STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1612–1625
Published online 17 September 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3048

Adaptive Dunnett tests for treatment selection

Franz Koenig[1], Werner Brannath[1], Frank Bretz[2] and Martin Posch[1,*,†]

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

SUMMARY

Clinical trials incorporating treatment selection at pre-specified interim analyses allow to integrate two clinical studies into a single, confirmatory study. In an adaptive interim analysis, treatment arms are selected based on interim data as well as external information. The specific selection rule does not need to be pre-specified in advance in order to control the multiple type I error rate. We propose an adaptive Dunnett test procedure based on the conditional error rate of the single-stage Dunnett test. The adaptive procedure uniformly improves the classical Dunnett test, which is shown to be strictly conservative if treatments are dropped at interim. The adaptive Dunnett test is compared in a simulation with the classical Dunnett test as well as with adaptive combination tests based on the closure principle. The method is illustrated with a real-data example. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: adaptive seamless design; flexible designs; conditional error rate; clinical trials, many-to-one comparisons

1. INTRODUCTION

Adaptive seamless designs (ASDs) are becoming increasingly popular in accelerating the drug development process [1–5]. ASDs are used to combine two clinical studies into a single, confirmatory study. In particular, they use accumulating data to decide during the conduct of the study how to modify aspects of the study without undermining the validity and integrity. A standard application is the combination of a clinical phase II study (focusing on treatment selection, for example) with a phase III study (confirmatory testing of the selected treatment).

Consider the classical concept of two separate studies. Suppose that four treatments (different dose levels, for example) are compared with a control in phase II. After finishing the phase II

*Correspondence to: Martin Posch, Section of Medical Statistics, Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Wien, Austria.
†E-mail: Martin.Posch@meduniwien.ac.at, martin.posch@univie.ac.at

Contract/grant sponsor: FWF; contract/grant number: P18698-N15

Received 3 April 2007
Accepted 16 July 2007

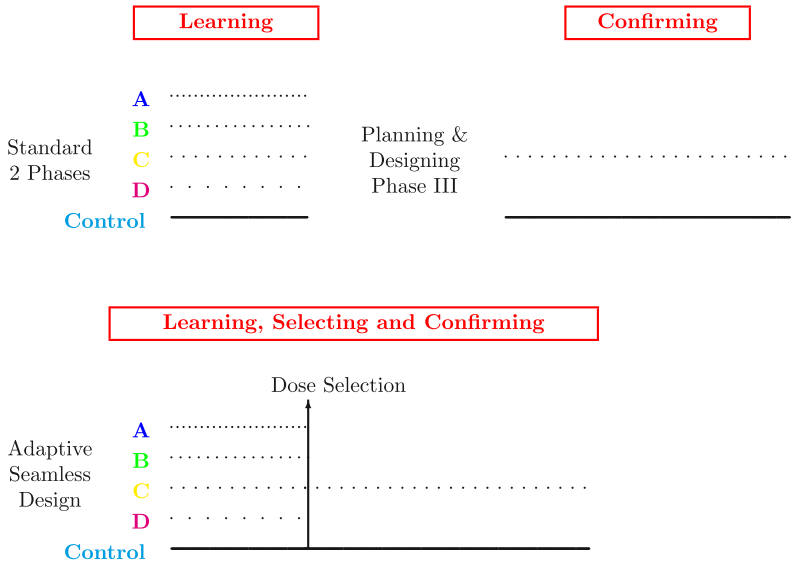

Figure 1. (Top): Classical approach involving standard phase II and phase III studies and (bottom): adaptive seamless design combining phase II and phase III studies.

study, it is decided whether to continue the drug development and which treatment to carry forward to phase III (Figure 1, top). If the decision is taken to carry out the phase III study, previously available information is mainly used to derive the parameter estimates necessary for sample size determination. The phase III study is then run independent of the phase II trials. In particular, the analysis of the confirmatory phase III study does not include the evidence from phase II, thus resulting in a loss of information during the drug development process.

ASDs aim at interweaving such studies by combining them into one single trial conducted in two stages: At the interim (after the first-stage), one treatment, for example, is selected and carried on together with the control arm in the second-stage (Figure 1, bottom). The final comparison of the selected treatment with the control arm includes patients of both stages and is performed such that the multiple type I error rate is controlled at a pre-specified level α. Ideally, ASDs thus reduce the intermediate decision time ('white space') and increase the value of the information by combining the evidence across different phases. Note that in a standard drug development program the white space between phases II and III often takes 6 months or more. This time is essentially reduced to a few weeks for the interim analysis in adaptive trials. ASDs thus require the study team to be better prepared for the decision part and at the same time relieve it from setting up a new study afterwards. As a potential downside, during the period when a decision is taken, recruitment must either be paused or additional patients, who are no longer may be assigned to treatment options pursued during the next stage.

A simple approach to analyse an ASDs involving treatment selection at the interim is to use the conventional Dunnett test [6] for comparing several treatments with a control. The conventional Dunnett test is based on pairwise comparisons of the individual treatments with the control. In the adaptive setting, the final test statistics of the deselected treatments can be set to $-\infty$ and the conventional Dunnett test can then be performed based on these modified test statistics. This

DOI: 10.1002/sim

approach controls the multiple level α, since the missing test statistics are imputed in a strictly conservative way. The type I error rate remains controlled at level α if the step-down Dunnett test based on the closed test procedure [7] is applied. In the following, we refer to this approach as the *classical* (*step-down*) *Dunnett test*. Note, however, that besides dropping treatments no other adaptations are possible with this approach. In addition, it will be shown in this paper that the classical Dunnett test can be uniformly improved.

Bauer and Kieser [8] proposed a new method for the analysis of ASD involving treatment selection at interim. Kieser *et al*. [9] used a similar approach in the context of multiple endpoints. Hommel [10] subsequently formulated a general framework to select and add hypotheses in an adaptive design. Posch *et al*. [1] derived simultaneous confidence intervals and multiplicity-adjusted p-values for ASD with treatment selection. All these approaches are based on applying combination functions [11] within a closed test procedure. At each stage and for each intersection hypothesis of the closed test procedure, one has to adjust for the number of treatments actually being considered. The multiplicity-adjusted stage-wise p-values are then combined using a suitable combination function.

In this paper, we consider a new method based on the closed test principle and the conditional error function approach suggested by Müller and Schäfer [12]. We apply the conventional Dunnett test [6] to each intersection hypothesis. At the interim analysis the conditional error rate of these Dunnett tests is calculated and used for the second-stage tests of the intersection hypotheses. The conditional error rate for a (intersection) hypothesis H is the conditional probability to reject H with the predefined Dunnett test given the interim data assuming that H is true. If no adaptations are performed at the interim, this approach leads to the conventional Dunnett test. In the case of adaptations, special test statistics are derived. This procedure allows further adaptations besides selecting treatments, such as sample size reassessments at interim, while controlling the multiple level α. The results are easily extended to more than two stages.

Accordingly, the paper is organized as follows. In Section 2 we review univariate tests based on the conditional error rate. In Section 3 we present the main methodological results. We give a general formulation of adaptive closed tests based on the conditional error function approach and the closure principle. In Section 4 we explore the different procedures in a simulation study. In Section 5 we apply our method to the data of the Zeymer *et al*. study [13] and conclude with some remarks.

2. ADAPTIVE TESTS BASED ON THE CONDITIONAL ERROR APPROACH FOR A SINGLE NULL HYPOTHESIS

Proschan and Hunsberger [14] proposed the following adaptive two-stage test for a single null hypothesis H at level α. Let X_1 denote the first-stage sample and let $A(X_1)$ denote a measurable function from the first-stage sample space to the unit interval $[0, 1]$ such that

$$E_H(A) \leqslant \alpha \tag{1}$$

The function A is referred to as the *conditional error function*. On the basis of interim data, the sample size and test statistics for the second-stage are planned, resulting in a second-stage p-value q (based only on data of the second-stage). At the end of the study, the null hypothesis H

DOI: 10.1002/sim

is rejected if

$$q \leqslant A(X_1) \tag{2}$$

This procedure controls the type I error rate as long as under H the conditional distribution of the second-stage p-value q, given the first-stage data, is stochastically larger than or equal to the uniform distribution (see Müller and Schäfer [12]). If independent sample units are recruited at the two stages and tests are applied that control the type I error probability for any pre-chosen significance level α, then this will apply. We will refer to such second-stage p-values as *p-clud* p-values [15]. Note that if $A=0$ (early acceptance) or $A=1$ (early rejection) no second-stage needs to be performed for the test decision.

Müller and Schäfer [12] proposed defining the conditional error function $A(X_1)$ *via* a (single or multi-stage) test φ with a pre-fixed group sample size n. Here, $\varphi=1$ denotes rejection and $\varphi=0$ denotes acceptance of the null hypothesis H, respectively. Then the corresponding conditional error function of φ conditioning on the first-stage observations for each treatment group is given by

$$A(X_1) = E_H(\varphi = 1 | X_1)$$

Note that with this choice of the conditional error function the original test φ can be applied if no adaptations are performed. That is, at the interim analysis one has the option to complete the trial as initially planned or to choose any other test for H at level $A(X_1)$ for the second-stage. If adaptations are performed, the null hypothesis H is rejected based on the second-stage p-value q whenever (2) is satisfied.

3. ADAPTIVE TREATMENT SELECTION AND THE CONDITIONAL ERROR FUNCTION

In the following we consider k experimental treatments $T_1, \ldots, T_k$, which are compared with a control T_0 in a parallel group design with n patients per treatment group. Let $H_i : \theta_i = \theta_0$, *versus* $H_i' : \theta_i > \theta_0$, $i \in \mathcal{T}_1 = \{1, \ldots, k\}$ denote the corresponding null and alternative hypotheses, where θ_i denotes the treatment effect parameter of treatment $i = 0, \ldots, k$. Although we concentrate on the one-sided hypotheses-testing problem, the results can be extended to the two-sided case by considering two one-sided tests.

3.1. The closure principle for many-to-one comparisons (without treatment selection)

The closure principle [7] is a general methodology to test multiple hypotheses while controlling the family-wise type I error rate in the strong sense. In the application to treatment–control comparisons, level α tests $\varphi_{\mathcal{S}}$ have to be defined for all intersection hypotheses $H_{\mathcal{S}} = \bigcap_{i \in \mathcal{S}} H_i$, $\mathcal{S} \subseteq \mathcal{T}_1$. An elementary null hypothesis H_j, $j \in \mathcal{T}_1$, is rejected at the multiple level α if all intersection hypotheses $H_{\mathcal{S}}$ with $j \in \mathcal{S} \subseteq \mathcal{T}_1$ are rejected through their level α tests $\varphi_{\mathcal{S}}$. As an example, consider the case of $k=2$ active treatments. Then an elementary hypothesis H_j is rejected at the multiple level α if both the test for the intersection hypothesis $H_1 \cap H_2$ and the test for the elementary hypothesis H_j, $j=1,2$, are significant.

3.2. Many-to-one comparisons and treatment selection

Assume now that after n_1 observations per treatment group an interim analysis is performed, where one or more active treatments can be dropped. Let $\mathcal{T}_2 \subseteq \mathcal{T}_1$ denote the set of the remaining active

DOI: 10.1002/sim

treatments. As long as active treatments are selected, the control group is assumed to be carried on to the second-stage as well. Note that we make no assumptions on the selection procedure at interim and allow the interim decision to depend on the interim data as well as on any external information. Thus we do not require, for example, that the treatment group with the largest observed treatment effect at the interim be carried on to the second-stage. This flexibility of deciding only at the interim on the number of treatment arms to be continued and how to select them is the major advantage of the class of adaptive tests considered in this paper. These methods thus allow a quick and proper reaction during the conduct of a trial to e.g. safety concerns or emerging information external to the study.

If treatments are dropped, the closure principle cannot be applied directly, since some of the $\varphi_{\mathcal{S}}$ may depend on the missing second-stage data from dropped treatments. The adaptive procedure tests all intersection hypotheses H_S such that $\mathcal{S} \subseteq \mathcal{T}_2$, with the originally planned tests $\varphi_{\mathcal{S}}$. If $\mathcal{S} \not\subseteq \mathcal{T}_2$ the original test is modified. First, we compute the conditional error rate $A_{\mathcal{S}}$ of $\varphi_{\mathcal{S}}$ given the interim data. Note that the conditional error rate can be computed also if $\mathcal{S} \not\subseteq \mathcal{T}_2$ since it depends only on the first-stage data. Next, we define a second-stage test for $H_{\mathcal{S}}$ with p-clud p-value $q_{\mathcal{S}}$, which is based on data from selected treatments only, setting $q_{\mathcal{S}} = q_{\mathcal{S} \cap \mathcal{T}_2}$, where $q_{\emptyset} = 1$ and $q_{\mathcal{S} \cap \mathcal{T}_2}$ denotes the p-value for the intersection hypothesis of all hypotheses included in $\mathcal{S} \cap \mathcal{T}_2$. In the final analysis $H_{\mathcal{S}}$ is rejected if $q_{\mathcal{S}} \leqslant A_{\mathcal{S}} = P_{H_{\mathcal{S}}}(\varphi_{\mathcal{S}} = 1 | X_1(\mathcal{S}))$, where $X_1(\mathcal{S})$ denotes the interim data for $H_{\mathcal{S}}$.

Remarks

(i) If no treatments are dropped, then $\mathcal{T}_1 = \mathcal{T}_2$ and all (intersection) hypotheses are tested with the original tests $\varphi_{\mathcal{S}}$ such that the above procedure is identical to the original closed test procedure defined in Section 3.1. (ii) The second-stage tests do not need to be pre-specified in the planning phase and may be chosen in the interim analysis based on first-stage data and/or external information.

Example

Consider again the case of two treatments. In the planning phase, define tests $\varphi_1, \varphi_2, \varphi_{\{1,2\}}$ for the elementary and intersection hypotheses. Assume that at the interim analysis it is decided to continue only with treatment 1. Let q_1 denote the p-value from the respective second-stage test for treatment 1 and let $A_{\{1,2\}}$ denote the conditional error rate for the test $\varphi_{\{1,2\}}$. Then the intersection hypothesis $H_{\{1,2\}}$ is rejected if $q_1 \leqslant A_{\{1,2\}}$. H_1 is rejected at the multiple level α if the intersection hypothesis is rejected and $q_1 \leqslant A_1$, i.e. φ_1 rejects. For H_2 no second-stage data of treatment 2 is available and thus H_2 can be rejected only if the test φ_2 rejects based on the interim data. If both treatments are continued to the second-stage, the originally planned test $\varphi_{\{1,2\}}$ is used for the final test of $H_{\{1,2\}}$. For example, if a Dunnett test has been proposed and no adaptations were made, the *a priori* planned Dunnett test is applied.

3.3. Conditional error of the adaptive Dunnett tests

Consider the one-sided step-down Dunnett test for normally distributed observations with known variance σ^2 and k balanced treatments groups and a control. Let n denote the pre-planned per group sample size and denote the final test statistics for the comparison of treatment i with control by $Z_i = (\bar{X}_i - \bar{X}_0)\sqrt{n/(2\sigma^2)}$, where $\bar{X}_i$ is the mean in treatment group i. We assume that $\bar{X}_i$ are independent across groups, since different patients are independently randomized to the treatment groups and the control. Assume that the experimenter decides to look at the data after

DOI: 10.1002/sim

n_1 observations in each treatment group. The conditional error for the test of $H_{\mathscr{S}}$ for all $\mathscr{S} \subseteq \mathscr{T}_1$ is then given by

$$\begin{aligned} A_{\mathscr{S}} &= P_{H_{\mathscr{S}}}(\varphi_{\mathscr{S}} = 1 | X_1(\mathscr{S})) \\ &= P_{H_{\mathscr{S}}}(\max_{i \in \mathscr{S}} Z_i \geqslant d_s | z_i^{(1)}, i \in \mathscr{S}) \\ &= 1 - \int_{-\infty}^{\infty} \left[\prod_{i \in \mathscr{S}} \Phi\left(d_s \sqrt{\frac{2n}{n - n_1}} - \sqrt{\frac{2n_1}{n - n_1}} z_i^{(1)} + x \right) \right] \phi(x)\, \mathrm{d}x \end{aligned} \tag{3}$$

where $\phi(x)$ and $\Phi(x)$ denote the density and the cumulative distribution function of the standard normal distribution, $z_i^{(1)}$ the standardized treatment–control difference for treatment i based on the first n_1 observations per group, s the number of treatment–control comparisons in $\mathscr{S}$ and d_s the (Dunnett) critical boundary [6] accounting for s treatment–control comparisons in hypothesis $H_{\mathscr{S}}$. Note that d_s is the $1 - \alpha$ equicoordinate quantile from a multivariate normal distribution with variance 1 and covariance $\frac{1}{2}$. For $s = 1$ the d_1 is $(1 - \alpha)$-quantile of the standard normal distribution.

3.4. Proposals for the second-stage p-values $q_{\mathscr{S}}$

As mentioned above, if $\mathscr{S} \subseteq \mathscr{T}_2$, the hypothesis $H_{\mathscr{S}}$ is tested with the originally planned test $\varphi_{\mathscr{S}}$. For all $\mathscr{S} \not\subseteq \mathscr{T}_2$ the intersection hypothesis $H_{\mathscr{S}}$ is tested with a second-stage test for $H_{\mathscr{S} \cap \mathscr{T}_2}$ at the level of the conditional error rate $A_{\mathscr{S}}$. In the following we discuss two proposals of such second-stage tests. In the first, the second-stage test statistics is based on the maximum of standardized mean differences from the second stage only, while in the second proposal the maximum of standardized mean differences from the pooled data of select treatments is used.

3.4.1. Separate second-stage Dunnett tests. In the second-stage, each intersection hypothesis $H_{\mathscr{S}}$, $\mathscr{S} \not\subseteq \mathscr{T}_2$, is tested with a separate Dunnett test accounting for the number of treatment–control comparisons in $\mathscr{S} \cap \mathscr{T}_2$. In this case, a new Dunnett test for $H_{\mathscr{S}}$ (using only the data of the second-stage) is performed at the level of the conditional error function $A_{\mathscr{S}}$. The second-stage Dunnett-adjusted p-value for $H_{\mathscr{S}}$ is given by

$$q_{\mathscr{S}}^* = 1 - \int_{-\infty}^{\infty} \left[\prod_{i \in \mathscr{S} \cap \mathscr{T}_2} \Phi(z_{\mathscr{S} \cap \mathscr{T}_2}^{(2),\max} \sqrt{2} + x) \right] \phi(x)\, \mathrm{d}x \tag{4}$$

where $z_{\mathscr{S} \cap \mathscr{T}_2}^{(2),\max} = \max_{i \in \mathscr{S} \cap \mathscr{T}_2} z_i^{(2)}$, $z_i^{(2)}$ denotes the standardized treatment–control difference for treatment i using only the second-stage data, and $q_S^* = 1$ if $\mathscr{S} \cap \mathscr{T}_2 = \emptyset$.

In principle, one could also test the hypotheses $H_{\mathscr{S}}$ with $\mathscr{S} \subseteq \mathscr{T}_2$ using the adaptive test based on the conditional error rate and the Dunnett-adjusted second-stage p-value (4). Such a procedure still controls the multiple level but has a different rejection region than the pre-planned test $\varphi_{\mathscr{S}}$ and thus leads to different test decisions. As an example, consider the comparison of two treatments with a control. Assume that in the first-stage the test statistics for treatment 1 is small and for treatment 2 is large. For the pre-planned Dunnett test a larger second-stage effect is required for treatment 1 to cross the rejection boundary than for treatment 2. If a separate second-stage Dunnett

DOI: 10.1002/sim

test is applied, for both treatments the same second-stage effect is necessary for a rejection of the intersection hypothesis. In contrast, with the conditional second-stage Dunnett test derived below, the two rejection regions coincide for all $\mathcal{S} \subseteq \mathcal{T}_2$ and the tests are identical.

3.4.2. Conditional second-stage Dunnett tests. As an alternative approach, one can test each $H_{\mathcal{S}}$, $\mathcal{S} \not\subseteq \mathcal{T}_2$, at the second-stage with the overall test statistic $\max_{i \in \mathcal{S} \cap \mathcal{T}_2} Z_i$, where Z_i is defined in Section 3.3. To obtain a second-stage test at the level of the conditional error rate, we choose a critical value $c_{\mathcal{S}}$ such that

$$P_{H_{\mathcal{S}}}\left(\max_{i \in \mathcal{S} \cap \mathcal{T}_2} Z_i \geqslant c_{\mathcal{S}} | z_i^{(1)}, i \in \mathcal{S} \cap \mathcal{T}_2\right) = A_{\mathcal{S}} \tag{5}$$

and reject $H_{\mathcal{S}}$ if $\max_{i \in \mathcal{S} \cap \mathcal{T}_2} Z_i \geqslant c_{\mathcal{S}}$. Equivalently, one can reject $H_{\mathcal{S}}$ in the final analysis if the second-stage p-value

$$\begin{aligned} q_{\mathcal{S}} &= P_{H_{\mathcal{S}}}\left(\max_{i \in \mathcal{S} \cap \mathcal{T}_2} Z_i \geqslant z_{\mathcal{S} \cap T_2}^{\max} | z_i^{(1)}, i \in \mathcal{S} \cap \mathcal{T}_2\right) \\ &= 1 - \int_{-\infty}^{\infty} \left[\prod_{i \in \mathcal{S} \cap \mathcal{T}_2} \Phi\left(z_{\mathcal{S} \cap T_2}^{\max} \sqrt{\frac{2n}{n-n_1}} - \sqrt{\frac{2n_1}{n-n_1}}\, z_i^{(1)} + x\right)\right] \phi(x)\, \mathrm{d}x \end{aligned} \tag{6}$$

falls below the value of the conditional error function $A_{\mathcal{S}}$, where $z_{\mathcal{S} \cap T_2}^{\max}$ denotes the actually observed value of $\max_{i \in \mathcal{S} \cap \mathcal{T}_2} Z_i$ and we set $q_S = 1$ if $\mathcal{S} \cap \mathcal{T}_2 = \emptyset$.

Remarks

(i) This second-stage test has the appealing property that for $\mathcal{S} \subseteq \mathcal{T}_2$ the pre-fixed test $\varphi_{\mathcal{S}}$ rejects $H_{\mathcal{S}}$ for the same sample points as the adaptive test does. This follows from the fact that the adaptive test rejects if $\max_{i \in \mathcal{S} \cap \mathcal{T}_2} Z_i \geqslant c_{\mathcal{S}}$ and in this case $c_{\mathcal{S}} = d_s$, as seen from (3) and (5).
(ii) By definition the p-value $q_{\mathcal{S}}$ in (6) is p-clud: $q_{\mathcal{S}}$ is defined as the conditional probability under the null to obtain in an independent experiment a larger test statistics than the observed one.
(iii) If only one treatment is selected, say treatment i, both, the second-stage p-values (6) and (4) are identical to the p-value of the z-test based on the second-stage data $1 - \Phi(z_i^{(2)})$. This holds since

$$z_{\mathcal{S} \cap T_2}^{(2),\max} = z_i^{(2)}, \quad z_{\mathcal{S} \cap T_2}^{\max} = z_i \quad \text{and} \quad z_i^{(2)} \sqrt{2} = z_i \sqrt{\frac{2n}{n-n_1}} - \sqrt{\frac{2n_1}{n-n_1}}\, z_i^{(1)}$$

In the following we show that the procedure based on the conditional second-stage Dunnett test is uniformly more powerful than the classical Dunnett test with treatment selection. As described in the Introduction, the classical Dunnett test rejects an intersection hypothesis $H_{\mathcal{S}}$ if

$$\max_{i \in \mathcal{S} \cap \mathcal{T}_2} Z_i \geqslant d_s \tag{7}$$

where d_s denotes the Dunnett boundary accounting for s comparisons. To show the uniform improvement, we rewrite the rejection region of the classical Dunnett test using its conditional error function. By a similar argument as in Remark (i) above, the classical Dunnett test rejects $H_{\mathcal{S}}$ if

DOI: 10.1002/sim

and only if $q_{\mathscr{S}} \leqslant P_{H_{\mathscr{S}}}(\max_{i\in\mathscr{S}\cap\mathscr{T}_2} Z_i \geqslant d_s | X_1(\mathscr{S}))$, where $q_{\mathscr{S}}$ is defined in (6). In contrast, the adaptive test with the conditional second-stage Dunnett p-values rejects if $q_{\mathscr{S}} \leqslant P_{H_{\mathscr{S}}}(\max_{i\in\mathscr{S}} Z_i \geqslant d_s | X_1(\mathscr{S}))$. Since $\max_{i\in\mathscr{S}\cap\mathscr{T}_2} Z_i \leqslant \max_{i\in\mathscr{S}} Z_i$, uniform improvement follows.

3.5. Other design modifications

So far we illustrated the adaptive Dunnett test when the selection of treatments is performed at a pre-fixed interim analysis after n_1 observations in each treatment group. The conditional error principle, however, allows for more flexibility in an ongoing clinical trial. In addition to dropped treatments in the interim analysis is the second-stage sample sizes may be adapted, for example, by distributing the sample size foreseen for dropped treatments to the continued treatments, or conditional power arguments [16]. Further modifications would be the addition of new treatment arms in an ongoing trial or the inclusion of unplanned interim analyses with associated design modifications. If in the interim analysis further design modifications are performed, adaptive tests based on the conditional error rates have to be performed for all (intersection) hypotheses $H_{\mathscr{S}}$ (including those where $\mathscr{S} \subset \mathscr{T}_2$) .

3.6. Multiplicity-adjusted overall p-values

To derive a multiplicity-adjusted p-value for all elementary hypotheses $H_1, \ldots, H_k$, we first need to define unadjusted p-values $p_{\mathscr{S}}$ for all (intersection) hypotheses. To this end, let $A^{\alpha'}_{\mathscr{S}}$ denote the conditional error rate (3) of a (single-stage) Dunnett test for $H_{\mathscr{S}}$ at level α', where $\alpha' \in [0, 1]$. Then $p_{\mathscr{S}}$ is defined as the smallest α' such that $q_{\mathscr{S}} \leqslant A^{\alpha'}_{\mathscr{S}}$, where $q_{\mathscr{S}}$ is defined by either (4) or (6) and $p_{\mathscr{S}} = 1$ if $\mathscr{S} \cap \mathscr{T}_2 = \emptyset$. Now, the multiplicity-adjusted p-value for hypothesis i is given by

$$p_i^{\text{adj}} = \max_{\mathscr{S}\subseteq\mathscr{T}_1, i\in\mathscr{S}} p_{\mathscr{S}}$$

If no adaptations are performed and the conditional second-stage Dunnett test is applied at the second-stage, this p-value coincides with the multiplicity-adjusted p-value of the step-down Dunnett test.

4. A SIMULATION STUDY

We illustrate the flexibility of the adaptive Dunnett test and compare its power characteristics with the classical Dunnett test with treatment selection as well as adaptive combination tests [1, 4, 8, 10]. Consider the comparison of $k = 2$ treatments with a control in the homoscedastic normal model with known variance σ^2. Let n denote the *a priori* planned group sample size. For simplicity, we choose n such that the individual treatment–control comparisons have a power of $1 - \beta = 0.80$ for a one-sided z-test with $\alpha = 0.025$ at a particular alternative $\delta_A = 0.5$, i.e. $n = 2(z_{1-\alpha} + z_{1-\beta})^2/\delta_A^2$, where z_c denotes the $c\,100$ per cent-quantile of the standard normal distribution and δ_A is the standardized treatment effect. For example, a common standard deviation $\sigma = 8$ and a treatment–control difference of 4 leads to $\delta_A = \frac{4}{8} = 0.5$ and a (rounded) total sample size of $n = 63$ per treatment. One mid-trial interim analysis is considered after $n/2$ observations per treatment group and no early stopping of the study is foreseen. The power values are computed *via* simulating 100 000 trials using SAS-IML.

DOI: 10.1002/sim

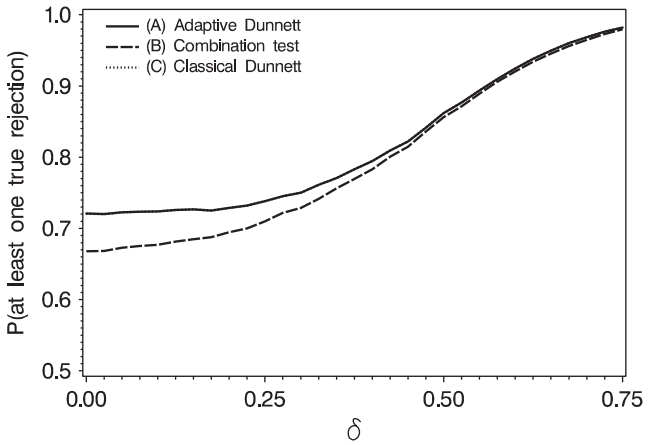

Figure 2. Power for selection rule (I): continue with both treatments in the second stage. The power is defined as the probability to reject correctly at least one of the hypotheses at the final analysis for the efficacy profile $(\mu_0/\sigma, \mu_1/\sigma, \mu_2/\sigma) = (0, \delta, 0.5)$, where $\delta \in [0, 0.75]$ is plotted on the abscissa.

We compute the probability of rejecting correctly at least one of the hypotheses under investigation at the final analysis (so-called minimum power, see [17] for a discussion of different power concepts in the context of multiple testing). We consider the following decision rules to be adopted in the interim analysis:

(I) Continue with both treatments in the second-stage (Figure 2).
(II) Select the better treatment based on the observed first-stage mean values (Figure 3).
(III) Continue with both treatments with probability $p_{\text{both}} = 0.5$ and continue with only one treatment with probability $1 - p_{\text{both}}$. Given that only one treatment is selected, the conditional probability to choose the treatment with the larger observed interim mean is $q_{\text{best}} = 0.5$ (Figure 4).

The motivation for rule (III) is to reflect the decision processes in clinical drug development more realistically. In practice, one often does not know at the beginning of the study, how many and in which way treatments will be selected at the interim. Considerations other than the observed efficacy results may influence this decision. For example, safety concerns may arise at the interim and suggest continuing the treatment having the smaller observed interim mean value. Thus, the pair $(p_{\text{both}}, q_{\text{best}})$ can be used to investigate realistic scenarios at the design stage of a clinical trial. Note that $p_{\text{both}} = 1$ is equivalent to rule (I) and $(p_{\text{both}}, q_{\text{best}}) = (0, 1)$ is equivalent to rule (II).

The following test procedures are investigated:

(A) Adaptive Dunnett test (solid line). Note that in the case of two treatments the separate second-stage Dunnett test defined in Section 3.4.1 is equivalent to the conditional second-stage Dunnett test (see Remark (iii) in Section 3.4.2).
(B) Adaptive combination test using Dunnett-adjusted p-values for the intersection hypotheses at each stage and combining the stage-wise p-values using the inverse normal method with equal weights (dashed line in Figures 2–4). For a detailed description of adaptive combination tests see [4, 8, 10]. Note that the Dunnett test for the second-stage reduces to the z-test if only one treatment is selected.

DOI: 10.1002/sim

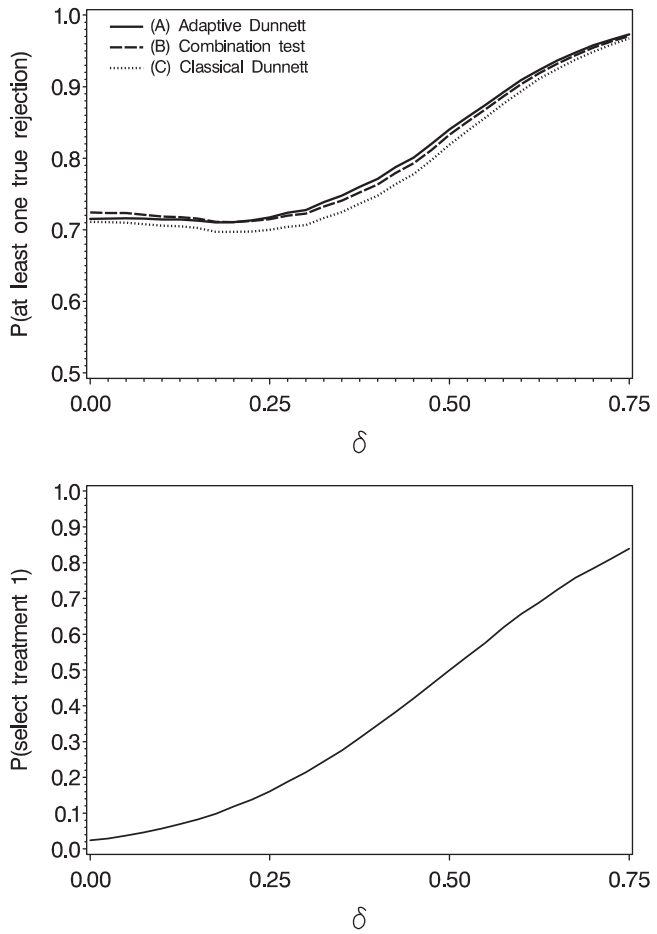

Figure 3. Selection rule (II): select the better treatment based on the observed first-stage mean values. Upper panel: the power is defined as the probability to reject correctly the hypotheses of the selected treatment at the final analysis for the efficacy profile $(\mu_0/\sigma, \mu_1/\sigma, \mu_2/\sigma) = (0, \delta, 0.5)$ where $\delta \in [0, 0.75]$ is plotted on the abscissa. Lower panel: the probability to select treatment 1 at the interim.

(C) Classical Dunnett test with treatment selection using Dunnett critical boundaries accounting for $k = 2$ comparisons, as defined in Section 1 (dotted line).

Let μ_i denote the mean of treatment group $i = 0, 1, 2$, where $i = 0$ denotes the control group. Each plot shows the power values for the efficacy profile $(\mu_0/\sigma, \mu_1/\sigma, \mu_2/\sigma) = (0, \delta, 0.5)$, where $\delta \in [0, 0.75]$ is plotted on the abscissa.

As shown in Section 3.4.2, procedure (A) is uniformly more powerful than (C). The amount of improvement depends on the scenario. Selection rule (I): Procedures (A) and (C) are identical in this situation since all treatments are selected and no adaptations are performed in the interim analysis (see Figure 2). Method (B) has lower power compared with (A) and (C) for small values of δ. Selection rule (II): When selecting always the treatment with the best interim result, the

DOI: 10.1002/sim

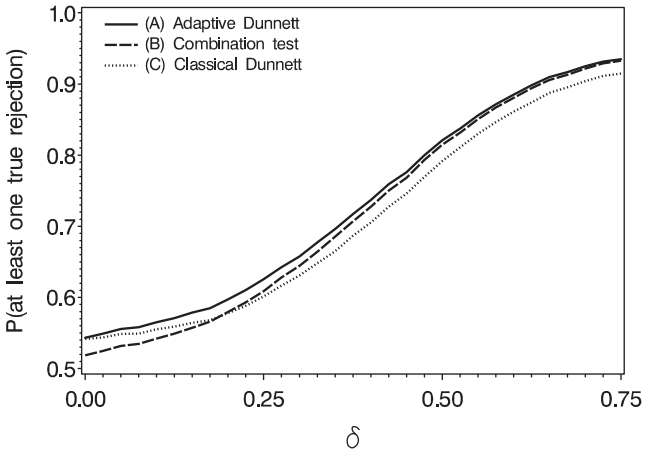

Figure 4. Power for selection rule (III): continue with both active treatment with probability $p_{\text{both}} = 0.5$ and continue with only active treatments with probability $1 - p_{\text{both}}$. If only one treatment is selected, the treatment with the larger observed mean at interim is selected with probability $q_{\text{best}} = 0.5$. The power is the probability to reject correctly at least one of the hypotheses at the final analysis for the efficacy profile $(\mu_0/\sigma, \mu_1/\sigma, \mu_2/\sigma) = (0, \delta, 0.5)$ where $\delta \in [0, 0.75]$ is plotted on the abscissa.

power curves are not monotone for the efficacy profiles investigated (see Figure 3, upper panel). When increasing the effect δ, the power first decreases and then increases. The treatment selection probabilities in Figure 3 (lower panel) clarify this behavior. For values around $\delta = 0$ the more efficient treatment $i = 2$ (with $\mu_2/\sigma = 0.5$) is almost always selected in the interim analysis. With increasing values of δ, the probability of selecting treatment one (which has a lower efficacy) increases. This reduces the power. When $\delta = 0.5$, both active treatments have the same means and the selection probability is exactly 50 per cent. If only one treatment is effective ($\delta = 0$) the combination test (B) has the highest power. If both treatments are effective the adaptive Dunnett test (A) is superior. Both (A) and (B) are superior to (C). Selection rule (III): The adaptive Dunnett procedure (A) has, for all values of δ, the highest power (Figure 4) compared with procedures (B) and (C). If only one treatment shows an effect, the combination test procedure (B) is inferior to the classical Dunnett test with treatment selection.

In summary, the adaptive Dunnett test has in most of the considered scenarios a higher power than the combination test approach and is by construction more powerful than the classical Dunnett test.

5. APPLICATION

Zeymer *et al*. [13] conducted an international, prospective, randomized, double-blind, placebo-controlled phase II dose-finding study applying a two-stage adaptive design. The primary efficacy endpoint was infarct size measured by the cumulative release of α-HDBH within 72 h after administration of the drug (area under the curve, α-HDBH AUC). The trial started with four dose levels (50, 100, 150 and 200 mg eniporiode) and a placebo group. The interim look leads to a dropping of the highest and the lowest dose group based on efficacy and safety arguments. For the

DOI: 10.1002/sim

remaining dose groups a sample size reassessment was performed based on conditional power arguments. Fisher's combination function was used to combine the adjusted p-values for all intersection hypotheses of a first-stage trend test and individual dose–control comparisons at the second-stage. The trial did not succeed in showing that the drug is superior to placebo at any of the investigated dose levels.

We re-analyze the data with the adaptive Dunnett test at multiple level $\alpha = 0.025$, assuming an *a priori* total sample size of $n = 278$ per treatment group. For simplicity, we assume that the interim analysis was conducted after $n_1 = 88$ observations in each treatment group (the actual first-stage sample sizes were very close to this number). For the second-stage we assume that the sample size associated with the dropped treatments is equally distributed between the two selected treatments (100 and 150 mg) and the control. This leads to a second-stage sample size of $\tilde{n}_2 = 320$ and a total per group sample size of 408 for the selected treatment groups and the control. (The actual second-stage sample sizes were again very close to this number.) We assume for simplicity that the common standard deviation is equal to the pooled estimate from the placebo and the treatment groups, that is $\sigma = 26$.

In the first-stage, the means (45.3, 40.2, 33.9, 43.9) were observed for the treatments, groups (50, 100, 150 and 200 mg eniporiode). In the placebo group the first-stage mean was 44.2. Thus, the largest effect (i.e. decrease of α-HDBH) was observed for the 150 mg dose. The conditional error rate of the adaptive Dunnett test for the global intersection hypothesis is $A_{\{1,2,3,4\}} = 0.128$. In the second stage, means of (43.0, 41.5) were observed for (100, 150 mg) and 41.2 for the placebo group. The conditional second-stage Dunnett p-value for the global intersection hypothesis is 0.60. Thus, the global intersection hypothesis cannot be rejected. Consequently, none of the elementary hypotheses can be rejected at the multiple level.

6. DISCUSSION

In this paper, we focused on interim treatment selection where statistical testing is performed only in the final analysis. No boundaries for early rejection or stopping for futility were considered. In principle, the presented method can be extended to include binding early stopping boundaries. The conditional error can also be computed for group sequential versions of the Dunnett test [18]. Another generalization would be to consider different sample sizes per-group. Let n denote the pre-planned per-group sample size in the control group and let $r_i n$ denote the pre-planned sample sizes in the treatment groups. Assume that an interim analysis is performed after observing n_1 patients in the control group and $r_i n_1$ patients in the treatment groups. Then the conditional error rate for the test of $H_{\mathcal{S}}$ for all $\mathcal{S} \subseteq \mathcal{T}_1$ is given by

$$A_{\mathcal{S}} = 1 - \int_{-\infty}^{\infty} \left[\prod_{i \in \mathcal{S}} \Phi\left(d_s \sqrt{\frac{(1+r_i)n}{n-n_1}} - \sqrt{\frac{(1+r_i)\,n_1}{n-n_1}}\, z_i^{(1)} + \sqrt{r_i}\, x \right) \right] \phi(x)\, \mathrm{d}x \tag{8}$$

Note that here d_s is the critical value for the unbalanced Dunnett test. The assumption that the allocation ratios in the interim analysis and the pre-planned final analysis are the same is essential. Otherwise the conditional error rate would depend on the unknown common mean. This, however, is not specific for the Dunnett test but holds also for two-arm studies comparing a single treatment with a control group.

DOI: 10.1002/sim

If nuisance parameters are present, the conditional error typically depends on them and approximations have to be applied [19]. The assumption of known variability still applies asymptotically, especially if a common variance is estimated from all treatment groups. The assumption of normal outcome variables can be considered as an asymptotic approximation for various distributional scenarios, which is commonly used for sequential decision methods.

The general formulation of adaptive closed tests based on the conditional error function approach and the closure principle in Section 3 can easily be applied to other testing procedures as e.g. hierarchical testing procedures [20], the Tukey test or the Hochberg test. Also, the combination test approach of Bauer and Kieser [8] and Hommel and Kropf [21] is covered by the procedure. Here the tests $\varphi_{\mathscr{J}}$ are combination tests combining multiplicity-adjusted p-values. Then, the test based on the conditional error function of $\varphi_{\mathscr{J}}$ is identical to the combination test (compare [22]).

If treatments are dropped in an interim analysis, the new adaptive Dunnett test (with the conditional second-stage Dunnett test) is a uniform improvement over the classical step-down Dunnett test. The conditional error of the final individual treatment–control comparison is not changed by the selection procedure. Therefore, the tests for individual treatment–control comparisons are identical to those in the classical Dunnett test. The improvement of the adaptive Dunnett test arises from the intersection tests, which are required according to the closure principle before addressing the elementary null hypotheses. If a treatment is dropped, the conditional error rate of the classical Dunnett test is strictly smaller than that of the adaptive Dunnett test. The latter borrows strength for the intersection tests from the data collected for the dropped treatment at the first-stage. Applying the classical Dunnett test, data from the dropped treatments are disregarded for the multiple test decisions, since the corresponding test statistics are set to $-\infty$. SAS-macros for the computation of conditional error rates and second-stage p-values can be obtained from the authors.

The new procedure also allows for sample size reassessment, e.g. to reallocate the designated sample size of dropped treatment arms. This leads to an improvement in power for the remaining comparisons. If no adaptations are performed and no treatments are dropped, the adaptive Dunnett test is identical to the classical step-down Dunnett test.

Another appealing feature of the procedure is that the selection rules do not have to be specified in advance and different decision tools (e.g. Bayesian methods) can be applied. As outlined in Section 3.5, even the timing of the interim analyses may be specified in a data-dependent way and need not be pre-planned. The only condition is that the conditional error for the pre-planned Dunnett test can be calculated. However, to maintain the integrity in confirmatory trials, it is recommended to specify possible adaptations and selection scenarios in the protocol at the study beginning. Nevertheless, even in a well-planned clinical trial, unanticipated, critical safety issues may arise, asking for the dropping of a treatment arm. The adaptive Dunnett test can be an important tool to deal with such unexpected situations.

ACKNOWLEDGEMENTS

We thank Ekkehard Glimm and the referees for their helpful remarks. The authors gratefully acknowledge the continuous support of Peter Bauer.

REFERENCES

1. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 2005; **24**:3697–3714.

DOI: 10.1002/sim

2. Gallo P. Operational challenges in adaptive design implementation. *Pharmaceutical Statistics* 2006; **5**:119–124.
3. Hung H, Wang S, O'Neill R. Methodological issues with adaptation of clinical trial design. *Pharmaceutical Statistics* 2006; **5**:99–107.
4. Bretz F, Schmidli H, Koenig F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts (with discussion). *Biometrical Journal* 2006; **48**:623–634.
5. 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.
6. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**:1096–1121.
7. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
8. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
9. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal* 1999; **41**:261–277.
10. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 2001; **43**:581–589.
11. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041.
12. 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.
13. 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.
14. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
15. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical Association* 2002; **97**:236–244.
16. Bauer P, Koenig F. The reassessment of trial perspectives from interim data—a critical view. *Statistics in Medicine* 2006; **25**:23–36.
17. Westfall P, Tobias R, Rom D, Wolfinger R, Hochberg Y. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Institute Inc.: Cary, NC, 1999.
18. Follmann D, Proschan M, NLGeller, Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* 1994; **50**:337–349.
19. 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.
20. Koenig F, Bauer P, Brannath W. An adaptive hierarchical test procedure after selecting safe and efficient treatments (with discussion). *Biometrical Journal* 2006; **48**:663–678.
21. Hommel G, Kropf S. Clinical trials with an adaptive choice of hypotheses. *Drug Information Journal* 2001; **35**:1423–1429.
22. Posch M, Bauer P. Adaptive two stage designs and the conditional error function. *Biometrical Journal* 1999; **41**:689–696.

DOI: 10.1002/sim