Research Article

Statistics in Medicine

Received 12 September 2011, Accepted 15 November 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5700

A group sequential Holm procedure with multiple primary endpoints

Yining Ye,[a*†] Ai Li,[b] Lingyun Liu[c] and Bin Yao[b]

We propose a group sequential Holm procedure when there are multiple primary endpoints. This method addresses multiplicities arising from multiple primary endpoints and from multiple analyses in a group sequential design. It has been shown to be a closed testing procedure and preserves the familywise error rate in the strong sense. When multiple endpoints are the only concern without an interim analysis, the method simplifies to the weighted Holm procedure. The proposed method is more powerful than the parallel group sequential method and avoids the need to anticipate the testing order as in the fixed sequence testing scheme.

Keywords: multiplicity adjustment; multiple endpoints; group sequential designs; Holm procedure; familywise type I error rate

1. Introduction

Clinical trials are commonly designed with one primary endpoint and multiple secondary or exploratory endpoints. In some clinical settings, multiple primary endpoints are of interest, and inferences are made within the context of group sequential designs.

For inferences on multiple endpoints without a group sequential design, authors widely used the closed testing principle and Bonferroni-based methods. For n endpoints with the corresponding n hypotheses, $H_0^1, \ldots, H_0^n$, the closed testing principle [1] allows the rejection of any one of the elementary hypotheses, for example, H_0^i, if all possible intersection hypotheses involving H_0^i can be rejected by using valid local level α tests. It strongly controls the familywise error rate (FWER) for all the n hypotheses at level α. The Bonferroni procedure compares each raw p-value to a critical value of $\frac{\alpha}{n}$ in order to maintain the FWER at level α [2]. Authors have proposed adaptations of the Bonferroni methods [3, 4] to increase power. When some endpoints are deemed more important than others, one may consider the weighted Bonferroni procedure or its extension, the weighted Holm procedure [3]. When the endpoints can be ordered hierarchically, authors often apply gatekeeping procedures [5–8]. Wiens [9, 10] introduced the fallback procedure to evaluate a family of hypotheses in a fixed sequence setting.

Group sequential trials have the flexibility to stop early because of overwhelming evidence of efficacy, harm, or futility. They provide important safeguards to ensure that subjects are not unnecessarily exposed to harmful or ineffective therapies. Such designs offer additional incentives to researchers because of the potential savings on subjects, resources, and time. The characteristics and benefits of group sequential trials are well understood when a single primary endpoint is involved [11–14].

More recent research extended inferences in group sequential trials to multiple endpoints. Tang and Geller [15] proposed applying the closed testing principle using a global test procedure for handling multiple primary endpoints. Glimm *et al.* [16] and Tamhane *et al.* [17] considered group sequential designs where a single primary endpoint and a secondary endpoint were tested in a hierarchical order.

In this paper, we consider multiple primary endpoints in the context of group sequential designs where the objective is to seek regulatory approvals on at least one of the primary endpoints. This is to be

[a] *Global Biostatistical Sciences, Amgen Inc., South San Francisco, CA, U.S.A.*
[b] *Global Biostatistical Sciences, Amgen Inc., Thousand Oaks, CA, U.S.A.*
[c] *Cytel Corporation, Cambridge, MA, U.S.A.*
**Correspondence to: Yining Ye, Global Biostatistical Sciences, Amgen Inc., South San Francisco, CA, U.S.A.*
[†] *E-mail: yye@amgen.com*

distinguished from trials in some disease settings where the objective is to succeed in all primary endpoints in order to satisfy the regulatory requirements [18]. We have also broadened the definition of multiple endpoints in this paper to include multiple hypotheses of the same endpoint associated with different populations. When multiple primary endpoints are involved in a group sequential design, it is possible to draw conclusions on some primary endpoints at the interim analysis, but the trial needs to continue in order to evaluate other primary endpoints. In this situation, the benefits to address critical questions related to other primary endpoints outweigh the immediate savings in resources as seen in the examples we will discuss. We compare planned and average sample sizes of the proposed approaches with traditional designs.

We present two examples in drug development settings that have motivated our interest and research in such designs. In oncology, overall survival (OS) is considered the gold standard endpoint by regulatory authorities for approval. However, endpoints such as progression-free survival (PFS) are also acceptable for approval depending on the disease setting and regulatory authority. A common approach is to have PFS as the primary endpoint followed by OS as the secondary endpoint where the study is designed to have adequate power for both endpoints. Such a strategy is not without risks. For example, a recently approved oncology product [19] had demonstrated OS benefits but did not demonstrate PFS benefits. Depending on the situations, it may be desirable to conduct a global oncology trial with both OS and PFS as primary endpoints [20] so testing of OS does not depend on the outcome of PFS. In practice, the PFS endpoint is expected to be realized earlier than the OS endpoint, and one does not stop the trial when significant benefit is shown only on PFS. The second example is in developing treatment with potential predictive biomarkers. There may be *a priori* belief that there is treatment effect in a biomarker subpopulation, but it is uncertain whether there is treatment effect in the broader population. One may design a trial to investigate as primary objectives the treatment effects in both the overall population and the biomarker subpopulation [21,22]. To fully characterize whether the biomarker is predictive, it is critical to assess the treatment effect not just in the biomarker subpopulation.

The goal of this paper is to propose methods that strongly control the FWER among multiple primary endpoints or hypotheses in a group sequential setting. The proposed group sequential Holm procedures do not require prespecification of the testing sequence among multiple primary endpoints and offer flexibility to reallocate α once a hypothesis is rejected. Section 2 describes the method. We demonstrate the connection to the weighted Holm procedure. Section 3 extends the method to more than two primary endpoints. In Section 4, we apply the proposed methods to an actual clinical trial with two primary objectives. Simulation results with regard to power and average sample sizes follow in Section 5. We provide additional remarks and discussion in Section 6.

2. Methodology

Consider a clinical trial in which there are two primary endpoints denoted by A and B. The intent of the trial is to assess the treatment effect on either A or B or both. Suppose that J interim analyses including the final analysis are planned. Let $X_j(Y_j)$ be the Wald statistics for testing endpoint A (B) based on cumulative data up to look $j(j = 1,\ldots,J)$. Let $H_A(H_B)$ denote the null hypothesis of no treatment effect on the endpoint A (B). Let w_A and w_B be the prespecified weights with $w_A + w_B = 1$. Let $\alpha_A = w_A\alpha$ and $\alpha_B = w_B\alpha$ be the significance levels initially allocated to endpoints A and B. Let c_j and c'_j be the group sequential boundaries for endpoint A derived from some prespecified error spending approach at significance level α_A and α, respectively, such that $c_j \geqslant c'_j(j = 1,\ldots,J)$. Let d_j and d'_j be the corresponding boundaries for endpoint B at significance level α_B and α, respectively, such that $d_j \geqslant d'_j(j = 1,\ldots,J)$. The boundaries satisfy the following equations:

$$
\begin{aligned}
P\left(\cup_{j=1}^{J}\{X_j > c_j\}\right) &= \alpha_A \\
P\left(\cup_{j=1}^{J}\left\{X_j > c'_j\right\}\right) &= \alpha \\
P\left(\cup_{j=1}^{J}\{Y_j > d_j\}\right) &= \alpha_B \\
P\left(\cup_{j=1}^{J}\left\{Y_j > d'_j\right\}\right) &= \alpha
\end{aligned}
$$

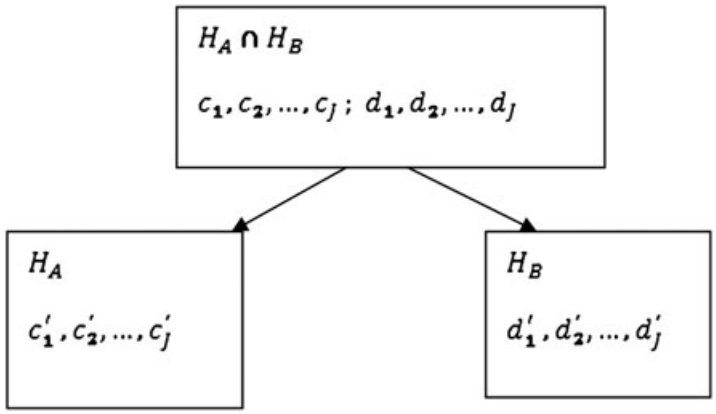
Consider the following group sequential design: monitor endpoint A using level α_A group sequential boundary c_j and endpoint B using level α_B group sequential boundary d_j. If either of the two endpoints crossed its corresponding boundary c_j or d_j, then the other endpoint can be tested using the full level α boundary. For example, if endpoint A crossed its level α_A boundary c_{j*} at some look j^*, then efficacy

with respect to endpoint A can be claimed and its type I error α_A can be reallocated to endpoint B so that endpoint B can be tested using full level α boundary d'_j where $j \geq j^*$. Similarly, if endpoint B crossed its level α_B boundary $d_{j^{**}}$ at some look j^{**}, then efficacy with respect to endpoint B can be claimed and its type I error α_B can be reallocated to endpoint A so that A can be tested using full level α boundary c'_j where $j \geq j^{**}$.

This type I error reallocation idea is similar to the α propagation idea [23]. The group sequential procedure strongly controls the type I error rate at level α because it is a closed test. To see this, consider the closed family $\{H_A \cap H_B, H_A, H_B\}$. The closed principle [1] states that the FWER will be strongly controlled if the following two requirements are met: (1) a local level α test is prespecified for each intersection hypothesis in the closed family, and (2) each individual hypothesis is rejected only if all the intersection hypotheses containing such particular individual hypothesis are rejected at their local α tests. For example, H_A is rejected only if both $H_A \cap H_B$ and H_A are rejected at local level α tests. For the proposed method, an α level global test for $H_A \cap H_B$ is as follows: reject $H_A \cap H_B$ if endpoint A crosses its level α_A group sequential boundary $c_j (j = 1, \ldots J)$ at any look or endpoint B crosses the level α_B group sequential boundary $d_j (j = 1, \ldots J)$ at any look. To see that this is a level α test for the intersection hypothesis $H_A \cap H_B$, note

$$\begin{aligned} P\,(\text{reject} H_A \cap H_B) &= P\left(\cup_{j=1}^{J}\left\{X_j > c_j\right\} \text{OR} \cup_{j=1}^{J}\left\{Y_j > d_j\right\}\right) \\ &\leqslant P\left(\cup_{j=1}^{J}\{X_j > c_j\}\right) + P\left(\cup_{j=1}^{J}\{Y_j > d_j\}\right) \\ &= \alpha_A + \alpha_B = \alpha \end{aligned}$$

The boundaries $c'_1, c'_2 \ldots, c'_J$ and $d'_1, d'_2, \ldots d'_J$ serve as local level α tests for H_A and H_B, respectively. We can illustrate this closed test by the following diagram:



The restriction on the boundaries $\left(c_j \geq c'_j \text{ and } d_j \geq d'_j, j = 1, \ldots, J\right)$ ensures that the closed test has a desirable property of being consonant [24]. This consonant property implies that if the intersection hypothesis $H_A \cap H_B$ is rejected, then at least one of the individual hypotheses H_A and H_B will be rejected by the closed test.

With regard to the boundary values for each endpoint, we can use different methods. For example, we can use O'Brien–Fleming boundaries for endpoint A and Pocock boundaries for endpoint B. After one hypothesis is rejected, one may continue using the predefined interim boundaries with $c'_j = c_j$ or $d'_j = d_j$ for $j < J$. In other words, the boundaries can be left unchanged at the interim analyses except at the final analysis J. Alternatively, c'_j or d'_j may be updated with completely different values after the α reallocation. We term the former group sequential Holm fixed (GSHf) and the latter group sequential Holm variable (GSHv). We will compare both methods with the naïve approach where α is split between the two endpoints each with independent group sequential procedures and no α reallocation. For ease of reference, we label the naïve approach as group sequential Bonferroni (GSB).

We can calculate the critical boundaries by using the Lan-DeMets error spending function [13]. Consider a simple situation, where one interim analysis and one final analysis are planned ($J = 2$). Let $\alpha(t)$, $\alpha_A(t)$, and $\alpha_B(t)$ be the nondecreasing function defined over $t \in [0, 1]$, such that $\alpha(0) = 0$,

$\alpha_A(0) = 0, \alpha_B(0) = 0, \alpha_A(1) = \alpha_A$, $\alpha_B(1) = \alpha_B$, and $\alpha(1) = \alpha$. Let t_{1A} be the information fraction for endpoint A at the interim analysis; t_{1B} be the information fraction for endpoint B at the interim analysis. Let $t_{2A} = t_{2B} = 1$ be the information fraction at the final analysis for endpoints A and B. We can calculate boundary values for endpoint A, c_1, c_1', c_2, and c_2', from the following two equations under the null hypotheses:

$$P(X_1 > c_1) = \alpha_A(t_{1A})$$

$$P\left(X_1 > c_1'\right) = \alpha(t_{1A})$$

$$P(X_1 > c_1) + \mathrm{P}\left(X_1 \leqslant c_1, X_2 > c_2\right) = \alpha_A(t_{2A}) = \alpha_A$$

$$P\left(X_1 > c_1'\right) + \mathrm{P}\left(X_1 \leqslant c_1',\ X_2 > c_2'\right) = \alpha(t_{2A}) = \alpha$$

If the O'Brien–Fleming boundary is desired, then an error spending function that approximates the O'Brien–Fleming boundary is given by [13]:

$$\alpha_{A_{OF}}(t_{1A}) = 2\Phi\left(-\frac{Z_{\alpha_A/2}}{\sqrt{t_{1A}}}\right) \tag{1}$$

where $\Phi(.)$ is the standard normal cdf and $Z_{\alpha_A/2}$ is the $\left(1 - \frac{\alpha_A}{2}\right)$ quantile of the standard normal distribution. Similarly, the error spending function that approximates the Pocock boundary is given by

$$\alpha_{A_{PO}}(t_{1A}) = \alpha_A ln\{1 + (e-1)t_{1A}\} \tag{2}$$

When $c_1' = c_1$, that is, the GSHf procedure is preferred, we can calculate the critical boundaries c_1, c_2, c_2' from the following equations under the null hypotheses:

$$P(X_1 > c_1) = \quad \alpha_A(t_{1A})$$

$$P(X_1 > c_1) + P\left(X_1 \leqslant c_1, X_2 > c_2\right) = \alpha_A(t_{2A}) = \alpha_A$$

$$P(X_1 > c_1) + P\left(X_1 \leqslant c_1, X_2 > c_2'\right) = \alpha(t_{2A}) = \alpha$$

For both the GSHv and GSHf procedures, similar calculations as mentioned previously can be performed to obtain boundaries d_1, d_1', d_2, and d_2'.

The proposed group sequential Holm methods simplify to the weighted Holm procedure [3] when only the final analysis is planned ($J = 1$). To see the connection, we can re-arrange the notation so that the boundary c_J corresponds to $\alpha_A = w_1\alpha$ and d_J corresponds to $\alpha_B = w_2\alpha$, where $w_1 + w_2 = 1$. The boundaries c_J' and d_J' correspond to significance level α. We further let $p_1^* = \frac{p_1}{w_1}$, $p_2^* = \frac{p_2}{w_2}$ be the weighted p-values, where p_1 and p_2 are the raw p-values for H_A and H_B, respectively. For simplicity and without loss of generality, we assume $p_1^* \leqslant p_2^*$. In the weighted Holm procedure, p_1^* is compared with α. Note that $\alpha = \frac{1}{w_1+w_2}\alpha$. If $p_1^* < \alpha$, then p_2^* is compared with $\frac{1}{w_2}\alpha$. In our proposed method, p_1 is compared with $w_1\alpha$. If $p_1 < w_1\alpha$, then p_2 is compared with α after the type I error reallocation. The two procedures are equivalent because $p_1^* < \alpha$ is equivalent to $p_1 < w_1\alpha$ and $p_2^* < \frac{1}{w_2}\alpha$ is equivalent to $p_2 < \alpha$. When $p_1^* \geqslant \alpha$, the weighted Holm procedure stops and will accept H_B. In the proposed method, despite $\mathrm{p}_1 \geqslant \mathrm{w}_1\alpha$, one can, in theory, continue testing H_B by comparing p_2 with $w_2\alpha$. Because $\frac{p_2}{w_2} \geqslant \frac{p_1}{w_1} \geqslant \alpha$, it is easily seen that $p_2 \geqslant w_2\alpha$, leading to acceptance of H_B as well.

3. Extension

We can extend the proposed method to the situation with more than two primary endpoints. Consider a clinical trial with K endpoints of interest. Denote the K hypotheses associated with the K endpoints by $H_1, H_2, \ldots, H_K$. Let $F^{(0)} = \{1, 2, \ldots, K\}$ be the index set for all the endpoints. Let $w_1^{(0)}, w_2^{(0)}, \ldots, w_K^{(0)}$ be the prespecified weights associated with these hypotheses such that $w_1^{(0)} +$

$w_2^{(0)} + \ldots + w_K^{(0)} = 1$. Let $\alpha_k^{(0)} = w_k^{(0)}\alpha (k = 1, \ldots, K)$ be the type I error allocated to H_k. Let $Z_{k1}, Z_{k2}, \ldots Z_{kJ}$ be the Wald statistics to test $H_k (k = 1, 2, \ldots, K)$ at each interim look. Let $c_{k1}^{(0)}, c_{k2}^{(0)}, \ldots, c_{kJ}^{(0)}$ be the group sequential boundary for $H_k (k = 1, 2, \ldots, K)$ at significance level $\alpha_k^{(0)}$ such that

$$P\left(\cup_{j=1}^{J}\{Z_{kj} > c_{kj}\}\right) = \alpha_k^{(0)} (k = 1, 2, \ldots, K)$$

Consider the following group sequential testing procedure.

Step 1. Test $H_k (k = 1, 2, \ldots, K)$ using the boundaries $c_{k1}^{(0)}, c_{k2}^{(0)}, \ldots, c_{kJ}^{(0)}$ with level $\alpha_k^{(0)}$. If none of the K endpoints crossed its boundary at any of the J looks, then retain all H_k $(k = 1, 2, \ldots, K)$ and stop testing. Otherwise, if any of the K endpoints crossed its boundary at one of the J looks, then efficacy with respect to this endpoint can be claimed. Let $j^{(1)}$ be the earliest interim look where at least one endpoint can be rejected; $k^{(1)}$ be the set of the m_1 endpoints that crossed their boundaries at look $j^{(1)}$; and $F^{(1)} = F^{(0)}\backslash\{k^{(1)}\}$ be the set for the remaining $K - m_1$ endpoints. Then, the significance level assigned to all the individual hypotheses $H_k\left(k \in F^{(1)}\right)$ will be updated as follows:

$$\alpha_k^{(1)} = \alpha_k^{(0)} + \sum_{i \in k^{(1)}} \alpha_i^{(0)} \times \frac{w_k^{(0)}}{1 - \sum_{i \in k^{(1)}} w_i^{(0)}} = \frac{w_k^{(0)}}{1 - \sum_{i \in k^{(1)}} w_i^{(0)}}\alpha, \quad \left(k \in F^{(1)}\right)$$

Let $w_k^{(1)} \equiv \frac{w_k^{(0)}}{1 - \sum_{i \in k^{(1)}} w_i^{(0)}}$. The updated significance level for the $H_k\left(k \in F^{(1)}\right)$ is simply $\alpha_k^{(1)} = w_k^{(1)}\alpha$. The boundaries for $H_k\left(k \in F^{(1)}\right)$ at significance level $\alpha_k^{(1)}$ will be updated using the error spending approach that satisfies the following equation:

$$P\left(\cup_{j=1}^{J}\left\{Z_{kj} > c_{kj}^{(1)}\right\}\right) = \alpha_k^{(1)}, \quad \left(k \in F^{(1)}\right)$$

Step 2. Test $H_k\left(k \in F^{(1)}\right)$ for any interim look $j (j \geq j^{(1)})$ using the updated boundary values $c_{k1}^{(1)}, c_{k2}^{(1)}, \ldots, c_{kJ}^{(1)}$. If none of the $K - m_1$ endpoints crossed the updated boundary values at any of the $j \geq j^{(1)}$ interim looks, then retain all $H_k\left(k \in F^{(1)}\right)$ and stop testing. Otherwise, if any of the $K - m_1$ endpoints crossed its boundary, then efficacy with respect to this endpoint can be claimed. Let $k^{(2)}$ be the set for the m_2 endpoints that crossed their boundaries at the earliest interim look $j^{(2)}$ after step 1, that is, $j^{(2)} \geq j^{(1)}$. Let $F^{(2)} = F^{(1)}\backslash\{k^{(2)}\}$. Then, the significance level assigned to all the individual hypotheses $H_k\left(k \in F^{(2)}\right)$ will be updated as follows:

$$\alpha_k^{(2)} = \alpha_k^{(1)} + \sum_{i \in K^{(2)}} \alpha_i^{(1)} \times \frac{w_k^{(1)}}{1 - \sum_{i \in k^{(2)}} w_i^{(1)}} = \frac{w_k^{(1)}}{1 - \sum_{i \in k^{(2)}} w_i^{(1)}}\alpha, \quad \left(k \in F^{(2)}\right)$$

Let $w_k^{(2)} \equiv \frac{w_k^{(1)}}{1 - \sum_{i \in k^{(2)}} w_i^{(1)}}$. The updated significance level for the $H_k\left(k \in F^{(2)}\right)$ is $\alpha_k^{(2)} = w_k^{(2)}\alpha$. The boundaries for $H_k\left(k \in F^{(2)}\right)$ at significance level $\alpha_k^{(2)}$ will be updated using the error spending approach, which satisfies the following equation:

$$P\left(\cup_{j=1}^{J}\left\{Z_{kj} > c_{kj}^{(2)}\right\}\right) = \alpha_k^{(2)}\left(k \in F^{(2)}\right)$$

Step i. Test $H_k\left(k \in F^{(i-1)}\right)$ for any interim look j $\left(\text{j} \geq j^{(i-1)}\right)$ using the updated boundary values, $c_{k1}^{(i-1)}, c_{k2}^{(i-1)}, \ldots, c_{kJ}^{(i-1)}$. If none of the $K - m_1 - m_2 - \cdots - m_{i-1}$ endpoints crossed its boundary, then retain all $H_k\left(k \in F^{(i-1)}\right)$ and stop testing. Otherwise, if any of the $K - m_1 - m_2 - \cdots - m_{i-1}$ endpoint crossed its boundary at some interim look $j^{(i)}$ $\left(j^{(i)} \geq j^{(i-1)}\right)$,

then efficacy with respect to this endpoint can be claimed. Let $k^{(i)}$ be the set for the m_i endpoints that crossed their boundary at the earliest interim look $j^{(i)}$, $j^{(i)} \geqslant j^{(i-1)}$. Let $F^{(i)} = F^{(i-1)} \backslash \{k^{(i)}\}$. Then, the significance level assigned to all the individual hypotheses $H_k\left(k \in F^{(i)}\right)$ will be updated as follows:

$$\alpha_k^{(i)} = \alpha_k^{(i-1)} + \sum_{i \in k^{(i)}} \alpha_i^{(i-1)} \frac{w_k^{(i-1)}}{1 - \sum_{i \in k^{(i)}} w_i^{(i-1)}} = \frac{w_k^{(i-1)}}{1 - \sum_{i \in k^{(i)}} w_i^{(i-1)}} \alpha, \quad \left(k \in F^{(i)}\right)$$

Let $w_k^{(i)} = \frac{w_k^{(i-1)}}{1 - \sum_{i \in k^{(i)}} w_i^{(i-1)}}$. The updated significance level for the $H_k\left(k \in F^{(i)}\right)$ is $\alpha_k^{(i)} = w_k^{(i)} \alpha$. The boundaries for $H_k\left(k \in F^{(i)}\right)$ at significance level $\alpha_k^{(i)}$ will be updated using the error spending approach. The updated set of boundaries satisfies the following equation:

$$P\left(\cup_{j=1}^{J} \left\{Z_{kj} > c_{kj}^{(i)}\right\}\right) = \alpha_k^{(i)} \quad \left(k \in F^{(i)}\right)$$

Continue the steps until all the endpoints are rejected or the final analysis is complete, whichever is earlier. Note that after type I error reallocation, the boundaries for a particular endpoint are calculated such that the following monotonicity condition are satisfied $c_{kj}^{(i_2)} \leqslant c_{kj}^{(i_1)}$ for $i_1 < i_2$ and all $j = 1, \ldots J$. Again, this monotonicity condition ensures the desired consonance property such that the group sequential test procedure with multiple endpoints admits such stepwise shortcut. For example, if $c_{kj}^{(i)}$ is fixed at $c_{kj}^{(0)}$ for $j < J$, then this is the GSHf method. Without the restrictions, it is the GSHv method.

4. Example

We apply the group sequential Holm methods proposed in Section 2 to an actual clinical trial. The MOtesanib Non-Small Cell Lung Cancer Efficacy and Tolerability (MONET1) study was a phase 3, placebo-controlled randomized oncology clinical trial [22]. The primary objectives of this study were to determine if motesanib in combination with chemotherapy would improve survival (1) in the overall study population and (2) in subjects with adenocarcinoma histology (adenocarcinoma subpopulation).

The type I error was split between the overall population (1.5%, one sided) and the adenocarcinoma subpopulation (1%, one sided). The study had 80% power requiring 742 deaths in the overall population to detect a hazard ratio of 1.25 and 80% power requiring 593 deaths in the adenocarcinoma subpopulation to detect a hazard ratio of 1.30. A total of 1060 subjects were enrolled including 70% with the adenocarcinoma histology. An interim analysis was planned when 50% of the total deaths occurred in the overall population. The number of deaths for patients with adenocarcinoma histology was also close to the 50% target in the subpopulation at the interim analysis. A negligible amount of type I error (0.00005, one sided) was assigned at the interim for each hypothesis in the original design.

To apply the GSHv method, we use the O'Brien–Fleming spending function as in Equation (1). The critical boundaries can be obtained by solving the following equations:

$$P(X_1 > c_1) = \alpha_{\mathrm{A}}(0.5) = 2\Phi\left(-\frac{\mathrm{Z}_{0.015}}{\sqrt{0.5}}\right)$$

$$P\left(X_1 > c_1^{'}\right) = \alpha(0.5) = 2\Phi\left(-\frac{\mathrm{Z}_{0.025}}{\sqrt{0.5}}\right)$$

$$P(X_1 > c_1) + P\left(X_1 \leqslant c_1,\ X_2 > c_2\right) = \alpha_{\mathrm{A}}(t_{2\mathrm{A}}) = \alpha_{\mathrm{A}} = 0.015$$

$$P\left(X_1 > c_1^{'}\right) + P\left(X_1 \leqslant c_1^{'},\ X_2 > c_2^{'}\right) = \alpha(t_{2\mathrm{A}}) = \alpha = 0.025.$$

It can be shown that $c_1 = 3.25, c_1^{'} = 2.96, c_2 = 2.18, c_2^{'} = 1.97$. Note that X_1 and X_2 are the log-rank statistics at interim and final analysis for the overall population. Similarly, $d_1 = 3.46, d_1^{'} = 2.96$, $d_2 = 2.33$, and $d_2^{'} = 1.97$ in the adenocarcinoma subpopulation.

Table I. Boundary values for the MONET1 trial.

	Overall population ($\alpha_A = 0.015$)		Adenocarcinoma subpopulation ($\alpha_B = 0.01$)	
Approach[a]	Interim	Final	Interim	Final
Original design	3.89	2.17	3.89	2.32
GSHv	3.25 (2.96[b])	2.18 (1.97[b])	3.46 (2.96[b])	2.33 (1.97[b])
GSHf	3.25	2.18 (1.96[b])	3.46	2.33 (1.96[b])
GSB	3.25	2.18	3.46	2.33

[a]Spending function approximating the O'Brien–Fleming boundary is used.
[b]Numbers in parentheses are the boundary values when α is reallocated.
GSHv, group sequential Holm variable method; GSHf, group sequential Holm fixed method; GSB, group sequential Bonferroni.

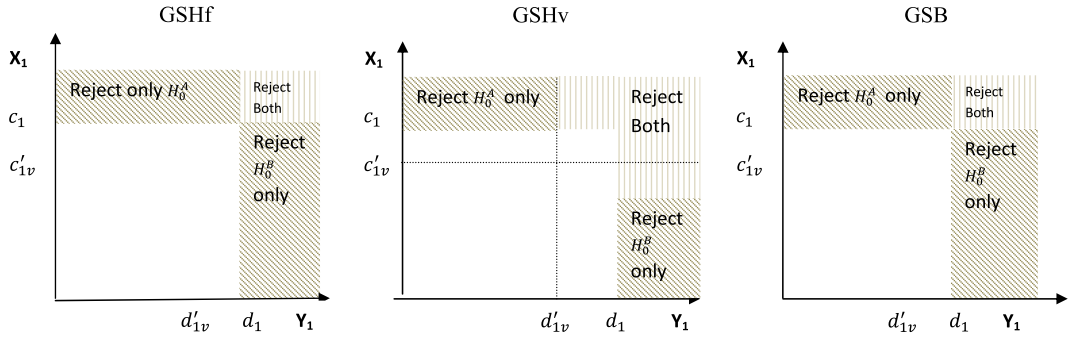


Figure 1. Rejection region at the interim analysis for MONET1. In MONET1 where α is split between the overall population and the adenocarcinoma subpopulation (0.015 and 0.01, one-sided respectively), one interim analysis is planned to occur when information fraction reaches 50% for overall population. The Lan-DeMets α spending function that approximates O'Brien–Fleming is used. $c_1 = 3.25$, $d_1 = 3.46$; $c'_{1v} = 2.96$, $d'_{1v} = 2.96$ for GSHv.

Similarly, if the GSHf is used with the same O'Brien–Fleming spending function, then critical boundaries can be obtained by solving

$$P(X_1 > c_1) = \alpha_A(0.5) = 2\Phi\left(-\frac{Z_{0.015}}{\sqrt{0.5}}\right)$$

$$P(X_1 > c_1) + P\left(X_1 \leqslant c_1,\ X_2 > c_2\right) = \alpha_A(t_{2A}) = \alpha_A = 0.015$$

$$P(X_1 > c_1) + P\left(X_1 \leqslant c_1,\ X_2 > c_2^{'}\right) = \alpha(t_{2A}) = \alpha = 0.025.$$

It can be shown that $c_1 = 3.25$, $c_2 = 2.18$, and $c_2^{'} = 1.96$ for the overall population. Similarly, we obtain $d_1 = 3.46$, $d_2 = 2.33$, and $d_2^{'} = 1.96$ in the adenocarcinoma subpopulation.

Table I summarizes the boundary values of the various methods. In addition, Figures 1 and 2 illustrate the boundaries and the rejection regions of GSHf, GSHv, and GSB. As shown in Figures 1 and 2C, the shaded areas representing rejection region for detecting significant effects for the overall population or subpopulation are identical for all methods. However, the rejection regions for detecting significant effects in both the overall and subpopulation are different, with the GSHf procedure having the largest area at the final analysis (Figure 2A–C) and the GSHv procedure having the largest area at the interim analysis (Figure 1).

The trial did not stop at the interim analysis. At the final analysis, the trial failed to demonstrate survival benefits of motesanib in either the overall population or the adenocarcinoma subpopulation [22].

5. Simulation

To evaluate the performance of the GSHf and GSHv methods, we conducted simulation studies to compare them with the GSB. We carried out the simulations within the contexts of each motivating example

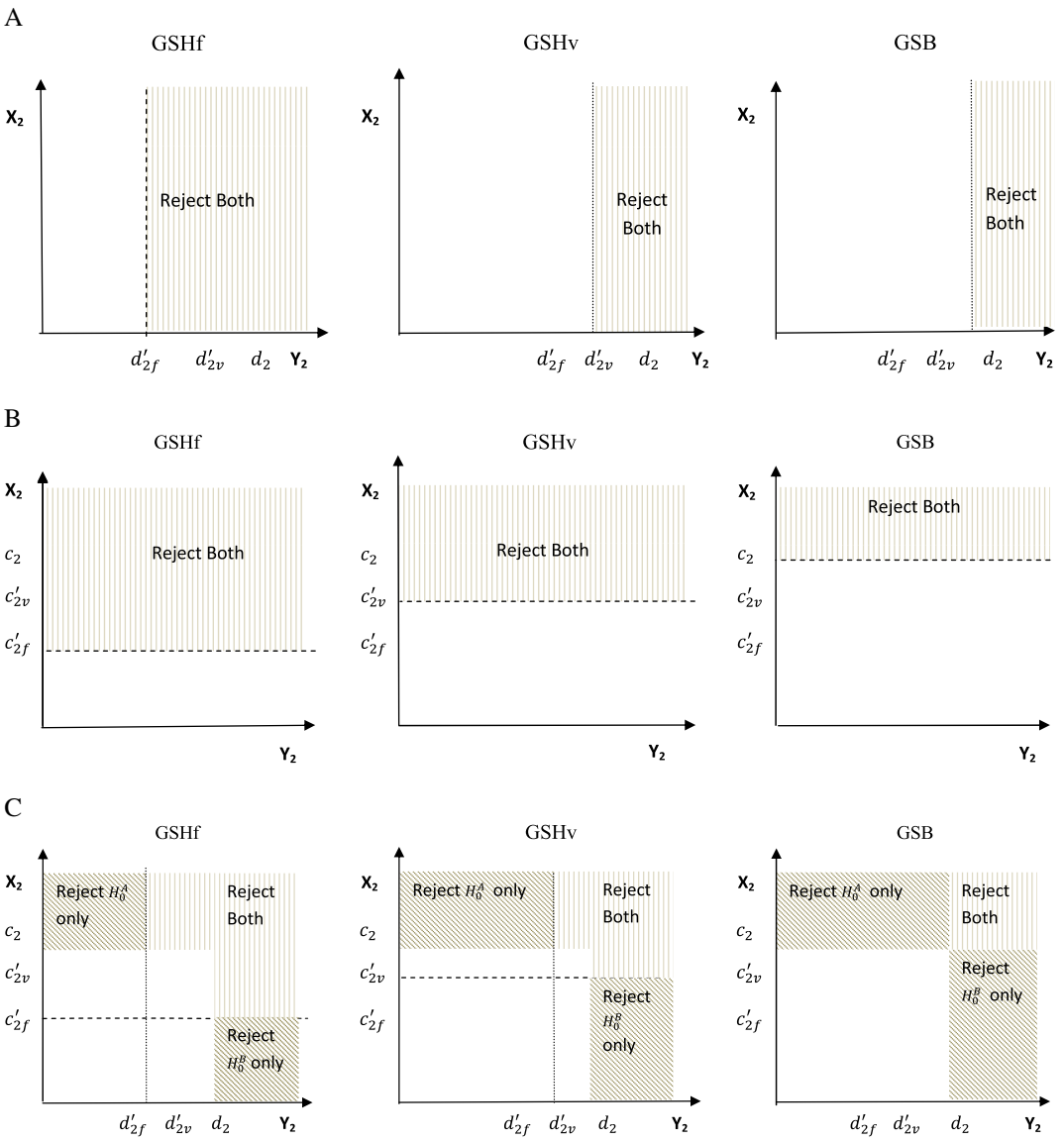


Figure 2. (A) Rejection region at the final analysis if H_0^{A} is rejected at the interim for MONET1. In MONET1, $d_2 = 2.33$, $d'_{2v} = 1.97$ for GSHv, $d'_{2f} = 1.96$ for GSHf for the subpopulation. (B) Rejection region at the final analysis if H_0^{B} is rejected at the interim for MONET1. In MONET1, $c_2 = 2.18$, $c'_{2v} = 1.97$ for GSHv, $c'_{2f} = 1.96$ for GSHf for the overall population. (C) Rejection region at the final analysis if neither H_0^{A} nor H_0^{B} is rejected at the interim for MONET1. In MONET1, $c_2 = 2.18$, $c'_{2v} = 1.97$ for GSHv, $c'_{2f} = 1.96$ for GSHf for the overall population; $d_2 = 2.33$, $d'_{2v} = 1.97$ for GSHv, $d'_{2f} = 1.96$ for GSHf for the subpopulation.

introduced in Section 1 where the overall α (two-sided in this section) was split equally between the two primary objectives. We used the α spending function that approximated Pocock boundaries to allow for a reasonable amount of rejections to occur at the interim. For the GSHv method, if one hypothesis was rejected at the interim, then the same Pocock-like spending function with the updated α was used to update the interim and final boundaries for the remaining hypothesis. We performed 10,000 simulations in each scenario.

In the first simulation study, we designed a hypothetical biomarker study similar to MONET1. We randomized a total of 1200 subjects to two treatment arms in a 1:1 ratio. The biomarker positive subpopulation accounted for 60% of the total population. Let A denote the biomarker positive subpopulation and $\bar{\mathrm{A}}$ denote the complementary subpopulation. We assumed the primary endpoint to have a normal distribution with details specified in Table II. Although the treatment effect was larger in the biomarker subpopulation, the effect was also present in the complementary population. We expected

Table II. Simulation settings for the hypothetical biomarker trial.

	Subpopulation A		Subpopulation $\bar{A}$	
Mean (SD)	H_1^A	H_0^A	$H_1^{\bar{A}}$	$H_0^{\bar{A}}$
Active	14 (27)	7 (27)	10.5 (36)	7 (36)
Control	7 (27)	7 (27)	7 (36)	7 (36)

H_1^A and $H_1^{\bar{A}}$ are the alternative hypotheses; H_0^A and $H_0^{\bar{A}}$ are the null hypotheses.

the study to have 85% power to detect the treatment difference in the subpopulation A and 82% power in the overall population based on the GSB. We separately tracked the power at the interim and final analyses for the subpopulation A and the overall population. The power to detect significant treatment effect in both populations and in either population is also provided. We provide the simulation results in Table III.

As expected, both GSHf and GSHv were more powerful than GSB in all situations. The power gains were most notable for demonstrating treatment effect in both populations. The power to detect significant treatment effect in at least one population was the same for all methods. These results are consistent with the observations from Figures 1 and 2 noted in Section 4. When comparing between the two Holm procedures, GSHv had more power at the interim analysis because α reallocation led to updated interim boundaries. But GSHv had less power at the final analysis and also less overall power (interim and final analyses combined). Because the hypothesis tests for the subpopulation and the overall population were correlated, with the correlation dependent on the proportion of the subpopulation to the overall population, we had conducted simulations under different assumptions of the correlation; the patterns described previously remained unchanged (data not shown).

We also evaluated the performance of the group sequential Holm methods by comparing the planned and average sample sizes with GSB under the alternative hypotheses specified in Table II. We added the fixed sample design (without interim analyses) as a benchmark in the simulations. For each method, we determined the planned sample size to ensure 85% power to detect treatment effect in the biomarker subpopulation. We derive the total sample size so that the biomarker subpopulation is at 60% of the total. We present the results in Table IV. As expected, the group sequential designs had larger planned sample sizes than the fixed sample design but lower average sample sizes because of the flexibility to stop at the interim. Compared with the GSB, the increase in the planned sample size was smaller for GSHv and GSHf because of the efficiency of the proposed methods. The average sample size for GSHv is the smallest as it is more likely to stop at the interim. The average sample size is also lower for GSHf compared with GSB. This is mainly due to the smaller planned sample size of the GSHf.

In the second simulation study, we designed a hypothetical oncology trial where OS and PFS were primary endpoints. We planned one interim analysis for OS when PFS had reached the final event goal. We randomized 420 subjects to two treatment arms with a 1:1 ratio. Under the GSB, we expected the trial to have 75% power ($\alpha = 0.025$) for PFS with 350 target events and 80% power ($\alpha = 0.025$) for OS with 320 target events. We assumed exponential distributions for the time to progression (TTP) and time to death (Table V).

We defined PFS as the time to progression or death, whichever was earlier. For comparison purposes, we included the GSHv, GSB, and the fixed sequence method where the PFS was tested first as a primary endpoint and OS was tested subsequently as a secondary endpoint (Table VI). It should be noted that in the fixed sequence method, $\alpha = 0.05$ for each endpoint, if applicable.

There were no surprises in the simulation results as GSHv dominated GSB. It is worth noting that when the treatment effect was present only for OS but not present for TTP (setting 2), there was no power gain for OS using GSHv (79%) as no additional alpha reallocation was expected because of the low power to reject PFS (TTP). However, the power for PFS, although low, was notably higher using GSHv compared with GSB (23% vs. 17%). This was because the rejection of OS at the interim analysis was possible, which could lead to updated more favorable boundaries for PFS. Compared with the fixed sequence method, GSHv had higher power on OS (settings 1 and 2) because of its ability to test OS irrespective of the PFS outcome. When the testing sequence was incorrectly specified (setting 2), the power loss for OS was dramatic for the fixed sequence (24%). The fixed sequence was the most powerful for PFS because $\alpha = 0.05$ was used. As a final note on the simulation results, the power for rejecting at least one hypothesis was the same for GSHv and GSB.

Table III. Simulation results for the hypothetical biomarker trial.

Settings	Approach[a]	Power for subpopulation[b]	Power for the overall population[b]	Power for subpopulation and overall population	Power for subpopulation or overall population
1. Treatment effect is present in	GSB	0.85 (0.51 + 0.34)	0.82 (0.48 + 0.35)	0.76	0.91
both the biomarker and the biomarker	GSHf	0.89 (0.51 + 0.39)	0.88 (0.48 + 0.4)	0.86	0.91
complementary populations	GSHv	0.88 (0.55 + 0.33)	0.86 (0.53 + 0.34)	0.83	0.91
2. Treatment effect is present	GSB	–	0.07 (0.04 + 0.03)[c]	–	–
only in the biomarker	GSHf	–	0.07 (0.04 + 0.04)[c]	–	–
complementary population	GSHv	–	0.07 (0.04 + 0.03)[c]	–	–
3. Treatment effect is present	GSB	0.85 (0.51 + 0.34)	0.54 (0.25 + 0.28)[c]	0.52	0.87
only in the biomarker	GSHf	0.86 (0.51 + 0.35)	0.67 (0.25 + 0.42)[c]	0.67	0.87
subpopulation	GSHv	0.86 (0.52 + 0.34)	0.64 (0.32 + 0.32)[c]	0.63	0.87

[a]Spending function approximating the Pocock boundary is used. $\alpha = 0.025$ for the biomarker subpopulation and the overall population, respectively, in all methods. GSB is the group sequential Bonferroni; GSHf is the group sequential Holm fixed method; and GSHv is the group sequential Holm variable method.

[b]Overall power (power at the interim analysis + power at the final analysis).

[c]Power for the overall population is provided given treatment effect in the biomarker subpopulation or the biomarker complementary subpopulation.

Table IV. Planned and average sample sizes for the hypothetical biomarker trial.

	Power[b]		Sample size	
Approach[a]	Biomarker subpopulation (%)	Overall population (%)	Planned	Average
Fixed sample design	85	82	1070	1070
GSB	85	82	1200	978
GSHf	85	83	1070	904
GSHv	85	83	1100	869

[a]GSB is the group sequential Bonferroni; GSHf is the group sequential Holm fixed method; and GSHv is the group sequential Holm variable method.
[b]The power for the biomarker subpopulation is fixed at 85% under the alternative hypotheses in Table II.

Table V. Simulation settings for the hypothetical oncology trial.

	Time to progression		Overall survival	
Median survival (months)	H_1^A	H_0^A	H_1^B	H_0^B
Active	12	9	18	12.5
Control	9	9	12.5	12.5

H_1^A and H_1^B are the alternative hypotheses; H_0^A and H_0^B are the null hypotheses.

Table VI. Simulation results for the hypothetical oncology trial.

Settings	Approach[a]	Power for PFS	Power for OS[b]	Power for PFS and OS	Power for PFS or OS
1. Treatment effect is present for both OS and TTP	Fixed sequence	0.84	0.76 (0.68 + 0.08)	0.76	0.84
	GSB	0.76	0.79 (0.67 + 0.13)	0.66	0.9
	GSHv	0.8	0.84 (0.73 + 0.11)	0.74	0.9
2. Treatment effect is present only for OS	Fixed sequence	0.25[c]	0.24 (0.23 + 0.01)	0.24	0.25
	GSB	0.17[c]	0.79 (0.63 + 0.16)	0.16	0.8
	GSHv	0.23[c]	0.79 (0.64 + 0.15)	0.23	0.8
3. Treatment effect is present only for TTP	Fixed sequence	0.3[c]	–	–	–
	GSB	0.21[c]	–	–	–
	GSHv	0.21[c]	–	–	–

[a]Spending function approximating the Pocock boundary is used. GSB is the group sequential Bonferroni; GSHv is the group sequential Holm variable method. $\alpha = 0.05$ for PFS followed by OS in the fixed sequence approach. $\alpha = 0.025$ for PFS and OS, respectively, in the GSB and GSHv.
[b]Overall power (power at the interim analysis + power at the final analysis).
[c]Power for progression-free survival (PFS) is provided given treatment effect in either time to progression (TTP) or overall survival (OS).

6. Discussion

In this article, we have proposed a general procedure to handle inferences related to multiple primary endpoints in group sequential designs. The group sequential Holm procedure is shown to be a closed testing procedure and controls the FWER in a strong sense when multiplicities arise from both multiple analyses over time and from multiple endpoints. The procedure simplifies to the weighted Holm procedure when there is no interim analysis. It is more efficient than the GSB because of the α reallocation after one hypothesis is rejected. The gains in power come primarily from being able to reject more than one hypothesis. This has important practical implications as illustrated in the oncology trial example where any gains in power to demonstrate treatment effect in OS after demonstrating an effect in PFS will be very desirable. It is worth noting that the method is not expected to have a power advantage for rejecting at least one hypothesis. The proposed method avoids the need to prespecify a test order as in the fixed sequence approach.

Similar to the traditional group sequential designs, the flexibility to stop a trial early leads to lower average sample size compared with the fixed sample designs. The tradeoff is the increased maximum or planned sample size in order to maintain power. For the group sequential Holm procedures, the study can only stop at the interim analysis when all primary hypotheses are rejected. When the GSHv procedure is used, the interim boundaries may be relaxed because of α reallocation. This procedure will lead to a higher probability of stopping at the interim analysis than the GSHf procedure.

The choice of the α spending function will also influence the probability of stopping a trial early. It is well known that O'Brien–Fleming-like boundaries require stronger evidence to stop a trial early compared with the Pocock-like boundaries in the traditional group sequential setting. The same principle applies in the settings we have discussed. In reality, when there are multiple primary endpoints, different spending functions may be chosen for different endpoints. We recommend prespecification of the spending functions for all endpoints prior to any analyses. When the information fraction cannot be determined for an endpoint at the interim analysis, interim boundaries based on the Bonferroni α split may be used instead of the α spending function. For example, when durable response rate is the primary endpoint in an oncology study, the final analysis is typically triggered after a specific follow-up time is reached. The total number of durable responders cannot be anticipated until the final analysis. As a result, the information fraction at the interim analysis may not be defined. Interim boundaries based on the Bonferroni method, while not efficient, presents no difficulties to the proposed method.

We have extended the proposed method to multiple primary endpoints and multilook designs. In our experience, it is more common to have trials with more than two looks than trials with more than two primary endpoints. When implementing the proposed methods for time-to-event endpoints, the timing of the analyses may be event driven causing asynchronous number of analyses for each endpoint. As seen in the oncology example with OS and PFS as primary endpoints, one PFS analysis and two OS analyses are planned. If an additional PFS interim analysis is desired, the design will have two PFS analyses and three OS analyses.

In the biomarker examples discussed in the paper, the hypotheses are set up to address the biomarker subpopulation and the overall population. A concern is that a significant treatment effect in the overall population is entirely driven by treatment effect in the biomarker subpopulation. An additional 'efficacy consistency requirement' may be stipulated so that the effect in the overall population is not allowed when the treatment effect in the biomarker and the biomarker complementary subgroups are inconsistent [25].

Finally, it should be noted that the proposed method ignores the correlation among the endpoints. Huque and Alosh [26] proposed a flexible fixed sequence procedure to take into account the correlation in group sequential trials. However, in most clinical settings, it is difficult to justify that a certain degree of correlation can be obtained reliably among endpoints. When it is possible to quantify the correlation (e.g., the biomarker subpopulation case) or when a bound on the correlation can be estimated, further gains on efficiency may be achieved. In these situations, additional research is needed to incorporate the correlation into the proposed procedure.

Acknowledgements

The authors would like to thank Dr. David Chang for stimulating discussions from the clinical and drug development perspectives at the time the paper was conceived. Drs. Ajit Tamhane and Steve Snapinn provided helpful reviews and comments during the development of the paper. Dr. Li Zhu provided careful reviews of the simulation code. We are also grateful to the anonymous referees for their helpful comments.

References

1. Marcus R, Peritx E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
2. Fisher RA. *The Design and Analysis of Experiments*. Oliver & Boyd: Edinburgh and London, 1935.
3. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
4. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**:800–802.
5. Maurer W, Hothorn L, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In *Biometrie in der Chemisch-Pharmazeutischen Industrie*, Vollmar J (ed.). Fischer Verlag: Stuttgart, 1995; 3–18.
6. Westfall PH, Krishen A. Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–40.

7. Dmitrienko A, Tamhane AC. Gatekeeping procedures with clinical trial applications. *Journal of Pharmaceutical Statistics* 2007; **6**:171–180.
8. Dmitrienko A, Tamhane AC. *Gatekeeping Procedures in Clinical Trials. Multiple Testing Problems in Pharmaceutical Statistics*. Taylor & Francis: Boca Raton, Florida, 2010; Chapter 5.
9. Wiens B. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* 2003; **2**:211–215.
10. Wiens B, Dmitrienko A. The fallback procedure for evaluating single family hypotheses. *Journal of Biopharmaceutical Statistics* 2005; **15**:929–942.
11. Pococok SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199.
12. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
13. Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **63**:655–660.
14. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987; **43**:193–200.
15. Tang DI, Geller NL. Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* 1999; **55**:1188–1192.
16. Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group-sequential trials. *Statistics in Medicine* 2009; **29**:219–228.
17. Tamhane AC, Mehta CR, Liu L. Testing a primary and a secondary endpoint in a group sequential design. *Biometrics* 2010; **66**:1174–1184.
18. Offen WW, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, Muirhead R, Stryszak P, Boddy A, Chen K, Copley-Merriman K, Dere W, Givens S, Hall D, Henry D, Jackson J, Krishen A, Liu T, Ryder SW, Sankoh AJ, Wang J, Yeh CH. Multiple co-primary endpoints: medical and statistical solutions: a report from the multiple endpoints expert team of the Pharmaceutical Research and Manufacturers of America. *Drug Information Journal* 2007; **41**:31–46.
19. Kantoff PW, Higano CS, Shore ND, Berger ER, Small EJ, Penson DF, Redfern CH, Ferrari AC, Dreicer R, Sims RB, Xu Y, Frohlich MW, Schellhammer PF. Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *The New England Journal of Medicine* 2010; **363**:411–422.
20. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R, Garbe C, Testori A, Maio M, Hogg D, Lorigan P, Lebbe C, Jouary T, Schadendorf D, Ribas A, O'Day SJ, Sosman JA, Kirkwood JM, Eggermont AMM, Dreno B, Nolop K, Li J, Nelson B, Hou J, Lee RJ, Flaherty KT, McArthur GA. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England Journal of Medicine* 2011; **364**:2507–2516.
21. Cappuzzo F, Ciuleanu T, Stelmakh L, Cicenas S, Szczesna A, Juhasz E, Esteban E, Molinier O, Brugger W, Melezinek I, Klingelschmitt G, Klughammer B, Giaccone G. Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: a multicentre, randomised, placebo-controlled phase 3 study. *The Lancet Oncology* 2010; **11**:521–529.
22. Scagliotti GV, Vynnychenko I, Park K, Ichinose Y, Kubota K, Blackhall F, Pirker R, Galiulin R, Ciuleanu TE, Sydorenko O, Dediu M, Papai-Szekely Z, Banaclocha NM, McCoy S, Yao B, Hei YJ, Galimi F, Spigel DR. International, randomized, placebo-controlled, double-blind phase III study of Motesanib plus Carboplatin/Paclitaxel in patients with advanced nonsquamous non–small-cell lung cancer: MONET1. *Journal of Clinical Oncology* 2012; **30**:2829–2836.
23. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; **28**:586–604.
24. Gabriel KR. Simultaneous test procedures – some theory of multiple comparisons. *The Annals of Mathematical Statistics* 1969; **40**:224–250.
25. Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* 2007; **26**:3535–3549.
26. Huque MF, Alosh MA. Flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference* 2008; **138**:321–335.