PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2007; **6**: 205–216
Published online 2 August 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.302

Multiplicity and flexibility in clinical trials

MAIN PAPER

Werner Brannath*,†, Franz Koenig and Peter Bauer
Medical University of Vienna, Vienna, Austria

Flexible designs offer a large amount of flexibility in clinical trials with control of the type I error rate. This allows the combination of trials from different clinical phases of a drug development process. Such combinations require designs where hypotheses are selected and/or added at interim analysis without knowing the selection rule in advance so that both flexibility and multiplicity issues arise. The paper reviews the basic principles and some of the common methods for reaching flexibility while controlling the family-wise error rate in the strong sense. Flexible designs have been criticized because they may lead to different weights for the patients from the different stages when reassessing sample sizes. Analyzing the data in a conventional way avoids such unequal weighting but may inflate the multiple type I error rate. In cases where the conditional type I error rates of the new design (and conventional analysis) are below the conditional type I error rates of the initial design the conventional analysis may, however, be done without inflating the type I error rate. Focusing on a parallel group design with two treatments and a common control, we use this principle to investigate when we can select one treatment, reassess sample sizes and test the corresponding null hypotheses by the conventional level α z-test without compromising on the multiple type I error rate. Copyright © 2007 John Wiley & Sons, Ltd.

Keywords: ***adaptive design; treatment selection; step-down Dunnett test; hierarchical test; multiple comparison; multiple type I error rate***

1. INTRODUCTION

Multiplicity and flexibility issues are among the major topics in the current discussion on clinical trial designs for pharmaceutical research [1]. In particular, trial designs with the possibility of selecting treatments, endpoints and population subgroups in the course of the trial at interim analyses have recently attracted much interest [2–12] and are controversially discussed by members from academia, industry and regulatory agencies [13–15]. The reason for this interest is

*Correspondence to: Werner Brannath, Medical University of Vienna, Vienna, Austria.
†E-mail: werner.brannath@meduniwien.ac.at

the possibility to combine trials from different clinical phases into a single trial which bears the potential of speeding up the development process of pharmaceutical products. The combination of phase IIb and III trials is the most frequently considered application [6–10, 12, 16], but combinations of other phases e.g. of phase IIa and IIb were considered as well [17]. An important issue in both contexts is control of the *multiple type I error rate* of the combined trial either to achieve approval of the drug by regulatory agencies or to obtain a proof of principle for the new pharmaceutical.

Another important issue is *flexibility* with regard to design modifications. Although pre-specification of the rule for the adaptations at interim analyses is helpful, in practice, investigators often feel uncomfortable to be bounded to fully pre-specified rules such that adaptation at interim becomes an automatic process. The reason for this uncomfort is the complexity in the planning of classical phase IIb or phase III trials. For instance, which treatments to pass over to the next clinical phase typically depends on the accumulated information on the primary and secondary efficacy variables as well as safety endpoints. Moreover, often information from several trials (some running in parallel) is used to guide the planning of later phase trials. Hence, in a combined phase IIb + III trial, for instance, one may expect that at the end of the internal phase IIb part new information from other external trials is available for planning the phase III part of the combined trial (see, e.g. [16]). This leads to a complex interplay between internal and external information in the internal decisions making process which can hardly be foreseen in all details at the beginning of the phase IIb + III trial. A design that guarantees control of the (multiple) type I error after design adaptations that are not fully laid down in advance is often denoted a *flexible* design.

It is important to notice the fundamental difference between *flexibility* and *adaptivity*. Adaptivity is often understood as the possibility to change design features during the ongoing trial based on interim data. Many designs have been suggested which incorporate adaptivity, however, are in no means flexible, since the rule of how the interim data determine the design of the second part of the trial is assumed to be completely specified in advance. In this case, we might better speak of *pre-specified adaptivity*. Recent definitions of adaptive designs [15, 18] seem not to make this clear distinguishing between flexibility and pre-specified adaptivity. Several designs with pre-specified adaptive treatment selection rules have been suggested in the recent years [6, 7]. These designs may be a starting point for a flexible design; however, additional efforts are required to allow the investigator to deviate from the pre-specified rule.

The focus of this paper is on methods that introduce flexibility to clinical trials with multiple testing. Two major approaches will be discussed. One approach is based on flexible closed tests suggested in [2–4]. Here, one utilizes the method of combination tests and the closed testing principle. Another approach is to start with either a fixed sample size design or a design with pre-specified adaptivity and to utilize its conditional type I error rates in order to achieve flexibility [19, 20]. One of the advantages of this approach is that in the case where no deviation from the pre-specified rules is required, one can just complete the trial as pre-specified. This approach has been recently applied to obtain confidence intervals following a flexible group sequential design [21] and to designs with interim selection of treatment arms and multiple testing based on the conditional type I error rates of fixed sample size Dunnett tests [12].

Whatever method we use, flexible designs have the disadvantage that in the final analysis patients of different stages may be weighted differently. Equal weights would be achieved by analyzing the altered design in a conventional way as if the new design were envisaged from the beginning. Such conventional analysis, however, may inflate the type I error rate. Moreover, the type I error rate of the conventional analysis depends on the unknown adaptation rule and hence cannot be quantified. In some cases, however, it is possible to analyze the data in a conventional way, although design features were changed in an unforeseen manner based on unblinded interim data. In [10],

DOI: 10.1002/pst

hierarchical tests are considered where the trial starts with two doses and the lower dose may be dropped at an interim analysis based on the unblinded interim data. The multiple level cannot be increased whatever rule is used for dropping the lower dose. In [10], the case where the higher dose may be dropped due to safety problems is also considered, and it is shown that we can test the lower dose with the conventional level α z-test if the toxicity and efficacy endpoints are non-negatively correlated. Often, however, it is not clear whether the correlation between toxicity and efficacy endpoints is indeed non-negative.

After giving a review on flexible designs with multiple testing, we shall consider in this paper another more general method than in [10] to decide whether a conventional analysis after design adaptations is possible or not. The method is based on conditional type I error rates and requires that conditional type I error rates of the original and altered design can be computed [20]. A conventional analysis is possible whenever the conditional type I error rate of the altered design is below the conditional type I error rate of the pre-specified design, since decreasing the conditional type I error rate cannot inflate the overall type I error rate (but may deflate it). We use this method to explore the case of a parallel group design with two treatments and a common control, and determine for which interim results we can select one treatment, possibly reassess sample sizes and test the corresponding null hypothesis with the conventional level α z-test.

2. REVIEW OF METHODS FOR DEALING WITH FLEXIBILITY AND MULTIPLICITY

2.1. Conditional invariance principle

Flexible designs follow a common principle which we may call *conditional invariance principle*. Assume a flexible trial with two sequential parts (e.g. phase IIb and phase III part) where the design features of the second part are chosen based on the data from the first part (called *interim data* below) as well as external information. We consider here the behavior of the trial under a specific elementary null hypothesis H. Let T_2 denote the statistics for H from the second part. Due to the data-driven choice of the design features, T_2 will in general depend on the interim data. However, we often can transform T_2 in a way that the *conditional* null distribution of T_2 given the interim data and the second-stage design equals a fixed pre-specified null distribution, and hence is *invariant* with respect to the interim data and mid-trial design adaptations. An invariant conditional distribution is typically achieved by transforming T_2 to a p-value q which is uniformly distributed (conditionally on the interim data and second-stage design) under H. Usually, the invariance of the conditional null distribution of q implies that q is statistically independent from the first-stage data. The currently most rigorous verification of this can be found in [22].

Since the common distribution of the interim data and q is known and invariant with respect to the unknown mid-trial adaptation rule, we can specify a level α rejection region in terms of the interim data and q. This gives a test with type I error rate α independently from the adaptation rule. In the case of nuisance parameters, the rejection region would need to be specified in terms of a pivotal first-stage test statistics and the second-stage p-value q.

Note that in the following we do not formally distinguish between random variables and it realizations. It should be clear from the context whether we speak of the observed value of the random variable or the random variable itself before its observation (i.e. its statistical properties).

2.2. Conditional error function approach and related methods

2.2.1. Conditional error function approach.

An invariant rejection region can be implemented via the so-called *conditional error function* approach. With the conditional error function approach,

DOI: 10.1002/pst

we pre-specify a function $0 \leqslant A(\text{interim data}) \leqslant 1$ of the interim data which meets the level condition:

$$E_H A(\text{interim data}) \leqslant \alpha$$

The function A(interim data) is called *conditional error function* and must be pivotal in the sense that the maximum of the expectation under H can be determined. The conditional error function A is used as conditional significance level for the second part of the trial, i.e. we reject the null H at stage 2 iff $q \leqslant A$. Since q is uniformly distributed and independent from A, we get that

$$\begin{aligned} P_H(\text{reject } H) &= P_H[q \leqslant A(\text{interim data})] \\ &\leqslant E_H A(\text{interim data}) \leqslant \alpha \end{aligned}$$

independently from the adaptation rule. Note that the function A(interim data) must be specified in advance in the planning phase of the trial. Note further that an interim rejection and acceptance rule can be implemented by defining $A(\text{interim data}) = 1$ and $A(\text{interim data}) = 0$, respectively. For instance, $A = 1$ implies rejection of H for any q, and hence, H can be rejected at stage 1.

2.2.2. Combination tests.

Proschan and Hunsberger [23] consider conditional error functions $A(p)$ which are non-decreasing functions of a first-stage p-value p for H. Since p is uniform under H, $A(p)$ is pivotal. In [24, 25], *combination tests* are suggested where the first- and second-stage p-values p and q for H are combined by some combination function as in meta-analysis p-values from different trials. Combination tests and conditional error functions $A(p)$ are just two different ways for specifying a rejection region in the two-dimensional (p, q)-plane. Hence, both methods are equivalent [26, 27]. We will see below that when testing the intersection of two or more hypotheses, then A may depend on two or more pivotal first-stage test statistics.

2.2.3. Starting with a conventional design.

In [19, 20], it is suggested to start with a conventional test design at level α (e.g. fixed sample size or group sequential design) and to use its conditional type I error rate as conditional error function at an interim look. Let φ be the test decision function of the initial design, i.e. $\varphi = 1$ if the initial test rejects and $\varphi = 0$ if it accepts. Assume that we perform an interim analysis after the recruitment of a fraction of the initially anticipated total patient number. Assume further that we learn from the interim data and/or external information that we should change design features like, e.g. the sample sizes. We can compute in this case the conditional rejection probability $A(\text{interim data}) = E_H(\varphi|\text{interim data})$ under the null and to choose a new test $\tilde{\varphi}$ for the altered design such that $\tilde{A} = E_H(\tilde{\varphi}|\text{interim data}) \leqslant A(\text{interim data})$. Should we decide from the interim data that no change of the original design is required, then we can just follow the pre-specified design and test φ. Following this principle, the conditional type I error rate of the second part of the trial, i.e. the part following the interim analysis, will never exceed $A \times$ (interim data), whether we alter or stay with the initial design. Hence, the overall type I error rate will be bounded by $E_H A(\text{interim data}) = E_H[E_H(\varphi|\text{interim data})] = E_H\varphi \leqslant \alpha$.

The method of Müller and Schäfer could, in principle, be applied to any design with pre-specified adaptivity to add flexibility to this design. To apply this method we must, however, be able to compute (or at least estimate) the conditional type I error rate of the initial test φ which can become difficult in the presence of nuisance parameters [28, 29].

2.3. Flexible closed tests for testing several hypotheses

Flexible closed tests [2–4] are based on the closed testing principle which is a general and simple method for obtaining multiple tests for k null hypotheses with strong control of the family-wise error rate.

DOI: 10.1002/pst

2.3.1. Closed testing principle.

With the closed testing principle one first defines for all intersection hypotheses $H_J = \bigcap_{j \in J} H_j$ with $J \subseteq I_1 = \{1, \ldots, k\}$ a level α test ψ_J ($\psi_J = 1$ if the test rejects and $\psi_J = 0$ if it accepts). The hypothesis H_i for $i \in \{1, \ldots, k\}$ is rejected with the closed testing principle if the level α tests ψ_J for all H_J with $i \in J \subseteq \{1, \ldots, k\}$ reject, i.e. $\min_{i \in J \subseteq \{1,\ldots,k\}} \psi_J = 1$. Hence, the test decision function for H_i in the closed test is $\phi_i = \min_{i \in J \subseteq \{1,\ldots,k\}} \psi_J$. One can easily see that the closed testing principle controls the family-wise error rate in the strong sense: let J_{true} denote the index set of the true null hypotheses. By the closed testing principle, we must reject the level α test for $H_{J_{\text{true}}}$ in order to reject any of the true null hypotheses $H_j, j \in J_{\text{true}}$. Hence, the probability to reject at least one true H_j is bounded by the level α of the test for $H_{J_{\text{true}}}$.

In the case of two hypotheses H_1 and H_2, for instance, we start with level α tests ψ_1 and ψ_2 for H_1 and H_2 and a level α-test $\psi_{1,2}$ for $H_{1,2} = H_1 \cap H_2$. The first hypothesis H_1, for instance, is rejected with the closed testing principle if $H_{1,2}$ and H_1 are rejected by their level α tests, i.e. $\phi_1 = \min(\psi_{1,2}, \psi_1) = 1$.

2.3.2. Flexible closed tests with conditional error functions.

In a flexible closed test, the local level α tests ψ_J are flexible tests [4]. As described in Section 2.2, the flexible test for H_J can be realized by a conditional error function A_J(interim data) and second-stage p-value q_J such that $\psi_J = 1$ if $q_J \leqslant A_J$(interim data) and $\psi_J = 0$ otherwise. In order to meet the level condition, the conditional error function A_J (interim data) must satisfy $E_{H_J} A_J \leqslant \alpha$. Each conditional error function A_J should be pivotal in the sense that we can determine its maximum expectation (or at least an upper bound) under H_J. Note that all conditional error functions A_J, $J \subseteq \{1, \ldots, k\}$, must be specified in advance.

At the first stage, we can decide to continue with a subset $I_2 \subseteq I_1$ of hypotheses (including the case $I_2 = I_1$ where all hypotheses are selected). We accept all H_i that are not selected, i.e. $i \notin I_2$. To test H_i for $i \in I_2$ with the closed test, we only need to consider the intersection hypotheses H_J where $i \in J$. Hence, we only need to perform flexible tests for all J where $J \cap I_2$ is non-empty. In the flexible test for H_J, we can use the second-stage p-value $q_{J \cap I_2}$ of $H_{J \cap I_2}$ since H_J implies $H_{J \cap I_2}$. In this case, the level α test for H_J (used in the closed test) has the rejection rule $q_{J \cap I_2} \leqslant A_J$.

2.3.3. Example: testing two treatments.

As an example assume that we start with two treatments and a control group in a parallel group design for testing the non-efficacy null hypotheses H_1 and H_2 of the treatments 1 and 2, respectively. The goal of the study is to demonstrate efficacy for at least one treatment. Let A_1, A_2 and $A_{1,2}$ be the conditional error function for H_1, H_2 and $H_{1,2}$, respectively. At the first stage, we can now decide which sample sizes and which tests we want to use at the second stage. Since we use flexible tests, the decision can be based on the data of the first part of the trial and any external information. When going with both treatments into the second part, we could, e.g. use t-tests for H_1 and the Dunnett test for $H_{1,2}$ leading to p-values q_1, q_2 and $q_{1,2}$. Note that q_1, q_2 and $q_{1,2}$ are computed from the data of the second patient cohort only. The flexible closed test then rejects, e.g. the hypothesis H_1 if $q_1 \leqslant A_1$ and $q_{1,2} \leqslant A_{1,2}$. We could, however, also decide to continue only with the first treatment and to terminate the second treatment arm. In this case, q_2 and $q_{1,2}$ cannot be computed due to the missing second treatment arm. Since we are not further interested in rejecting H_2, the dropping of this dose is equal to the acceptance of H_2. Consequently, we need not compute a second-stage p-value q_2. However, we need to do the flexible test for $H_{1,2}$. Here, we can use as second-stage p-value q_1 which is also conservative under $H_{1,2}$ because the intersection hypothesis implies H_1. Deciding after the first stage to use q_1 also for $H_{1,2}$, we would finally reject H_1 iff $q_1 \leqslant A_1$ (test for H_1) and $q_1 \leqslant A_{1,2}$ (test for $H_{1,2}$), i.e. iff $q_1 \leqslant \min(A_1, A_{1,2})$.

DOI: 10.1002/pst

2.3.4. On the choice of the conditional error functions A_J.

In [2–4], all conditional error functions $A_J = A(p_J)$ are a function of the first-stage p-value p_J for H_J whereby the function $A(\cdot)$ is the same for all $J \subseteq I_1$. For instance, $A_J = c/p_J$ (with c the critical value of Fisher's product test) results in using Fisher's combination tests for all H_J [2, 25]. For a control of the multiple type I error rate, it is, however, not required to use a function $A(p_J)$ of the first-stage p-values p_J for A_J. In [12], for instance, the conditional type I error rate of a fixed sample size Dunnett test (assuming a common known variance) is used in a trial where k parallel treatments are compared with a control group with regard to one sided null hypotheses $H_i : \mu_i \leqslant \mu_0$ where μ_i and μ_0 are the means of the treatment $i = 1, \ldots, k$ and control group, respectively. For the elementary hypotheses H_i, the null conditional rejection probability of the z-test is used. Assume, for example, that $k = 2$. Then the conditional type I error rate for H_i from the z-test is

$$A_i(z_i^{(1)}) = P_{H_i}(Z_i \geqslant z_{1-\alpha} | z_i^{(1)}) = 1 - \Phi\left(\frac{\sqrt{n}\, z_{1-\alpha} - \sqrt{n_1} z_i^{(1)}}{\sqrt{n - n_1}}\right) \quad (1)$$

and for the Dunnett test of $H_{1,2}$

$$A_{1,2}(z_1^{(1)}, z_2^{(1)}) = P_{H_{1,2}}\left(\max_{i \in \{1,2\}} Z_i \geqslant d | z_i^{(1)}, i \in \{1,2\}\right) = 1 - \int_{-\infty}^{\infty} \left[\prod_{i=1}^{2} \Phi\left(d_2 \sqrt{\frac{2n}{n - n_1}} - \sqrt{\frac{2n_1}{n - n_1}} z_i^{(1)} + x\right)\right] \phi(x)\, \mathrm{d}x \quad (2)$$

where $\phi(x)$, $\Phi(x)$ and $z_{1-\alpha}$ denote the density, the cumulative distribution function and the $1 - \alpha$ quantile of the standard normal distribution, Z_i is the final test statistic for H_i based on all n observation per group, $z_i^{(1)}$ is the standardized treatment-control difference for treatment i based on the first n_1 observations per group, and d is the Dunnett critical boundary accounting for two treatment-control comparisons [12]. Using the conditional error functions (1) and (2) provide a method for selecting treatments and reassessing sample sizes in a fixed sample size Dunnett test without pre-specifying the selection and reassessment process. As shown in [12], this method has preferable properties in terms of power when compared with other flexible closed tests.

2.3.5. Extensions.

Flexible closed tests can be extended to allow rejection and acceptance of hypotheses at the interim analysis by defining $A_J = 1$ and 0 for specific interim data [2–4]. The hypotheses H_i can be rejected at the first stage iff $A_J = 1$ for all $J \subseteq I_1$ with $i \in J$. If $A_J = 0$ for at least one $J \in i$, then H_i must be accepted. One can further extend the method for the possibility of adding hypotheses which have not been considered at stage one to be considered at stage two [4, 8].

3. FLEXIBLE SWITCHING BETWEEN CLASSICAL TESTS WHEN SELECTING TREATMENTS

The flexible design methodology of the previous section is a general and powerful method to analyze flexible designs where, e.g. sample sizes are reassessed and treatments are selected mid-trial. This method, however, has the disadvantage to lead to unusual tests where, in case of sample size reassessments, the patients from the different stages are weighted unequally [26, 30]. In practice, we may therefore tend to follow a simpler and more intuitive approach. For instance, when selecting one of two treatments at the interim analysis, we may prefer to test the selected treatment by the conventional level α z-test. It is well known that using the conventional z-test in case of treatment selection may lead to an inflation of the multiple type I error rate. However, we can use conditional type I error rates to decide whether such a strategy has the potential to increase the type I error rate or not. Type I error inflations are avoided if the conditional type I error rate of the new design (z-test for selected treatment) is below

DOI: 10.1002/pst

or equal to the conditional type I error rates of the initial design (multiple test for both treatments) since decreasing the conditional type I error rates cannot inflate the overall type I error rate.

More specific, consider the comparison of two treatments with a control in a parallel group design for testing the null hypothesis $H_i : \mu_i \leqslant \mu_0$ against the alternative hypothesis $H_i^A : \mu_i > \mu_0$ (with $i = 1, 2$) in the homoscedastic normal model with known variance σ^2 and unknown means μ_i, $i = 0, 1, 2$. Let n denote the *a priori* planned group sample size. In the initially planned design, an elementary null hypothesis H_i is rejected if the z-test for dose i is rejected at level α and the intersection hypothesis $H_{1,2} = H_1 \cap H_2$ is rejected by a pre-defined level α test. According to the closed testing principle, this controls the multiple type I error rate. Assume that at an interim look we come to the decision to drop a dose and possibly change the total sample size $\tilde{n}$ for the selected dose without specifying this in advance. Let s be the index of the selected dose. If a dose is dropped, then we compare the conditional error $\tilde{A}_s$ of the conventional level α z-test for the selected dose s (where $\tilde{A}_s$ is defined as in (1) with the new total sample size $\tilde{n}$ instead of n and s instead of i) with the conditional errors for H_s and $H_{1,2}$ of the initial planned design. If $\tilde{A}_s \leqslant \min(A_s, A_{1,2})$, then we can do the adaptation and use the conventional z-test because the z-test for H_s then gives a conservative test for both H_s and $H_{1,2}$.

We consider two different initial test designs, the hierarchical test procedure with a pre-defined order and the step-down Dunnett test. We identify for a range of the standardized interim effects and specific second-stage sample sizes whether we can select one treatment and test the corresponding null with the conventional z-test without inflating the conditional type I error rates for H_s and $H_{1,2}$, thereby not inflating the multiple type I error rate. We chose the initial sample size n per group such that the individual treatment-control comparisons have power $1 - \beta = 0.80$ with a one-sided z-test at $\alpha = 0.025$ for the alternative $\delta_A = (\mu_i - \mu_0)/\sigma = 1$, i.e. $n = 2(z_{1-\alpha} + z_{1-\beta})^2/\delta_A^2$. Two different timings of the interim analysis are considered, namely after $n/4$ (Figure 1) and $n/2$ (Figure 2) observations per

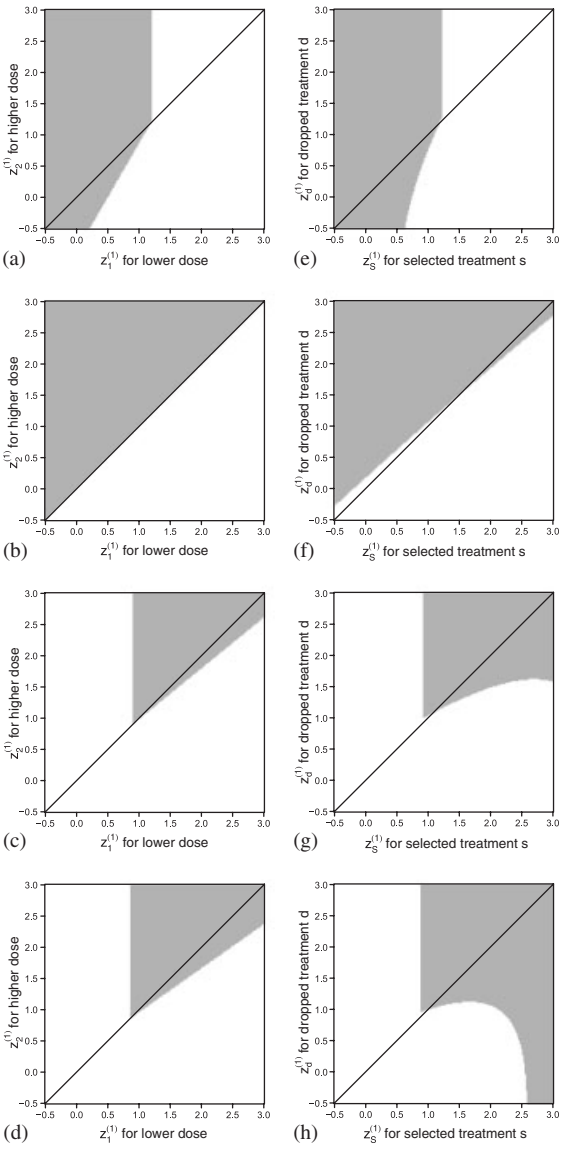

Figure 1. In the gray area, one can switch from the initially planned design to the conventional level α z-test for the selected treatment depending on the standardized interim effects $z_i^{(1)}$ without compromising the type I error rate; interim look after $n_1 = 0.25n$ observations per group; hierarchical test (a)–(d), step-down Dunnett test (e)–(h); sample size reassessments with a new second-stage sample size $\tilde{n}_2 = \tilde{n} - n_1$ for the selected treatment and the control: (a) and (e): $\tilde{n}_2 = 0.25n$ (decrease), (b) and (f): $\tilde{n}_2 = 0.75n$ (as pre-planned), (c) and (g): $\tilde{n}_2 = 1.125n$ (reallocation of the sample size of the dropped treatment), (d) and (h): $\tilde{n}_2 = 1.5n$ (increase).

DOI: 10.1002/pst

Pharmaceutical STATISTICS

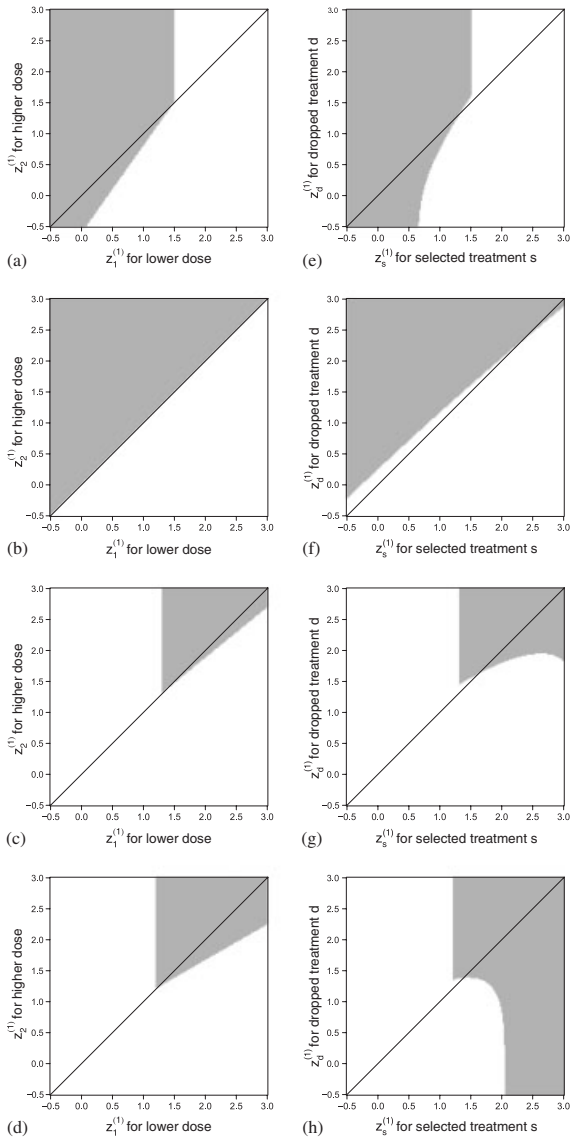

Figure 2. In the gray area, one can switch from the initially planned design to the conventional level α z-test for the selected treatment depending on the standardized interim effects $z_i^{(1)}$ without compromising the type I error rate; interim look after $n_1 = 0.5n$ observations per group; hierarchical test (a)–(d), step-down Dunnett test (e)–(h); sample size reassessments with a new second-stage sample size $\tilde{n}_2 = \tilde{n} - n_1$ for the selected treatment and the control: (a) and (e): $\tilde{n}_2 = 0.25n$ (decrease), (b) and (f): $\tilde{n}_2 = 0.5n$ (as pre-planned), (c) and (g): $\tilde{n}_2 = 0.75n$ (reallocation of the sample size of the dropped treatment), (d) and (h): $\tilde{n}_2 = 1.5n$ (increase).

treatment group. No interim efficacy testing is foreseen. Four strategies to adapt the sample size per group are considered: a decrease, staying with pre-planned samples size, reallocating the sample size of the dropped treatment to the selected one and the control and a major increase.

3.1. Hierarchical test procedure

In the hierarchical test procedure without dose selection, the null hypotheses are ordered in the sequence H_2 followed by H_1. The procedure starts with testing the null hypothesis H_2. The null hypothesis for the lower dose H_1 is rejected if itself and the null hypothesis of the higher dose is rejected both at level α. Hence, the test for the higher dose is identical with the test for the intersection hypothesis $H_{1,2}$. By the closed testing principle, this procedure controls the multiple type I error rate at level α [31]. Note that the conditional error for the intersection $H_{1,2}$ is given by the conditional error for the higher dose, $A_{1,2} = A_2$, because the underlying tests are identical.

The lower dose can always be dropped for any reasons. As long as the sample size for the higher dose remains unchanged, the multiple type I error is controlled since the test for the intersection hypothesis is as pre-specified [10]. In the case of a sample size reassessment for the higher dose and control group, the multiple type I error may be inflated in the same way as in a single dose trial [23]. In [30], interim sample points and sample sizes are identified where there is a zero probability that the conventional z-test accepts but the flexible test (based on the conditional type I error rate of a classical z-test with sample size n) rejects (see also [32]). It can be shown that these are exactly those sample points and sample sizes where the conditional type I error rate with the reassessed sample size $\tilde{n}$ is below the conditional type I error rate with initially planned sample size n. It can been seen from Figure 5 of [30] that an increase in sample size is possible for large treatment effects and a decrease is possible for small interim effects.

We now consider the case where the higher dose is dropped during the trial, e.g. due to safety

DOI: 10.1002/pst

problems. If the lower dose has a smaller first-stage standardized mean than the larger dose, $z_1^{(1)} \leqslant z_2^{(1)}$, then the conditional error function of the usual z-test is smaller for the lower dose than for the larger dose, $A_1 \leqslant A_2 = A_{1,2}$. Hence, without a sample size adjustment we can drop the higher dose and test the lower with the usual z-test, since this gives the conditional error function $\tilde{A}_1 = A_1 \leqslant A_{1,2}$. This case corresponds to the gray area above the diagonal in panel (b) of Figures 1 and 2. In other words, dropping the higher dose and testing directly the lower dose (ignoring the pre-specified hierarchical order) is possible if the interim effect of the lower dose is smaller than the interim effect of the higher dose because in this case we just switch from a less to a more conservative gate-keeping test. Obviously, the gray area in (b) of Figures 1 and 2 does not depend on the timing n_1 of the interim analysis. Panels (a), (c) and (d) in Figures 1 and 2 illustrate the case of sample size reassessment. As seen from panels (a), we can decrease the sample size only for $z_1^{(1)}$ below a specific threshold (vertical edge of the gray area). This threshold depends on the timing and the new total sample size $\tilde{n} < n$. Increasing the sample size and dropping the higher dose (panels (c) and (d)) is only possible with large interim effects for both treatments. Note that in (c) and (d) there is also an area where dropping the higher dose is possible if both interim effects are promising and the higher dose has a smaller interim effect. The reason is that for large interim effects an increase in sample size results in $\tilde{A}_1 < \min(A_1, A_{1,2})$, because in the new test the promising first-stage data are down weighted. A similar phenomenon is observed in (a) with a decrease in sample size and low interim effects.

3.2. Dunnett test

Performing the step-down Dunnett test, $H_{1,2}$ is rejected if $\max(Z_1, Z_2) > d$, where d is the Dunnett critical boundary accounting for two treatment-control comparisons. After a rejection, one can test an elementary null hypothesis H_i with the corresponding z-test at full level α. The conditional error $A_{1,2}$ for the intersection hypothesis $H_{1,2}$ for the Dunnett test is given in (2).

The gray areas in panels (e)–(h) of Figures 1 and 2 show where one can switch from the step-down Dunnett test to the conventional z-test for the selected dose without inflation of the multiple type I error rate. The standardized mean of the selected dose s (whether it is dose 1 or dose 2) is given on the x-axis. Without sample size reassessments (panels (f)) the border of the gray area is not identical to the diagonal so that the conditional error rates from the new and initial design differ along the diagonal. There are two reasons for this: (i) the difference between the critical boundaries of the Dunnett and the elementary level α z-test and (ii) the higher chance to reject with two treatments in the Dunnett test compared with the single test with the selected treatment. The first reason (i) tends to increase the conditional error rate for $H_{1,2}$ when selecting a single treatment (and testing also the intersection with the smaller critical boundary $z_{1-\alpha}$). Reason (ii) tends to decrease the conditional error rate when selecting one treatment since we leave out the chance to reject with the dropped treatment. Which of the two reasons (i) and (ii) dominates the other depends on the first-stage interim effects. If both interim treatment effects are small and similar in size then (i) dominates (ii) and $A_{1,2} < A_s = \tilde{A}_s$. If both interim treatment effects are very promising and similar in size then (ii) dominates (i) and $A_{1,2} > A_s = \tilde{A}_s$.

With sample size reassessments we observe similar tendencies as with the hierarchical test, however, different in magnitude. When the sample size is increased (panels (g) and (h) in Figures 1 and 2), there is now a larger area where it is possible to select the better treatment without the potential of inflating the type I error rate. There is a region in the right upper corner with both interim effects above specific thresholds, where an increase in the sample size of the selected treatment is possible. Moreover, for sufficiently large interim effects of the selected dose, an increase in sample size is possible irrespective of the observed interim effect of the dropped treatment.

DOI: 10.1002/pst

Pharmaceutical STATISTICS

4. NUMERICAL EXAMPLE

We will illustrate the issue discussed in the previous section by a hypothetical example which is similar in spirit to the phase II clinical trial reported in [17]. Assume a clinical trial to investigate the effect of two different doses of eniporide for patients with acute ST-elevation myocardial infarction. The trial is a prospective, multi-center, randomized, double-blind, placebo-controlled design with three parallel groups (placebo, dose 1 and dose 2). The primary endpoint is infarct size measured by the area under the curve of the cumulative release of alpha-hydroxybutyrate dehydrogenase (alpha-HDBH) measured from 0 to 72 h. The initially planned total sample size is $n = 400$ patients per treatment group (which is about the sample size in [17]). It is planned to test efficacy of the two doses by the hierarchical z-test, where we first compare dose 2 to placebo by the z-test at one-sided level $\alpha = 0.025$ and, in case of a significant effect in the dose 2 group, also dose 1 to placebo by the z-test at level 0.025.

Now assume that after $n_1 = 100$ patients per treatment group an un-blinded safety analysis is done. (In [17], the interim analysis was done at about the same sample size.) At this point serious (and unexpected) toxicity problems are observed with the higher dose. Therefore, it is decided to stop recruitment for dose 2 and to continue with dose 1 and placebo only. Furthermore, reshuffling of the unused sample size (300 patients) to the dose 1 and placebo group would increase power and the chance to detect potential toxicity problems with dose 1. Hence, the question arises whether one can drop dose 2, recruit additionally 450 patients (i.e. increase the total to 550 patients) for each of the remaining treatment groups (dose 1 and placebo), and do the usual z-test at the end of the trial without inflating the multiple type I error rate. As we have seen in the previous section, the answer to this questions depends on the effects observed at the interim safety look. To utilize Figure 1 we express the interim effects in terms of the z-scores, although efficacy testing is not envisaged at the interim look. Assume that the z-score for dose 1 is $z_1^{(1)} = 1.1$ and for dose 2 is $z_2^{(1)} = 1.2$. (these are about the interim z-scores of the 100 and 150 mg groups in [17].) We can see from Figure 1(c) that we can indeed reshuffle the unused sample size and compare dose 1 with placebo by the usual z-test at the end of the trial. In detail, the figure indicates that the conditional type I error rate of the z-test for dose 1 with the reassessed sample sizes is below the conditional type I rate error of dose 1 and of dose 2 with the initial sample size, namely, $\tilde{A}_1(z_1^{(1)}) = 0.0496 < \min(A_1(z_1^{(1)}), A_2(z_2^{(1)})) = \min(0.0518, 0.0582)$. Hence, the desired adaptation results in a switch from a less to a more conservative test for both hypotheses H_1 and $H_1 \cap H_2$, thus cannot inflate the multiple type I error rate. (Note that acceptance of H_2 at an interim look can anyhow not inflate the type I error rates.)

5. DISCUSSION

We have reviewed the concepts for achieving fully flexible designs which do not compromise on the type I error rate and allow for multiple testing without a completely pre-specified adaptation rule. We have particularly focused on designs with a single adaptive interim look. In the second part of the paper, we considered the specific adaptation of treatment selection together with sample size reassessment. As an example we took the situation of comparing two treatments with a common control in a parallel group design where a treatment may be dropped mid-trial based on interim data and sample sizes may be reassessed (without fixed binding rules for the treatment selection and sample size reassessment).

Flexible designs have been criticized because, in case of sample size reassessment, they lead to an unequal weighting of observations from different stages in the final test. To follow the philosophy 'one patient one vote', we investigated what happens in the above many-one comparison scenario if we use the conventional level α z-test for the selected treatment. We considered two strategies for the initially planned test procedure, the hierarchical test with *a priori* fixed ordering of

DOI: 10.1002/pst

the hypotheses and the step-down Dunnett test. A hierarchical ordering may be considered in the case where the two treatments represent two different doses of a drug.

With the hierarchical test and without a sample size reassessment, we can always drop the lower dose since this does not change the test for the intersection hypothesis. We further found the intuitive behavior that without sample size reassessment we can select the lower dose if it has a smaller interim effect than the higher dose. Such a decision may be driven by safety problems with the higher dose or by a high interim efficacy of the lower dose, e.g. indicating that the lower dose is already on the dose–response plateau. This may be particularly useful in a phase II context. Increasing sample sizes allows to select the lower dose even if it has the larger interim effect in the case where both interim treatment effects are very promising and of similar magnitude. As expected from single treatment comparison scenarios, increasing sample sizes in case of moderate or small effects may, however, inflate the multiple type I error rate. Such an adaptation is frequently considered as major achievement of flexible designs. For the control of the type I error rates, we then have to use flexible tests breaking the 'one patient one vote' principle.

With the step-down Dunnett test, the situation is more complicated. Without sample size reassessment, there is no big difference in the hierarchical test: nearly always we can select the treatment with the smaller effect but there are exceptions; if both interim effects are very similar in size and not large, then we can select the treatment with the smaller effect only if there is a certain amount of difference to the superior dose at interim, otherwise, we must stay with both treatments or use a flexible test. With a decrease in sample size, the findings are also similar to the hierarchical test; a single treatment can be selected only if a small-to-moderate interim effect is observed for this treatment. With an increase in the sample size, the treatment with the smaller effect can be selected only if both interim treatment effects are large. We can also select the treatment with larger effect if its interim effect is substantial. It is questionable if a strategy to increase the sample size in case of large observed effects is advisable in many situations. More general, following the 'one patient one vote' philosophy and analyzing data arising from fully flexible designs by conventional tests are limited to types of adaptations which seem to be rather the exception than the rule in practice.

The fact that with the Dunnett test we cannot go for all types of interim data with the inferior treatment (even in the case of no sample size reassessment) shows that arguing with a specific selection rule can be misleading: obviously, following the fixed rule where we always go with the inferior treatment (and stay with the pre-planned sample size) cannot inflate the type I error rate. However, utilizing full flexibility, we have the options to sometimes select the inferior and sometimes to go with both treatments. Switching to the (slightly) inferior treatment in the case of two similar but small effects and staying with the initial two treatments otherwise will inflate the multiple type I error rate. The reason is that we switch only in cases where the chance to reject with one treatment is higher (due to the smaller critical boundary) than to reject with both treatments. This once more shows the fundamental difference between flexibility and pre-specified adaptivity.

ACKNOWLEDGEMENTS

We would like to thank the referees for their helpful comments. This work was supported by the Austrian FWF, grant number P-18698-N15.

REFERENCES

1. Gallo P, Krams M. PhRMA working group on adaptive designs: introduction to the full white paper. *Drug Information Journal* 2006; **40**:421–423.
2. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**: 1833–1848.
3. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal* 1999; **41**: 261–277.

4. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 2001; **43**:581–589.
5. Posch M, König F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Adaptive treatment selection. *Statistics in Medicine* 2005; **24**:3697–3714.
6. Kelly PJ, Stallard N, Todd S. An adaptive group sequential design for phase II/III clinical trials that involve treatment selection. *Journal of Biopharmaceutical Statistics* 2005; **15**:641–658.
7. Todd S, Stallard N. A new clinical trial design combining phases II and III: sequential designs with treatment selection and a change of endpoint. *Drug Information Journal* 2005; **39**:109–118.
8. Bretz F, Schmidli H, Koenig F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biometrical Journal* 2006; **48**:623–634.
9. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal* 2006; **48**:635–643.
10. Koenig F, Bauer P, Brannath W. An adaptive hierarchical test procedure for selecting safe and efficient treatments. *Biometrical Journal* 2006; **48**:663–678.
11. Jennison C, Turnbull BW. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: opportunities and limitations. *Biometrical Journal* 2006; **48**:650–655.
12. Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. 2006, submitted.
13. Hung HMJ, O'Neill RT, Wang SJ, Lawrence J. A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal* 2006; **48**:565–573.
14. Koch A. Confirmatory clinical trials with an adaptive design. *Biometrical Journal* 2006; **48**:574–585.
15. CMP. *Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan* (*draft*). The European Agency for the Evaluation of Medicinal Products, London, 2006.
16. Zuber E, Brannath W, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A. Phase II/III seamless adaptive designs with bayesian decision tools for an efficient development of a targeted therapy in oncology. *Technical Report*, *TM 2006-05*, Department of Statistics and Decision Support Systems, University of Vienna.
17. Zeymer U, Suryapranata H, Monassier JP, Opolski G, Davies J, Rasmanis G, Linssen G, Tebbe U, Schroder R, Tiemann R, Machnig T, Neuhaus KL. The Na^+/H^+ exchange inhibitor Eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction. *Journal of the American College of Cardiology* 2001; **38**:1644–1651.
18. Dragalin V. Adaptive designs: terminology and classification. *Drug Information Journal* 2006; **40**:425–435.
19. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **57**:886–891.
20. Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497–2508.
21. Mehta CR, Bauer P, Posch M, Brannath W. Repeated confidence intervals for adaptive group sequential trials. 2006, submitted.
22. Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association* 2002; **97**: 1034–1041.
23. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
24. Bauer P. Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie* 1989; **20**:130–148.
25. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**: 1029–1041.
26. Posch M, Bauer P. Adaptive two stage designs and the conditional error function. *Biometrical Journal* 1999; **41**:689–696.
27. Wassmer G. *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien*. Verlag Alexander Mönch: Köln, Germany, 1999.
28. Posch M, Timmesfeld N, König F, Müller HH. Conditional rejection probabilities of Student's *t*-test and design adaptations. *Biometrical Journal* 2004; **46**:389–403.
29. Timmesfeld N, Schäfer H, Müller HH. Increasing the sample size during clinical trials with *t*-distributed test statistics without inflating the type I error rate. *Statistics in Medicine* 2006; **26**:2449–2464.
30. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Statistics in Medicine* 2003; **23**: 953–969.
31. Bauer P. Multiple testing in clinical trials. *Statistics in Medicine* 1991; **10**:871–889.
32. Chen YHJ, DeMets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine* 2004; **23**: 1023–1038.

DOI: 10.1002/pst