Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: General Concepts

Frank Bretz[*,1], **Heinz Schmidli**[1], **Franz König**[2], **Amy Racine**[1], and **Willi Maurer**[1]

[1] Novartis Pharma AG, Lichtstrasse 35, 4002 Basel, Switzerland
[2] Section of Medical Statistics, Core Unit of Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

Received 7 July 2005, revised 21 December 2005, accepted 15 March 2006

Summary

Traditional drug development consists of a sequence of independent trials organized in different phases. Full development typically involves (i) a learning phase II trial and (ii) one or two confirmatory phase III trial(s). For example, in the phase II trials several doses of the new compound might be compared to a control and/or placebo with the goal of deciding whether to stop or continue development and, in the latter case, selecting one or two "best" doses to carry forward into the confirmatory phase. The phase III trials are then conducted as stand-alone confirmatory studies, not incorporating in their statistical analyses data collected in the previous phases.

Seamless phase II/III designs are aimed at interweaving the two phases of full development by combining them into one single, uninterrupted study conducted in two stages. In the dose-finding example above, one (or more) dose(s) are selected after the first stage based on the available data at interim, and are then observed further in the second stage. The final analysis of the selected dose(s) includes patients from both stages and is performed such that the overall type I error rate is controlled at a prespecified level regardless of the dose selection rule used at interim. The adequacy of the dose selection at interim is obviously a critical step for the success of a seamless phase II/III trial. In this paper we focus on the description of flexible test procedures allowing for adaptively selecting hypotheses at interim and thus allowing the combination of learning and confirming in a single seamless trial. We review the statistical background, introduce different test procedures and compare them in a power study. In a subsequent paper (Schmidli et al., 2006) we give several applications from our daily practice and discuss related implementation issues in conducting adaptive seamless designs.

Key words: Adaptive seamless design; Adaptive tests; Closure principle; Combination tests; Multiple testing; Hypotheses selection.

1 Introduction

The development of biopharmaceutical products is becoming increasingly challenging, inefficient and costly. In March 2004 the US Food and Drug Administration (FDA) released a white paper entitled "Stagnation/Innovation: Challenge and Opportunity on the Critical Path to New Medical Products" (Anonymous, 2004). The document acknowledges that today's revolution in biomedical science has raised new hope for the treatment of many diseases, but points out that the number of new drug and biologic applications submitted to the FDA has declined considerably in the last decade and discusses several potential causes for this decline. The white paper concludes that if the drug development processes do not become more efficient and effective, innovation may continue to stagnate and the biomedical revolution may fail to achieve its full potential.

[*] Corresponding author: e-mail: frank.bretz@novartis.com, Phone: +41 61 324 4064, Fax: +41 61 324 3039

There are many ways that statistics and biometrics in general can contribute to improve the drug development cycle. The FDA document does not directly mention the role statisticians could play in this process. However, O'Neill (2004) recently reported on feedback the FDA received from stakeholders in the pharmaceutical industry and academia regarding its Critical Path initiative. He in particular proposed adaptive designs as one innovative statistical approach worth to be investigated.

Classical drug development consists of a sequence of independent trials in different phases. In a typical phase II trial one would compare several treatments (for example, different dose levels of a new compound) with a control. After the completion of this trial it is then decided whether to continue the drug development and which treatment(s) to carry forward to the phase III. The phase III trials are then evaluated as stand-alone confirmatory trials, ignoring information from previous phases.

Seamless phase II/III designs aim at interweaving these trials by combining them into one single study conducted in two stages. In the example above, one (or more) treatment(s) are selected after the first stage based on the available data at interim, and observed further in the second stage. The final analysis of the selected treatment includes the patients of both stages and is performed such that the overall type I error rate is controlled at pre-specified level regardless of the adaptation rule used at interim. Such flexibility particularly allows the use of Bayesian decision tools for interim adaptation without affecting the frequentist significance level. Flexibility is of upmost importance for efficient and ethical conduct of clinical trials with a continuous monitoring of efficacy and safety. Ideally, adaptive seamless designs (ASD) thus (i) reduce the time to decide on, plan and implement the next clinical phase (reduction of the "white space" between the two studies), (ii) save costs through the combination of evidences across two studies and thus the need for fewer patients (or, equivalently, increase the information value and the reliability of decision making while maintaining the same sample sizes), and (iii) get long-term safety data earlier as a direct consequence of following-up the phase II patients.

This paper discusses some statistical aspects related to the design and analysis of ASD. Based on our experience from real applications we introduce several innovative designs and point to relevant methodological problems. We review adaptive designs and closed test procedures and describe how these concepts can be combined for adaptive choices of hypotheses at an interim stage. We also investigate alternative strategies which are based on well known one-stage multiple test procedures. In a subsequent paper (Schmidli et al., 2006) we introduce applications of innovative adaptive and seamless designs for different experimental questions and also discuss related practical implementation issues.

This paper serves as an introduction to the general concepts of ASD and addresses any interested statistician, including those who have never conducted such a study so far. The subsequent paper is devoted to a broader audience, including clinicians, project management, and related personnel, who are faced with practical planning issues and implementation problems inherent to ASD.

2 Principles of Adaptive Designs

For simplicity, we assume in this section that a single directional (i.e., one-sided) null hypothesis H is tested against the alternative K in a two-stage design, i.e., with one single interim analysis. Based on the first-stage data (the unblinded data collected up to interim) it is decided whether to continue the study (conducting the second stage) or not (early stopping: either due to futility or due to early rejection of H). In the case that one continues with the second stage, the final analysis at the end of the study combines the results of both stages.

Let p_i denote the p-values for stage $i = 1, 2$. The *adaptive test procedure*, as proposed by Bauer and Köhne (1994), is specified as follows:

1. Define a test procedure for stage 1, determine the stopping rules for the interim decision and pre-specify the combination function C of p_1 and p_2 for the final analysis.
2. Conduct stage 1 of the study, resulting in p_1.
3. Based on p_1, decide whether to stop at interim (either rejecting or retaining H) or to continue the study.

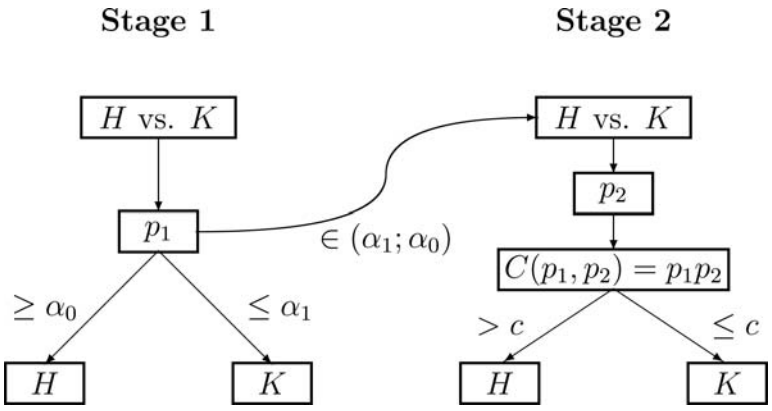

Figure 1 Two-stage adaptive design to test a single null hypothesis H.

4. If the study is continued, use all information (also external to the study, if available) to design the second stage, for example, re-assess the second stage sample size.
5. Conduct stage 2 of the study, resulting in p_2 being independent from p_1 under H.
6. Combine p_1 and p_2 using $C = C(p_1, p_2)$ and decide for or against H by comparing C with an appropriate critical value.

Note that adaptive designs offer a high level of flexibility during the conduct of the trial. They require the least amount of pre-specified decision rules prior to the study among multistage designs so that the total information available at interim can be used in designing the second stage.

To make the ideas concrete, assume that Fisher's combination test is used, i.e., H is rejected at the final stage if (Bauer and Köhne, 1994)

$$C(p_1, p_2) = p_1 p_2 \leq c = \exp\left(-\chi^2_{4,1-\alpha}/2\right),$$

where $\chi^2_{\nu,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with ν degrees of freedom. Consider the early stopping boundaries α_0 and α_1, such that (i) if $p_1 \leq \alpha_1$ the trial stops after the interim analysis with an early rejection of H, (ii) if $p_1 \geq \alpha_0$ the trial stops after the interim analysis for futility (H is not rejected). Note that if $\alpha_0 = 1$ no stopping for futility is foreseen and if $\alpha_1 = 0$ no early rejection of H is possible. In order to maintain the type I error rate at pre-specified level α simultaneously across both stages, α_1 is computed by solving $\alpha_1 + c(\ln \alpha_0 - \ln \alpha_1) = \alpha$ for given α and α_0. The flow chart in Figure 1 depicts the associated decision process.

Note that other combination functions than Fisher's product test can be used, see among others Proschan and Hunsberger (1995) and Cui, Hung and Wang (1999). A common choice is the weighted inverse normal method (Lehmacher and Wassmer, 1999)

$$C(p_1, p_2) = 1 - \Phi[w_1 \Phi^{-1}(1-p_1) + w_2 \Phi^{-1}(1-p_2)],$$

where $0 < w_i < 1$, $i = 1, 2$, are arbitrary weights subject to $w_1^2 + w_2^2 = 1$ and Φ denotes the standard normal cdf. If the weights $w_i, i = 1, 2$, are properly chosen, this combination function corresponds to the classical two-stage group sequential test, i.e., the squared weights are proportional to the sample size or information fractions at both stages assuming that no adaptation is done.

An alternative approach to the use of combination functions is to consider *conditional error functions* (CEF; Proschan and Hunsberger, 1995)

$$A(p_1) = P_H(\text{reject } H \mid p_1) = \begin{cases} 1 & \text{if } p_1 \leq \alpha_1 \\ 0 & \text{if } p_1 \geq \alpha_0 \\ \max\{p_2 \mid C(p_1, p_2) \leq c\} & \text{if } p_1 \in (\alpha_1; \alpha_0). \end{cases}$$

The CEF is the probability of rejecting H in the final analysis given the first-stage p-value p_1, i.e., H is rejected if $p_2 < A(p_1)$. For Fisher's product combination test, the CEF is given by c/p_1. Note that any adaptive procedure based on combination functions can be written in terms of a CEF. The CEF principle can also be applied to any test statistics used for the first stage and foreseen for the second. We refer to Brannath, Posch, and Bauer (2002) and Posch and Bauer (2003) for further details on adaptive designs.

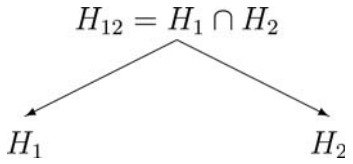

Figure 2 Closure principle for two null hypotheses H_1 and H_2.

3 The Closure Principle

In this section we review the closure principle (CP; Marcus et al., 1976), which is an important multiple testing concept and which will be used extensively in the subsequent sections. We do not consider adaptive testing in this section.

Assume that n directional null hypotheses H_i, $i = 1, \ldots, n$, are to be tested (for example, the comparison of n treatments with a control). In multiple testing situations the goal is to control strongly the familywise error rate (FWER) at pre-specified level α, where the strong FWER control is defined as the probability of rejecting at least one true null hypothesis irrespective of the configuration of null hypotheses. The CP considers all intersection hypotheses constructed from the initial hypotheses set. A null hypothesis H_i is rejected at FWER α, if all hypotheses implying H_i are rejected. More formally, the CP is defined as follows:

1. Define a set of elementary hypotheses $H_1, \ldots, H_n$ of interest.
2. Construct all possible $m \geq n$ intersection hypotheses $H_I = \bigcap_{i \in I} H_i$, $I \subseteq \{1, \ldots, n\}$.
3. For each of the m hypotheses find a suitable local level-α test.
4. Reject H_i at FWER α, if all hypotheses H_I with $i \in I$ are rejected, each at (local) level α.

Note that the choice of the tests for the m hypotheses is free and that different tests can be used for different hypotheses. This is a crucial property for the adaptive hypotheses tests in the subsequent sections.

Consider Figure 2 for an example of the CP with $n = 2$ (assuming, for example, that two treatments are compared with a control, thus resulting in two primary hypotheses of interest). In this situation

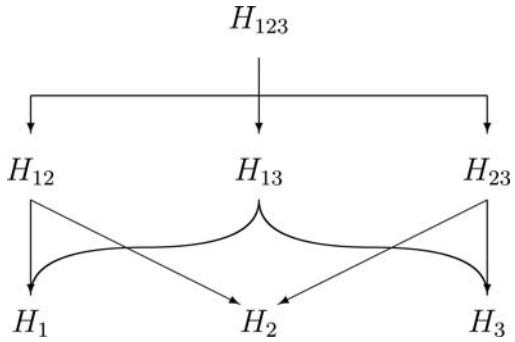

Figure 3 Closure principle for three null hypotheses H_1, H_2, and H_3.

there is only one non-trivial intersection hypothesis $H_{12} = H_1 \cap H_2$, so that the set $\mathcal{H} = \{H_1, H_2, H_{12}\}$ is closed under intersection with $m = |\mathcal{H}| = 3$. According to the CP one rejects H_1 (at FWER α) if both H_1 and H_{12} are rejected, each at (local) level α. Conversely, one rejects H_2 if both H_2 and H_{12} are rejected.

Consider Figure 3 for an example of the CP with $n = 3$. In this situation, four additional intersection hypotheses have to be considered to obtain a closed hypotheses set with $m = 7$. Now, H_1 (say) is rejected, if H_{123}, H_{12}, H_{13} and H_1 are all rejected at level α, where $H_{ij} = H_i \cap H_j, 1 \leq i, j \leq 3$ and $H_{123} = H_1 \cap H_2 \cap H_3$ is the global intersection hypothesis.

4 Multiple Testing in Adaptive Designs

In this section we show how to test adaptively multiple hypotheses by combining the techniques from the previous two sections. The following results are rather general and allow a very flexible adaptation of hypotheses at interim, as illustrated later in Section 5 and in Schmidli et al. (2006).

Assume that we are now interested in testing n hypotheses $H_1, \ldots, H_n$ using a two-stage design. The general rule is to apply the CP by constructing all intersection hypotheses and to test each of them with a suitable combination test (Hommel, 1997, 2001; Bauer and Kieser, 1999; Kieser et al., 1999). Following the CP, a null hypothesis H_i is rejected if all hypotheses implying H_i are also rejected. In the sequence we call this general principle "adaptive combination test" or simply the "Hommel procedure".

Consider Figure 4 for an example of testing adaptively $n = 2$ hypotheses. As before, let H_1, H_2, and H_{12} denote the hypotheses to be tested according to the CP. Let further $p_{i,j}$ denote the one-sided p-value for hypothesis H_j, $j \in \{1, 2, 12\}$ at stage $i = 1, 2$. Finally, let $C(p_{1,j}, p_{2,j})$, $j \in \{1, 2, 12\}$, denote the combination function C applied to the p-values $p_{i,j}$ from stage $i = 1, 2$. Note that different combination functions as well as different stopping boundaries could be used within the closed hypotheses set (for simplicity we omit this generalization here). According to the CP, H_1 (say) is rejected at FWER α, if H_1 and H_{12} are both rejected at level α, i.e., if $C(p_{1,1}, p_{2,1}) \leq c$ and $C(p_{1,12}, p_{2,12}) \leq c$.

The rejection rule can also be expressed in terms of CEF. Following this alternative approach, H_1 is rejected if $p_{2,1} \leq A(p_{1,1})$ and $p_{2,12} \leq A(p_{1,12})$. Note that the rejection rules are simplified if early rejection or non-rejection of any hypothesis is achieved at interim. If, for example, $p_{1,12} \leq \alpha_1$, then $A(p_{1,12}) = 1$, so that the condition $p_{2,12} \leq A(p_{1,12}) = 1$ is always satisfied. Consequently, H_1 is re-

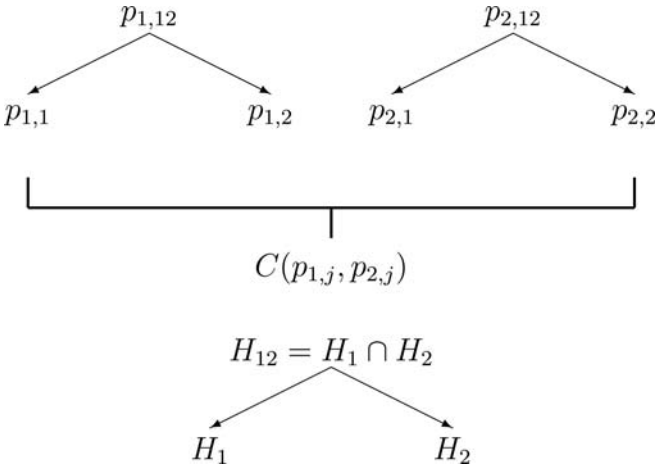

Figure 4 Closure principle for testing adaptively $n = 2$ null hypotheses H_1 and H_2.

jected if $p_{2,1} \leq A(p_{1,1})$ as long as $p_{1,12} \leq \alpha_1$. Similarly, if $p_{1,12} \geq \alpha_0$, then $A(p_{1,12}) = 0$ and H_1 cannot be rejected at the FWER α, meaning that the study can be stopped for futility at interim. Even if one decided to continue the study, H_1 would never be rejected whichever the results at the second stage were. Note that in this particular case H_2 can not be rejected either. Finally, if $p_{1,1}$, $p_{1,12} \leq \alpha_1$, H_1 is already rejected at interim (no need to continue testing H_1). Kieser et al. (1999) gave a flow chart to depict the complete decision process for $n = 2$ hypotheses.

Note that alternative test strategies are available, some of which are described in Section 6 and compared with the Hommel procedure.

5 Generic Examples

In this section we consider two generic examples of adaptively modifying multiple hypotheses after an interim analysis. The first example illustrates the selection of a treatment at interim, similar to the example described in Section 1. The second example considers a treatment switch at interim. Further generic examples and applications are described in Schmidli et al. (2006) and Posch and Bauer (2003).

5.1 Treatment selection

Assume that we have two treatments to be compared with a control. At interim we decide which of the two treatments to carry forward into the second stage. The final analysis of the selected treatment includes the patients of both stages by applying the results of the previous sections. Assume that one decides at interim to continue with treatment 1 and let H_1 be the related null hypothesis. No data is therefore available for treatment 2 after the second stage. Consequently, the intersection hypothesis H_{12} for the second stage is equal to H_1 and its test is performed using only the test of H_1. Figure 5 depicts the CP associated with the two null hypotheses H_1 and H_2 together with the related stagewise p-values as well as the resulting combination of both stages in terms of CEF.

From the CP it follows that we have to reject H_1 and H_{12} to be able to declare treatment 1 as significantly different from the control. From Figure 5 it is clear that we thus require $p_{2,1} < A(p_{1,1})$ and $p_{2,1} < A(p_{1,12})$. Equivalently, H_1 is rejected at FWER α if $p_{2,1} < \min\{A(p_{1,1}), A(p_{1,12})\}$. Note that the approach above also applies to other adaptive selection problems involving two hypotheses, such as subgroup or endpoint selection, for example.

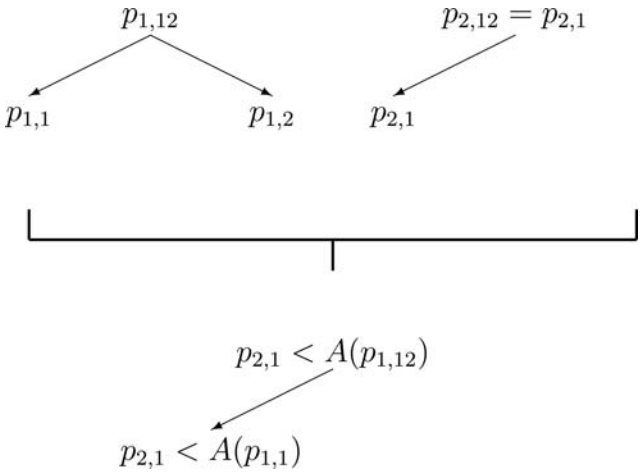

Figure 5 Closure principle for treatment selection at interim.

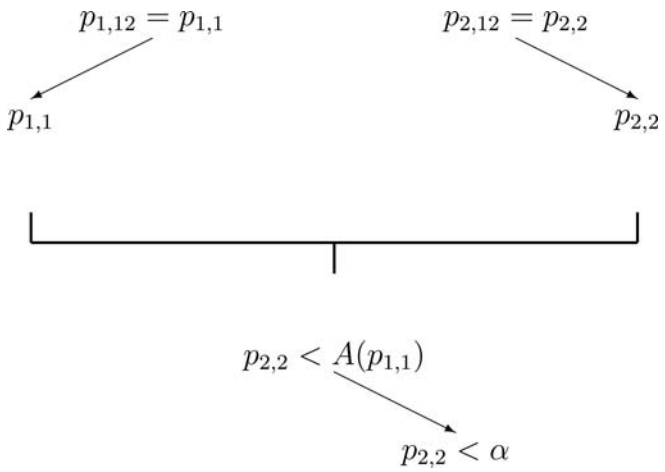

Figure 6 Closure principle for a treatment switch at interim.

5.2 Treatment switch

Assume that a study is planned to investigate the behavior of a single treatment (1, say) in comparison with a control. Assume further that at interim safety problems are detected and it is decided to discontinue the present treatment arm. Instead, it is decided to continue the study with a new treatment (2, say, which could be a lower dose of treatment 1, for example) being investigated at the second stage. Figure 6 depicts the CP associated with the two null hypotheses H_1 and H_2 being tested in the course of the study.

Since at stage 1 no data is available for treatment 2, and vice versa at stage 2 no data is available for treatment 1, the related stagewise p-values for the intersection hypothesis H_{12} are just the corresponding p-values from the elementary hypotheses H_1 and H_2. As concluded from Figure 6, H_2 is rejected if $p_{2,2} < \min\{A(p_{1,1}), \alpha\}$, i.e., if the second stage p-value $p_{2,2}$ associated with treatment 2 is less than α and less than the CEF resulting from the first stage p-value $p_{1,1}$ associated with treatment 1. In practice, the latter condition is not severe, since in most cases of practical relevance the CEF will be larger than α.

Note also that in practice the treatment switch example will probably never be applied as described here. One would rather stop the entire study after interim and start a second (seemingly independent) study investigating treatment 2 at full level α. The above considerations are not only instructive for illustrating ASD, but they also put the current statistical practice into a new perspective.

6 Power Study for Treatment Selection

The power is one of the key operation characteristics needed for designing a clinical trial. We consider in this section the evaluation of power for a two-stage ASD with treatment selection at interim. We compare the power of adaptive combination tests described in the previous sections with those of standard one-stage tests. We restrict the investigation to the comparison of $n = 2$ and $n = 3$ treatments with a control in the homoscedastic normal model with known variance $\sigma^2 = 1$. The sample size s per group is fixed so that the single treatment control comparison for one dose provides a power of $1 - \beta = 0.80$ for a one-sided z-test with $\alpha = 0.025$ at a particular alternative $\delta_A = 1$, so that $s = 2(z_{1-\alpha} + z_{1-\beta})^2/\delta_A^2$, where z_γ is the γ-quantile of the standard normal distribution. One mid-trial interim analysis is considered after $\frac{s}{2}$ observations per treatment group and no early stopping of the study is foreseen. The power values are computed by simulating 100 000 trials.

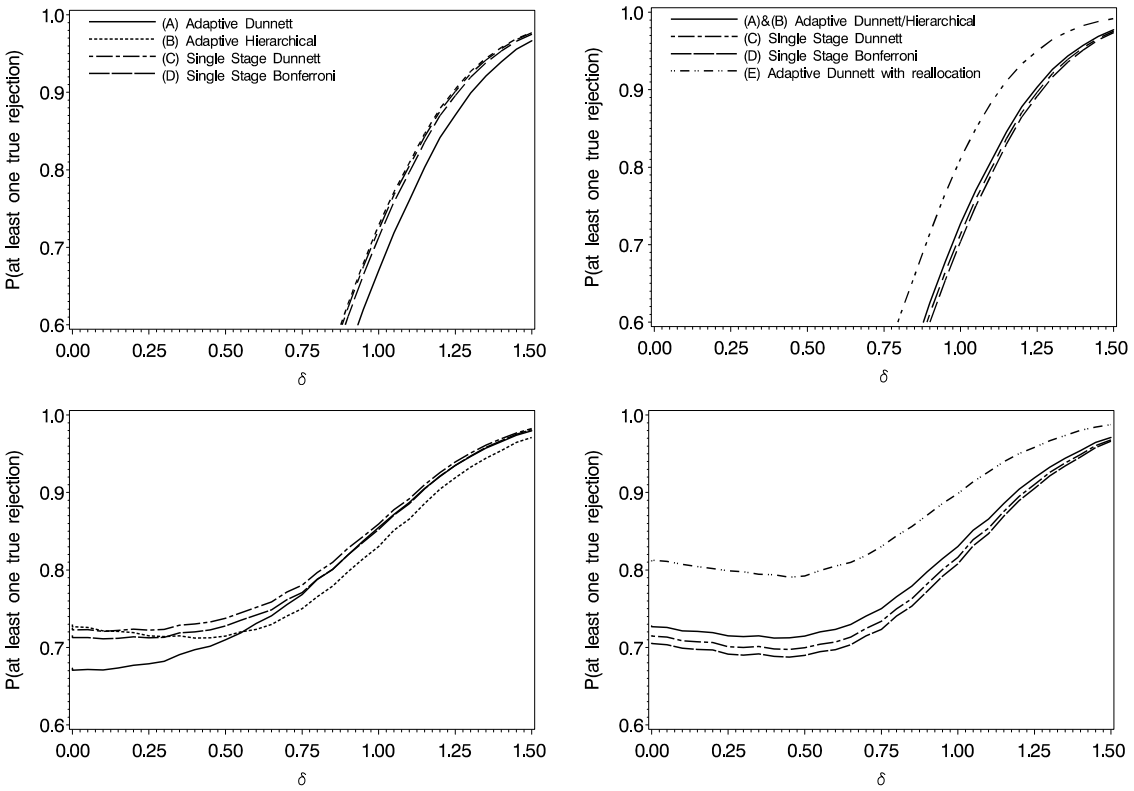

Figure 7 Two active treatments: Power for different tests when continuing with all treatments (left column) or selecting the best treatment at interim (right column). Detailed information about the tests and parameter configuration are given in the text.

We compute the probability to reject correctly at least one of the hypotheses under investigation at the final analysis (so-called minimum power, see Westfall et al., 1999, for a discussion of different power concepts in multiple testing situations). Figure 7 shows the results for comparing $n = 2$ treatments to control and Figure 8 for comparing $n = 3$ treatments to control. We consider the two cases that at interim it is decided (I) to continue with all treatments in the second stage (left column in Figures 7 and 8) and (II) to select the best treatment based on the observed first stage mean value (right columns). The following test procedures are investigated:

(A) *Adaptive Dunnett* Adaptive combination test using the many-to-one test of Dunnett (1955) for the intersection hypotheses at each stage and combining the stagewise p-values using the inverse normal method with equal weights (solid line in Figure 7 and 8). Note that the Dunnett test reduces to the t-test if only one treatment is selected.

(B) *Adaptive Hierarchical* Adaptive combination test using the many-to-one test of Dunnett for the first stage intersection hypothesis. Based on the interim data, the most promising treatment with the largest mean is selected and a fixed sequence test procedure (Westfall and Krishen, 2001) starting with the selected treatment is applied for the second stage intersection hypothesis. The Dunnett test at interim is also to provide the p-value for the intersection hypothesis, irrespective of whether one continues with both treatments. If it is decided to continue only with one treatment in the second stage, this procedure is the same as procedure (A). The inverse normal method with equal weights is used for combining the stagewise p-values (dotted line).

(C) *Single Stage Dunnett* Irrespective of whether the second stage is conducted with one or both treatments, the final analysis uses the multiplicity adjustment according to Dunnett (1955) (dotted-dashed line).

(D) *Single Stage Bonferroni* Irrespective of whether the second stage is conducted with one or both treatments, the final analysis uses the Bonferroni adjustment, that is, the resulting test statistics are compared with the standard normal $(1-\alpha/2)$-quantile. This test is uniformly less powerful than procedure (C) and is included for reference purposes only (dashed line).

Let μ_j denote the mean of treatment group $j = 0, 1, 2(, 3)$, where $j = 0$ denotes the control group. In Figures 7 and 8, each row shows the power values for two different efficacy patterns (i) $\{\mu_0, \mu_1, \mu_2(, \mu_3)\} = \{0, \delta, 0(, 0)\}$ (first row in Figures 7 and 8) and (ii) $\{\mu_0, \mu_1, \mu_2(, \mu_3)\} = \{0, \delta, 1(, 1)\}$ (second row) where $\delta \in [0, 1.5]$ is plotted on the abscissa.

When selecting all treatments for the second stage, procedure (A) has less power than the competing methods (case I, left column), especially than both single stage procedures. This behavior is even more pronounced in the situation of comparing $n = 3$ treatments to control (see Figures 7 and 8, left upper panel). For the efficacy pattern (i), where all other active treatments are ineffective $(\mu_2(= \mu_3) = 0)$ procedure (B) outperforms (A) by constantly more than 5% power for $n = 2$ (resp. 8–9% for $n = 3$) over a wide range of δ. For the efficacy pattern (ii), where $\mu_2(= \mu_3) = 1$, the additional Dunnett test at the second stage in procedure (A) uses the information of all treatments being effective for large values of δ. Consequently, procedure (B) is only more powerful for small values of δ. In case (II) of selecting one treatment at interim the single-stage tests pay too high a price for

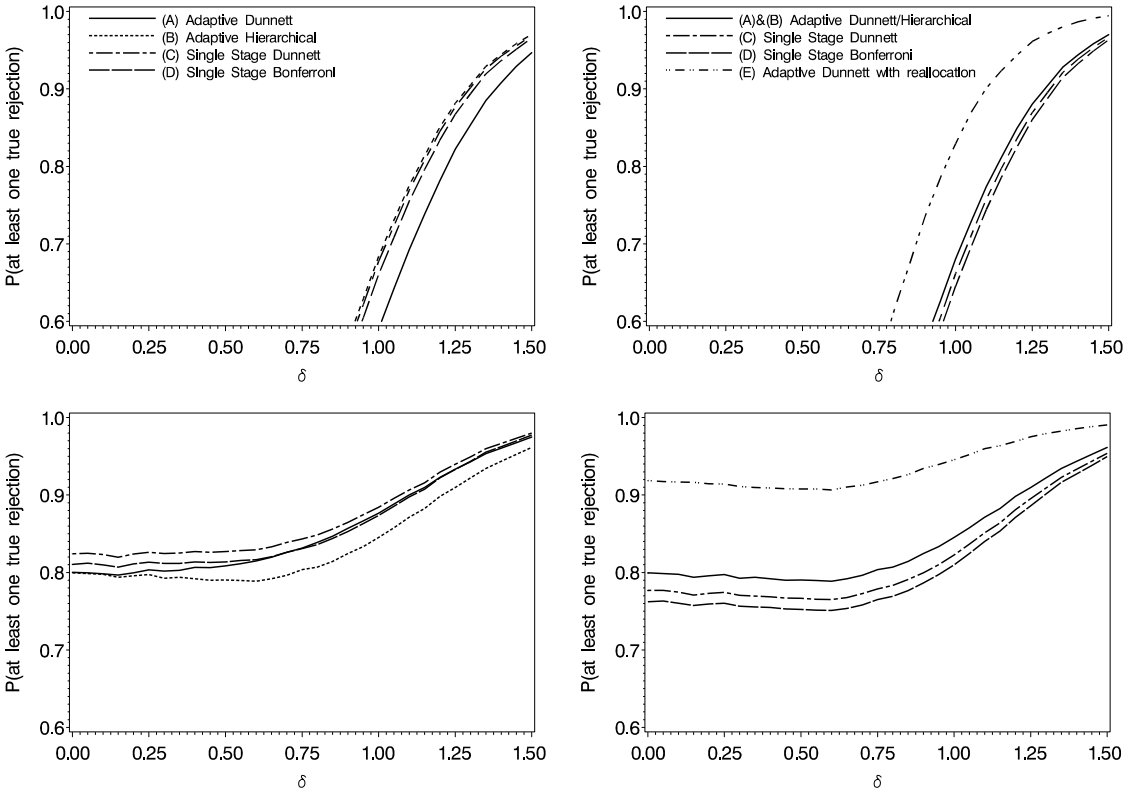

Figure 8 Three active treatments: Power for different tests when continuing with all treatments (left column) or selecting the best treatment at interim (right column). Detailed information about the tests and parameter configuration are given in the text.

multiplicity and thus perform inferior to the adaptive combination tests. As expected, the Bonferroni test is always less powerful than the Dunnett test. As mentioned before, procedures (A) and (B) are identical in this situation and the two curves are lying upon each other in the right panels of Figures 7 and 8.

For the adaptive procedures (A) and (B) a sample size reassessment can be performed in the interim analysis in contrast to the single stage procedures of (C) and (D). Here we investigated a simple sample size reallocation strategy for case (II) by fixing the total sample size of the trial over all treatment groups. For sample size reallocation we distribute the unused number of observations of the dropped treatments evenly over the selected treatments (including the control group). This test procedure is denoted by (E) *Adaptive Dunnett with reallocation* in Figures 7 and 8 (dotted-dotted-dashed line). Again, if it is decided to continue only with one treatment in the second stage, the adaptive procedures of (A) and (B) with sample size reallocation are equal. Note that the power of procedure (E) is substantially increased for all efficacy scenarios considered in comparison with the methods without sample size reallocation. Clearly, a higher increase of power is achieved for the comparison of $n = 3$ treatments to control in contrast to $n = 2$, since more observations can be reallocated. For example, the power for the effect from the planning $\delta_d = 1$ is increased from 68% to 82% for $n = 3$ and efficacy pattern (i).

In practice the decision whether to continue with one, two, ..., or all treatments is mostly taken at interim and thus unknown in the design phase of an adaptive seamless study. The true power curves will thus lie between those presented in the left and the right columns of Figure 7 $(n = 2)$ and Figure 8 $(n = 3)$, respectively.

7 An Approximate Sample Size Comparison

When planning an adaptive design, e.g., for treatment selection, it would be helpful to have a simple rule of thumb for the sample size calculation. In particular we want to derive an approximate relative efficiency of an ASD as compared to a classical independent phase II/III program. Since using single stage tests has similar operation characteristics compared with adaptive designs using a combination function (see Section 6), it can be used as a first approximation for the power and sample size calculations. We consider the case where exactly one active treatment is selected at interim. Let s_{ij} denote the planned sample size for treatment $j = 0, 1, \ldots, n$ at stage $i = 1, 2$. Let $s(\alpha, \beta) = 2(z_{1-\alpha} + z_{1-\beta})^2/\Delta^2$ denote the sample size that provides for a particular design a power of $1 - \beta$ to detect a standardized treatment difference Δ with a level-α test in a two-arm comparison. Assume further that in stage 1 the sample sizes per arm s_{1j} are chosen to be $s(\alpha, \beta_1)$. The aim of this stage (or, equivalently, of a separate phase II study) is primarily to select the right dose(s) to be investigated in the second stage (or, equivalently, an independent phase III study).

In order to achieve a power $1 - \beta_2$ for the confirmatory part in a classical design, the necessary sample size is given by $s(\alpha, \beta_2)$. The total sample size across both stages and all treatment arms is then $s_{\text{class}} = (n + 1)\, s(\alpha, \beta_1) + 2s(\alpha, \beta_2)$. For an ASD we assume that the information of both stages is combined using a single stage test at the end of the second stage. In order to obtain a conservative sample size estimate we use a Bonferroni correction to account for the inherent multiplicity. The total sample size across both stages for the control group and the continued treatment arm is given by $s(\alpha/n, \beta_2)$. The total sample size for an ASD is then approximated by $s_{\text{ASD}} = (n - 1)\, s(\alpha, \beta_1) + 2s(\alpha/n, \beta_2)$.

If we further assume $\beta_1 = \beta_2 = \beta$ and asymptotically normal distributed tests we obtain with $s(\alpha, \beta) = 2(z_{1-\alpha} + z_{1-\beta})^2/\Delta^2$ the relative efficiency

$$r = \frac{s_{\text{ASD}}}{s_{\text{class}}} = \frac{n - 1 + 2\left(\dfrac{z_{1-\alpha/n} + z_{1-\beta}}{z_{1-\alpha} + z_{1-\beta}}\right)^2}{n + 3}.$$

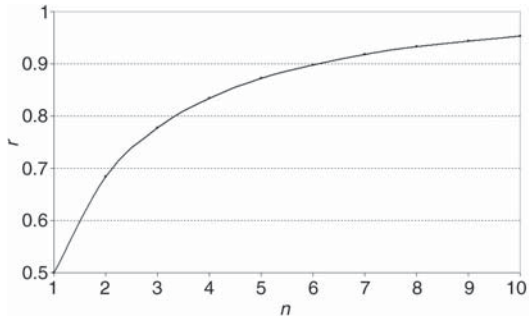

Figure 9 Relative efficiency.

Note that this formula overestimates the true ratio of required sample sizes since (i) in reality more powerful multiple comparison procedures than the Bonferroni procedure may be used, (ii) the above derivation implicitly assumes that the treatment selection at interim is done independent of the observed effect sizes at interim (e.g., via a random selection or pre-determining the selected treatment), whereas in practice a "good" treatment would rather be selected, and (iii) if in fact only one treatment is selected for the second stage then a combination function approach will have an even higher power than the single-stage approach considered here. Figure 9 shows the relative efficiencies for several values of n with $\alpha = 0.025$ and $\beta = 0.1$, indicating that the classical independent phase II/III designs are particularly inefficient when the number n of hypotheses is small.

8 Conclusions

The statistical methodology for adaptive designs developed in the last 15 years allows sufficient flexibility to be applied to the two-stage seamless phase II/III studies discussed here. Extensions to more than two stages are also possible, using the same concept of combining p-values over the different stages. More recently, methods controlling the type I error were developed which do not even require the number of interim analyses to be pre-specified (Müller and Schäfer, 2001). For the confirmatory trial we have in mind here, some of this flexibility may not be made use of. A study where the design is changed at any time during the course of a trial is probably neither convincing to health authorities nor to the medical community – except perhaps in rare cases of revolutionary drugs in indications with great unmet medical need. For adaptive seamless phase II/III studies, more modest adaptations are appropriate such as dose selection, treatment regimen selection (e.g., once daily vs twice daily), selection of a pre-specified subgroup (e.g., restriction of inclusion criteria at interim to patients with severe disease), adaptation of the sample size, and selection of the testing strategy for the second stage. It is likely that statistical methods can be developed for this more narrow range of adaptations that are even more efficient than those currently available.

Acknowledgement We thank Werner Brannath, Paul Gallo, Jeff Maca and Martin Posch for many helpful discussions.

References

Anonymous (2004). "Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products." FDA report from March 2004, available at http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.

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

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

Brannath, W., Posch, M. and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97**, 236–244.

Cui, L., Hung, H. M. J. and Wang, S. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 321–324.

Dunnett, C. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096–1121.

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.

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

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

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**, 655–660.

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

O'Neill, R. T. (2004). A Perspective on Contributions of Biostatistics to the Critical Path Initiative. Presentation given at Workshop on "Model-based drug development – A cornerstone of the FDA's Critical Path Initiative", Basel Biometric Society, December 2004, available at http://www.psycho.unibas.ch/BBS/slides/ONeill2.ppt.zip

Posch, M. and Bauer, P. (2003). Dealing with the unexpected. Proceedings of the ROeS meeting, St. Gallen.

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

Schmidli, H., Bretz, F., Racine, A. and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim. Part II: Applications and Practical Considerations. *Biometrical Journal*, **48**, in press.

Westfall, P., Tobias, R., Rom, D., Wolfinger, R. and Hochberg, Y. (1999). Multiple Comparisons and Multiple Tests Using the SAS System. SAS Institute INC., Cary, NC.

Westfall, P. and Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference* **99**, 25–40.

