropped after the interim analysis for purely logical reasons. Switching the alternative from non-inferiority to superiority is not a possibility unique to adaptive designs, but may as well be performed after conclusion of any trial provided that (i) it has been properly designed and carried out according to the strict requirements of a non-inferiority trial and (ii) the intention-to-treat analysis receives the greatest emphasis (CPMP/EWP, 2000). However, within an adaptive design, the alternative can be strengthened after an (early) interim analysis and then combined with an adequate re-calculation of the sample size to ensure high power. General closure procedures (step-down/up) for establishment of superiority/non-inferiority of a new treatment compared with several standard treatments were investigated by Dunnett and Tamhane (1997) and may well be applied within adaptive designs following the strategies presented in Section 4.

A tested substance and an active control are deemed equivalent regarding their efficacy if the 95% confidence interval of their difference (or ratio, see Hauschke and Kieser, 2001) θ in mean efficacy is completely covered by a pre-specified equivalence margin $(-\delta,+\delta)$, $\delta > 0$ (CPMP/EWP, 2000). Alternatively,

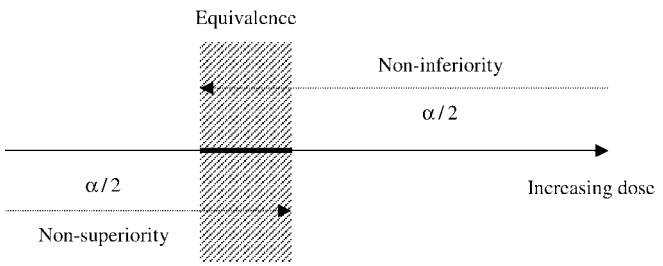

Figure 4: Hierarchical test procedure with equal α-splitting for multi-dose equivalence studies assuming a monotone increasing dose-response relationship (Channon, 2000; Bauer et al., 2001b).

equivalence is demonstrated if both one-sided hypotheses of relevant inferiority ($H_0^{\text{inf}} : \theta \leq -\delta$ vs. $H_a^{\text{inf}} : \theta > -\delta$) and relevant superiority ($H_0^{\text{sup}} : \theta \geq \delta$ vs. $H_a^{\text{sup}} : \theta < \delta$) can be rejected at level $\alpha/2 = 2.5\%$. Channon (2000) focussed on multi-dose equivalence studies (with multiple doses of the test substance and one dose of the reference) and proposed a hierarchical (closed) test procedure for monotone (increasing) dose-response relationships which was extended by Bauer et al. (2001b). The two subfamilies of inferiority and superiority hypotheses are subjected to separate MCPs that strongly control the FWE at 2.5% (equal α-splitting). Hence, any test dose for which both one-sided hypotheses can be rejected is assumed to be equivalent to the reference dose, see Figure 4. Again, the application of these closure procedures within adaptive designs is straightforward.

6. Discussion

Adaptive designs which permit extensive re-design in consequence of interim analyses whilst not compromising (strong) type I error control can successfully alleviate/circumvent the difficulties entailed by multiplicity as present in virtually all clinical trials. Software supporting adaptive study planning is becoming available, including SA2D (2000), ADDPLAN 2.1 (2003) and East 3.0 (2003).

Within closure procedures any statistical level α test can be used for intersection null hypotheses. Specifically, local Bonferroni tests can be improved upon by accounting for the correlations between the test statistics which are either known or can be estimated using resampling methods (Westfall and Young, 1993). Moreover, logical interrelations between null hypotheses can be exploited in order to gain in power (Hommel and Bernhard, 1999). Since confidence sets are regrettably not generally available for closure procedures, either unadjusted confidence intervals (controlling only the CER) or more conservative simultaneous confidence intervals from corresponding single-step procedures (e.g., according to Dunnett or Tukey) may be presented.

The new flexibility introduced by adaptive designs to clinical trials entails a possibly high danger of malpractice. To prevent any fraudulent use, specific regulatory guidance is required on the prerequisites of any contingent modification of the original study plan. Key requirements certainly include that (i) only prospective modifications fully detailed in and communicated by study protocols or amendments are acceptable and (ii) the trial investigator has to establish that following

an adaptation the same trial is still being conducted. If the latter cannot sufficiently be demonstrated, both subtrials (ante/post adaption) must not be interpreted as delivering evidence against the same null hypothesis (but the intersection of two different ones), resulting in a probably dramatic loss in power.

Acknowledgment

The authors thank the editors and anonymous referees for their comments which helped to improve the presentation of the paper.

References

ADDPLAN 2.1 (2003). Adaptive Designs – Plans and Analyses. ADDPLAN GmbH, Robert-Perthel-Straße 77a, 50739 Köln. See http://www.addplan.org/.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika* **44**, 9–56. MR85685

Armitage, P. (1975). *Sequential medical trials*. New York: John Wiley & Sons, 2nd ed. MR370997

Bauer, P. (1989). Multistage testing with adaptive designs. *Biom. Inform. Med. Biol.* **20**, 130–148.

Bauer, P., Brannath, W. and Posch, M. (2001a). Flexible two-stage designs: an overview. *Methods Inf. Med.* **40**, 117–121.

Bauer, P., Brannath, W. and Posch, M. (2001b). Multiple testing for identifying effective and safe treatments. *Biom. J.* **5**, 605–616. MR1863490

Bauer, P. and Budde, M. (1994). Multiple testing for detecting efficient dose steps. *Biom. J.* **36**, 3–15.

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Stat. Med.* **18**, 1833–1848.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041. Correction 1996; 52: 380.

Bauer, P. and Röhmel, J. (1995). An adaptive method for establishing a dose-response relationship. *Stat. Med.* **14**. 1595–1607.

Bauer, P., Röhmel, J., Maurer, W. and Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Stat. Med.* **17**, 2133–2146.

Birkett, M. A. and Day, S. J. (1994). Internal pilot studies for estimating sample size. *Stat. Med.* **13**, 65–72.

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *J. Amer. Statist. Assoc.* **97**, 236–244. MR1947283

Channon, E. J. (2000). Equivalence testing in dose-response studies. *Drug Inf. J.* **34**, 551–562.

CPMP/EWP (2000). Points to consider on switching between superiority and non-inferiority, CPMP/EWP/482/99. Available at http://www.emea.eu.int/.

CPMP/EWP (2003). Points to consider on adjustment for baseline covariates, CPMP/EWP/2863/99. Available at http://www.emea.eu.int/.

CPMP/ICH (1994). Topic E4 Step 5: Note for guidance on dose response information to support drug registration, CPMP/ICH/378/95. Available at http://www.emea.eu.int/.

Dunnett, C. W. and Tamhane, A. C. (1997). Multiple testing to establish superiority/equivalence of a new treatment compared with k standard treatments. *Stat. Med.* **16**, 2489–2506.

East 3.0 (2003). Superior software for the design, simulation and interim monitoring of flexible clinical trials. Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139-3309. See http://www.cytel.com/.

Follmann, D. A., Proschan, M. A., and Geller, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* **50**, 325–336. MR1294682

Friede, T. and Kieser, M. (2002). On the inappropriateness of an em algorithm based procedure for blinded sample size re-estimation. *Stat. Med.* **21**, 165–176.

Gould, A. L. and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Comm. Statist. Theory Methods* **21**, 2833–2853.

Hauschke, D. and Kieser, M. (2001). Multiple testing to establish noninferiority of k treatments with a reference based on the ratio of two means. *Drug Inf. J.* **35**, 1247–1251.

Hellmich, M. (2001). Monitoring clinical trials with multiple arms. *Biometrics* **57**, 892–898. MR1863452

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70. MR538597

Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika* **33**, 321–336. MR868042

Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biom. J.* **43**, 581–589. MR1863488

Hommel, G. and Bernhard, G. (1999). Bonferroni procedures for logically related hypotheses. *J. Statist. Plann. Inference* **82**, 119–128. MR1736436

Hommel, G. and Kropf, S. (2001). Clinical trials with an adaptive choice of hypotheses. *Drug Inf. J.* **35**, 1423–1429.

Hsu, J. C. (1996). *Multiple Comparisons. Theory and methods*. Boca Raton: Chapman & Hall/CRC. MR1629127

Jennison, C. and Turnbull, B. W. (1989). Interim analysis: the repeated confidence interval approach. *J. Roy. Statist. Soc. Ser. B* **51**, 305–361. MR1017201

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: Chapman & Hall/CRC. MR1710781

Kieser, M., Bauer, P. and Lehmacher, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analysis. *Biom. J.* **41**, 261–277.

Kropf, S., Hommel, G., Schmidt, U., Brickwedel, J. and Jepsen, M. S. (2000). Multiple comparisons of treatments with stable multivariate tests in a two-stage adaptive design, including a test for non-inferiority. *Biom. J.* **42**, 951–965. MR1847815

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663. MR725380

Lehmacher, W., Kieser, M., and Hothorn, L. (2000). Sequential and multiple testing for dose-response analysis. *Drug Inf. J.* **34**, 591–597.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286–1290.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 487–497. MR468056

Müller, H. H. and Schäfer, H. (2001). Adaptive group sequential designs for clincal trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**, 886–891. MR1859823

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

Peritz, E. (1970). A note on multiple comparisons. Unpublished manuscript, Hebrew University, Israel.

Pocock, S. J. (1977). Group sequential methods in the design and anaylsis of clinical trials. *Biometrika* **64** 191–199.

Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biom. J.* **41**, 689–696.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extensions of studies based on conditional power. *Biometrics* **51**, 1315–1324.

Rom, D. M., Costello, R. J., and Connell, L. T. (1994). On closed test procedures for dose-response analysis. *Stat. Med.* **13**, 1583–1596.

SA2D (2000). Statistical Software for Simulation of Adaptive Two Stage Designs. Written by M. Bauer; Department of Medical Statistics, University of Vienna. Available at http://www.univie.ac.at/medstat/sa2d.html.

Tamhane, A. C., Hochberg, Y. and Dunnett, C. W. (1996). Multiple test procedures for dose finding. *Biometrics* **52**, 21–37.

Tang, D. and Geller, N. L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* **55**, 1188–1192.

Tsiatis, A. A., Rosner, G. L. and Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**, 797–803. MR775387

Wald, A. (1947). *Sequential analysis*. New York: John Wiley & Sons. MR20764

Wassmer, G. (1999). *Statistische Testverfahren für gruppensequentielle Pläne in klinischen Studien*. Köln: Alexander Mönch.

Wassmer, G., Eisebitt, R. and Coburger, S. (2001). Flexible interim analyses in clinical trials using multistage adaptive test designs. *Drug Inf. J.* **35**, 1131–1146.

Westfall, P. H., Krishen, A. and Young, S. S. (1998). Using prior information to allocate significance levels for multiple endpoints. *Stat. Med.* **17**, 2107–2119.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests. Using the SAS® System*. Books by Users. Cary: SAS® Institute Inc.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*. New York: John Wiley & Sons.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat. Med.* **9**, 65–72.