Published in final edited form as:

J Biopharm Stat. 2016 ; 26(2): 202–216. doi:10.1080/10543406.2014.972509.

A simple and flexible graphical approach for adaptive group-sequential clinical trials

Toshifumi Sugitani¹, Frank Bretz^{2,3}, and Willi Maurer²

¹Section for Medical Statistics, Medical University of Vienna, Austria ²Novartis Pharma AG, Basel, Switzerland ³Institute for Biometry, Hannover Medical School, Hannover, Germany

Abstract

Adaptive clinical trial designs incorporating treatment selection at pre-specified interim analyses have recently attracted considerable attention. In this paper, we introduce a graphical approach to testing multiple hypotheses in group-sequential clinical trials allowing for mid-term design modifications. It is intended for structured study objectives in adaptive clinical trials and extends the graphical group-sequential designs from Maurer and Bretz (Statistics in Biopharmaceutical Research 2013; 5: 311–320) to adaptive trial designs. The resulting test strategies can be visualized graphically and performed iteratively. We illustrate the methodology with two examples from our clinical trial practice. First, we consider a three-armed gold-standard trial with the option to reallocate patients to either the test drug or the active control group, while stopping the recruitment of patients to placebo, after having demonstrated superiority of the test drug over placebo at an interim analysis. Second, we consider a confirmatory two-stage adaptive design with treatment selection at interim.

Keywords

Adaptive design; Closed test procedure; Group-sequential design; Inverse normal combination test; Sample size reassessment

1 Introduction

Clinical trial designs are becoming increasingly more complex in modern drug development. Trial sponsors are often interested in investigating multiple structured study objectives, such as comparing several treatments to a common control, assessing non-inferiority and superiority, exploring multiple endpoints, and studying several populations. Ethical aspects include the obligation that only a minimum number of patients should be enrolled into the trial. To this end, group-sequential or adaptive clinical trials are commonly conducted that aim at efficient interim decision making. Such complexity of clinical trial designs inevitably forces statisticians to employ more sophisticated multiple test procedures. Among many multiple test procedures developed in the recent years are graphical approaches (Bretz et al.,

2009; Burman et al., 2009; Bretz et al., 2011b; Maurer and Bretz, 2013a), which include fixed-sequence (Maurer et al., 1995), fallback (Wiens, 2003) and several gatekeeping (Dmitrienko et al., 2011; Huque et al., 2013) procedures as special cases. Graphical approaches are a viable tool to address complex clinical trial designs by visualizing multiple test procedures with nodes, weights, and directed edges: Using graphical approaches, nodes with associated weights denote the individual null hypotheses and their local significance levels. Directed edges between the nodes specify how the local significance levels are propagated in case of significant results. The resulting procedures control the family wise error rate (FWER) in the strong sense across all primary and secondary hypotheses. That is, the probability of erroneously rejecting at least one true null hypothesis is controlled at a pre-specified significance level $\alpha \in (0,1)$ under any configuration of true and false null hypotheses. Such graphical visualizations reflect the underlying decision strategies and thus facilitate the discussion with clinical teams to choose a suitable multiple test procedure tailored to the study objectives.

In the meantime, the graphical approach has been extended to more complex applications, including group-sequential trials (Maurer and Bretz, 2013b) and (non group-sequential) two-stage adaptive clinical trials (Sugitani et al., 2013). Extending these previous works, in this paper, we introduce a graphical approach to testing multiple hypotheses in group-sequential clinical trials allowing for mid-term design modifications, such as hypotheses selection or sample size reestimation. We refer to such trials as adaptive group-sequential clinical trials. The adaptive group-sequential clinical trials have been investigated in various practical settings, such as for comparing several treatments to a common control (Stallard and Friede, 2008; Di Scala and Glimm, 2011), combined testing of non-inferiority and superiority (Wang et al., 2001; Öhrn and Jennison, 2010; Gao et al., 2013), exploring primary and secondary endpoints (Tamhane et al., 2012), studying population enrichment designs (Brannath et al., 2009; Magnusson and Turnbull, 2013), and adaptively analyzing microarray experiments (Zehetmayer et al., 2008). Roughly speaking, in such trials, closed test procedures (Marcus et al., 1976) are used in connection with either combination tests (Bauer and Kieser, 1999; Hommel, 2001; Bretz et al., 2006) or conditional error rates (Müller and Schäfer, 2001). This approach, however, is not originally intended for testing hierarchically structured study objectives and therefore cannot easily be applied to more complex adaptive clinical trials. Thus, we introduce in this paper a marginal p -value combination approach, which has the advantage of leading to sequentially rejective test procedures. Thus, the proposed approach can be used in connection with graphical approaches employing essentially the same iterative algorithm as for fixed designs such that standard programs for the computation of the rejection boundaries can be used.

The remainder of this paper is organized as follows. In Section 2, we introduce the marginal p -value combination approach. In Section 3 we extend the graphical approach by Maurer and Bretz (2013b) to adaptive group-sequential designs based on the marginal p -value combination approach. In Section 4 we illustrate the proposed method with two examples. In Section 5 we provide a numerical example to illustrate the proposed method. In Section 6 we conclude with some remarks.

2 Marginal p -value combination approach

In this section, we introduce the marginal p -value combination approach. The standard closed combination test approach (Bauer and Kieser, 1999; Hommel, 2001; Bretz et al., 2006) combines p -values of tests for intersection hypotheses. To illustrate the difference between this and the marginal p -value combination approach, consider two hypotheses H_1 and H_2 that are tested with a weighted Bonferroni test at levels $w_1\alpha$ and $w_2\alpha$, $w_1 + w_2 = 1$, resulting in p -values p_1 and p_2 at an interim analysis. In the standard approach, the p -value of the intersection hypothesis $H_{12} = H_1 \cap H_2$ is the multiplicity adjusted p -value $p_{12} = \min(p_1/w_1, p_2/w_2)$ that is to be combined with the p -value for the intersection hypothesis of a further stage and tested at level α . In order to reject H_1 (say), this test and the combination test of H_1 alone at level α have to be significant. In the marginal p -value combination approach, on the contrary, the p -values p_1 and p_2 of the different stages are combined to levels q_1 and q_2 , respectively. Then, H_1 can be rejected if $q_1 \leq w_1\alpha$ or if $q_2 \leq w_2\alpha$, and if $q_1 \leq \alpha$. This approach protects the FWER as well, can be slightly less powerful than the standard approach, but on the other hand is consonant and allows shortcuts.

The original idea dates back to Kieser et al. (1999), who implicitly used the marginal p -value combination approach to extend the hierarchical test procedure (Maurer et al., 1995) and the Bonferroni-Holm procedure (Holm, 1979) to two-stage adaptive trial designs. They used Fisher's combination test (Bauer and Köhne, 1994) for combining stagewise p -values. In this paper, we use the inverse normal combination test (Lehmacher and Wassmer, 1999; Cui et al., 1999) because it allows one to use the same rejection boundaries as standard group-sequential clinical trials and extend the applicability of the marginal p -value combination approach to any graphical multiple test procedure.

2.1 Bonferroni-based closed test procedures

We first review the basic concepts of closed test procedures (Marcus et al., 1976). Assume that we are interested in testing one-sided null hypotheses H_i , $i \in I$, where I denotes some index set. In this paper we consider Bonferroni-based closed test procedures, which apply weighted Bonferroni tests to each intersection hypothesis $H_J = \bigcap_{i \in J} H_i$, $J \subseteq I$. To this end, we define for each $J \subseteq I$ a collection of weights $w_i(J)$ such that $0 \leq w_i(J) \leq 1$ and $\sum_{i \in J} w_i(J) = 1$. We then reject H_J if $p_i \leq w_i(J)\alpha$ for at least one $i \in J$, where p_i denotes the unadjusted p -value for H_i . Applying the closure principle, we reject H_i , $i \in I$, if all intersection hypotheses H_J with $i \in J$ are rejected. This strongly controls the FWER at level α as long as each intersection hypothesis is tested at level α .

If in addition the weights $w_i(J)$ are chosen such that they satisfy the monotonicity condition

$$w_i(J) \leq w_i(U) \quad \text{for all } U \subseteq J \text{ and } i \in U, \quad (1)$$

consonance is ensured. That is, rejection of an intersection hypothesis H_J implies rejection of at least one elementary null hypothesis H_i , $i \in J$, which in turn allows one to perform the closed test procedure in a sequential way (Hommel et al., 2007). Many common multiple

test procedures satisfy condition (1) and are thus sequentially rejective, including the procedures based on the graphical approach by Bretz et al. (2009).

2.2 Two-stage adaptive designs without early efficacy stopping

To illustrate the main ideas in this paper, we initially consider two-stage adaptive designs without early efficacy stopping, but with the possibility to drop certain hypotheses at an interim analysis. With the marginal p -value combination approach, we use the inverse normal combination function

$$C(p_{i,1}, p_{i,2}) = 1 - \Phi \left\{ \sqrt{\eta_1} \Phi^{-1}(1 - p_{i,1}) + \sqrt{\eta_2} \Phi^{-1}(1 - p_{i,2}) \right\}, \quad (2)$$

to combine the p -values from both stages, where η_1 and η_2 denote pre-specified positive weights such that $\eta_1 + \eta_2 = 1$, Φ^{-1} the inverse of the standard normal cumulative distribution function, $p_{i,t}$ the unadjusted p -values for H_i , $i \in I$, from stage $t = 1, 2$. Usually, the weights η_t are chosen proportional to the planned sample size per stage. See also the remark in Section 5, regarding the recommended choice of η_t in case of treatment selection designs.

Let $D_1 \subsetneq I$ denote the index set of hypotheses dropped at the interim analysis ($t = 1$). At the final analysis ($t = 2$), the marginal p -value combination approach rejects an intersection hypothesis $H_J = \cap_{i \in J} H_i$, $J \subseteq I$, if $C(p_{i,1}, p_{i,2}) < w_J(J)\alpha$ for at least one $i \in J_1 = J \cap D_1$. According to the Bonferroni inequality, it can be shown that this rejection rule controls the Type I error rate at level α under H_J . That is, $\Pr(\text{Reject } H_J) = \Pr(\text{Reject } H_i \text{ for at least one } i \in J_1) = \Pr(\cup_{i \in J_1} \{p_i < w_J(J)\alpha\}) \leq \sum_{i \in J_1} \Pr(p_i < w_J(J)\alpha) = \sum_{i \in J_1} w_J(J)\alpha \leq \alpha$. Therefore, based on the closure principle, the marginal p -value combination approach rejects H_J while strongly controlling the FWER at level α if all intersection hypotheses H_J with $i \in J$ are rejected at level α . Note that we can apply sequentially rejective Bonferroni-based graphical approaches if the weights $w_J(J)$ satisfy the monotonicity condition (1), as in Section 2.1.

2.3 Extension to adaptive group-sequential designs

We now extend the ideas from Section 2.2 to general adaptive group-sequential designs with interim rejection of null hypotheses, sample size reassessment and possibly more than two stages. To begin with, we consider testing a single one-sided null hypothesis H in a h -stage adaptive group-sequential design using the error spending approach (Lan and DeMets, 1983). Let $\gamma \in (0, 1)$ denote a pre-specified significance level, and ψ_t the information time or information fraction at certain analysis time point t , i.e., the cumulative weights η_t for the data from stages 1 to t , summing to 1 over all h stages. Let further $a(\gamma, \psi_t)$ denote a family of error spending functions that is non-decreasing in ψ_t with $a(\gamma, \psi_t) \leq a(\gamma, \psi_{t'})$, $a(\gamma, 0) = 0$, and $a(\gamma, 1) = \gamma$ for any γ , $0 < \gamma < 1$ and $0 \leq t < t' \leq 1$. Moreover, we assume that the nominal significance levels $\alpha_t^*(\gamma)$ are non-decreasing in γ for any number of analyses h and information fraction ψ_t . Liu and Anderson (2008) called a family of rejection boundaries with this property well ordered. As shown in Maurer and Bretz (2013b), examples of families of spending functions that lead to well ordered rejection boundaries include the Pocock type $a(\gamma, \psi_t) = \gamma \log\{1 + (e - 1)\psi_t\}$ (Pocock, 1977) and the O'Brien-Fleming type

$a(\gamma, \psi_t) = 2 - 2\Phi\left(z_{\gamma/2} / \sqrt{\psi_t}\right)$ (O'Brien and Fleming, 1979) for $\gamma < 0.318$, where z_α denotes the $100(1-\alpha)\%$ quantile of the standard normal distribution. As described later, the well ordered rejection boundaries play an important role when considering consonance for the marginal p -value combination approach in adaptive group sequential designs.

Let p_t denote the p -value for H from stage $t = 1, 2, \dots, h$ and

$Z_t = \left\{ \sqrt{\eta_1} \Phi^{-1}(1 - p_1) + \dots + \sqrt{\eta_t} \Phi^{-1}(1 - p_t) \right\} / \sqrt{E_t}$, where $E_t = \sum_{s=1}^t \eta_s$. The inverse normal combination test at time point $t = 1, \dots, h$ uses

$$q_t = C(p_1, \dots, p_t) = 1 - \Phi(Z_t). \quad (3)$$

Note that $q_1 = p_1$. The error spending function then needs to satisfy

$$\gamma = 1 - \Pr\left(\bigcap_{t=1}^h \{q_t > \alpha_t^*(\gamma)\}\right) \quad (4)$$

and

$$a(\gamma, \psi_t) - a(\gamma, \psi_{t-1}) = \Pr\left(\{q_t \leq \alpha_t^*(\gamma)\} \cap \bigcap_{s=1}^{t-1} \{q_s > \alpha_s^*(\gamma)\}\right) \quad (5)$$

under the null hypothesis H in order to control the Type I error rate at level γ . That is, for any $t = 1, \dots, h$, H can be rejected at time point t if $q_t \leq \alpha_t^*(\gamma)$. Note that for $t = 1$, equation (5) reduces to $a(\gamma, \psi_1) = \Pr(q_1 \leq \alpha_1^*(\gamma)) = \alpha_1^*(\gamma)$. Here, the nominal rejection levels can be calculated in the same manner as in conventional group-sequential designs, since the joint distribution of $(Z_1, Z_2, \dots, Z_h)$ is equivalent to that of the conventional group-sequential test statistics (Lehmacher and Wassmer, 1999). For example, using standard statistical software packages one can calculate the nominal rejection levels $\alpha_t^*(\gamma)$ satisfying equations (4) and (5), using $\psi_t = E_t/E_h$. Note that ψ_t are not adapted to the actual sample sizes, in particular not if at interim analysis the sample size is reestimated based on interim information. This ensures the control of the Type I error rate (Lehmacher and Wassmer, 1999).

Next, we consider testing multiple one-sided null hypotheses H_i , $i \in I$, in a h -stage adaptive group-sequential design using the error spending approach, with the option to adaptively select hypotheses in addition to mid-course sample size reassessment. Let D_t ($D_t \subseteq I$) denote a set of dropped hypotheses at time point $t = 1, 2, \dots, h-1$, e.g., due to safety concerns. In what follows, we introduce the marginal p -value combination approach using the same notations as for the case of a single hypothesis testing problem, except that we use the subscripts i where needed. As in Section 2.2, at time point t the marginal p -value combination approach rejects an intersection hypothesis

$H_J = \bigcap_{i \in J} H_i$, $J \subseteq I$, if $q_{i,t} < \alpha_{i,t}^*(w_i(J) \alpha)$ for at least one $i \in J_t = J \setminus D_t$. Based on the

results from Maurer and Bretz (2013b), it can be shown that this rejection rules defines a group-sequential test at level α for H_J , with the distinction that we use $q_{i,t}$ instead of the standard group-sequential p -values and that we also allow for hypotheses selection. Therefore, based on the closure principle, the marginal p -value combination approach rejects H_i while strongly controlling the FWER at level α if all intersection hypotheses H_J with $i \in J$ are rejected in addition to H_i .

Moreover, if the weights $w_i(J)$ are chosen so as to satisfy the monotonicity condition (1), then consonance is ensured (Maurer and Bretz, 2013b), since we assume well-ordered rejection boundaries as described above, such that $\alpha_{i,t}^*(w_i(U)\alpha) \geq \alpha_{i,t}^*(w_i(J)\alpha)$ holds for $w_i(U) \leq w_i(J)$ and $U \subseteq J$. Therefore, we can use the Bonferroni-based graphical approaches in adaptive group-sequential designs. In the next section, we introduce a general algorithm for the proposed graphical approach.

3 Graphical approaches for adaptive group-sequential trials

The graphical approach from Bretz et al. (2009) allows one to implicitly define weighted Bonferroni tests for each intersection hypothesis H_J , $J \subseteq I = \{1, 2, \dots, m\}$. The null hypotheses H_i , $i \in I$ are represented by nodes with associated weights $w_i(I)$ defining the local significance levels $w_i(I)\alpha$. The transition weight g_{ij} , $i, j \in I$ associated with a directed edge between any two vertices H_i and H_j indicates the fraction of the (local) significance level that is propagated to the hypothesis H_j at the head once the hypothesis H_i at the tail has been rejected. The g_{ij} 's have to be specified upfront in the trial protocol. They can be freely chosen subject to the regularity conditions $0 \leq g_{ij} \leq 1$, $g_{ii} = 0$ and $\sum_{i=1}^m g_{ii} \leq 1$ for all $i, j \in I$. Based on the results from Section 2, we can apply the algorithm from Maurer and Bretz (2013b) to adaptive group-sequential designs. More specifically, the following algorithm determines sequentially rejective graphical testing procedures based on consonant closed weighted Bonferroni tests using group-sequential boundaries in an adaptive clinical trial. In what follows, we use the notation from the preceding sections.

Algorithm

0. Set $t = 1$ and $I = \{1, 2, \dots, m\}$.
1. At interim analysis t , if $D_t \subsetneq I$, continue the trial, set $I \setminus D_t \rightarrow I$ and go to Step 2; otherwise stop. Compute combined p -values $q_{i,t}$ with the inverse normal combination function (3) and nominal significance levels $\alpha_{i,t}^*(w_i(I)\alpha)$ for $i \in I$.
2. Select a $j \in I$ such that $q_{j,t} \leq \alpha_{j,t}^*(w_j(I)\alpha)$, reject H_j and go to Step 3.
If no such j exists and $t < h$, the trial can be continued with $t \rightarrow t + 1$ and go to Step 1; otherwise stop.
3. Update the graph:

$$\begin{aligned}
 I &\rightarrow I \setminus \{j\} \\
 w_\ell(I) &\rightarrow \begin{cases} w_\ell(I) + w_j(I) g_{j\ell}, & \text{for } \ell \in I, \\ 0, & \text{otherwise,} \end{cases} \\
 g_{\ell k} &\rightarrow \begin{cases} \frac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \text{for } \ell, k \in I, \ell \neq k, g_{\ell j} g_{j\ell} < 1, \\ 0, & \text{otherwise.} \end{cases}
 \end{aligned}$$

4. If $|I| = 1$, go to Step 1; otherwise stop.

Note that, as shown in Bretz et al. (2009), the test decisions are independent of the rejection sequence. That is, if at any point more than one hypothesis could be rejected, the choice of the hypothesis does not influence the total set of hypotheses that eventually can be rejected. The initial graph and the algorithm unequivocally define the multiple test strategy. The algorithm above also allows for sample size reassessment without compromising the FWER, because the inverse normal combination method (3) is used for hypothesis testing. For example, Tamhane et al. (2012) discussed the case of one primary and one secondary hypothesis where the latter can only be tested after rejection of the primary hypothesis at one of the interim or final analyses, allowing also for sample size reestimation. The algorithm above reduces to this procedure if one would set $m = 2$, with the local significance level weights $w_1(\{1,2\}) = 1$ and $w_2(\{1,2\}) = 0$, and the transition weights $g_{12} = 1$ and $g_{21} = 0$.

4 Applications

In this section we illustrate the proposed graphical approach with examples from our clinical trial practice.

4.1 Example 1: Three-armed gold standard non-inferiority trials

For certain disease areas, such as asthma, panic disorder, or migraine, three-armed clinical trial designs are often used that include a test drug, an active control, and placebo. These trials are sometimes referred to as gold standard non-inferiority trial designs. Schlömer and Brannath (2013) considered a group-sequential design to test superiority of the test drug over placebo as well as non-inferiority of the test drug against the active control, while controlling the FWER in the strong sense. They called this design “full group-sequential design”. In some applications, structured primary and secondary endpoints may be evaluated in addition to the combined non-inferiority and superiority test problem. Among others, some authors (Hung and Wang, 2010; Lawrence, 2011; Guilbaud, 2011; Sugitani et al., 2013) considered tailored graphical approaches to appropriately address the associated multiplicity issues.

In the following, we show that the full group-sequential design by Schlömer and Brannath (2013) can be extended to more general situations when using the approach from Section 3. That is, their approach can be extended to incorporate testing of structured primary and secondary endpoints as well as sample size reallocation. In particular the possibility for sample size reallocation could be attractive in the context of testing non-inferiority: If, for example, the superiority of a test drug over placebo is demonstrated early during the course of a trial, then the trial sponsor has an option to reallocate the patients to either the test drug

and/or the active control group, while stopping the recruitment of patients to placebo. This would enable the trial sponsor to increase the power for testing non-inferiority of the test drug against the active control, while keeping the total sample size unchanged.

Let H_1 and H_3 denote the null hypotheses corresponding to the superiority of the test drug over placebo for a primary and a secondary endpoint, respectively. Similarly, let H_2 and H_4 denote the null hypotheses corresponding to the non-inferiority of the test drug against the active control for a primary and a secondary endpoint, respectively. Assume that hypothesis testing is performed at a single interim analysis and the final analysis based on some suitable error spending functions. Further assume that the initial graph from Figure 1 is used. This graph is based on the following considerations:

- (I) H_1 is of most importance: If superiority over placebo for the primary endpoint cannot be established, the trial is considered a failure.
- (II) Once H_1 has been rejected, H_2 and H_3 are of equal importance, and after H_2 has been rejected, H_3 and H_4 are of equal importance. In particular, H_4 is of least importance: Establishing non-inferiority against the active control for the secondary endpoint is nice to have, but not critical to the success of the trial.
- (III) Significance for a secondary endpoint should not be declared (either for the superiority or the non-inferiority claim) until significance for the primary hypothesis is declared for the same claim (successiveness property; Maurer et al., 2011).
- (IV) The multiple test procedure should be α -exhaustive. That is, at any step of the sequentially rejective test procedure the entire significance level α should be allocated across the remaining hypotheses.

Based on consideration (I), we initially allocate the entire significance level α to H_1 , i.e., $w_1(\{1, 2, 3, 4\}) = 1$ and $w_2(\{1, 2, 3, 4\}) = w_3(\{1, 2, 3, 4\}) = w_4(\{1, 2, 3, 4\}) = 0$. Based on consideration (II), we set the transition weights $g_{12} = g_{13} = 0.5$ and $g_{23} = g_{24} = 0.5$. Based on consideration (III), we have no directed edge from H_3 to H_4 and we set the transition weights $g_{32} = 1$, $g_{34} = 0$. Otherwise, if $g_{34} > 0$, we could not exclude the possibility to reject H_1 and H_3 (in this sequence), thus propagating some significance level to H_4 without having rejected H_2 before. Finally, based on consideration (IV), we set the transition weight $g_{43} = 1$, yielding the directed edge from H_4 to H_3 .

Assume now that at the interim analysis both H_1 and H_3 can be rejected. That is, we can conclude already at the interim analysis that the test drug is superior over placebo for both the primary and the secondary endpoint. In this case, recruitment of patients to placebo is stopped and patients recruited at the second stage are randomized to either test drug or active control. After rejecting both H_1 and H_3 the graph has to be updated as shown in Figure 2. At the final analysis we can then test both H_2 and H_4 at increased local significance levels without increasing the total sample size; see Figure 3.

4.2 Example 2: Two-stage adaptive design with dose selection at interim

The INHANCE study (Lawrence et al., 2014) was a randomized, double-blind, two-stage adaptive, parallel group study design to assess the efficacy of four doses of indacaterol (75, 150, 300, and 600 μg) against two active controls (formoterol and tiotropium) and placebo in patients with chronic obstructive pulmonary disease (COPD). The study design is shown in Figure 4. After the first stage, two of the four indacaterol doses were selected at an interim analysis based on pre-specified efficacy and safety criteria. The primary objective was to establish the superiority of at least one of the two selected indacaterol doses over placebo on trough Forced Expiratory Volume in 1 second (FEV_1) after 12 weeks of treatment. The key secondary objective was to demonstrate the non-inferiority of at least one of the two selected indacaterol doses to tiotropium for trough FEV_1 after 12 weeks treatment. Also, the study included an important secondary objective to show the superiority of at least one of the two selected indacaterol doses over placebo for days of poor control (symptomatic patient reported endpoint based on dairy data). In the following, however, for the sake of simplicity, we do not include it as a part of the multiple test strategy nor consider the comparisons with formoterol.

In this study, a Bonferroni adjustment with a significance level $\alpha/4$ was used for comparing each of the two selected doses against placebo, since the study started with four indacaterol doses. Here, α denotes the usual studywise significance level for confirmatory trials (i.e. $\alpha = 0.025$ for one-sided hypotheses testing). The primary and key secondary study objectives were tested sequentially at level $\alpha/4$ in the pre-specified hierarchy for each of the two selected doses (Maurer et al., 1995). Figure 5 visualizes the multiple test procedure employed in the INHANCE study using the graphical approach, where in the final analysis only the two test sequences associated with the selected doses A and B would be performed. This procedure controls the FWER in the strong sense, but is conservative because the local significance levels for the key secondary hypotheses are never propagated.

The proposed graphical approach from Section 3 enables us to employ a “loop back” strategy after treatment selection. That is, the local significance levels of rejected hypotheses can be recycled to test not yet rejected null hypotheses. Figure 6 visualizes two possible loop-back strategies, where r in graph (b) denotes a transition weight with $0 < r \leq 1$. The test strategy in (a) is a special case of (b) by setting $r = 1$. Strategy (a) is uniformly more powerful than the original proposal from Figure 5 because it offers the possibility to propagate the local significance levels from the secondary to the primary hypotheses. Strategy (b) is not uniformly more powerful than the original proposal, but could lead to higher power for the primary hypotheses for small values of r by recycling a large fraction of the significance level among the primary hypotheses before propagating them further to the secondary hypotheses. Note that the transition weights in Figure 6 are intentionally chosen such that the levels of higher doses are propagated to lower doses except for the lowest one. Also, using the results from Section 3, the proposed graphical approach allows the possibility for group-sequential testing of hypotheses or sample size reassessment, which was not possible with the original proposal implemented in the INHANCE study.

5 Numerical example

In this section we illustrate the proposed method with a hypothetical data example motivated by the INHANCE study. We use test strategy (b) from Figure 6 with $r = 0.5$ in a two-stage group-sequential design with dose selection at interim and one-sided significance level $\alpha = 0.025$. At the interim analysis, two of the four doses are selected based on pre-specified efficacy and safety criteria and are tested using the proposed graphical approach in connection with the associated rejection boundaries. All rejection boundaries are determined based on the Pocock-type error spending function. Also, the second-stage sample size is reassessed based on the observed interim treatment differences. At final analysis the selected doses are tested similarly using the updated graph and with the applicable rejection boundaries. Let n_t , $t = 1, 2$, denote the pre-planned sample size per stage. The interim analysis is performed after $N_1 = n_1 = 100$ patients per group and the final analysis is initially planned after $N_2 = n_1 + n_2 = 400$ patients per group. That is, the analyses are performed at information fractions $(\psi_1, \psi_2) = (100/400, 1) = (0.25, 1)$.

We assume that the true difference in trough FEV₁ versus placebo is (0.15, 0.18, 0.18, 0.16, 0.14) for the four doses of indacaterol and tiotropium, respectively, with a common standard deviation $\sigma = 0.3$. Hence, the observed treatment effects in trough FEV₁ are asymptotically normally distributed with the standardized effect size (0.50, 0.60, 0.60, 0.53, 0.47). The non-inferiority margin for the difference in trough FEV₁ is set as 0.04. The stagewise p -values are calculated based on Student's t -tests. The stagewise p -values are then combined using

the inverse normal combination function (3) with weights $\sqrt{\eta_1} = \sqrt{100/400} = 0.50$ and

$\sqrt{\eta_2} = \sqrt{300/400} \simeq 0.87$. Note that in treatment arm selection designs it is not always clear what sample size should be taken to define the weights. Wassmer (2011), however, showed that it is reasonable to use the planned sample sizes per treatment arm, that is,

$\sqrt{\eta_t} = \sqrt{n_t / (n_1 + n_2)}$, $t = 1, 2$, as also applied in the INHANCE study.

We now illustrate the proposed method with a hypothetical data example. Let initially $I = \{1, \dots, 8\}$. At the interim analysis $t = 1$, we select the two doses 150 μg and 300 μg . Thus, $D_1 = \{1, 4, 5, 8\}$ and we set $I \setminus D_1 \rightarrow I = \{2, 3, 6, 7\}$. The selected hypotheses are then tested using the updated graph displayed in the left panel of Figure 7. Assume that the observed unadjusted first-stage p -values are $(p_{2,1}, p_{3,1}, p_{6,1}, p_{7,1}) = (0.001, 0.001, 0.038, 0.004)$. The associated nominal levels from the Pocock-type error spending function are

$(\alpha_{2,1}^*(\alpha/4), \alpha_{3,1}^*(\alpha/4), 0, 0) = (0.003, 0.003, 0, 0)$ at information fraction $\psi_1 = 0.25$ for $\alpha = 0.025$. Thus, we can reject H_3 and the graph is update as shown in the middle panel of Figure 7. The updated nominal levels are

$(\alpha_{2,1}^*(3\alpha/8), -, 0, \alpha_{7,1}^*(\alpha/8)) = (0.005, -, 0, 0.002)$. Thus, we can also reject H_2 , resulting in the updated graph shown in the right panel of Figure 7. The updated nominal levels are

$(-, -, \alpha_{6,1}^*(3\alpha/16), \alpha_{7,1}^*(\alpha/8)) = (-, -, 0.003, 0.002)$ and no further hypotheses can be rejected. However, before proceeding to the second stage we assume that a sample size reassessment is performed. Based on the observed interim treatment effects it is likely that

the initially planned sample size $N_2 = 400$ makes the study overpowered and we assume that the actual second-stage sample size is reduced to $\tilde{N}_2 = 200$.

At the final analysis $t = 2$, no dose selection is performed so that we set $D_2 = \emptyset$ and thus $I \setminus D_2 \rightarrow I = \{2, 3, 6, 7\}$. Assume that the observed second-stage unadjusted p -values are $(-, -, p_{6,1}, p_{7,1}) = (-, -, 0.006, 0.001)$. Using the inverse normal method we thus obtain the combined p -values $q_{i,2} = 1 - \Phi\{0.50\Phi^{-1}(1 - p_{i,1}) + 0.87\Phi^{-1}(1 - p_{i,2})\}$, resulting in $(-, -, q_{6,2}, q_{7,2}) = (-, -, 0.002, 0.001)$. The nominal levels for the final analysis are $(-, -, \alpha_{6,2}^*(3\alpha/16), \alpha_{7,2}^*(\alpha/8)) = (-, -, 0.003, 0.002)$ at information time $\psi_2 = 1$; see the left panel in Figure 8. Thus, we reject H_7 , update the graph and finally also reject H_6 at the updated nominal levels $(-, -, \alpha_{6,2}^*(\alpha/4), -) = (-, -, 0.003, -)$; see the right panel in Figure 8. Note that the final analysis cannot be performed using an updated information fraction to account for the sample size reassessment, such as $\tilde{\psi}_2 = 200/400 = 0.50$, because this may lead to a Type I error rate inflation, as indicated in Section 2.3. The Appendix contains the R code used for our calculations. We used the `gsDesign` package (Anderson, 2014) in R for the calculation of the nominal significance levels; see, for example, Bretz et al. (2011a), for more details.

6 Discussion

In this paper, we discussed an adaptive group-sequential graphical approach, extending the work of Maurer and Bretz (2013b). The main advantage of the proposed approach is its simplicity and flexibility in practical situations. The extension is based on the marginal p -value combination approach. In Section 2, we have shown that the marginal p -value combination approach leads to consonant and thus sequentially rejective test procedures. This makes it easier to handle structured hypotheses in adaptive designs. For example, other than the group-sequential graphical approach by Maurer and Bretz (2013b), the marginal p -value combination approach would also enable the extension of other Bonferroni-based multiple test procedures, including fixed-sequence, fallback and several gatekeeping procedures. One possible improvement of the marginal p -value combination approach is to replace the Bonferroni test by Simes test. This topic deserves future research. The use of the Simes inequality for fixed designs in general does not guarantee consonance. This is also the case for group sequential or adaptive designs. Thus, we do not expect the graphical approach proposed in this paper to extend directly to other than Bonferroni-based multiple test procedures.

Alternative adaptive closed test procedures based on combination tests and conditional error rates are not so flexible in situations of structured hypotheses because they are in general not consonant; see, for example, Brannath and Bretz (2010). For example, one could consider an α -exhaustive graphical approach in two-stage adaptive designs, as described in Klinglmueller et al. (2014). While their approach leads to a slight gain in power, it loses the advantage of being a consonant multiple test procedure and thus an algorithm as described in Section 3 is not available.

In Section 4, we illustrated the methodology using two examples from our clinical trial practice. Other important applications would include, for example, adaptive enrichment designs (see, for example, Wassmer and Dragalin, 2014), involving multiple pre-specified subgroups, and/or primary and secondary endpoints. Furthermore, we focused in this paper on hypothesis testing although treatment effect estimation is also important. For example, research for interval estimation seems important, because no paper is applicable so far for constructing simultaneous confidence intervals for testing structured study objectives in confirmatory adaptive designs, while several publications, such as Posch et al. (2005) and Magirr et al. (2013), deal with simultaneous confidence intervals for testing unstructured study objectives in adaptive designs. Hence, developing simultaneous confidence intervals that are compatible with the proposed graphical approach is a topic for future research.

Acknowledgments

We thank two referees and the editor for the comments, which helped improve the presentation of the paper. Toshifumi Sugitani was supported by the Austrian Science Fund FWF – Project P23167.

Appendix

```
> library(mvtnorm)

> library(gsDesign)

>

> set.seed(5478)

>

> #- Setting design parameters -#

> diff_SUP <- c(0.18,0.18)

> diff_dose_tiotropium <- c(0.04,0.04)

> margin <- c(0.04,0.04)

> diff_NI <- diff_dose_tiotropium + margin

> sigma <- 0.3

> effsize_SUP <- diff_SUP/sigma

> effsize_NI <- diff_NI/sigma

>

> #- variance-covariance matrix -#
```

```

> sigma <- diag(2)

> sigma[1,2]<-sigma[2,1]<-0.5

>

> #- Stage 1 data -#

> n1 <- 100

> stagel_prim_data <- rmvnorm(1,mean=sqrt(n1/2)*effsize_SUP,sigma=sigma)

> stagel_sec_data <- rmvnorm(1,mean=sqrt(n1/2)*effsize_NI,sigma=sigma)

>

> p_prim <- 1-pnorm(stagel_prim_data)

> sprintf("%.3f",p_prim) # stage 1 p-value for H2 and H3

[1] "0.000" "0.000"

> p_sec <- 1-pnorm(stagel_sec_data)

> sprintf("%.3f",p_sec) # stage 1 p-value for H6 and H7

[1] "0.038" "0.004"

>

> #- Hypothesis testing at interim analysis -#

> boundary_H3 <- gsDesign(k=2,test.type=1,sfu="Pocock",
alpha=0.025/4,timing=0.25)

> boundary_H3

One-sided group sequential design with

90 % power and 0.625 % Type I Error.

Sample

Size

Analysis Ratio* Z Nominal p Spend

1 0.277 2.72 0.0033 0.0033

```

```

      2   1.110 2.72 0.0033 0.0030

      Total              0.0062

> # H3 rejected

> boundary_H2 <- gsDesign(k=2, test.type=1, sfu="Pocock", alpha=(0.025*3)/8,
timing=0.25)

> boundary_H2

One-sided group sequential design with

90 % power and 0.9375 % Type I Error.

      Sample

      Size

      Analysis Ratio* Z   Nominal p   Spend

      1   0.279 2.58 0.0049 0.0049

      2   1.116 2.58 0.0049 0.0044

      Total              0.0094

> # H2 rejected

> boundary_H7 <- gsDesign(k=2, test.type=1, sfu="Pocock", alpha=(0.025)/8,
timing=0.25)

> boundary_H7

One-sided group sequential design with

90 % power and 0.3125 % Type I Error.

      Sample

      Size

      Analysis Ratio* Z   Nominal p   Spend

      1   0.275 2.94 0.0016 0.0016

      2   1.099 2.94 0.0016 0.0015

```

```

Total                                0.0031

> # H7 not rejected

>

> #- Stage 2 data -#

> n2 <- 100

> stage2_sec_data <- rmvnorm(1, mean=sqrt(n2/2)*effsize_NI,sigma=sigma)

> inv_norm <- 0.5*stage1_sec_data + 0.81*stage2_sec_data

> p_comb <- 1-pnorm(inv_norm)

> sprintf("%.3f",p_comb) # Combined p-value for H6 and H7

[1] "0.002" "0.000"

>

> #- Hypothesis testing at final analysis -#

> boundary_H7

One-sided group sequential design with

90 % power and 0.3125 % Type I Error.

Sample

Size

Analysis Ratio* Z Nominal p Spend

1 0.275 2.94 0.0016 0.0016

2 1.099 2.94 0.0016 0.0015

Total                                0.0031

> # H7 rejected

> boundary_H6 <- gsDesign(k=2, test.type=1, sfu="Pocock", alpha=(0.025)/4,
timing=0.25)

> boundary_H6

```


One-sided group sequential design with

90 % power and 0.625 % Type I Error.

Sample

Size

Analysis Ratio* Z Nominal p Spend

1 0.277 2.72 0.0033 0.0033

2 1.110 2.72 0.0033 0.0030

Total 0.0062

> # H6 rejected

References

- [1]. Anderson, K. gsDesign: Group Sequential Design. 2014. R package version 2.8-8. URL <http://cran.r-project.org/web/packages/gsDesign/index.html>
- [2]. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*. 1999; 18:1833–1848. [PubMed: 10407255]
- [3]. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994; 50:1029–1041. [PubMed: 7786985]
- [4]. Brannath W, Bretz F. Shortcuts for locally consonant closed test procedures. *Journal of the American Statistical Association*. 2010; 105:660–669.
- [5]. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*. 2009; 28:1445–1463. [PubMed: 19266565]
- [6]. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*. 2009; 28:586–604. [PubMed: 19051220]
- [7]. Bretz, F.; Hothorn, T.; Westfall, P. *Multiple Comparisons Using R*. Chapman & Hall; Boca Raton, FL: 2011a.
- [8]. Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal*. 2011b; 53:894–913. [PubMed: 21837623]
- [9]. 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. [PubMed: 16972714]
- [10]. Burman C-F, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine*. 2009; 28:739–761. [PubMed: 19142850]
- [11]. Cui L, Hung HMJ, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics*. 1999; 55:853–857. [PubMed: 11315017]
- [12]. Di Scala L, Glimm E. Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine*. 2011; 30:3067–3081. [PubMed: 21898523]
- [13]. Dmitrienko A, Kordzakhia G, Tamhane AC. Multistage and mixture parallel gatekeeping procedures in clinical trials. *Journal of Biopharmaceutical Statistics*. 2011; 21:726–747. [PubMed: 21516566]

- [14]. Gao P, Liu L, Mehta C. Adaptive designs for noninferiority trials. *Biometrical Journal*. 2013; 55:310–321. [PubMed: 23494872]
- [15]. Guilbaud O. Note on simultaneous inferences about non-inferiority and superiority for a primary and a secondary endpoint. *Biometrical Journal*. 2011; 53:927–937. [PubMed: 22002473]
- [16]. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979; 6:65–70.
- [17]. Hommel G. Adaptive modifications of hypothesis after an interim analysis. *Biometrical Journal*. 2001; 43:581–589.
- [18]. Hommel G, Bretz F, Maurer W. Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine*. 2007; 26:4063–4073. [PubMed: 17348083]
- [19]. Hung HMJ, Wang SJ. Challenges to multiple testing in clinical trials. *Biometrical Journal*. 2010; 52:747–756. [PubMed: 20589856]
- [20]. Huque MF, Dmitrienko A, D'Agostino R. Multiplicity issues in clinical trials with multiple objectives. *Statistics in Biopharmaceutical Research*. 2013; 5:321–337.
- [21]. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal*. 1999; 41:261–277.
- [22]. Klinglmueller F, Posch M, Koenig F. Adaptive graph-based multiple testing procedures. *Pharmaceutical Statistics*. to appear.
- [23]. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70:659–663.
- [24]. Lawrence, D.; Bretz, F.; Pocock, SJ. INHANCE: An Adaptive Confirmatory Study with Dose Selection at Interim. In: Trifilieff, A., editor. *Indacaterol - The First Once-daily Long-acting Beta2 Agonist for COPD*. Springer; Basel: 2014. p. 77-92. Chapter 5
- [25]. Lawrence J. Testing non-inferiority and superiority for two endpoints for several treatments with a control. *Pharmaceutical Statistics*. 2011; 10:318–324. [PubMed: 20949636]
- [26]. Liu Q, Anderson M. On adaptive extensions of group sequential trials for clinical investigations. *Journal of the American Statistical Association*. 2008; 102:1621–1630.
- [27]. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999; 55:1286–1290. [PubMed: 11315085]
- [28]. Magirr D, Jaki T, Posch M, Klinglmueller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika*. 2013; 100:985–996. [PubMed: 27019516]
- [29]. Magnusson BP, Turnbull BW. Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine*. 2013; 32:2695–2714. [PubMed: 23315698]
- [30]. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976; 63:655–660.
- [31]. Maurer W, Bretz F. Memory and other properties of multiple test procedures generated by entangled graphs. *Statistics in Medicine*. 2013a; 32:1739–1753. [PubMed: 23280780]
- [32]. Maurer W, Bretz F. Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research*. 2013b; 5:311–320.
- [33]. Maurer W, Glimm E, Bretz F. Multiple and repeated testing of primary, co-primary and secondary hypotheses. *Statistics in Biopharmaceutical Research*. 2011; 3:336–352.
- [34]. Maurer, W.; Hothorn, L.; Lehmacher, W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In: Vollmar, J., editor. *Biometrie in der Chemisch-Pharmazeutischen Industrie*. Vol. 6. Fischer Verlag; Stuttgart: 1995. p. 3-18.
- [35]. 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. [PubMed: 11550941]
- [36]. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 5:549–556. [PubMed: 497341]
- [37]. Öhrn F, Jennison C. Optimal group-sequential designs for simultaneous testing of superiority and non-inferiority. *Statistics in Medicine*. 2010; 29:743–759. [PubMed: 19941286]

- [38]. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977; 64:191–199.
- [39]. 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. [PubMed: 16320264]
- [40]. Schlömer P, Brannath W. Group sequential designs for three-arm 'gold standard' non-inferiority trials with fixed margin. *Statistics in Medicine*. 2013; 32:4875–4889. [PubMed: 23970488]
- [41]. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine*. 2008; 27:6209–6227. [PubMed: 18792085]
- [42]. Sugitani T, Hamasaki T, Hamada C. Partition testing in confirmatory adaptive designs with structured objectives. *Biometrical Journal*. 2013; 55:341–359. [PubMed: 23576221]
- [43]. Tamhane AC, Wu Y, Mehta CR. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (II): sample size re-estimation. *Statistics in Medicine*. 2012; 31:2041–2054. [PubMed: 22733687]
- [44]. Wang SJ, Hung HMJ, Tsong Y, Cui L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine*. 2001; 20:1903–1912. [PubMed: 11427948]
- [45]. Wassmer G. On sample size determination in multi-armed confirmatory adaptive designs. *Journal of Biopharmaceutical Statistics*. 2011; 21:802–817. [PubMed: 21516570]
- [46]. Wassmer G, Dragalin V. Designing issues in confirmatory adaptive population enrichment trials. *Journal of Biopharmaceutical Statistics*. Published online: 6 Jun 2014.
- [47]. Wiens BL. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*. 2003; 2:211–215.
- [48]. Zehetmayer S, Bauer P, Posch M. Optimized multi-stage designs controlling the false discovery or the family-wise error rate. *Statistics in Medicine*. 2008; 27:4145–4160. [PubMed: 18444249]

Figure 1.
Initial graph for a three-armed gold standard non-inferiority trial.

Figure 2. Sequentially rejective graphical procedure at the interim analysis, after rejecting H_1 , H_3 and with the possibility of a sample size reallocation.

Figure 3. Sequentially rejective graphical procedure at final analysis, after rejecting H_1 , H_3 and sample size reallocation at the interim analysis.

Figure 4.

INHANCE study design (adapted from Lawrence et al. (2014)). Two selected indacaterol doses in stage 2 (which could be any two of the four indacaterol doses from stage 1) are denoted as A and B.

Figure 5.
Hierarchical test procedure used in the INHANCE study.

Figure 6.
Two possible loop-back strategies.

Figure 7.
Resulting graph after dose selection at the interim analysis for the hypothetical data example.

H_7 rejected

Figure 8.
Final analysis for the hypothetical data example.