Adaptive Designs with Discrete Test Statistics and Consideration of Overrunning*

R. Schmidt1; B. Burkhardt2; A. Faldum3
1Institute of Biostatistics and Clinical Research, University of Muenster, Muenster, Germany; 2University Children’s Hospital Muenster, Pediatric Hematology and Oncology, Muenster, Germany

Keywords
Adaptive design, discrete test statistic, interim analysis, interim patients, multiple endpoints, overrunning

Summary
Background: Interim analyses are used in clinical trials in order to enable early decisions for medical, ethical, and economic reasons. However, it appears unfeasible to stop a trial during such an interim analysis. New patients will thus enter the trial while the interim analysis is ongoing. Moreover, depending on the event kinetics of the specific disease, the trial design, and the corresponding endpoints, some patients might still be unevaluable at the interim analysis due to not yet completed follow-up. Occurrence of these types of patients is characteristic for sequentially analyzed trials. Such patients are referred to as interim patients. In trials with multiple primary endpoints, another type of interim patients occurs. If some but not all null hypotheses can be rejected at the interim analysis, the trial might be continued to a second stage in order to answer the remaining questions. These second stage patients, however, provide new data to all trial questions including the already rejected ones and thus formally act as interim patients regarding the already rejected null hypotheses. Although all kinds of interim patients are not part of the interim analysis, the data collected on those patients have to be sent to the office of regulatory affairs and will be analyzed. If a smaller or contrasting treatment effect is observed in interim patients, this might lead to a withdrawal of an earlier superiority proof.

Objectives: Presently, interim patients and their data are usually not considered in the confirmatory test. We offer a strategy to deal with interim patients in sequentially analyzed trials with discrete test statistics. The method covers sequentially analyzed single- and multi-arm trials with one or multiple primary endpoints.

Methods: When planning adaptive designs, it is common practice to assume that the stage-wise p-values are independent and standard uniformly distributed under the null hypothesis. In the context of discrete test statistics, this implies conservative tests. We provide an algorithm which iteratively optimizes an initially given design while adjusting for both discreteness of test statistics and interim patients. The algorithm is described verbally, graphically and formally to facilitate immediate implementation in computer software.

Results: The optimized design exploits the expected significance level better and is more powerful than the initial one. The algorithm applies to fixed sample and planned flexible adaptive designs for single- and multi-arm trials with one or multiple primary endpoints. The benefit increases with the number of interim patients.

Conclusions: When planning a trial with interim analyses, the rules for decisions must be adjusted to interim patients. Otherwise, the test procedure is conservative resulting in loss of power. This is essential in situations where the number of interim patients is important compared to the first stage, particularly in trials with multiple primary endpoints.

1. Introduction

Using a group sequential or adaptive design within a clinical trial, the investigator can perform planned interim analyses in order to draw conclusions from the data collected so far. However, data collection and analysis need time. New patients will enter the trial while the interim analysis is ongoing. Moreover, depending on the event kinetics of the specific disease, the trial design, and the corresponding endpoints, some patients might not yet be evaluable at the date of the interim analysis. For example, if the endpoint is a response rate at t₀ months after the start of treatment, all those patients are not yet evaluable at the date of the interim analysis who were enrolled during the last t₀ months before the interim analysis. These newly recruited patients as well as those patients with still unknown outcome at the date of the interim analysis are referred to as interim patients of the first kind (F-patients). An-
other type of interim patients naturally occurs in sequentially analyzed trials with multiple endpoints, for example, in a trial which assess both toxicity and efficacy of a new treatment. If the toxicity question can already be answered at the interim analysis but the efficacy question not, it might be decided to continue the trial to a second stage in order to answer both questions. Then the patients recruited in the second stage will usually also provide additional data on toxicity and thus play a hybrid role. Whereas they are regular stage two patients for the efficacy question, they act at the same time as interim patients for the toxicity question. These patients will be called *interim patients of the second kind (S-patients)*. In a trial both kinds of interim patients can occur and will therefore be called interim patients whenever no confusion is possible.

The number of interim patients essentially depends on duration of the interim analysis and the endpoints chosen. Particularly, in a trial with multiple endpoints, the number of interim patients can be important. Disregarding the additional information provided by these interim patients would be unsatisfactory. Those problems arising with patients recruited during interim analyses are addressed in the EMA reflection paper on methodological issues in confirmatory clinical trials with an adaptive design [1]. A practicable solution might be to perform a repeated interim analysis with interim patients included and to declare early superiority only if the null hypothesis $H_0$ is rejected twice: In the original interim analysis and in the repeated interim analysis with interim patients included. Whitehead used the term “overrunning” for the occurrence of data on interim patients. He stated that the stopping boundaries establish when the trial should be terminated. The decision if significance is reached or not is made in the second interim analysis [2]. Traditional group sequential [3–5] and adaptive designs [6–12], however, neither foresee nor take into account repeated analysis with interim patients, that is, the underlying stage-wise rejection boundaries are not adjusted for interim patients. This results in a conservative test procedure since there are now two hurdles for rejection of $H_0$. To avoid conservativeness and thus loss in power, interim patients have to be considered already in the planning phase and the rejection boundary has to be adjusted accordingly for interim patients.

General strategies for including patients recruited during adaptive interim analyses were proposed and studied by Faldum and Hommel [13]. Their focus was on adaptive designs with independent and standard uniformly distributed $p$-values under the null hypothesis. In this paper, we explicitly consider the setting where the stage-wise $p$-values arise from discrete data. Then, there are only finitely many observable $p$-values, that is, the stage-wise $p$-values are no longer standard uniformly distributed, but usually the $p$-lcloud condition [11] still holds true. As a consequence of Brannath et al. [11], the adaptive test controls the type I error rate, but is in general conservative. Strategies of improving adaptive designs in the presence of discrete $p$-values were proposed by Ruckes et al. [14] and Englert [15–17], however, without accounting for interim patients. The strategies described in [15–17] are based on the exact binomial test. Here, we propose strategies of planning and analyzing adaptive designs while explicitly accounting for both discreteness of $p$-values and interim patients. The strategy of the paper is as follows. In section 2.1, we start with a brief summary of strategies for including interim patients in adaptive designs following [13]. On this basis, we present an algorithm for the construction of adaptive designs with interim patients in the context of discrete $p$-values (section 2.2). We start with a verbal and graphical illustration of the algorithm in the fixed sample setting in section 2.2.1 followed by a mathematical description in section 2.2.2. The latter facilitates immediate implementation of the algorithm in computer software. Finally, in section 2.2.3, we consider the adaptive setting. We concentrate on two-stage adaptive designs with a one-sided test problem in the sequel. This illustrates the underlying ideas more easily. In section 3, we illustrate performance of our procedure using the example of the exact binomial test. Our numerical example is motivated by requirements of the B-NHL 2013 trial (Eudra-CT number: 2013–003253–21). We conclude with a discussion of our findings in section 4.

2. Methods

2.1 General Aspects

2.1.1 F-patients

In the sequel, random variables will be denoted by upper case letters whereas realizations will be denoted by the corresponding lower case letters. Moreover, let us assume that we test a single one-sided null hypothesis $H_0$, say, using a two-stage adaptive design. We assume that the two-stages are defined by disjoint cohorts of patients, that is, each patients is only analyzed once either in the first or second analysis. Let $P_1$ and $P_2$ denote the $p$-values corresponding to these two analyses.

A two-stage adaptive design with consideration of interim patients is defined by the following algorithm. Let $\alpha$ denote the significance level of the adaptive test. Additionally fix two bounds $\alpha_1$, $\alpha_0$ with $0 \leq \alpha_1 < \alpha < \alpha_0 \leq 1$. The first stage is planned with power $1 - \beta_1$ and a sample size $n_1$ corresponding to an a priori effect estimator. At the interim analysis, the $p$-value $P_1$ is calculated. We assume that patients are accrued to the trial without interruptions in the time interval from $\tau_0$ to $\tau_1$, where $\tau_0$ denotes the calendar date when the trial is opened and $\tau_1$ the calendar date of the end of the interim analysis when the first-stage $p$-value $P_1$ is certain. In particular, accrual to the trial is not stopped while the interim analysis is ongoing. Decisions regarding accrual to the trial and handling of interim patients will for the first time be made at time $\tau_1$. These actions depend on the observed value of $P_1$ as specified below. We distinguish three cases: i) $P_1 \leq \alpha_1$, ii) $\alpha_1 < P_1 \leq \alpha_0$, and iii) $P_1 > \alpha_0$.

In case of $P_1 > \alpha_0$, the trial will immediately be stopped for futility and no further patient is recruited beyond time $\tau_1$. The interim patients accumulated so far do not have any impact on the decision regarding the null hypothesis $H_0$: If $P_1 > \alpha_0$, the null hypothesis $H_0$ will be accepted in any case irrespective of the results of the analysis of interim patients.
If \( p_1 \) falls into the continuation region \((\alpha_1, \alpha_0]\), the trial will go on with a regular second stage. In this case a new sample size \( n_2' \) is calculated for the second stage to achieve the conditional power \( 1 - \beta_2 \) (e.g. based on the observed effect from the sampled data of the first stage). Assume that \( n_{int} \) interim patients have been accumulated in the meantime. Recall that these are either patients recruited during the interim analysis or patients recruited before the start of the interim analysis due to not yet completed follow-up (e.g. if the endpoint of interest is outcome at a fixed point of time after the start of treatment). Then, we define the sample size of the regular second stage as \( n_2 := \max(n_2', n_{int}) \). That is, if \( n_2' \leq n_{int} \) accrual to the trial is immediately stopped. In this case the patients of the regular second stage are exactly the interim patients. If \( n_{int} < n_2' \), accrual to the trial goes on up to until further \( n_2' - n_{int} \) patients have been accrued. In the latter case, the regular second stage is formed by all interim patients plus \( n_2' - n_{int} \) newly recruited patients. The analysis of the regular second stage will be performed as soon as these \( n_2 \) patients have completed their follow-up as foreseen in the trial protocol. With this prescription, notice that the second analysis does not generate new interim patients.

With an ordinary adaptive design [6, 11], the trial would be stopped after the first stage with rejection of \( H_0 \) if \( p_1 \leq \alpha_1 \). But in our setting, if \( p_1 \leq \alpha_1 \), an additional analysis of interim patients is performed as proposed by [13]. More precisely, if \( p_1 \leq \alpha_1 \), the trial will immediately be stopped and no further patient will be recruited. But \( H_0 \) will not yet be rejected at that point of time. Instead, a second analysis of interim patients will be performed, as soon as all interim patients have completed their follow-up as foreseen in the trial protocol and become evaluable. The null hypothesis \( H_{0/} \) is rejected if and only if this second analysis of interim patients is "significant" as well. Since accrual to the trial was already stopped (at calendar time \( \tau_1 \)), notice that this second analysis does not generate new interim patients.

So, following the interim analysis and provided that no stop for futility occurred, there will always be a second analysis. Therefore, let \( P_2 \) denote the \( p \)-value of this second analysis, which either denotes the \( p \)-value corresponding to the analysis of interim patients if \( p_1 \leq \alpha_1 \) or the \( p \)-value corresponding to the analysis of the regular second stage if \( \alpha_1 < p_1 \leq \alpha_0 \). After finishing this second analysis, the observed \( p \)-values \( p_1 \) and \( p_2 \) from both analyses have to be combined to get a study result. This combination is performed by a conditional error function \( a(p_1) : [0, 1] \rightarrow [0, 1] \), which is non-increasing [8, 18]. The null hypothesis is rejected if and only if \( p_2 \leq a(p_1) \). This yields an adaptive level-\( \alpha \) test, provided that

\[
P_{int}(P_2 \leq a(P_1)) \leq \alpha \]

(1)

In the special case \( a(p_1) \equiv 1 \) for \( p_1 \leq \alpha_1 \), we are reduced to an ordinary adaptive design without consideration of interim patients. When interim patients are explicitly considered as proposed by [13], the conditional error function is decreased in \([0, \alpha_1]\).

2.1.2 S-patients

Above, we focused on a trial with a single primary null hypothesis. Let us now consider a traditional group-sequential or adaptive trial with multiple primary null hypotheses, \( m \) say, with synchronized interim analysis of all null hypotheses after \( n_1 \) observations. Assume that \( m_1 < m \) null hypotheses \( H_{1}, \ldots, H_{m} \) can already be rejected after the interim analysis, whereas further \( n_2 \) patients are recruited for a regular second stage in order to assess the remaining \( m_2 := m - m_1 \) non-rejected null hypotheses \( H_{m_1+1}, \ldots, H_{m} \) in a second analysis. In general, these \( n_2 \) patients will provide additional data regarding the \( m_1 \) already rejected null hypotheses \( H_{1}, \ldots, H_{m_1} \). Disregarding this additional information would be unsatisfactory, particularly since the sample size \( n_2 \) might be of important size as compared to the stage one sample size \( n_1 \). In particular, this additional information will have to be sent to authorities and will be analyzed as well. If this additional stage two data shows a smaller or contrasting treatment effect, this might lead to a withdrawal of an earlier superiority proof of the null hypotheses \( H_{1}, \ldots, H_{m_1} \). Thus, the stage stage two patients play a hybrid role in a trial with multiple endpoints. Whereas the \( n_2 \) patients are regular stage two patients for the not yet rejected null hypotheses \( H_{m_1+1}, \ldots, H_{m} \), they act at the same time as interim patients for the already rejected null hypotheses \( H_{1}, \ldots, H_{m_1} \).

Again, a practicable solution might be to perform a repeated analysis of \( H_{1}, \ldots, H_{m} \) with interim patients included. That is, a null hypothesis \( H_i, 1 \leq i \leq m_1 \), is rejected if and only if \( H_i \) is "rejected" twice (in the original and in a repeated interim analysis with interim patients included) with the conditional error function explicitly being adjusted for interim patients to avoid a conservative test. Notice that \( F \)-patients (see section 2.1.1) are relevant here, too, if all null hypotheses are rejected at the interim analysis. However, with both kinds of interim patients, we are in general led to consider adaptive designs with decreased conditional error function in \([0, \alpha_1]\).

2.2 Interim Patients with Discrete Test Statistics

When planning adaptive designs, it is common practice to assume that the \( p \)-values \( P_1 \) and \( P_2 \) are independent and standard uniformly distributed under the null hypothesis [11]. So, our starting point is an initial conditional error function (CEF) \( a(p_1) \) which exactly controls the type I error rate \( \alpha \) in case of independent and standard uniformly distributed \( p \)-values. In the sequel, let

\[
a(p_1) = \begin{cases} f_m(p_1), & \text{if } p_1 \leq \alpha_1, \\ f_1(p_1), & \text{if } \alpha_1 < p_1 \leq \alpha_0, \\ 0, & \text{if } p_1 > \alpha_0, \end{cases}
\]

(2)

for some non-decreasing functions \( f_1 \) : \([0, 1] \rightarrow [0, 1]\) with \( f_1(\alpha_1) \geq f_1(\alpha) \) such that

\[
\alpha = \int_0^{\alpha_0} f_1(p_1) dp_1 + \int_{\alpha_1}^{\alpha_0} f_m(p_1) dp_1.
\]

(3)

Analysis of interim patients is implicitly considered here, since we do not demand \( f_m(p_1) \equiv 1 \). Regarding the choice of \( a(p_1) \) different philosophies may be followed. Depending on the underlying trial situation, one might either favor fast decreasing con-
ditional error functions (like those proposed by [7, 9]) which put more weight on small $p$-values of the first stage, or one might favor stable conditional error functions (as proposed by [20]) which attribute equal weight to the first stage $p$-values. Examples of suitable choices for $\alpha(p_1)$ in the presence of interim patients are discussed in [13]. We therefore assume that an initial CEF $\alpha(p_1)$ according to Equation 2 is given with Equation 3 being fulfilled. Equation 3 ensures that the adaptive test exactly controls the significance level $\alpha$ in case of independent and standard uniformly distributed $p$-values $P_1$ and $P_2$.

However, in the sequel, we assume that the stage-wise $p$-values $P_1$ and $P_2$ arise from discrete data and thus have discrete range instead of being standard uniformly distributed. More precisely, we assume that for all $0 \leq c \leq 1$, $P_{ik}(P_{1k} \leq c) = c$ if $c$ is the location of an observable $p$-value, and $P_{ik}(P_{1k} < c) < c$ otherwise, $i = 1, 2$, and additionally that $P_{1k}(P_{2k} < c \mid P_{1k}) \leq c$. Notice that only those (data-based) modifications of the design are admissible which are compatible with these properties of the stage-wise $p$-values. Then the p-cloud condition holds true and, as a consequence of Brannath et al. [11], the adaptive test based on $\alpha(p_1)$ preserves the significance level $\alpha$. However, the adaptive test will in general be conservative, that is, we may have $P_{ik}(P_{2k} \leq \alpha(P_{1k})) < \alpha$, because the discreteness of $p$-values is not considered in Equation 3. In this section, we propose an algorithm which iteratively improves the initial CEF $\alpha(p_1)$ by explicitly accounting for the discreteness of the stages. The algorithm yields a final CEF $\alpha'(p_1)$ which is adapted to discreteness and thus corresponds to a less conservative and more powerful test.

In a first step, the algorithm will be described for a priori fixed sample sizes $n_1$, $n_2$ and $n_{\text{int}}$ for the first stage, the regular second stage and the analysis of interim patients (fixed sample design). The adaptive case will be discussed later in section 2.2.3. In view of the discreteness let

$$R_{i,n_{\text{int}}} = \{p_{i1}^{(1)}, \ldots, p_{im_{\text{int}}}^{(1)}\}$$

with $p_{i1}^{(1)} < \ldots < p_{im_{\text{int}}}^{(1)}$ denote the set of observable $p$-values $P_i$ of the first stage. Likewise let

$$R_{i,n_{\text{int}},n_2} = \{p_{i1}^{(2)}, \ldots, p_{im_{\text{int}},n_2}^{(2)}\}$$

with $p_{i1}^{(2)} < \ldots < p_{im_{\text{int}},n_2}^{(2)}$ denote the set of observable $p$-values in interim patients (range of $P_2$ conditional on $P_1 \leq \alpha_1$), and let

$$R_{2,n_2} = \{p_{21}^{(2)}, \ldots, p_{2m_{\text{int}},n_2}^{(2)}\}$$

with $p_{21}^{(2)} < \ldots < p_{2m_{\text{int}},n_2}^{(2)}$ denote the set of observable $p$-values of a regular second stage (range of $P_2$ conditional on $P_1 \leq \alpha_2$), and let $n_{\text{int}} = n_1 + 1$, $m_{\text{int}} = n_{\text{int}} + 1$ and $m_2 = n_2 + 1$ (for examples, see Appendix A1). The latter, however, is not required in the sequel in order to cover more general situations. Notice that $R_{1,n_1}, R_{\text{int},n_{\text{int}}}$ and $R_{2,n_2}$ depend on $n_1$, $n_{\text{int}}$ and $n_2$, respectively. In the sequel, we demand that the condition

$$p_{i1}^{(1)} \leq p_{i2}^{(1)}$$

is fulfilled. The latter condition appears natural, because if a regular second stage is performed the interim patients are always part of the second stage cohort. That is, we always have $n_2 \geq n_{\text{int}}$ and may thus expect that the set $R_{2,n_2}$ is denser than the set $R_{\text{int},n_{\text{int}}}$. For the test statistics considered exemplary in Appendix A1 and in our numerical example below, condition 7 is obviously fulfilled.

Before describing our algorithm in full generality, we describe and motivate the algorithm verbally and graphically. This illustrates the idea underlying our algorithm.

In a next step, in section 2.2.2, we give a formal description of the algorithm. This provides an immediate basis for an implementation of the algorithm in computer software. A formal proof that the algorithm is well defined and that the final CEF $\alpha'(p_1)$ preserves the significance level $\alpha$ is given in Appendices A2 and A3, respectively. Finally, in section 2.2.3, we proceed to the adaptive setting. A numerical example is given in section 3.

2.2.1 Verbal and Graphical Illustration of the Algorithm (Fixed $n_1$, $n_{\text{int}}$ and $n_2$)

Assume an initial conditional error function (CEF) with interim patients $\alpha(p_1)$ as defined in Equation 2 and Equation 3 and indicated in Figure 2A. Throughout this section, we assume that the sample sizes $n_1$, $n_2$, and $n_{\text{int}}$ are fixed and remain so during the trial. Let $\alpha'_1$ denote the biggest observable first-stage $p$-value $p_{i1}$ below $\alpha_1$. Let $p_{i2}^{(\text{\textsc{int}}} - 1)$ denote the smallest observable first-stage $p$-value $p_{i2}$ above $\alpha_1$, and let $p_{i2}^{(\text{\textsc{int}})*}$ denote the next bigger one (Figure 2B). Since $\alpha(p_1)$ is chosen without accounting for discreteness of $p$-values, relevant portion of the area below the conditional error function might be spent in regions without probability mass. The test corresponding to $\alpha(p_1)$ will therefore in general be conservative, if the $p$-values are discrete. The non-exploited level is illustrated in Figure 2B by the light grey area below the CEF and will be denoted by $\Delta$. In order to improve the initial design $\alpha(p_1)$, we aim for redistributing the light grey area in a suitable way in order to achieve that non-vanishing probability mass is covered by the light grey area after reallocation of the area while taking care that the resulting modified CEF is still non-increasing. The final design $\alpha'(p_1)$ is displayed in Figure 2F. Notice that the dark grey, light grey and overall area below the CEFs $\alpha(p_1)$ and $\alpha'(p_1)$ from Figure 2B and Figure 2F are equal. However, as evident from Figure 2F, the final CEF $\alpha'(p_1)$ is more suitably adapted to the given discrete nature of the $p$-values, since with $\alpha'(p_1)$ the light grey area now carries non-vanishing probability mass (atoms of the distribution being indicated by dots in Figure 2). We will now describe the intermediate steps from Figure 2A to Figure 2F underlying our algorithm in detail. Since the light grey area in Figure 2B has probability mass zero, the probability mass below $\alpha(p_1)$ and the probability mass of the dark grey area below $\alpha(p_1)$ are equal. Our starting point is therefore the CEF $\alpha^{(0)}(p_1)$ defined by the dark grey area below $\alpha(p_1)$ displayed in Figure 2C. If the CEF $\alpha(p_1)$ is stage-wise constant as in Figure 2 and in our numerical example below, the same holds true for $\alpha^{(0)}(p_1)$, that is, the functions $f_1^{\text{int}}$ and $f_2^{\text{int}}$ defined below in Equation 10 are constants $f_1^{\text{int}}(p_1) = \alpha_{\text{int}}$ and $f_2^{\text{int}}(p_1) = \alpha_2$ in this special case (Figure 2C). We are now free to redistribute the area $\Delta$ and we do so by successively lifting the second stage.
boundary $\alpha^{(0)}(p)$. Notice that we pursue a “left-to-right” strategy, that is, we start lifting of the second stage boundary above the smallest observable first-stage \( p \)-value bigger than $\alpha_1$ and continue with lifting above successively bigger first-stage \( p \)-values. The rationale is achieving maximal benefit for the second stage in case of promising first-stage results. This concept is formalized in Step 3 of the algorithm described in section 2.2.2.

**Figure 1** Graphical illustration (not true to scale) of an adaptive design with interim patients based on independent and uniformly distributed \( p \)-values (as described in section 2.1). In a traditional adaptive design (see Figure 1A), a null hypothesis $H_0$ is rejected after the interim analysis if $p_1 \leq \alpha_1$, that is, the conditional error function equals one for all $p_1 \leq \alpha_1$. In presence of interim patients, we additionally require a “significant” \( p \)-value $p_2$ for the analysis of interim patients, e.g. rejection of $H_0$ after the interim analysis if $p_1 \leq \alpha_1$ and $p_2 \leq \alpha_{\text{int}}$ as indicated in Figure 1B. This results in a conservative test with the light grey area $\Delta$ corresponding to the non-exploited level (Figure 1B). Thus, the second stage boundary may be lifted by the amount $\Delta$ if interim patients are considered already in the planning phase (Figure 1C). This yields an improved design with consideration of interim patients which fully exploits the level (Figure 1D). Potential discreteness of \( p \)-values is not considered so far. Then the design might be improved further (see Figures 2 and 3).
Figure 2  Graphical illustration (not true to scale) of the algorithm applied to a two-stage fixed sample design (as described in section 2.2.1). Atoms of the discrete distribution of $p$-values under the null hypothesis are indicated by dots.
Here, we give a graphical illustration. So, in a first step (Figure 2D), we allocate the whole area $\Delta$ in a rectangle above the interval $[\alpha_{n}^{*}, \alpha_{1}]$. The rectangle in Figure 2D has height $\Delta/(p^{(n-1)}_1 - \alpha_{1})$ and reaches up to some point $x_1$. Now we lift the CEF over $[\alpha_{1}^{*}, p^{(n-1)}_1]$ in the maximal sensible way such that the resulting CEF $\alpha^{(1)}(p_1)$ is non-increasing, that is, at most up to the biggest observable second-stage $p$-value $\gamma$ below $\alpha^{(1)}(\alpha^*_1)$. In our numerical example from section 3, we have $x_1 > \gamma$ (as indicated in Figure 2D), that is, we lift the CEF over $[\alpha_{1}^{*}, p^{(n-1)}_1]$ up to $\gamma$ (Figure 2E). The remnant area $\Delta'$ is allocated as rectangle above the interval $[\alpha_{1}^{*}, p^{(n-1)}_1]$ (Figure 2F). The rectangle has height $\Delta/(p^{(n-2)}_1 - p^{(n-1)}_1)$ and reaches up to some point $x_2$. Now, we lift the CEF over $[\alpha_{1}^{*}, p^{(n-2)}_1]$ up to $x_2$. Since the whole remnant length is spent at this step, the procedure stops after this step with final CEF $\alpha^{(2)}(p_1) := \alpha^{(2)}(p_1)_\gamma$. In case of $x_2 > \alpha^{(2)}(p_1)_\gamma$, we have continued as described above up until the whole area $\Delta$ had been reallocated at some step. The CEF $\alpha^{(2)}(p_1)$ represents a design with more suitable allocation of the light grey area as compared to the initial design $\alpha(p_1)$ against the background of the given discreteness of the $p$-values. Notice that lifting of the CEF may in principle be done in manifold ways provided that the resulting CEF in non-increasing. In our algorithm we advocate a “left-to-right” strategy when reallocating the area $\Delta$ in order to achieve maximal benefit from promising first stage results for the second stage.

In the next section, we will formalize the algorithm described above. This facilitates implementation of the algorithm in computer software.

### 2.2.2 Algorithm (for Fixed $n_1$, $n_{\text{int}}$ and $n_2$)

Assume an initial conditional error function (CEF) $\alpha(p_1)$ as defined according to Equations 2 and 3 such that $0 < P(P_1 \leq \alpha_1) < P(P_1 \leq \alpha_0)$ and $P(P_2 \leq \alpha_0 | P_1 = p_1) > 0$ for all observable first-stage $p$-values $p_1 \leq \alpha_0$, that is, for all $p_1 \in R_{\gamma, \alpha_0} \cap [0, \alpha_0]$. Recall that we also require Condition 7 to be fulfilled.

**Step 1. Considering discreteness of stage 1.**

Define $\alpha^{*}_1$ and $\alpha^{*}_0$ as the biggest observable $p$-value $p_1$ of the first stage below or equal to $\alpha_1$ and $\alpha_0$, respectively, that is,

$$
\alpha^{*}_1 := \max \{ R_{\alpha_1} \cap [0, \alpha_1] \}, \quad \alpha^{*}_0 := \max \{ R_{\alpha_0} \cap [0, \alpha_0] \}.
$$

We thus have $x_1 > \gamma$ (as indicated in Figure 2D), that is, we lift the CEF over $[\alpha_{1}^{*}, p^{(n-1)}_1]$ up to $\gamma$ (Figure 2E). The remnant area $\Delta'$ is allocated as rectangle above the interval $[\alpha_{1}^{*}, p^{(n-1)}_1]$ (Figure 2F). The rectangle has height $\Delta/(p^{(n-2)}_1 - p^{(n-1)}_1)$ and reaches up to some point $x_2$. Now, we lift the CEF over $[\alpha_{1}^{*}, p^{(n-2)}_1]$ up to $x_2$. Since the whole remnant length is spent at this step, the procedure stops after this step with final CEF $\alpha^{(2)}(p_1) := \alpha^{(2)}(p_1)_\gamma$. In case of $x_2 > \alpha^{(2)}(p_1)_\gamma$, we have continued as described above up until the whole area $\Delta$ had been reallocated at some step. The CEF $\alpha^{(2)}(p_1)$ represents a design with more suitable allocation of the light grey area as compared to the initial design $\alpha(p_1)$ against the background of the given discreteness of the $p$-values. Notice that lifting of the CEF may in principle be done in manifold ways provided that the resulting CEF in non-increasing. In our algorithm we advocate a “left-to-right” strategy when reallocating the area $\Delta$ in order to achieve maximal benefit from promising first stage results for the second stage.

In the next section, we will formalize the algorithm described above. This facilitates implementation of the algorithm in computer software.

---

© Schattauer 2015

Methods Inf Med 5/2015
as final CEF, that is, \( \alpha^*(p_1) := \alpha^{0}_1(p_1) \). Else, we go on with the algorithm and we may assume that there is a \( j \geq 1 \) such that \( x_j > y \) and \( c_j + l < c_{o} \) for all \( l \leq j \), where \( x_j \) is defined by

\[
\sum_{i=1}^{j-1} \left( y - f_1^*(p_1^{(i+1)}) \right) \cdot P_{\alpha i} (p_i = p_1^{(i+1)}) + \left( x_j - f_1^*(p_1^{(j)}) \right) \cdot P_{\alpha j} (p_i = p_1^{(j)}) = \Delta.
\] (17)

In order to settle the inductive step, let \( x_{j+1} \) be defined by

\[
\sum_{i=1}^{j} \left( y - f_1^*(p_1^{(i+1)}) \right) \cdot P_{\alpha i} (p_i = p_1^{(i+1)}) + \left( x_{j+1} - f_1^*(p_1^{(j+1)}) \right) \cdot P_{\alpha j+1} (p_i = p_1^{(j+1)}) = \Delta.
\] (18)

If \( x_{j+1} \leq y \), then stop and choose \( \alpha^{(j+1)}_i(p_1) \) from \( \alpha^{(j)}_i(p_1) \) as final CEF, that is, \( \alpha^*(p_1) := \alpha^{(j+1)}_1(p_1) \). Notice that \( \alpha^{(j+1)}_i(p_1) \) is non-increasing, since we have \( x_{j+1} > f_1^*(p_1^{(j+1)}) \) by assumption on \( x_j \). If \( x_{j+1} > y \) and \( c_j + j + 1 = c_{o} \), then stop and choose \( \alpha^{(j+1)}_i(p_1) \) from \( \alpha^{(j+1)}_i(p_1) \) as final CEF, that is, \( \alpha^*(p_1) := \alpha^{(j+1)}_1(p_1) \). Else, we have \( x_{j+1} > y \) and \( c_j + j + 1 < c_{o} \), and we go on with the above algorithm with \( j \) replaced by \( j + 1 \) everywhere. After finitely many steps, the algorithm stops and we arrive at a final CEF \( \alpha^*(p_1) \).

2.2.3 The Adaptive Case

So far, we assumed a fixed sample size setting, where \( n_1, n_2, \) and \( n_{int} \) are a priori fixed and remain so during the trial. In practice, one might be interested in the adaptive case where the sample size \( n_1 \) is adapted after the interim analysis based on the observed \( p \)-value \( p_1 \) of the first stage. Following Bauer [19], we distinguish two types of adaptive designs: i) Planned flexible adaptive designs, and ii) Fully flexible adaptive designs. With planned flexible designs, flexibility follows a strict predefined adaptation rule. In particular, the sample size adaptation rule \( n_2(p_1) \) is fixed in advance and may thus be used already in the planning phase. With fully flexible designs no fixed adaptation rule is given in advance. In particular, \( n_2(p_1) \) is a priori unknown. However, both fully flexible and planned flexible designs, require that the final conditional error function (CEF) \( \alpha^*(p_1) \) underlying the design is fixed in advance and remains so throughout the trial, in order to guarantee that the adaptive test preserves the level.

We first consider the case of planned flexible adaptive designs. Then the number of second stage patients \( n_2(p_1) \) depends on \( p_1 \) in an a priori determined manner. The algorithm described in section 2.2.2 for the fixed sample setting canonically extends to the setting of planned flexible adaptive designs, because in this case the second stage sample size \( n_2(p_1) \) conditional on \( p_1 \) is again a priori known. Then, the discrete distribution of the \( p \)-values in the regular second stage is a priori known as well, and, in analogy to the fixed sample setting, the remnant level \( \Delta \) from (13) can be redistributed in order to improve the second stage boundary. With planned flexible adaptive designs, the essential modification in our algorithms is due to the fact that the set \( R_{k,n_2} \) becomes dependent on \( p_1 \) via \( n_2 = n_2(p_1) \). Thus, the description of our algorithm becomes more extensive as compared to fixed sample setting. Special care has to be taken that the resulting CEF is non-increasing. The functional structure of the algorithm, however, remains essentially unchanged. We therefore confine ourselves to giving a brief verbal and graphical illustration of the algorithm in the setting of planned flexible adaptive designs. Again assume a CEF \( \alpha(p_1) \) fulfilling \( \alpha^{(0)}_i(p_1) \) as final CEF, that is, \( \alpha^*(p_1) := \alpha^{(0)}_1(p_1) \). Notice that \( \alpha^{(i+1)}_i(p_1) \) is non-increasing, since we have \( x_{i+1} > f_1^*(p_1^{(i+1)}) \) by assumption on \( x_i \). If \( x_{i+1} > y \) and \( c_i + i + 1 = c_{o} \), then stop and choose \( \alpha^{(i+1)}_i(p_1) \) from \( \alpha^{(i+1)}_i(p_1) \) as final CEF, that is, \( \alpha^*(p_1) := \alpha^{(i+1)}_1(p_1) \). Else, we have \( x_{i+1} > y \) and \( c_i + i + 1 < c_{o} \), and we go on with the above algorithm with \( i \) replaced by \( i + 1 \) everywhere. After finitely many steps, the algorithm stops and we arrive at a final CEF \( \alpha^*(p_1) \).

\[
\alpha^{(1)}(p_1) := \begin{cases} f_1^*(p_1), & \text{if } p_1 \leq p_1^{(c_1)}, \\
x_1, & \text{if } p_1 = p_1^{(c_1+1)}, \\
f_1^*(p_1), & \text{if } p_1^{(c_1+1)} < p_1 \leq p_1^{(c_0)}, \\
0, & \text{if } p_1 > p_1^{(c_0)}. \end{cases}
\]

\[
\alpha^{(1)}(p_1) := \begin{cases} f_1^*(p_1), & \text{if } p_1 \leq p_1^{(c_1)}, \\
x_1, & \text{if } p_1 = p_1^{(c_1+1)}, \\
f_1^*(p_1), & \text{if } p_1^{(c_1+1)} < p_1 \leq p_1^{(c_0)}, \\
0, & \text{if } p_1 > p_1^{(c_0)}. \end{cases}
\]

\[
\alpha^{(1)}(p_1) := \begin{cases} f_1^*(p_1), & \text{if } p_1 \leq p_1^{(c_1)}, \\
x_1, & \text{if } p_1 = p_1^{(c_1+1)}, \\
f_1^*(p_1), & \text{if } p_1^{(c_1+1)} < p_1 \leq p_1^{(c_0)}, \\
0, & \text{if } p_1 > p_1^{(c_0)}. \end{cases}
\]
Figure 6  Graphical illustration (not true to scale) of the algorithm applied to a two-stage planned flexible adaptive design (as described in section 2.2.3). Atoms of the discrete distribution of \( p \)-values under the null hypothesis are indicated by dots.
Let $\Delta$ denote the difference of the areas below $\alpha_1(p_1)$ and $\alpha_0(p_1)$. We now improve the CEF $\alpha_0(p_1)$ and redistribute the area by lifting the second stage boundary of $\alpha_0(p_1)$. Again, we follow a “left-to-right” strategy while adopting the notation from section 2.2.1. In a first step, we allocate the whole area $\Delta$ in a rectangle above $[\alpha_1^{(0)}, p_1^{(c+1)}]$ (Figure 6D). The rectangle of area $\Delta$ reaches up to some point $x_1$, say (Figure 6D). We lift the CEF above $[\alpha_1^{(0)}, p_1^{(c+1)}]$ in the maximal sensible way such that the resulting CEF $\alpha_0^{(0)}(p_1)$ is non-increasing, that is, at most up to the biggest observable $p$-value below $\alpha_0^{(0)}(x_1)$ (Figure 6E). In our graphical example, we have $x_1 > \alpha_0^{(0)}(x_1)$. Therefore, we lift the CEF above $[\alpha_1^{(0)}, p_1^{(c+1)}]$ and continue with allocating the remaining area $\Delta - \alpha_0^{(0)}(x_1)$ (Figure 6E). The remaining rectangle of area $\Delta$ reaches up to some point $x_2$, say (Figure 6F). Again, we lift the CEF above $[p_1^{(0)}, p_1^{(c+2)}]$ in the maximal sensible way such that the resulting CEF in non-increasing, that is, at most up to the biggest observable $p$-value below $\alpha_0^{(0)}(p_1)$. In our graphical example, we have $x_2 > \alpha_0^{(0)}(p_1)$. Thus, we lift the CEF above $[p_1^{(0)}, p_1^{(c+2)}]$ up to $x_2$, and the procedure stops at this step with final CEF $\alpha_0^{(0)}(p_1)$, because the whole area $\Delta$ has been reallocated. Otherwise, that is, in case of $x_2 > \alpha_0^{(0)}(p_1)$, we would have continued with the procedure as described above, up until the whole area $\Delta$ had completely been reallocated at some step. The CEF $\alpha_0^{(0)}(p_1)$ represents a design with more suitable allocation of the light grey area as compared to the initial design $\alpha(p_1)$ against the background of the given discreteness of the $p$-values. This settles the algorithm in the setting of planned flexible adaptive designs.

In practice, unplanned changes of sample size might become necessary, too, for reasons that could not be foreseen in advance. This requires fully flexible adaptive designs. With fully flexible adaptive designs, the functional relation between the actual second stage sample size $n_2$ and $p_1$ is no longer specified in advance. Consequently, no CEF $\alpha_0^{(0)}(p_1)$ exists in this setting which optimally exploits the level in general, because the CEF has to be fixed in advance based on an a priori estimate $n_2^{(0)}$ for the second stage sample size. That is, we are in general no longer optimal whenever $n_2 \neq n_2^{(0)}$. If full flexibility is required, a practicable solution might be a hybrid approach which considers discreteness of $p$-values in stage one, but ignores discreteness in stage two (treating second stage $p$-values as uniformly distributed under the null hypothesis). The hybrid design will still be less conservative than the initial design which ignores discreteness of $p$-values in both stages.

### 3. Numerical Example: A Two-stage Design for the Exact Binomial Test with Interim Patients

The numerical example is based on the B-NHL 2013 trial (Eudra-CT number: 2013-003253-21). The B-NHL 2013 trial is a clinical trial for the treatment of children and adolescents younger than 18 years of age with aggressive mature B-cell Non-Hodgkin lymphoma or leukemia. Amongst others, this trial is comparing the event rate at a fixed point of time to a historic control. For this purpose, we consider a single-arm design with binary primary outcome measure, e.g., tumor response, which is defined by whether a response to treatment is observed after a fixed time or not. Let $\pi$ be the (unknown) true response rate of the new treatment to be assessed. Let $n_0$ denote a lower bound for $\pi$ below which the treatment is considered inadequate and is abandoned. Thus, we are interested in testing the null hypothesis $H_0: \pi \leq n_0$. Power is calculated under the planning alternative hypothesis $H_1: \pi = n_1$. Hypothesis testing will be based on the binomial test statistic discussed in Appendix A1.

Our starting point is an initial two-stage design $\alpha(p_1)$ according to the Modified Simes Test (MST) [20] with interim patients:

$$
\alpha(p_1) = \begin{cases} 
\alpha_{\text{int}}, & \text{if } p_1 \leq \alpha_1, \\
\alpha_2, & \text{if } \alpha_1 < p_1 \leq \alpha_0, \\
0, & \text{if } p_1 > \alpha_0,
\end{cases}
$$

where $\alpha_1 < \alpha_0$ and $\alpha_2 \leq \alpha_{\text{int}}$ are chosen such that

$$
\alpha_2 = \frac{\alpha - \alpha_1 - \alpha_{\text{int}}}{\alpha_0 - \alpha_1}.
$$

Equation 22 ensures that the significance level $\alpha$ is fully exploited for independent and standard uniformly distributed $p$-values. This design enables early stopping for efficacy (with first stage level $\alpha_1$) as well as stopping for futility (with boundary $\alpha_0$). We make the choice $\alpha_1 = 0.025$, $\alpha_0 = 0.5$ and $\alpha_{\text{int}} = 0.5$. This implies $\alpha_2 = 0.07895$. Regarding the choice of $\alpha_1$ and $\alpha_{\text{int}}$, notice that a $p$-value above 0.5 corresponds to an observed effect in the opposite direction. The stage-wise $p$-values $p_1$ and $p_2$ are based on the exact binomial test and thus have discrete range.

To obtain a numerical example, we assume the following frame conditions. As usual, we let $n_1$, $n_2$ and $n_{\text{int}}$ denote the number of patients from stage one, from a regular second stage, and the number of interim patients, respectively. Moreover, let:

- $\pi = \text{True response rate at six months after the start of treatment.}$
- $n_0 = 0.5$, $n_1 = 0.65$.
- Nominal one-sided significance level: $\alpha = 5\%$.
- Accrual rate: 4 evaluable observations per month for 20 months, that is, $n_1 + n_2 = 80$ observations.
- Interim analysis at information rate 0.5 with an expected duration of one month. Thus:
  - Number of evaluable observations at date of interim analysis: $n_1 = 0.5 \cdot 80 = 40$.
  - Number of not yet evaluable observations at the date of the interim analysis: $6 \cdot 4 = 24$.
  - Number of patients accrued during the interim analysis: $1 \cdot 4 = 4$.
  - Total number of interim patients: $n_{\text{int}} = 28$. 
Based on the normal approximation of the binomial test, the above choice of sample size $n_1 = 40$, $n_{int} = 28$, $n_2 = 40$ represents a one-sided test of the null hypothesis $H_0$: $\pi \leq \pi_0 = 0.5$ to the level $\alpha = 5\%$ according to the MST from (21) with $\alpha = 0.025$, $\alpha_{int} = 0.5 = \alpha_0$ and with power of $80\%$ under the planning alternative hypothesis $H_1$: $\pi = \pi_1 = 0.65$. The expected sample sizes under null and planning alternative hypothesis are $EN(\pi_0) = 59.0$ and $EN(\pi_1) = 60.1$, respectively, and thus both below the single-stage sample size $n_{seq} = 69$. Notice that the above design already accounts for occurrence of interim patients, but is based on the normal approximation of the binomial test and thus does not account for discreteness of the test statistic. Therefore, the value for the aspired level to be derived under the normal approximation are overoptimistic. The actually aspired level is $4.03\%$ instead of $5\%$, and the actually achieved power is $77\%$ instead of $80\%$ under the given planning alternative hypothesis. In order to illustrate performance of our procedure, we exemplarily apply our algorithm from section 2.2.2 to the above CEF $\alpha(p_i)$ from (21) with $\alpha = 0.025$, $\alpha_{int} = 0.5 = \alpha_0$, in order to adjust for discreteness.

In a first step, we determine the sets $R_{1,n_1}$, $R_{2,n_2}$ and $R_{int,n_{int}}$.

Since $p_i^{(k)} = \Pi(n_j - k + 1, n_j, \pi_0)$ with $\Pi$ from App. A1, we find $n_1 = 40$, $n_{int} = 28$, $n_2 = 40$ the results displayed in Figure 7 (Equation 23) that

$$\alpha^* = \left\{ \begin{array}{ll} \alpha_{int} & \text{if } p_i \leq \alpha^* = p_i^{(14)}, \\
\gamma & \text{if } p_i = p_i^{(15)}, \\
x_2 & \text{if } p_i = p_i^{(16)}, \\
\alpha_{int} & \text{if } p_i = p_i^{(17)}, p_i^{(18)}, p_i^{(19)}, p_i^{(20)}, \\
0 & \text{if } p_i > \alpha_{int} = p_i^{(20)}. \end{array} \right.$$  

This settles the CEF $\alpha^{(0)}(p_i)$ from Equation 10. $\gamma$ as defined in Equation 11 amounts to $\gamma = p_2^{(10)} = \Pi(22, 40, \pi_0) = 0.318$. (26)

Step 3: Improving the boundary of stage 2. The actually exploited level of the initial design $\alpha(p_i)$ is $4.03\%$, that is, the non-exploited level amounts to

$$\Delta = \alpha - 0.0403 = 0.010. \quad (27)$$

We now start the inductive procedure for improving the boundary of stage 2. Recall that $P_{1,n_1}(p_i^{(j)}) = p_i^{(j)} - p_i^{(j-1)}$ for all $1 \leq j \leq n_1$, $p_i^{(0)} := 0$. To settle the base case, let $x_1$ be defined by

$$(x_1 - \alpha_{int}^*) \cdot (p_i^{(14)} - p_i^{(1)}) = \Delta. \quad (28)$$

This yields $x_1 = 0.534 > \gamma$ and $c_1 + 1 = 15 < 20 = c_\gamma$. Thus, we go on and define $x_2$ by

$$(x_2 - \alpha_{int}^*) \cdot (p_i^{(15)} - p_i^{(14)}) = \Delta. \quad (29)$$

Thus, $x_2 = 0.202 \leq \gamma$, and the procedure stops with final CEF $\alpha^*(p_i)$ given in Equation 30 (Figure 8).

With $n_1 = 40, n_{int} = 28, n_2 = 40$, a two-stage fixed sample test of $H_0$: $\pi \leq \pi_0$ based on CEF $\alpha^*(p_i)$ achieves a power of $82\%$ under the planning alternative hypothesis $H_1$: $\pi = \pi_1$, while exploiting a one-sided significance level of $4.75\%$.

4. Discussion

Many trials with interim analyses are concerned with those problems with interim patients. A practicable solution is performance of repeated interim analysis with interim patients included. Traditional group-sequential and adaptive designs do not take into account repeated analysis with interim patients. This results in conservative tests since there are now two hurdles for rejection. Moreover, when planning adaptive designs, it is common practice to assume that the $p$-values $P_1$ and $P_2$ are independent and standard uniformly distributed under the null hypothesis [11]. This implies...
Further conservativeness, if the $p$-values are in truth discrete, thus yielding two potential sources of conservativeness. Strategies for adaptive designs with discrete test statistics were proposed by [14–17]. The focus of [16] on single-arm two-stage phase II designs with binary endpoint based on the binomial distribution. In [17], an extension is proposed that allows handling of situations when the attained sample size deviates from the planned one. However, interim patients are not addressed in [14–17]. Conversely, strategies for considering interim patients were proposed by [13], but restricted on the setting of continuously distributed $p$-values. In this paper, we aim to fill this gap. We propose strategies for planning and analyzing adaptive designs which account for both interim patients and discreteness of $p$-values. The algorithm proposed in section 2.2 defines a procedure which successively modifies an initially given design $\alpha(p_1)$ according to the requirements of discrete data while accounting for interim patients.

The methods presented in this paper apply to any trial with interim analyses and provide less conservative and more powerful tests. The magnitude of gain in power depends on the specific setting, and essentially on the number of interim patients. In a trial with a single categorical (or metric) primary endpoint, interim patients refer to those subjects who are recruited during the interim analysis or who are not yet evaluable at the date of the interim analysis. The methods presented here offer a way to deal with the problem. However, we emphasize that the number of interim patients should be sufficiently large in order to control the probability of false negative results from the analysis of interim patients. That is, $n_{\text{int}}$ should be sufficiently large to ensure adequate conditional power

$$P_{1}\left(P_{2} \leq \alpha \left(P_{1}\right) \mid P_{1} = p_{1}\right) \text{ under the planning alternative hypothesis } H_{1}.$$  

Otherwise, it appears recommendable to resort to an ordinary adaptive design without interim patients.

The methods also apply in survival trials. At the first glance, it appears that the number of interim patients in survival trials is smaller as compared to the setting with categorical (or metric) endpoint. Indeed, whereas patients recruited during interim analysis still remain $F$-patients, the remaining patients recruited up to data closure for the interim analysis are all analysed with event-free patients being censored at the date of the interim analysis. On closer inspection, one notices that further follow-up information is gained for all patients without event while the interim analysis is ongoing. New events might even occur during the interim analysis. This shows that in a survival setting all patients accumulate interim data while the interim analysis is ongoing which might require consideration.

In a trial with multiple primary endpoints, the impact of interim patients might be particularly pronounced. If only some but not all null hypotheses are rejected at the interim analysis, the trial might be continued to a second stage in order to answer the remaining questions. Nevertheless, the patients recruited at stage two will in general also provide additional data regarding the already answered trial questions. Disregarding this additional information would be unsatisfactory. This implies performance of a second analysis of the already answered trial questions based on the additional information. Thus, the second stage patients formally act as interim patients regarding the already rejected null hypotheses ($S$-patients). Since the sample size $n_{2}$ of stage two might be of important size as compared to the sample size $n_{1}$ of the first stage, the number of interim patients might be of important size in trials with multiple endpoints. In order to deal with this problem, we might consider local adaptive tests for each intersection hypothesis which explicitly account for interim patients using the techniques described here and by [13].

Regarding handling of interim patients, we follow a strict approach in this paper: Following the interim analysis, there is always a second analysis including interim patients, and the underlying null hypothesis is rejected if and only if both analyses are "significant". Thus, in case of a futility stop at the interim analysis, the futility stop cannot be overruled by a "significant" result in interim patients, that is, the null hypothesis in accepted irrespective of the result in interim patients. Conversely, a "significant" result at the interim analysis will only result in rejection of the null hypothesis provided that the analysis of interim patients is "significant" as well. Therefore, particular care has to be taken in the planning phase that the analysis of interim patients is well-powered. In principle, less strict strategies of handling of interim patients might be possible as well. For example, one might decide to follow the philosophy of an "eased" futility stop. Then we would proceed as follows: If $p_{1} \leq \alpha_{1}$ or $\alpha_{1} < p_{1} \leq \alpha_{0}$, we proceed with interim patients as described in this paper. However, in case of $p_{1} > \alpha_{0}$, we still immediately stop accrual to the trial, but (deviating from the expositions in this paper) we do not require that the conditional error function (CEF) $\alpha(p_{1})$ equals zero when $p_{1} > \alpha_{0}$. This implies, that a futility stop can be overruled later if the data of interim patients show a clear treatment effect after all. The latter concept of an "eased" futility stop might appear justifiable in early phase II trial settings when the focus is still on identifying active and discarding inactive treatments, and will be contents of future research. Here, however, we focused on the stricter view to require two "significant" analyses before rejecting the null hypothesis.

The central idea underlying our algorithm is adjustment of an initially given design $\alpha(p_{1})$ with interim patients for discreteness of $p$-values by lifting the second-stage boundary of the CEF $\alpha(p_{1})$. This way, conservativeness resulting from ignoring interim patients and discreteness of $p$-values may be attenuated. As usual, we let $p_{1}^{\text{obs}}$ denote the smallest observable first-stage $p$-value which results in performance of a regular second stage. Regarding lifting the second-stage boundary, we followed a "left-to-right" strategy. That is, the CEF is first lifted above $p_{1}^{\text{obs}}$ in the maximal sensible way such that the modified CEF is still non-increasing. Then, we successively proceed in this way above the next larger observable first-stage $p$-value $p_{1}^{\text{obs}}$ and continue up until no significance level remains to be reallocated at some step. This strategy implies maximal benefit for a potential regular second stage in case of promising results after stage one. Of course, one might follow many different strategies of reallocating the non-exploited significance level. Instead of following our "left-to-
right” strategy, one could for example decide to lift the second stage boundary according to the proviso to achieve a uniform benefit for the whole second stage. Then, instead of lifting the CEF maximally above \( p_1^{(\alpha+1)} \), one would lift the CEF uniformly above all observable \( p \)-values \( p_1 \) with \( \alpha_1 < p_1 \leq \alpha_0 \). A comparison of different strategies, particularly regarding power, average sample number (ASN), maximum sample size and properties of the ASN in case of deviations from the planning assumptions, will be contents of future research as well. However, regarding different strategies of significance level reallocation, the present paradigm of non-increasing CEF’s should be respected.

In the present paper, we described our algorithm verbally as well as graphically. For two-stage fixed sample designs (with option for futility and efficacy stopping), the algorithm was also described formally in order to facilitate implementation into computer software. This yields a tool for optimizing designs according to the philosophy described above. The algorithm naturally extends to the setting of planned flexible adaptive designs. However, with fully flexible adaptive designs, no CEF exists which optimally exploits the level in general, because the CEF has to be fixed in advance based on an a priori estimate \( n_2^{(0)} \) for the second stage sample size. That is, we are in general no longer optimal whenever \( n_1 \neq n_2^{(0)} \). This restriction, however, appears acceptable, since fully flexible adaptive sample size adjustments will usually be performed in case of important new discoveries that could not be foreseen in advance (e.g. severe deviations of observed effects from planning assumptions). Then, a main focus is on setting the trial back on the right course without violating the integrity of the trial, and a loss in efficacy might be considered as acceptable trade-off for preventing a failure of the trial.

For future research, we aim for an implementation of the algorithm described in section 2.2.2 in computer software. On this basis, the size of increase in power may be studied in a variety of settings. For given initial design \( \alpha(p) \) and for each sample size constellation, the algorithm yields a unique optimized CEF \( \alpha^*(p) \). Due to discreteness of the \( p \)-values, there are thus only finitely many such designs with \( n_1, n_2, n_{\text{int}} \leq n_{\text{max}} \) for some a priori fixed maximal sample size \( n_{\text{max}} < \infty \). In analogy to the proceeding in [21], these designs may explicitly be determined and compared using a computer search algorithm running over the set of all sample size constellations. On this basis, future research might help to work out recommendations for optimal two-stage designs in the setting of discrete \( p \)-values while explicitly accounting for those problems arising with interim patients. In particular, we aim for a detailed elaboration of our method in the context of trials with multiple primary endpoints. Here, the impact of interim patients appears particularly striking.

References