Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

Group Sequential Holm and Hochberg Procedures

Ajit C. Tamhane[1] | Dong Xi[2] | Jiangtao Gou[3]

[1]Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL, USA
[2]Statistical Methodology, Novartis Pharmaceuticals, East Hanover, NJ, USA
[3]Department of Mathematics and Statistics, Villanova University, Villanova, PA, USA

Correspondence
Ajit C. Tamhane, Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60208, USA. Email: atamhane@northwestern.edu

The problem of testing multiple hypotheses using a group sequential procedure (GSP) often arises in clinical trials. We review several group sequential Holm (GSHM) type procedures proposed in the literature and clarify the relationships between them. In particular, we show which procedures are equivalent or, if different, which are more powerful and what are their pros and cons. We propose a step-up group sequential Hochberg (GSHC) procedure as a reverse application of a particular step-down GSHM procedure. We conducted an extensive simulation study to evaluate the familywise error rate (FWER) and power properties of that GSHM procedure and the GSHC procedure and found that the GSHC procedure controls FWER more closely and is more powerful. All procedures are illustrated with a common numerical example, the data for which are chosen to bring out the differences between them. A real case study is presented to illustrate application of these procedures. R programs for applying the proposed procedures, additional simulation results and the proof of the FWER control of the GSHC procedure in a special case are provided in Supplementary Material.

KEYWORDS:
Closed procedure, Error spending function, Familywise error rate, Lan and DeMets flexible boundary, Multiple hypothesis testing, O'Brien-Fleming boundary, Pocock boundary

1 | INTRODUCTION

Modern phase 3 clinical trials often involve multiple testing problems arising due to multiple endpoints, multiple looks at data in group sequential trials, multiple treatment arms or doses of a drug, multiple subgroup analyses, etc. The FDA draft guidance on multiplicity[1] recommends controlling the type I familywise error rate (FWER) at a designated level α with respect to all sources of multiplicity. This paper focuses on an important class of problems at the interface of multiple testing and group sequential trials, in particular, on group sequential extensions of the Holm (1979) and Hochberg (1988) procedures.[2,3] This class of problems is of significant practical interest as these extensions offer more powerful alternatives to Bonferroni-type adjustments. The early papers that addressed multiple testing problems in group sequential trials include Tang, Gnecco and Geller (1989)[4], Follmann, Proschan and Geller (1994)[5] and Tang and Geller (1999).[6]

Ye et al. (2013),[7] Maurer and Bretz (2013)[8] and Fu (2018)[9] have proposed different group sequential extensions of the Holm step-down (GSHM) procedures. It is not immediately clear if these different GSHM procedures are equivalent or, if different, which ones are more powerful and what are their pros and cons. The main purpose of this paper is to present these procedures in a unified framework and notation in order to clarify the relationships between them, indicate any improvements and provide mathematical proofs of key results. We also propose a group sequential Hochberg (GSHC) step-up procedure by reversing a particular GSHM procedure.

The outline of the paper is as follows. Section 2 reviews the fixed sample Holm and Hochberg procedures. Section 3 does the same for group sequential procedures. Section 4 is the core of the paper. It presents the algorithms for the GSHM procedures proposed in the literature, studies relationships between them and offers some modifications. Section 5 proposes the GSHC procedure. Section 6 presents a simulation study of FWER and power properties of a particular GSHM procedure and the GSHC procedure. In Section 7 we apply the procedures to the data from a recent clinical trial. Section 8 describes the R programs for applying the proposed procedures. Section 9 discusses possible extensions of procedures discussed in the paper. Finally, Section 10 gives conclusions and recommendations. Proofs of some key results are given in Appendix. R programs for applying the proposed procedures, additional simulation results and the proof of the FWER control of the GSHC procedure in a special case are provided in Supplementary Material.

2 | FIXED SAMPLE HOLM AND HOCHBERG PROCEDURES

Consider testing $n \geq 2$ null hypotheses, $H_1, \ldots, H_n$, while strongly controlling[10] the FWER at a designated level α:

$$\text{FWER} = P\{\text{Reject at least one true } H_i\} \leq \alpha \tag{1}$$

under all possible configurations of the true and false H_i's. We assume that the hypotheses are non-hierarchical, i.e., there are no implication relations between them. Holm (1979)[2] refers to this as the free combinations condition.

Let $p_1, \ldots, p_n$ denote the p-values associated with $H_1, \ldots, H_n$, respectively. Further denote the ordered p-values by $p_{(1)} \leq \cdots \leq p_{(n)}$ and the corresponding hypotheses by $H_{(1)}, \ldots, H_{(n)}$. Each p_i is assumed to be marginally distributed as uniform $[0, 1]$ under H_i.

Both the Holm and Hochberg procedures use the same set of critical constants, but they test the ordered hypotheses in opposite directions. The Holm procedure operates in a so-called step-down sequence beginning by testing $H_{(1)}$ and rejecting it if $p_{(1)} < \alpha/n$ and continuing to test $H_{(2)}$; otherwise it retains all hypotheses and stops testing. In general, it tests $H_{(i)}$ and rejects it if and only if (iff) all $H_{(j)}$ for $j < i$ are rejected and $p_{(i)} < \alpha/(n - i + 1)$ and continues to test $H_{(i+1)}$; otherwise it retains all remaining hypotheses, $H_{(i)}, H_{(i+1)}, \ldots, H_{(n)}$, and stops testing.

The Hochberg procedure operates in a step-up sequence beginning by testing $H_{(n)}$ and retaining it if $p_{(n)} \geq \alpha$ and continuing to test $H_{(n-1)}$; otherwise it rejects all hypotheses including $H_{(n)}$ and stops testing. In general, it tests $H_{(i)}$ and retains it iff all $H_{(j)}$ for $j > i$ are retained and $p_{(i)} \geq \alpha/(n - i + 1)$ and continues to test $H_{(i-1)}$; otherwise it rejects all remaining hypotheses, $H_{(i)}, H_{(i-1)}, \ldots, H_{(1)}$, and stops testing. It is evident that the Hochberg procedure rejects all hypotheses rejected by the Holm procedure and possibly more, and so is uniformly more powerful.

3 | GROUP SEQUENTIAL PROCEDURES

Next we briefly review some basic background on group sequential procedures (GSPs). Whitehead (1997),[11] Jennison and Turnbull (2000)[12] and Wassmer and Brannath (2016)[13] provide comprehensive coverage of this topic. We will adopt the flexible boundary approach of Lan and DeMets (1983).[14] This approach employs an error spending function (e.s.f.) $f(t, \alpha)$, which is a monotone non-decreasing function defined on $t \in [0, 1]$, where t is the information time or fraction, with $f(0, \alpha) = 0$ and $f(1, \alpha) = \alpha$. A GSP tests a single null hypothesis H_0 based on the p-values p_k observed at successive information times t_k where $0 = t_0 < t_1 < \cdots < t_K = 1$. The information times t_k as well as the number of looks K are not assumed to be fixed in advance, only the planned total amount of information at the end of the trial. In the case of normally distributed outcomes the information time is proportional to the sample size and in the case of time-to-event outcomes it is proportional to the number of events. The null hypothesis H_0 is rejected at the first stage k (the terms "stage" and "look" are used interchangeably in this paper) when $p_k < \alpha_k$, where the critical boundary $(\alpha_1, \ldots, \alpha_K)$ satisfies the type I error probability requirement:

$$P_{H_0}\{\text{Reject } H_0\} = P_{H_0}\left\{\cup_{k=1}^{K}\left(p_k < \alpha_k\right)\right\} \leq \alpha. \tag{2}$$

Given the e.s.f. $f(t, \alpha)$ and the information times t_k, the "nominal levels" α_k can be determined by recursively solving the equations:

$$P_{H_0}\left\{\cap_{j=1}^{k-1}\left(p_j \geq \alpha_j\right) \cap \left(p_k < \alpha_k\right)\right\} = f\left(t_k, \alpha\right) - f\left(t_{k-1}, \alpha\right) \quad (k = 1, \ldots, K). \tag{3}$$

For example, $P_{H_0}\left\{p_1 < \alpha_1\right\} = \alpha_1 = f\left(t_1, \alpha\right)$. Given α_1, we can calculate α_2 from the equation $P_{H_0}\left\{\left(p_1 \geq \alpha_1\right) \cap \left(p_2 < \alpha_2\right)\right\} = f\left(t_2, \alpha\right) - f\left(t_1, \alpha\right)$ if we know the joint distribution of (p_1, p_2) under H_0, and so on. To evaluate the probability on the left

hand side of (3) at each k, we need to know the joint distribution of $(p_1, \ldots, p_K)$. To model this distribution, it is convenient to assume the normal theory setup. In this setup, the sequential test statistics Z_k are assumed to be $N(0,1)$ under H_0 with $\text{corr}\left(Z_i, Z_j\right) = \sqrt{t_i/t_j}$ for $i < j$. The one-sided p-values are given by $p_k = 1 - \Phi\left(Z_k\right)$, where $\Phi(\cdot)$ is the standard normal c.d.f. The null hypothesis H_0 is rejected at the first look k when $Z_k > c_k$. Analogous to (3), the critical constants c_k can be determined by recursively solving the equations:

$$P_{H_0}\left\{\cap_{j=1}^{k-1}\left(Z_j \le c_j\right) \cap \left(Z_k > c_k\right)\right\} = f\left(t_k, \alpha\right) - f\left(t_{k-1}, \alpha\right) \quad (k = 1, \ldots, K). \tag{4}$$

Then the c_k can be transformed to the α_k by letting $\alpha_k = 1 - \Phi\left(c_k\right)$. The critical boundary $\left(\alpha_1, \ldots, \alpha_K\right)$ or equivalently $\left(c_1, \ldots, c_K\right)$ is said to be induced by the e.s.f. $f(t, \alpha)$.

The classical O'Brien and Fleming (1979) (OBF)[15] and Pocock (1977) (POC)[16] GSPs use fixed boundaries $\left(c_1, \ldots, c_K\right)$. They are called fixed because the information times t_k and the number of looks K are assumed to be fixed in advance and consequently the critical constants c_k. The critical constants c_k are expressed parametrically in terms of a single constant $c > 0$ (different for OBF and POC). For the OBF boundary $c_k = c/\sqrt{t_k}$ while for the POC boundary $c_k = c$ for all $k = 1, \ldots, K$. The unknown $c > 0$ for each boundary can be determined to satisfy the type I error probability requirement. Then the c_k can be transformed to the α_k as indicated before. Approximations to the OBF and the POC boundaries are given by the following e.s.f.s:

$$\text{OBF: } f(t, \alpha) = 2\Phi\left(-z_{\alpha/2}/\sqrt{t}\right) \quad \text{and} \quad \text{POC: } f(t, \alpha) = \alpha \ln\{1 + (e-1)t\}. \tag{5}$$

A desirable property for the e.s.f. $f(t, \alpha)$ to have is that the critical constants α_k induced by it are monotonically nondecreasing (or the c_k are monotonically nonincreasing) in the significance level α. Maurer and Bretz (2013)[8] have given the conditions on $f(t, \alpha)$ to have this property. We will call such an e.s.f. as monotone. It can be shown that the OBF and POC e.s.f.s given above are monotone. We will assume that all e.s.f.s considered in the sequel are monotone.

4 | GROUP SEQUENTIAL HOLM (GSHM) PROCEDURES

We now consider the problem of testing multiple hypotheses in the group sequential setting and review three GSHM procedures that have been proposed in the literature: (i) the Ye et al. (2013)[7] procedure (referred to herein as the YLLY procedure), (ii) the Maurer and Bretz (2013)[8] graphical procedure (referred to herein as the MB procedure), and (iii) the Fu (2018)[9] procedure (referred to herein as the F procedure).

A different e.s.f. can be used to test each hypothesis H_i $(i = 1 \cdots, n)$, but for the sake of simplicity, we assume that the same e.s.f. $f(t, \alpha)$ is used for all GSPs. Furthermore, all hypotheses are tested at the same information times t_k $(1 \le k \le K)$. Let p_{ik} denote the p-value of hypothesis H_i at Stage k $(1 \le i \le n, 1 \le k \le K)$. We assume that the p_{ik} are marginally distributed as uniform $[0, 1]$ under H_i. Let $I = \{1, \ldots, n\}$ and $J \subseteq I$ denote the index set of the as yet unrejected hypotheses at any step of testing.

When testing hypotheses $H_i,\ i \in J$, the weights on hypotheses H_i are denoted by $w_i^{(J)} \ge 0$, which represent the relative importance of the hypotheses in the set J of hypotheses under test. The weights $w_i^{(J)}$ are normalized to sum to 1 and are assumed to satisfy the condition $w_i^{(J')} \ge w_i^{(J)}$ if $J' \subseteq J$. Bretz, Maurer and Hommel (2011)[17] have shown that this monotonicity condition ensures consonance[18] when using sequentially rejective weighted Bonferroni procedures. In the equal weights case, this condition is obviously satisfied.

Let $\{\alpha_{ik}^{(J)}, k = 1, \ldots, K\}$ denote a group sequential boundary of level $w_i^{(J)}\alpha$ to test any hypothesis $H_i,\ i \in J$. This boundary satisfies

$$P_{H_J}\left\{\cup_{k=1}^{K}\left(p_{ik} < \alpha_{ik}^{(J)}\right)\right\} = w_i^{(J)}\alpha, \tag{6}$$

where $H_J = \cap_{i \in J} H_i$. Analogous to (3), this boundary can be computed by recursively solving the equations:

$$P_{H_J}\left\{\cap_{j=1}^{k-1}\left(p_{ij} \ge \alpha_{ij}^{(J)}\right) \cap \left(p_{ik} < \alpha_{ik}^{(J)}\right)\right\} = f\left(t_k, w_i^{(J)}\alpha\right) - f\left(t_{k-1}, w_i^{(J)}\alpha\right) \quad (k = 1, \ldots, K). \tag{7}$$

Since we assume that the e.s.f. $f(t, \alpha)$ is monotone and weights $w_i^{(J)}$ are monotone in J, it follows that the critical constants $\alpha_{ik}^{(J)}$ are monotonically nondecreasing in J, i.e., $\alpha_{ik}^{(J')} \ge \alpha_{ik}^{(J)}$ for all k if $J' \subseteq J$. Therefore as the set of retained hypotheses becomes smaller, the critical boundary becomes more relaxed. In the case of equal weights and common e.s.f. $f(t, \alpha)$, the critical constants $\alpha_{ik}^{(J)}$ do not depend on i, so we denote them simply by $\alpha_k^{(J)}$. Furthermore they depend on J only through its cardinality $|J|$.

4.1 | Ye et al. (2013) GSHM Procedure

Ye et al. (2013)[7] proposed two variants of the GSHM procedure (GSHv and GSHf), which are based on the idea of recycling α from rejected hypotheses to unrejected ones (Bretz et al., 2009).[19] The GSHv procedure recycles α from rejected hypotheses to all stages of the group sequential boundaries of the unrejected hypotheses, thus modifying their entire boundaries. The GSHf procedure recycles α from rejected hypotheses only to the final stages of the group sequential boundaries of the unrejected hypotheses, thus modifying only their final stage critical constants. Xi and Tamhane (2015)[20] have generalized this to recycling α to unrejected hypotheses from any prespecified stage onward. In this paper we shall restrict to the GSHv procedure. The corresponding GSHM procedure (referred to as YLLY-1) is given in Algorithm 1. YLLY-1 is not feasible to implement in practice because it looks ahead to future p-values; so we propose its modification (referred to as YLLY-2) in Algorithm 2 (see Remarks 1 and 2).

Algorithm 1 (Ye et al. (YLLY-1) GSHM Procedure)

Initialization: Assign weights $w_i^{(I)} > 0$ to all hypotheses $H_i, i \in I$ such that $\sum_{i\in I} w_i^{(I)} = 1$. Let the set of unrejected hypotheses $J \leftarrow I$ and stage $k_{\text{current}} = k_{\text{new}} \leftarrow 1$.

Step 1: Check if the p-values vector $(p_{i1}, \ldots, p_{iK})$ for each $H_i, i \in J$ crosses its $w_i^{(J)}\alpha$-level critical boundary (6) at some stage $k \geq k_{\text{current}}$, i.e., whether $p_{ik} < \alpha_{ik}^{(J)}$ for $k \geq k_{\text{current}}$. If none of the p-values crosses its boundary at any stage $k \geq k_{\text{current}}$, retain all hypotheses in J and stop testing.

Step 2: Let $k_{\text{new}} \leftarrow \min\{k : p_{ik} < \alpha_{ik}^{(J)}, i \in J, k \geq k_{\text{current}}\}$ and the index set of rejected hypotheses be $R = \{i : p_{ik} < \alpha_{ik}^{(J)}, i \in J, k = k_{\text{new}}\}$. If $R = J$ or if $k_{\text{new}} = K$ then stop testing; otherwise go to Step 3.

Step 3: Let

$$w_i^{(J\setminus R)} \leftarrow \frac{w_i^{(J)}}{\sum_{i\in\{J\setminus R\}} w_i^{(J)}} \quad \text{for} \quad i \in \{J \setminus R\},$$

$J \leftarrow \{J \setminus R\}$, $k_{\text{current}} \leftarrow k_{\text{new}}$ and return to Step 1.

Remark 1. There is a discrepancy in Ye et al.'s (2013)[7] algorithm (given above as Algorithm 1 using a different notation). They define $\mathbf{R}$ to be the index set of the hypotheses rejected at look $k = k_{new}$, but at the beginning of each step of their algorithm they state that if any of the endpoints crossed its boundary at some stage $k \geq k_{current}$ (which could be later than k_{new}) then efficacy with respect to those endpoints can be claimed; in other words, those hypotheses can be rejected. We realize that a hypothesis whose p-values vector crosses the boundary at $k > k_{new}$ is not included in $\mathbf{R}$ at k_{new} because it is guaranteed to be rejected at a future stage since the critical values $\alpha_{ik}^{(J)}$ are nondecreasing in $\mathbf{J}$ as noted before at the end of Section 3. Yet, this may be confusing to a practitioner. In any case, YLLY-1 requires all future p-values to be available to implement it. We consider this to be a significant drawback since we lose the benefit of early stopping of a group sequential trial. The only situation where this could be useful is when it is desired to carry out a post-hoc analysis of an already conducted group sequential trial.

Remark 2. YLLY-1 is essentially a hypothesis-by-hypothesis testing procedure as each hypothesis is tested separately by looking ahead to its future p-values. In Algorithm 2 we modify it as a stage-by-stage testing procedure to make it practically implementable. To obviate the undesirable feature of k_{new} defined by looking ahead at future p-values beyond $k_{current}$, we propose an equivalent iterative procedure that if no hypothesis can be rejected at stage $k_{current}$, the procedure moves on to the next stage and retests the remaining hypotheses. This stage-by-stage testing nature of YLLY-2 is made clear in Step 2 of Algorithm 2. Note that Algorithm 2 does not require k_{new} since $k_{current}$ is updated in Step 2 to the earliest stage at which any of the remaining hypotheses can be rejected, which is k_{new} in Algorithm 1. Thus the final rejection decision regarding any hypothesis is the same using either algorithm. Nevertheless, we prefer the YLLY-2 since it does not involve looking ahead to future p-values. The following example illustrates how YLLY-1 and YLLY-2 give the same results but operate differently.

Algorithm 2 (Modified Ye et al. (YLLY-2) GSHM Procedure)

Initialization: Assign weights $w_i^{(I)} > 0$ to all hypotheses $H_i, i \in I$ such that $\sum_{i\in I} w_i^{(I)} = 1$. Let the set of unrejected hypotheses $J \leftarrow I$ and stage $k_{\text{current}} \leftarrow 1$.

Step 1: Check if the p-values vector $(p_{ik}, \ldots, p_{ik})$ for each $H_i, i \in J$ crosses its $w_i^{(J)}\alpha$-level critical boundary (6) at stage $k = k_{\text{current}}$, i.e., whether $p_{ik} < \alpha_{ik}^{(J)}$ for $k = k_{\text{current}}$.

Step 2: Let the index set of rejected hypotheses be $R = \{i : p_{ik} < \alpha_{ik}^{(J)}, i \in J, k = k_{\text{current}}\}$. If $R \neq \emptyset$ and $R \neq J$ then go to Step 3; if $R = \emptyset$ and $k_{\text{current}} < K$ then let $k_{\text{current}} \leftarrow k_{\text{current}} + 1$ and return to Step 1; otherwise stop testing.

Step 3: Let

$$w_i^{(J\setminus R)} \leftarrow \frac{w_i^{(J)}}{\sum_{i\in\{J\setminus R\}} w_i^{(J)}} \quad \text{for} \quad i \in \{J \setminus R\},$$

$J \leftarrow \{J \setminus R\}$ and return to Step 1.

Example 1 (YLLY-1 and YLLY-2 GSHM Procedures)
Consider testing three hypotheses, H_1, H_2 and H_3 at a familywise level $\alpha = 0.05$ using the same two-stage GSP (POC or OBF) for each hypothesis. Assume that the hypotheses are equally weighted. Further assume that the interim look is at the midpoint, i.e., $t_1 = 1/2$ and $t_2 = 1$. The p-values for the three hypotheses at the two looks are given in Table 1.

TABLE 1 p-Values for Examples 1,2 and 3

H_i	$k = 1$	$k = 2$
H_1	0.0005	0.0200
H_2	0.0050	0.0500
H_3	0.0120	0.0150

Since the hypotheses are assumed to be equally weighted, the $w_i^{(J)}\alpha = \alpha/|J|$ are independent of i. Table 2 gives the POC and OBF boundaries for levels $\alpha/|J|$ for $|J| = 1, 2$ and 3. Furthermore, since the critical boundaries are assumed to be the same (OBF or POC) for all three hypotheses, the critical constants $\alpha_{ik}^{(J)}$ do not depend on i but only on J (through $|J|$) and the Stage number k and hence are denoted as $\alpha_k^{(J)}$.

TABLE 2 Critical constants $\alpha_k^{(J)}$ for YLLY-1, YLLY-2, MBF and GSHC procedures using the POC and OBF boundaries (5) with $K = 2$, equispaced information times and equal weights ($\alpha = 0.05$)

$\lvert J\rvert$	$\alpha/\lvert J\rvert$	POC		OBF	
		$k = 1$	$k = 2$	$k = 1$	$k = 2$
1	$\alpha/1$	0.0310	0.0297	0.0056	0.0482
2	$\alpha/2$	0.0155	0.0139	0.0015	0.0245
3	$\alpha/3$	0.0103	0.0089	0.0007	0.0164

Let us first use the POC boundary. (We denote the iterations within the same step by letters a, b, etc., e.g., within Step 1 we have Step 1a, Step 1b, etc.)

YLLY-1 Step 1a: We have $|J| = 3$ and $k_{\text{current}} = 1$. The p-value vectors for H_1 and H_2 both cross the boundary for $\alpha/3$ at $k = 1$ since $p_{11} = 0.0005 < 0.0103$ and $p_{21} = 0.0050 < 0.0103$. The p-values vector for H_3 does not cross the boundary for $\alpha/3$ at either stage since $p_{31} = 0.0120 > 0.0103$ and $p_{32} = 0.0150 > 0.0089$. Hence $R = \{1, 2\}$, $J = \{3\}$ and $k_{\text{current}} = k_{\text{new}} = 1$.

YLLY-1 Step 1b: We have $|J| = 1$. The p-values vector for H_3 crosses the boundary for $\alpha/1$ at $k = 1$ since $p_{31} = 0.0120 < 0.0310$. Hence H_3 is rejected and testing stops. Thus all three hypotheses are rejected.

The YLLY-2 procedure gives the same results as does the YLLY-1 procedure in this case since H_1 and H_2 are both rejected at $k = 1$ resulting in $R = \{1, 2\}$ and $k_{\text{current}} = 1$. So the steps in YLLY-2 are the same as those in YLLY-1.

Next let us use the OBF boundary.

YLLY-1 Step 1a: We have $|J| = 3$ and $k_{\text{current}} = 1$. The p-values vector for H_1 crosses the boundary for $\alpha/3$ at $k = 1$ since $p_{11} = 0.0005 < 0.0007$ and the p-values vector for H_3 crosses the boundary for $\alpha/3$ at $k = 2$ (by looking ahead) since $p_{32} = 0.0150 < 0.0164$. The p-values vector for H_2 does not cross the boundary for $\alpha/3$ at either stage since $p_{21} = 0.0050 > 0.0007$ and $p_{22} = 0.0500 > 0.0164$. Only H_1 is rejected at this step; H_3 is temporarily retained but it is guaranteed to be rejected at a later step when it will be tested at a higher significance level, e.g., $\alpha/2$. Thus $k_{\text{new}} = 1$, $R = \{1\}$ and $J = \{2, 3\}$.

YLLY-1 Step 1b: We have $|J| = 2$ and $k_{\text{current}} = 1$. The p-values vector for H_3 crosses the $\alpha/2$ level boundary at $k = 2$ (by looking ahead) since $p_{32} = 0.0150 < 0.0245$. Again, H_3 is temporarily retained but it is guaranteed to be rejected at Stage 2. The p-values vector for H_2 does not cross the $\alpha/2$ level boundary either at $k = 1$ or at $k = 2$ since $p_{21} = 0.0120 > 0.0015$ and $p_{22} = 0.0500 < 0.0245$. So $k_{\text{new}} = 2$.

YLLY-1 Step 2a: We have $|J| = 2$ and $k_{\text{current}} = k_{\text{new}} = 2$. H_3 is rejected at this step but H_2 is not rejected as noted above. So $R = \{3\}$ and $J = \{2\}$ and we go the next step.

YLLY-1 Step 2b: We have $|J| = 1$ and $k_{\text{current}} = 2$. H_2 still cannot be rejected at this step as its p-value does not cross the boundary for $\alpha/1$ at $k = 2$ since $p_{22} = 0.0500 > 0.0482$ and testing stops.

Next let us use the YLLY-2 procedure.

YLLY-2 Step 1a: We have $|J| = 3$ and $k_{\text{current}} = 1$. As noted above, the p-value for H_1 crosses the boundary for $\alpha/3$ at Stage 1. The p-values for H_2 and H_3 do not cross the boundary for $\alpha/3$ at Stage 1. So YLLY-2 only rejects H_1 at this step. As a result, it sets $R = \{1\}$, $J = \{2, 3\}$ and $k_{\text{current}} = 1$.

YLLY-2 Step 1b: We have $|J| = 2$. The p-values for H_2 and H_3 do not cross the boundary for $\alpha/2$ at Stage 1 since $p_{21} = 0.0050 > 0.0015$ and $p_{31} = 0.0120 > 0.0015$. Hence no hypothesis can be further rejected at Stage 1 and testing proceeds to Stage 2.

YLLY-2 Step 2a: We have $|J| = 2$ and $k_{\text{current}} = 2$. The p-value for H_2 does not cross the boundary for $\alpha/2$ at Stage 2 since $p_{22} = 0.0500 > 0.0245$. The p-value for H_3 crosses the boundary for $\alpha/2$ at Stage 2 since $p_{32} = 0.0150 < 0.0245$. Hence we reject H_3 and update $J = \{2\}$.

YLLY-2 Step 2b: We have $|J| = 1$ and $k_{\text{current}} = 2$. H_2 still cannot be rejected at this step as its p-value does not cross the boundary for $\alpha/1$ at $k = 2$ since $p_{22} = 0.0500 > 0.0482$ and testing stops.

Thus, YLLY-1 and YLLY-2 reject the same hypotheses using either the POC or the OBF boundary, but operate differently as occurred for the OBF procedure in this example. Because YLLY-1 "looks ahead" and temporarily retains hypotheses that are only rejected at future stages, it is more complex to implement than YLLY-2. Therefore we prefer YLLY-2 in practice.

Remark 3. It is possible to modify YLLY-1 in another way by reconciling the discrepancy in the definition of R *noted in Remark 1 by including in* R *the hypotheses whose p-value vectors cross the boundary at any stage* $k \geq k_{current}$*. The resulting procedure (referred to as YLLY-3) will be strictly "look-ahead" and can be shown to control the FWER and more powerful than YLLY-1 = YLLY-2. The reason for its higher power is that it rejects more hypotheses in earlier stages (which would have been otherwise rejected in later stages) and hence smaller sets* J *of retained hypotheses. Therefore it uses more relaxed critical constants* $\alpha_{ik}^{(J)}$ *(because of the monotonicity property assumed for the e.s.f.* $f(t, \alpha)$*) and so rejects even more hypotheses in later stages. For instance, in the above example, in Step 1a of the YLLY-1 procedure using the OBF boundary, YLLY-3 would reject both* H_1 *and* H_3 *at* $k = 1$ *although the p-values vector for* H_3 *crosses the boundary at* $k = 2$*. Thus it would set* $R = \{1, 3\}, J = \{2\}$ *and* $|J| = 1$*. Then in Step 1b,* H_2 *would be tested at level* $\alpha/1$ *and would be rejected since* $p_{21} = 0.0050 < 0.0056$*. Thus YLLY-3 would reject all three hypotheses at Stage 1, while both YLLY-1 and YLLY-2 fail to reject* H_2 *at either stage. Even though YLLY-3 is more powerful, we do not recommend it because of its "look ahead" feature and hence skip other details such as the formal proofs of its FWER control and higher power than YLLY-1 = YLLY-2.*

4.2 | Maurer and Bretz (2013) and Fu (2018) GSHM Procedures

Maurer and Bretz (2013)[8] proposed a graphical procedure (referred to as the MB procedure) to test multiple hypotheses in a group sequential setting. The MB procedure operates according to Algorithm 3. The critical constants $\alpha_{ik}^{(J)}$ are the same as those used by the YLLY procedure and in the case of equal weights they can be taken from Table 2 for example. Both the MB procedure and the Fu (2018)[9] procedure discussed later in this section use stage-by-stage testing, so their algorithms are presented in a stagewise manner. Within each stage, testing is done in steps.

Algorithm 3 (Maurer and Bretz (MB) GSHM Procedure)

Initialization: Assign weights $w_i^{(I)} > 0$ to all hypotheses $H_i, i \in I$ such that $\sum_{i\in I} w_i^{(I)} = 1$. Also assign transition parameters $g_{ij}^{(I)} \geq 0, g_{ii}^{(I)} = 0$ for all $i, j \in I$ such that $\sum_{j\neq i} g_{ij}^{(I)} = 1$ for all $i \in I$. Let the index set of unrejected hypotheses $J \leftarrow I$ and stage $k \leftarrow 1$.

Stage k:

Step 1: Reject any $H_i,\ i \in J$ for which $p_{ik} < \alpha_{ik}^{(J)}$ and go to Step 3.

Step 2: If no $H_i,\ i \in J$ is rejected and $k < K$ then let $k \leftarrow k + 1$ and go to Step 1; otherwise stop testing.

Step 3: Update the graph:

$$w_j^{(J\setminus\{i\})} \leftarrow \begin{cases} w_j^{(J)} + g_{ij}^{(J)} w_i^{(J)} & j \in \{J \setminus \{i\}\} \\ 0 & \text{otherwise} \end{cases}$$

and

$$g_{j\ell}^{(J\setminus\{i\})} \leftarrow \begin{cases} [g_{j\ell}^{(J)} + g_{ji}^{(J)} g_{i\ell}^{(J)}]/[1 - g_{ji}^{(J)} g_{ij}^{(J)}] & g_{ji}^{(J)} g_{ij}^{(J)} < 1, j, \ell \in \{J \setminus \{i\}\} \\ 0 & \text{otherwise.} \end{cases}$$

Step 4: Let $J \leftarrow \{J \setminus \{i\}\}$. If $J \neq \emptyset$ then return to Step 1; otherwise stop testing.

Fu (2018)[9] proposed a GSHM procedure for testing multiple endpoints that adjusts the critical constants $\alpha_{ik}^{(J)}$ for the correlations between the endpoints by up-scaling them by a constant $\xi_I > 1$ determined numerically. The idea of up-scaling the critical constants was originally introduced by Xi et al. (2017).[21] Wolbers et al. (2019)[22] argued that Fu's adjustment does not necessarily control the FWER. In the following we ignore this adjustment and assume $\xi_I = 1$. The resulting procedure is described in Algorithm 4. The critical constants $\alpha_{ik}^{(J)}$ are the same as those used by the YLLY procedure and hence can be obtained from Table 2 in the case of equal weights.

Proposition 1. In the special case where all initial transition parameters $g_{ij}^{(I)}$ are equal to $1/(|I| - 1)$, the MB procedure is equivalent to the F procedure. Furthermore, both procedures are equivalent to the YLLY-2 procedure.

Remark 4. In this paper we shall only be concerned with the special case of equal transition parameters. Therefore we will treat the MB and the F procedures as equivalent and refer to them together as the MBF procedure. While the MB procedure rejects one hypothesis at a time, the F procedure rejects hypotheses in groups consisting of all those hypotheses whose p-values cross the boundary. However, the final rejection decisions are the same for both which makes them equivalent. Note that, as long as at least one hypothesis is rejected (i.e., $R \neq \emptyset$), both procedures continue testing any unrejected hypotheses within the same stage after updating the weights. Only after no hypotheses are rejected ($R = \emptyset$), they proceed to the next stage, where the unrejected (retained) hypotheses have another chance of being tested and rejected. This is the case with all algorithms given in this paper. In other words, rejection of a hypothesis is permanent but nonrejection (retention) is temporary until the final stage.

The F procedure can be equivalently formulated in terms of the ordered p-values; the resulting procedure is analogous to the fixed sample Holm procedure given in Section 2. Consider a set of hypotheses $H_i,\ i \in J$ to be tested at Stage k. Assume that the weights are equal, $w_i^{(J)}\alpha = \alpha/|J|$ for all $i \in J$, and so the critical boundaries are the same for all hypotheses. Order the p-values of the $|J|$ hypotheses as $p_{(1)k} \leq \cdots \leq p_{(|J|)k}$ and denote the corresponding ordered hypotheses by $H_{(1)k}, \ldots, H_{(|J|)k}$. If at least

Algorithm 4 (Fu (F) GSHM Procedure)

Initialization: Assign weights $w_i^{(I)} > 0$ to all hypotheses $H_i, i \in I$ such that $\sum_{i\in I} w_i^{(I)} = 1$.
Stage k:

Step 1: Reject all hypotheses $H_i,\ i \in J$ for which $p_{ik} < \alpha_{ik}^{(J)}$.

Step 2: Let R be the index set of the rejected hypotheses. If $R \neq \emptyset$ go to Step 3; if $R = \emptyset$ and $k < K$ then let $k \leftarrow k + 1$ and return to Step 1; otherwise stop testing.

Step 3: Let

$$w_j^{(J\setminus R)} \leftarrow \frac{w_j^{(J)}}{\sum_{i\in\{J\setminus R\}} w_i^{(J)}}, \quad j \in \{J \setminus R\}.$$

Step 4: Let $J \leftarrow \{J \setminus R\}$. If $J \neq \emptyset$ then return to Step 1; otherwise stop testing.

one p_{ik} crosses the $(\alpha/|J|)$-level boundary (6) at Stage k then clearly it will be the smallest $p_{(i)k}$, namely $p_{(1)k}$, i.e., $p_{(1)k} < \alpha_k^{(J)}$ where $\alpha_k^{(J)}$ is the kth component of the $(\alpha/|J|)$-level critical boundary and so we reject $H_{(1)k}$. Therefore the cardinality of the unrejected hypotheses set $\{(2), \ldots, (|J|)\}$ becomes $|J| - 1$. Next, we test the second smallest $p_{(i)k}$, namely $p_{(2)k}$, to see if it crosses the $(\alpha/(|J| - 1))$-level boundary at Stage k , i.e., if $p_{(2)k} < \alpha_k^{(\{(2),\ldots,(|J|)\})}$, where $\alpha_k^{(\{(2),\ldots,(|J|)\})}$ is the kth component of the $(\alpha/(|J| - 1))$-level critical boundary. Continuing this process as long as rejection occurs, $H_{(i)k}$ will be rejected if $p_{(i)k}$ crosses the $(\alpha/(|J| - i + 1))$-level boundary at Stage k, i.e., if $p_{(i)k} < \alpha_k^{(\{(i),\ldots,(|J|)\})}$ where $\alpha_k^{(\{(i),\ldots,(|J|)\})}$ is the kth component of the $(\alpha/(|J| - i + 1))$-level critical boundary. This procedure is described in Algorithm 5.

Note that the levels of the critical boundaries, $\alpha/(|J| - i + 1)$, are the same as the levels of the critical constants with which the $p_{(i)}$ are compared in a fixed sample Holm procedure. The critical constants $\alpha_k^{(\{(i),\ldots,(|J|)\})}$ needed for this procedure can be obtained from the same table as used for the YLLY-1, YLLY-2 and MBF procedures, e.g., Table 2, but the table must be entered with $|J| - i + 1$ equal to $|J|$ in that table. In particular, if $(i, |J|) \in \{(1, 1), (2, 2), (3, 3)\}$ then $|J| = 1$, if $(i, |J|) \in \{(1, 2), (2, 3)\}$ then $|J| = 2$ and if $(i, |J|) \in \{(1, 3)\}$ then $|J| = 3$. For example, when testing $p_{(1)1}$ for $|J| = 3$, we have $i = 1$ and $|J| - i + 1 = 3$, so we enter Table 2 for $\alpha/3$ and $k = 1$. When testing $p_{(2)2}$ for $|J| = 3$, we have $i = 2$ and $|J| - i + 1 = 2$, so we enter Table 2 for $\alpha/2$ and $k = 2$.

Algorithm 5 (Fu (F) GSHM Procedure Using Ordered p-Values)

Initialization: Assign equal weights $w_i^{(I)} = 1/|I|$ to all hypotheses $H_i,\ i \in I$. Let the index set of unrejected hypotheses $J \leftarrow I$ and stage $k \leftarrow 1$.
Stage k:

Step 1: Order the p-values $p_{(1)k} \leq \cdots \leq p_{(|J|)k}$. Denote the corresponding hypotheses by $H_{(1)k}, \ldots, H_{(|J|)k}$.

Step 2: Begin testing with $H_{(1)k}$ and continue testing as long as each $H_{(i)k}$ is rejected, i.e., as long as $p_{(i)k} < \alpha_k^{(\{(i),\ldots,(|J|)\})}$ for $i = 1, \ldots, |J|$, where $\alpha_k^{(\{(i),\ldots,(|J|)\})}$ is the kth component of the critical boundary of level $\alpha/(|J| - i + 1)$. If testing stops with $H_{(m)k}$ being the first hypothesis retained for $m \geq 1$ then retain all remaining hypotheses $H_{(m)k}, \ldots, H_{(|J|)k}$ and let the index set of rejected hypotheses be $R = \{(1), \ldots, (m - 1)\}$. If no hypothesis is rejected then let $R = \emptyset$. Go to Step 3.

Step 3: If $k = K$ or $R = J$ then stop testing. Otherwise, let $J \leftarrow \{J \setminus R\}$ and $k \leftarrow k + 1$ and go to Step 1.

Example 2 (MBF GSHM Procedure Using Ordered *p*-Values (Algorithm 5))
The *p*-values for this example are given in Table 1. The MBF procedure uses the same critical constants $\alpha_k^{(J)}$ given in Table 2 for the YLLY procedure.

Let us first use the POC boundary.
Stage 1: We have $|J| = 3$, $\alpha/|J| = \alpha/3$ and $k = 1$. The ordered *p*-values are

$$p_{(1)1} = p_{11} = 0.0005 < p_{(2)1} = p_{21} = 0.0050 < p_{(3)1} = p_{31} = 0.0120.$$

First H_1 is rejected since $p_{11} = 0.0005 < 0.0103$. So $|J| = 2$. Therefore H_2 is next rejected since and $p_{21} = 0.0050 < 0.0155$. Then we have $|J| = 1$, $\alpha/|J| = \alpha/1$ and $k = 1$. So H_3 is rejected since $p_{31} = 0.0120 < 0.0310$. Thus all three hypotheses are rejected and testing stops.

Next we use the OBF boundary.
Stage 1: We have $|J| = 3$, $\alpha/|J| = \alpha/3$ and $k = 1$. So H_1 is rejected since $p_{11} = 0.0005 < 0.0007$. Then we have $|J| = 2$, $\alpha/|J| = \alpha/2$ and $k = 1$. Neither H_2 nor H_3 can be rejected since $p_{21} = 0.0050 > 0.0015$ and $p_{31} = 0.0120 > 0.0015$. So we go to Stage 2.
Stage 2: We have $|J| = 2$, $\alpha/|J| = \alpha/2$ and $k = 2$. The ordered *p*-values are

$$p_{(1)2} = p_{32} = 0.0150 < p_{(2)2} = p_{22} = 0.0500.$$

So H_3 is rejected since $p_{32} = 0.0150 < 0.0245$. We now have $|J| = 1$, $\alpha/|J| = \alpha/1$ and $k = 2$. H_2 still cannot be rejected since $p_{22} = 0.0500 > 0.0482$ and testing stops.

Thus the MBF procedure rejects all three hypotheses using the POC boundary but rejects only H_1 and H_3 using the OBF boundary. Comparing with the results from Example 1, we see that MBF and YLLY-2 make the same rejection decisions using either the POC or the OBF boundary since the two procedures are equivalent as shown in Proposition 1.

4.3 | MBF Procedure with "Look-Back" Option

A "look-back" option can be added to the MBF procedure which allows testing and rejecting of hypotheses that had previously failed to be rejected in the previous stages. As an example, consider testing two hypotheses at $\alpha = 0.05$ using a two-stage GSP with POC boundaries having an interim look at the midpoint. The POC boundaries for $\alpha/1$ and $\alpha/2$ taken from Table 2 are:

$$(\alpha/1)\text{-level boundary} : (0.00310, 0.0297), \quad (\alpha/2)\text{-level boundary} : (0.0155, 0.0139).$$

Suppose that the *p*-value vectors for H_1 and H_2 are:

$$H_1 : \left(p_{11} = 0.0300, p_{12} = 0.0325\right), \quad H_2 : \left(p_{21} = 0.0200, p_{22} = 0.0125\right).$$

At Stage 1, neither hypothesis can be rejected since both $p_{11} = 0.0300$ and $p_{21} = 0.0200$ are > 0.0155. At Stage 2, H_2 can be rejected since $p_{22} = 0.0125 < 0.0139$. If we use the "look-back" option then we can also reject H_1 using the $(\alpha/1)$-level boundary at Stage 1 since $p_{11} = 0.0300 < 0.0310$.

Formally, we can implement the "look-back" option by modifying Step 1 in Algorithm 3 to "Reject any H_i, $i \in J$ for which $p_{ij} < \alpha_{ij}^{(J)}$ for any $j \leq k$ and go to Step 3."

Proposition 2. The "look-back" MBF procedure strongly controls the FWER in (1).

The proof of this proposition is based on the fact that it is a shortcut to the closed procedure[23] that uses a group sequential Bonferroni procedure to test each intersection hypothesis $H_J = \cap_{i \in J} H_i$ and this closed procedure is consonant.

Remark 5. Even though the "look-back" MBF procedure is more powerful than its competitors, it has an undesirable feature that it can reject a hypothesis based on less data at an earlier stage, but not based on more data at a later stage. Regulatory agencies are unlikely to accept such a result. For this reason Maurer and Bretz (2013)[8] ruled out the "look-back" option for what they called as "practical and interpretational reasons." However, they noted that Hampson amd Jennsion (2013)[24] have shown that the probabilities of such contradictory results are quite small for all situations of practical interest. In a sense the "look-back" option is opposite of the "look-ahead" option employed in the YLLY-1 procedure. Both options are more powerful, but have practical drawbacks and are not recommended.

5 | GROUP SEQUENTIAL HOCHBERG (GSHC) PROCEDURE

The MBF GSHM procedure using ordered p-values given in Algorithm 5 can be reversed to obtain a GSHC procedure, which uses a step-up testing sequence. For convenience, we assume equal weights. The resulting procedure is given in Algorithm 6.

Algorithm 6 (GSHC Procedure)

Initialization: Assign equal weights $w_i^{(I)} = 1/|I|$ to all hypotheses $H_i,\ i \in I$. Set $J \leftarrow I = \{1, \ldots, n\}$ and $k = 1$.

Step 1: Order the p-values $p_{(1)k} \leq \cdots \leq p_{(|J|)k}$. Denote the corresponding hypotheses by $H_{(1)k}, \ldots, H_{(|J|)k}$.

Step 2: Begin testing with $H_{(|J|)k}$ and continue testing each $H_{(i)k}$ for $i = |J|, |J|-1, \ldots, 1$ as long as each hypothesis is retained, i.e., as long as $p_{(i)k} \geq \alpha_k^{(\{(J),\ldots,(i)\})}$ for $i = |J|, |J|-1, \ldots, 1$, where $\alpha_k^{(\{(J),\ldots,(i)\})}$ is the kth component of the critical boundary of level $\alpha/(|J|-i+1)$. If testing stops with $H_{(m)k}$ being the first hypothesis rejected for $m \leq |J|$ then reject all remaining hypotheses $H_{(m)k}, \ldots, H_{(1)k}$ and let the index set of rejected hypotheses be $R = \{(m), \ldots, (1)\}$. If no hypotheses are rejected then let $R = \emptyset$. Go to Step 3.

Step 3: If $k = K$ or $R = J$ then stop testing. Otherwise, let $J \leftarrow \{J \setminus R\}$ and $k \leftarrow k+1$ and go to Step 1.

Example 3 (GSHC Procedure)

Let us continue with Example 2. First we use the POC boundary.

Step 1: We have $|J| = 3$. We begin testing with $p_{(3)1} = p_{31} = 0.0120$; here $i = 3, |J| - i + 1 = 1$ and so we use the $(\alpha/1)$-level boundary. Since $0.0120 < 0.0310$ we reject all three hypotheses and stop testing.

Next we use the OBF boundary.

Step 1: As before, we use the $(\alpha/1)$-level boundary. Since $p_{(3)1} = p_{31} = 0.0120 > 0.0056$, retain H_3. Next test $p_{(2)1} = p_{21} = 0.0050$ using the $(\alpha/2)$-level boundary. Since $0.0050 > 0.0015$, retain H_2. Next test $p_{(1)1} = p_{11} = 0.0005$ using the $(\alpha/1)$-level boundary. Since $0.0005 < 0.0056$, reject H_1.

Step 2: We have $|J| = 2$. We begin by testing $p_{(2)2} = p_{22} = 0.050$ using the $(\alpha/1)$-level boundary. Since $0.0500 > 0.0482$, retain H_2. Next test $p_{(1)2} = p_{32} = 0.0150$ using the $(\alpha/2)$-level boundary. Since $0.0150 < 0.0245$, reject H_3 and stop testing by retaining H_2.

To summarize, both the MBF and GSHC procedures reach the same decisions for this example. Both procedures reject all three hypotheses using the POC boundary and reject only H_1 and H_3 using the OBF boundary. Normally, the GSHC procedure may reject more hypotheses than the MBF procedure just as the fixed sample Hochberg procedure rejects more hypotheses than the Holm procedure.

We have proposed the GSHC procedure based on intuitive grounds as the reverse of the GSHM procedure, analogous to the relationship between the two procedures for the fixed sample case as discussed in Section 2. We don't yet have an analytical proof that it controls the FWER except in the special case $n = 2$ and $K \geq 2$, which is provided in Supplementary Material. However, that proof is not easily generalizable to $n > 2$. To assess whether the GSHC procedure controls the FWER in the general case and also to evaluate its power performance relative to GSHM, we conducted an extensive simulation study reported in the next section.

6 | SIMULATION STUDY

Consider testing $n \geq 2$ null hypotheses $H_i : \delta_i = 0$ against upper one-sided alternatives, where δ_i is a noncentrality parameter (n.c.p.) of interest defined as follows. Assume that the data for testing H_i are i.i.d. observations from a $N(\mu_i, \sigma_i^2)$ distribution and we are using a K-stage GSP with a cumulative sample size N_k at the kth stage $(k = 1, \ldots, K)$ common to all H_i. The n.c.p. δ_i for H_i equals $\delta_i = \delta_{iK} = \mu_i\sqrt{N_K}/\sigma_i$, which is its n.c.p. at the final look. Then the n.c.p. at the kth look is $\delta_{ik} = \mu_i\sqrt{N_k}/\sigma_i = \delta_i\sqrt{t_k}$, where the information time $t_k = N_k/N_K$ $(k = 1, \ldots, K)$. The test statistics Z_{ik} for testing H_i at stage k were generated from the

$N(\delta_{ik}, 1)$ distribution with $\text{corr}(Z_{ik}, Z_{i\ell}) = \sqrt{t_k/t_\ell}$ for $k < \ell$. In addition, we allowed a common correlation $\rho \geq 0$ between all pairs of test statistics (Z_{ik}, Z_{jk}) at any stage k from hypotheses H_i and H_j, namely $\text{corr}(Z_{ik}, Z_{jk}) = \rho$, which in turn implies that $\text{corr}(Z_{ik}, Z_{j\ell}) = \rho\sqrt{t_k/t_\ell}$ for $i \neq j, k < \ell$. All simulations were performed for equispaced information times in which case $t_k = k/K$ $(k = 1, \ldots, K)$ and so $\sqrt{t_k/t_\ell} = \sqrt{k/\ell}$. Finally, the p-values were obtained as $p_{ik} = 1 - \Phi(Z_{ik})$, where $\Phi(\cdot)$ is the standard normal c.d.f.

We investigated various scenarios involving the following combinations: $n = 2, 3, 4$, $K = 2, 3, 4, 5$, OBF and POC boundaries using the e.s.f.s (5) at level $\alpha = 0.05$, $\rho = 0.01(0.01)0.99$ and m true null hypotheses with $\delta_i = 0$ and $n - m$ false null hypotheses with $\delta_i = 2$, where $m = 0, 1, \ldots, n$. For example, for $n = 3$, we consider the configurations $(0, 0, 0)$, $(0, 0, 2)$, $(0, 2, 2)$, and $(2, 2, 2)$. For each combination, we report the simulation estimates of the FWER and the power to reject at least one false null hypothesis. For example, for the configuration $(0, 0, 2)$, the FWER is the probability to reject at least one of H_1 or H_2, while the power is the probability to reject H_3. Similarly, for the configuration $(0, 2, 2)$, the FWER is the probability to reject H_1, while the power is the probability to reject at least one of H_2 or H_3. Each FWER and power estimate is based on 10^6 independent replications. Due to space restrictions here we present the results only for $n = 3$ and for the OBF boundary.

The simulation estimates of FWER and power of both the MBF GSHM procedure (Algorithm 5) and the GSHC procedure (Algorithm 6) using the OBF boundary are plotted vs. ρ in Figure 1 and Figure 2, respectively. Note that there is no FWER plot for the configuration $(2, 2, 2)$ as there is no true null hypothesis. Similarly, there is no power plot for the configuration $(0, 0, 0)$ as there is no false null hypothesis. From Figure 1 it can be seen that the FWER is controlled at level $\alpha = 0.05$ in all cases for both GSHM and GSHC procedures. In addition, the FWER of GSHC is closer to the designated $\alpha = 0.05$ in all cases; i.e., GSHM controls FWER more conservatively. In Figure 2 the power advantage of GSHC over GSHM can be seen to increase with ρ. In the configuration $(0, 2, 2)$, the maximum difference in power is 0.028 when $\rho = 0.99$. In the configuration $(2, 2, 2)$, the maximum difference in power is 0.122 when $\rho = 0.99$. The power difference between GSHC and GSHM appears to increase as the number of false null hypotheses increases.

The results for other cases ($n = 2, 4, 5$) as well as for the POC boundary are included in the Supplementary Material. They are consistent with those in Figures 1 and 2. The FWER is controlled at level $\alpha = 0.05$ for both GSHM and GSHC with GSHC controlling the FWER more closely to the designated α. GSHC is at least as powerful as GSHM in all cases. In terms of the comparison between OBF and POC, the results show a power advantage of OBF over POC in all cases. A similar result for testing a single null hypothesis was given in Theorem 1 of Tamhane et al. (2018)[25]. The FWER of OBF is closer to $\alpha = 0.05$ than that of POC in a majority of cases.

7 | CASE STUDY: CANTOS CLINICAL TRIAL

Ridker et al. (2017)[26] report the results of a double blind clinical trial (called the CANTOS trial) of canakinumab for cardiovascular disease. The primary efficacy endpoint was the first occurrence of nonfatal myocardial infarction, nonfatal stroke or cardiovascular death in a time-to-event analysis. There were two secondary endpoints but we will not consider them in the present analysis. There were three dose groups: 50 mg, 150 mg and 300 mg, which were compared to the placebo group. Randomization was done in the ratio of 1.5 (placebo):1:1:1. The trial was planned to enroll 17,200 patients in order to accrue 1400 events. Later, the sample size was reduced to 10,000 patients but the follow-up period was extended by one year to meet the target number of events.

One-sided Dunnett (1955)[27] comparisons were made at the two interim looks when 50% and 75% of the targeted events had occurred but none of the doses were found to be significantly more effective than placebo at $\alpha = 0.025$. (Ridker et al. (2017)[26] report two-sided p-values and make comparisons at $\alpha = 0.05$ as per the journal policy. The one-sided p-values and the corresponding critical values are obtained by dividing by 2.) Hence the trial was continued to the final look when the comparisons were repeated. The total $\alpha = 0.025$ was split with $\alpha_1 = 0.0001, \alpha_2 = 0.0004$ and $\alpha_3 = 0.0245$ between the three looks. Furthermore, at each look the available α was divided unequally between the three doses using weights 0.4 for each of the 50 mg and 150 mg vs. placebo comparisons and 0.2 for the 300 mg vs. placebo comparison. Denote the three null hypotheses by H_1, H_2 and H_3, respectively. Thus, at the final look, H_1 and H_2 were tested at $\alpha = 0.4 \times 0.0245 = 0.0098$ and H_3 was tested at $\alpha = 0.2 \times 0.0245 = 0.0049$. At the final look, no significant effect was observed in the 50 mg group (hazard ratio vs. placebo, 0.93; $p = 0.150$). A significant effect was observed in the 150 mg group (hazard ratio vs. placebo, 0.85; $p = 0.0104$, with a threshold p-value of 0.0106). In the 300 mg group, the p-value did not meet the prespecified threshold for significance

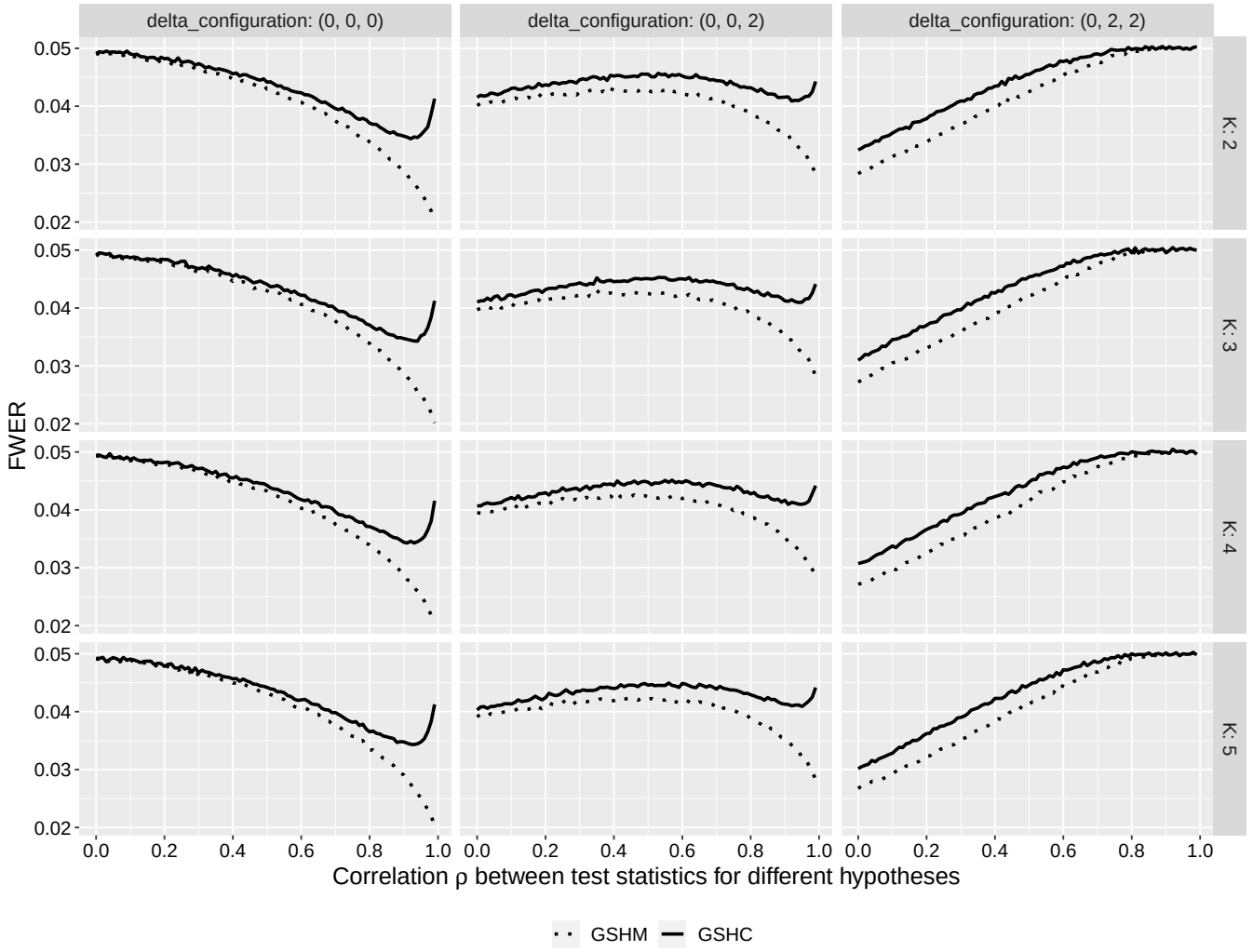

FIGURE 1 FWER for GSHM (Algorithm 5) and GSHC (Algorithm 6) using OBF with $n = 3$ hypotheses and $K = 2, 3, 4, 5$ stages.

even though the hazard ratio was slightly higher than the 150 mg dose group (hazard ratio vs. placebo, 0.86; $p = 0.0157$, with a threshold p-value of 0.0106). Thus only the 150 mg dose was declared significantly more effective compared to placebo.

We will use this case study to illustrate the application of the GSHM with ordered p-values using Algorithm 5 and GSHC procedures. We recognize that these procedures do not exploit the structure of the doses vs. placebo comparisons employed in the Dunnett procedure. However, we did not find a suitable case study, e.g., one involving multiple endpoints.

We made a few additional changes in the trial's setting. First, instead of the Bonferroni split of α among the three looks, we will use a more powerful O'Brien-Fleming error spending function (5). Second, instead of allocating the available α at each look unequally between the three dose vs. placebo comparisons, we will use equal allocation. One reason for this is that the weighted version of the GSHC procedure is not yet available although the weighted version of the GSHM procedure is available in Algorithms 3 and 4. Finally, Ridker et al. (2017)[26] did not provide the interim p-values. In order to illustrate the proposed procedures, we substituted for these missing interim p-values. The one-sided p-values for the three looks are given in Table 3. Note that the p-values at the final look ($k = 3$) are the same as those given in Ridker et al. (2017).[26] We now proceed with the analysis of the resulting data.

The OBF boundaries for three looks using equal split of α among the unrejected hypotheses are given in Table 4.

MBF GSHM Procedure Using Ordered p-Values (Algorithm 5)

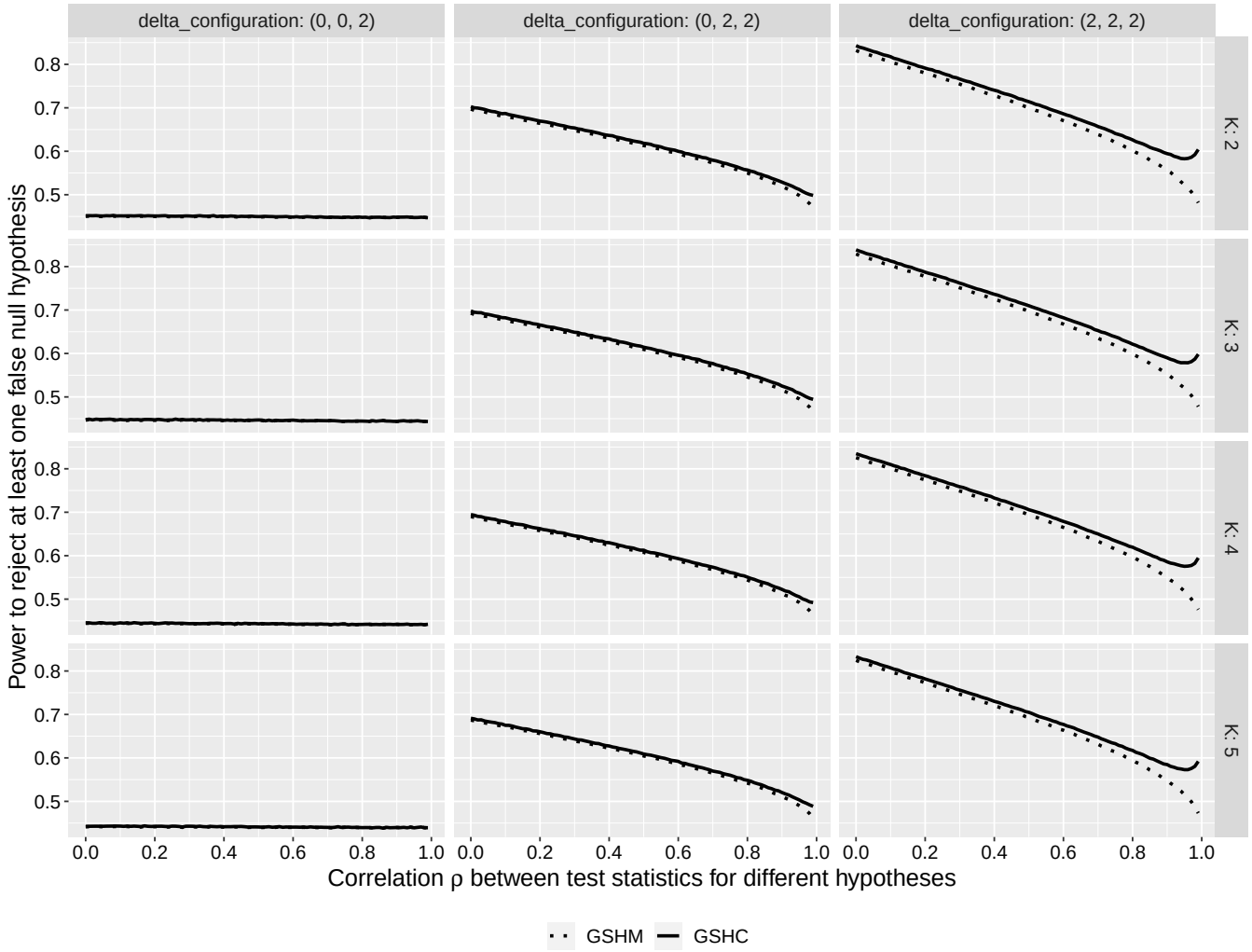

FIGURE 2 Power for GSHM (Algorithm 5) and GSHC (Algorithm 6) using OBF with $n = 3$ hypotheses and $K = 2, 3, 4, 5$ stages.

TABLE 3 p-Values for the CANTOS trial

H_i	$k = 1$	$k = 2$	$k = 3$
H_1	0.0100	0.0150	0.1500
H_2	0.00025	0.0020	0.0104
H_3	0.0003	0.0040	0.0157

Stage 1: At $k = 1$, the ordered p-values are $p_{21} = 0.00025 < p_{31} = 0.0003 < p_{11} = 0.0100$. Since $p_{21} = 0.00025 >$ the critical value 0.0002 for the $\alpha/3$ boundary at $k = 1$, we do not reject any hypothesis and go to Stage 2.
Stage 2: At $k = 2$, the ordered p-values are $p_{22} = 0.0020 < p_{32} = 0.0040 < p_{12} = 0.0150$. Since $p_{22} = 0.0020 <$ the critical value 0.0023 for the $\alpha/3$ boundary at $k = 2$, we reject H_2. However, we cannot reject H_3 since $p_{32} = 0.0040 >$ the critical value 0.0038 for the $\alpha/2$ boundary at $k = 2$, so we go to Stage 3 with $J = \{1, 3\}$.
Stage 3: At $k = 3$, the ordered p-values are $p_{33} = 0.0157 < p_{13} = 0.1500$. Since $p_{33} = 0.0157 >$ the critical value 0.0113 for the $\alpha/2$ boundary at $k = 3$, we cannot reject H_3 and hence also not H_1. Thus testing stops with only H_2 rejected.

GSHC Procedure

TABLE 4 Critical constants using the OBF boundary (5) with $K = 3$ and interim looks at 50% ($k = 1$) and 75% ($k = 2$) information times ($\alpha = 0.025$)

$\|J\|$	$\alpha/\|J\|$	$k = 1$	$k = 2$	$k = 3$
1	$\alpha/1$	0.0015	0.0092	0.0220
2	$\alpha/2$	0.0004	0.0038	0.0113
3	$\alpha/3$	0.0002	0.0023	0.0076

Stage 1: At $k = 1$, the ordered p-values are $p_{21} = 0.00025 < p_{31} = 0.0003 < p_{11} = 0.0100$. Since $p_{11} = 0.0100 >$ the critical value 0.0015 for the $\alpha/1$ boundary at $k = 1$, we do not reject H_1. Next, since $p_{31} = 0.0003 <$ the critical value 0.0040 for the $\alpha/2$ boundary at $k = 1$ we reject H_3 and hence also H_2 and go to Stage 2 with $J = \{1\}$.
Stage 2: At $k = 2$, since $p_{12} = 0.0150 >$ the critical value 0.0092 for the $\alpha/1$ boundary at $k = 2$ we do not reject H_1 and we go to Stage 3.
Stage 3: At $k = 3$, since $p_{13} = 0.1500 >$ the critical value 0.0220 for the $\alpha/1$ boundary, we do not reject H_1 and testing stops.

Thus in this case study, the GSHC procedure is able to reject both H_2 and H_3, while the MBF GSHM procedure using ordered p-values rejects only H_2. The Dunnett procedure used in Ridker et al.[26] also rejects only H_2, but it uses different error allocations. Note that the assumption of positive dependence among the test statistics at each stage, required by the GSHC procedure, is satisfied in this example because they involve comparisons of doses with a common placebo.

8 | R PROGRAMS

In Supplementary Material, we have provided R programs for applying the proposed procedures in the simulation study and in the case study. In particular, the function `GSP_function` implements the GSHM procedure using ordered p-values in Algorithm 5 and the GSHC procedure in Algorithm 6. Calculations under the multivariate normal distribution are done using the `mvtnorm` package[28] based on the algorithm by Miwa et al.[29] Four miscellaneous functions are provided to facilitate `GSP_function`. The function `cr_function` calculates the correlation structure of test statistics based on information fractions. The functions `esf_POC_function` and `esf_OBF_function` are e.s.f.s of the Lan-DeMets spending functions[14] (5) for Pocock and O'Brien-Fleming boundaries, respectively. The function `solver_boundary_esf_function` solves for the nominal level at each stage using (7).

9 | EXTENSIONS

There are various extensions of the procedures considered here that are worth pursuing. One extension is unequal information times on different endpoints. For example, in oncology trials generally three efficacy endpoints are of interest: overall survival (OS), progression-free survival (PFS) and overall response rate (ORR). The data on these three endpoints become available at different rates, so the information times for them are unequal at interim and final looks. These unequal information times will need to be prespecified. Gou and Xi (2019)[30] have considered this problem when there are single primary and secondary endpoints with different information times on each. In practice, issues with overrunning or underrunning due to more than or less than anticipated accrual rates may lead to information times different from those prespecified. The actual information times used to derive the interim and final boundaries should be documented before unblinding.

Another extension is the weighted version of the GSHC procedure analogous to the fixed sample weighted Hochberg procedures studied in Tamhane and Liu (2008).[31] Still another extension of interest arises when there are multiple primary and secondary endpoints, which are to be tested in a group sequential setting subject to a gatekeeping restriction (Dmitrienko, Tamhane and Bretz (2009), Chapter 5).[32] The case of a single primary and a secondary endpoint was studied in Tamhane, Mehta and Liu (2010)[33] for $K = 2$ looks and in Tamhane et al. (2018)[25] for $K > 2$ looks. These procedures can be extended to the present setting. If the primary endpoints are of the coprimary type then a group sequential extension of a fixed sequence procedure (Maurer, Hothorn and Lehmacher, 1995)[34] should be used instead of the the GSHM or GSHC procedures since all of them must be significant before testing the secondary endpoints which corresponds to serial gatekeeping.

Some alternative approaches to the present work include multi-arm multi-stage parametric methodologies[35,36] and adaptive designs.[37].

10 | CONCLUSIONS

We reviewed and compared the following GSHM procedures: YLLY-1, YLLY-2, YLLY-3, MB and F. YLLY-1 is the original hypothesis-by-hypothesis testing procedure proposed by Ye et al.(2013)[7] and YLLY-2 is its stage-by-stage testing implementation. YLLY-1 and YLLY-2 are equivalent and since YLLY-2 is more practically implementable, we dropped YLLY-1 from consideration because of its drawback noted in Remark 1. YLLY-3 is another modification of YLLY-1. Even though it is more powerful than YLLY-1=YLLY-2, because of its strict "look-ahead" nature, which causes the same drawback as YLLY-1, we also dropped it from consideration.

We reviewed and compared the following GSHM procedures: YLLY-1 and YLLY-2, MB and F. We showed that the MB and F procedures are equivalent when the transition parameters in the former are equal and labeled the two procedures together as MBF, which in turn is equivalent to the YLLY-2 procedure. The YLLY-1 procedure is essentially the MBF procedure with the "look-ahead" implementation. But it has a practical drawback noted in Remark 1 that the *p*-values for all hypotheses at all stages need to be available prior to its application. A "look-back" option can be added to the MBF procedure but it also has interpretational drawbacks and hence is unlikely to be accepted by practitioners and regulators. Thus both these more powerful options are not recommended.

We showed that the MB and F procedures are equivalent when the transition parameters in the former are equal and labeled the two procedures together as MBF, which in turn is equivalent to the YLLY-2 procedure. A "look-back" option can be added to the MBF procedure but it has interpretational drawbacks and hence is unlikely to be accepted by practitioners and regulators, so it is also not recommended.

In the case of equal weights, a practical way of applying the MBF procedure is using ordered *p*-values similar to the fixed sample Holm procedure. The GSHC procedure, derived by reversing the steps in the ordered *p*-values version of the MBF procedure, is more powerful than the MBF procedure, but requires the assumption of positive dependence among the test statistics. An analytical proof of the FWER control by GSHC in the general case remains a topic for future research.

In the case of equal weights, a practical way of applying the MBF procedure is using ordered *p*-values, as described in Algorithm 5, similar to the fixed sample Holm procedure. The GSHC procedure described in Algorithm 6 is derived by reversing the testing steps in Algorithm 5. It is more powerful than the MBF procedure, but requires the assumption of positive dependence among the test statistics. An analytical proof of the FWER control by GSHC in the general case remains a topic for future research.

Thus our overall recommendation is if an analytically guaranteed proof of the FWER control is required and if the test statistics are not necessarily positively dependent then to use the MBF procedure (Algorithm 4) in the case of weighted hypotheses and Algorithm 5 in the case of equally weighted hypotheses. Algorithm 3 should be used in the case of unequal transition parameters. If an empirical proof of the FWER control is acceptable, the test statistics are positively dependent and the hypotheses are not weighted then use the GSHC procedure.

ACKNOWLEDGMENTS

We are extremely grateful to an Associate Editor and two referees for their insightful and detailed comments and suggestions which greatly helped us in producing a much improved and expanded revision.

DATA SHARING

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. Food and Drug Administration (FDA) . *Multiple endpoints in clinical trials guidance for industry (draft guidance).* : U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER); 2017. FDA-2016-D-4460.

2. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics.* 1979;6:65–70.

3. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika.* 1988;75(4):800–802. doi: 10.1093/biomet/75.4.800

4. Tang DI, Gnecco C, Geller NL. Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association.* 1989;84(407):776–779.

5. Follmann DA, Proschan MA, Geller NL. Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics.* 1994;50(2):325–336.

6. Tang DI, Geller NL. Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics.* 1999;55(4):1188–1192. doi: 10.1111/j.0006-341X.1999.01188.x

7. Ye Y, Li A, Liu L, Yao B. A group sequential Holm procedure with multiple primary endpoints. *Statistics in Medicine.* 2013;32(7):1112–1124. doi: 10.1002/sim.5700

8. Maurer W, Bretz F. Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research.* 2013;5(4):311–320. doi: 10.1080/19466315.2013.807748

9. Fu Y. Step-down parametric procedures for testing correlated endpoints in a group-sequential trial. *Statistics in Biopharmaceutical Research.* 2018;10(1):18–25. doi: 10.1080/19466315.2017.1369898

10. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. New York, New York: John Wiley and Sons . 1987.

11. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. New York, New York: John Wiley and Sons. 2 ed. 1997.

12. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. New York, New York: Chapman and Hall/CRC . 2000.

13. Wassmer G, Brannath W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Switzerland: Springer International Publishing . 2016.

14. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika.* 1983;70(3):659–663. doi: 10.1093/biomet/70.3.659

15. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics.* 1979;35(3):549–556.

16. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika.* 1977;64(2):191–199. doi: 10.1093/biomet/64.2.191

17. Bretz F, Maurer W, Hommel G. Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine.* 2011;30(13):1489–1501. doi: 10.1002/sim.3988

18. Gabriel KR. Simultaneous test procedures–Some theory of multiple comparisons. *The Annals of Mathematical Statistics.* 1969;40(1):224–250. doi: 10.1214/aoms/1177697819

19. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine.* 2009;28(4):586–604. doi: 10.1002/sim.3495

20. Xi D, Tamhane AC. Allocating recycled significance levels in group sequential procedures for multiple endpoints. *Biometrical Journal.* 2015;57(1):90–107. doi: 10.1002/bimj.201300157

21. Xi D, Glimm E, Maurer W, Bretz F. A unified framework for weighted parametric multiple test procedures. *Biometrical Journal.* 2017;59(5):918–931. doi: 10.1002/bimj.201600233

22. Wolbers M, Glimm E, Xi D. "Step-down parametric procedures for testing correlated endpoints in a group-sequential trial" by Yiyong Fu. *Statistics in Biopharmaceutical Research.* 2019;11(1):104–105. doi: 10.1080/19466315.2018.1529614

23. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 1976;63:655–660.

24. Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2013;75(1):3–54.

25. Tamhane AC, Gou J, Jennison C, Mehta CR, Curto T. A gatekeeping procedure to test a primary and a secondary endpoint in a group sequential design with multiple interim looks. *Biometrics.* 2018;74(1):40–48. doi: 10.1111/biom.12732

26. Ridker PM, Everett BM, Thuren T, et al. Antiinflammatory therapy with canakinumab for atherosclerotic disease. *New England Journal of Medicine.* 2017;377(12):1119–1131.

27. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association.* 1955;50(272):1096–1121.

28. Genz A, Bretz F, Miwa T, et al. *mvtnorm: Multivariate Normal and t Distributions*. : ; 2020. R package version 1.1-1.

29. Miwa T, Hayter AJ, Kuriki S. The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2003;65(1):223–234.

30. Gou J, Xi D. Hierarchical Testing of a primary and a secondary endpoint in a group sequential design with different information times. *Statistics in Biopharmaceutical Research.* 2019;11(4):398–406. doi: 10.1080/19466315.2018.1546613

31. Tamhane AC, Liu L. On weighted Hochberg procedures. *Biometrika.* 2008;95(2):279–294. doi: 10.1093/biomet/asn018

32. Dmitrienko A, Tamhane AC, Bretz F. *Multiple Testing Problems in Pharmaceutical Statistics*. Boca Raton, Florida: Taylor & Francis . 2009.

33. Tamhane AC, Mehta CR, Liu L. Testing a primary and a secondary endpoint in a group sequential design. *Biometrics.* 2010;66(4):1174–1184. doi: 10.1111/j.1541-0420.2010.01402.x

34. Maurer W, Hothorn LA, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. *Biometrie in der chemisch-pharmazeutischen Industrie.* 1995;6:3–18.

35. Wason J, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research.* 2016;25(2):716–727. doi: 10.1177/0962280212465498

36. Ghosh P, Liu L, Mehta C. Adaptive multiarm multistage clinical trials. *Statistics in Medicine.* 2020;39(8):1084–1102. doi: https://doi.org/10.1002/sim.8464

37. Sugitani T, Bretz F, Maurer W. A simple and flexible graphical approach for adaptive group-sequential clinical trials. *Journal of Biopharmaceutical Statistics.* 2016;26(2):202–216. doi: 10.1080/10543406.2014.972509

APPENDIX

Proof of Proposition 1:

First we show that the MB and the F procedures are equivalent if the transition parameters are equal. Let us assume that the initial transition parameters $g_{ij}^{(I)}$ are equal to $1/(|I|-1)$ (since each hypothesis node is connected to $|I|-1$ other hypothesis nodes) for all $i, j \in I, i \neq j$. Then we show that they will continue to be equal in all subsequent steps. Assume that at any step where $J \subseteq I$, $g_{j\ell}^{(J)} = 1/(|J|-1)$ for all $j, \ell \in J, j \neq \ell$. Let $J \leftarrow J \setminus \{i\}$ at the next step. Then using the updating formula for $g_{i\ell}^{(J)}$ we get

$$g_{j\ell}^{(J\setminus\{i\})} = \frac{\frac{1}{|J|-1} + \frac{1}{(|J|-1)^2}}{1 - \frac{1}{(|J|-1)^2}} = \frac{1}{|J|-2}.$$

Thus the MB procedure recycles $w_j(J)\alpha$ for any rejected hypothesis $H_i,\ i \in J$ equally among all unrejected hypotheses $H_j,\ j \in J \setminus \{i\}$ as in the F procedure.

Next we show that the updated weights are normalized to sum to 1 in each step of the MB procedure, as they are in the F procedure. Thus both procedures use the same set of weights if the initial weights are the same and sum to 1. Assume that at any step where $J \subseteq I$, $\sum_{i\in J} w_i^{(J)} = 1$. The updating formula for the weights in the MB procedure in the case of equal transition parameters is

$$w_j^{(J\setminus\{i\})} = w_j^{(J)} + \frac{1}{|J|-1} w_i^{(J)}.$$

Summing both sides of the above equation over $j \in J \setminus \{i\}$) we get

$$\sum_{j\in J\setminus\{i\}} w_j^{(J\setminus\{i\})} = \sum_{j\in J\setminus\{i\}} w_j^{(J)} + \left(\frac{|J|-1}{|J|-1}\right) w_i^{(J)} = \sum_{j\in J} w_j^{(J)} = 1.$$

The MB and F procedures also differ on how they update the index set J of unrejected hypotheses. Whereas the MB procedure updates J after rejection of each individual hypothesis H_i, the F procedure does so after rejection of a group of hypotheses $H_i, i \in R$ satisfying $p_{ik} < \alpha_{ik}^{(J)}$. However, it can be readily seen that the final rejection decisions will be the same for both procedures. This is because if a hypothesis H_j satisfies $p_{jk} < \alpha_{jk}^{(J)}$ but it is not rejected by the MB procedure since there is another hypothesis H_i, which it rejects since $p_{ik} < p_{jk} < \alpha_{jk}^{(J)}$. On the other hand, the F procedure rejects both H_i and H_j. But at a later step of the MB procedure, the new index set of unrejected hypotheses $J' = J \setminus \{i\} \subset J$ and hence $\alpha_{jk}^{(J')} > \alpha_{jk}^{(J)}$. Therefore $p_{jk} < \alpha_{jk}^{(J')}$ and so H_j will be rejected within the same Stage k.

Since both the MB and the F procedures use the same weights and use the same critical constants, they are equivalent. So we may consider them together and refer to them as the MBF procedure.

Finally, we show that the MBF procedure is equivalent to the YLLY-2 procedure. Toward this end, compare the algorithms for the YLLY-2 procedure in Algorithm 2 (obtained by modifying Algorithm 1 as explained in Remark 2). At Step 1, the YLLY-2 procedure and the F procedure start with Stage $k = 1$ and updates $k \leftarrow k+1$ until the first rejection occurs. Both procedures recycle within Stage $k = k_{\text{current}}$ by updating $J \leftarrow J \setminus R$ and weights $w_i^{(J)}$ in the same way. Further rejections and updates are done within Stage $k = k_{\text{current}}$, until no rejection can be made. This process is repeated at all subsequent steps. Thus, the YLLY-2 procedure and the F procedure proceeds reach identical decisions.

Proof of Proposition 2

The proof follows along the lines of that given for the MB procedure by Maurer and Bretz (2013). Let $J \subseteq I$ denote any index set of unrejected hypotheses. Consider a closed procedure[23] that rejects H_J if there exists $H_i,\ i \in J$ such that $p_{ik} < \alpha_{ik}^{(J)}$ for some $k = 1, \ldots, K$. The following calculation shows that this local test of H_J is of level α:

$$\begin{aligned} P_{H_J}\{\text{Reject } H_J\} &= P_{H_J}\{\cup_{i\in J} \cup_{k=1}^{K} (p_{ik} < \alpha_{ik}^{(J)})\} \\ &\leq \sum_{i\in J} P_{H_J}\{\cup_{k=1}^{K} (p_{ik} < \alpha_{ik}^{(J)})\} \\ &= \sum_{i\in J} w_i^{(J)}\alpha \quad \text{from} \quad (6) \\ &= \alpha. \end{aligned}$$

Therefore this closed procedure controls the FWER at level α. Furthermore, the closed procedure is consonant since if H_J is rejected because H_i is rejected, i.e., $p_{ik} < \alpha_{ik}^{(J)}$, then all intersection hypotheses $H_{J'}$ with $i \in J' \subseteq J$ will be rejected since

$\alpha_{ik}^{(J')} \geq \alpha_{ik}^{(J)}$ because $w_i^{(J')} \geq w_i^{(J)}$. So H_i will be rejected. This shortcut to the closed procedure is the MBF procedure with "look-back option" which therefore controls the FWER at level α.

Supplementary Material for "Group Sequential Holm and Hochberg Procedures"

Ajit C. Tamhane[1], Dong Xi[2], and Jiangtao Gou[3]

[1]Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL, USA
[2]Statistical Methodology, Novartis Pharmaceuticals, East Hanover, NJ, USA
[3]Department of Mathematics and Statistics, Villanova University, Villanova, PA, USA

May 2021

In this supplementary material, we provide all results from the simulation study in Section 1. In addition, we provide R programs in Section 2.1 and results from the CANTOS clinical trial in Section 2.2. Finally, we provide the proof of the familywise error rate (FWER) control of the group sequential Hochberg (GSHC) procedure in a special case in Section 3.

1 Simulation Results

In this section, we provide additional results from the simulation study in Section 6. We investigated various scenarios involving the following combinations: $n = 2, 3, 4$, $K = 2, 3, 4, 5$, O'Brien-Fleming (OBF) and Pocock (POC) boundaries at level $\alpha = 0.05$, $\rho = 0.01(0.01)0.99$ and m true null hypotheses with $\delta_i = 0$ $(i = 0, \ldots, m)$ and $n - m$ false null hypotheses with $\delta_i = 2$, where $m = 0, 1, \ldots, n$. For example, for $n = 3$, we consider the configurations $(0, 0, 0), (0, 0, 2), (0, 2, 2)$, and $(2, 2, 2)$. For each combination, we report the simulation estimates of the FWER and the power to reject at least one false null hypothesis. For example, for the configuration $(0, 0, 2)$, the FWER is the probability to reject at least one of H_1 or H_2, while the power is the probability to reject H_3. Each FWER and power estimate is based on 10^6 replications.

In each plot, we compare the performance of a group sequential Holm (GSHM) procedure using Algorithm 5 and a group sequential Hochberg (GSHC) procedure using Algorithm 6 for $\alpha = 0.05$, $K = 2, 3, 4, 5$, $\rho = 0.01(0.01)0.99$ and various configurations of δ. For $n = 2$, Figures 1 and 2 present the FWER and power for OBF boundaries; Figures 3 and 4 present for POC boundaries. For $n = 3$, Figures 5 and 6 present thr FWER and power for OBF boundaries; Figures 7 and 8 present for POC boundaries. For $n = 4$, Figures 9 and 10 present thr FWER and power for OBF boundaries; Figures 11 and 12 present for POC boundaries.

In all cases, the FWER is controlled at level $\alpha = 0.05$ for both GSHM and GSHC; GSHC is at least as powerful as GSHM. In terms of the comparison between OBF and POC, the results show a power advantage of OBF over POC in all cases and the FWER of OBF is closer to $\alpha = 0.05$ than that of POC in the majority of cases.

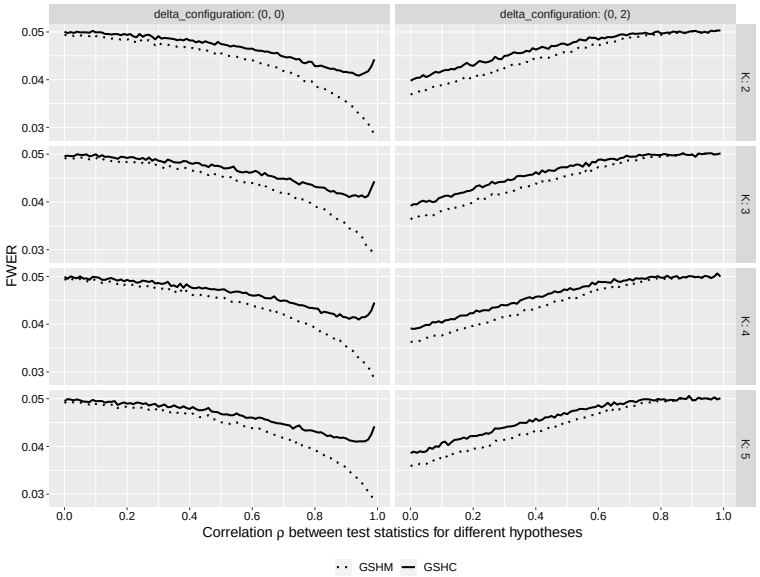

Figure 1: FWER for GSHM (Algorithm 5) and GSHC (Algorithm 6) using OBF with $n = 2$ hypotheses and $K = 2, 3, 4, 5$ stages.

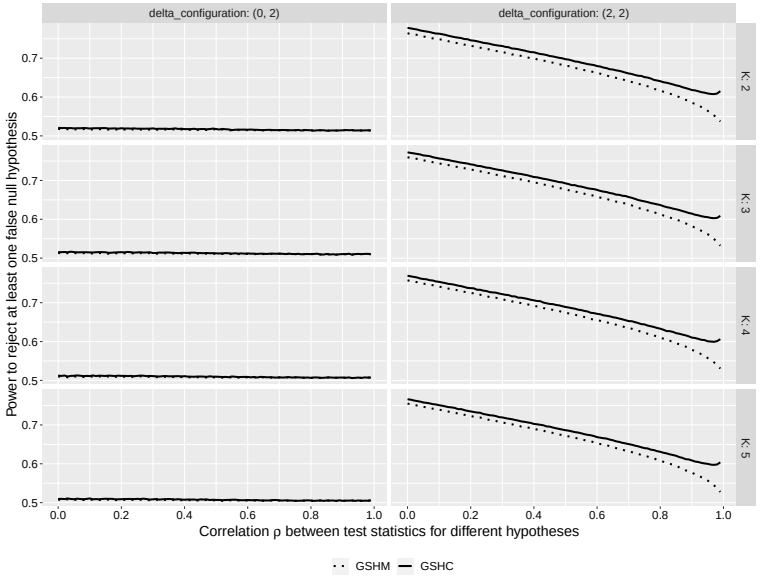

Figure 2: Power for GSHM (Algorithm 5) and GSHC (Algorithm 6) using OBF with $n = 2$ hypotheses and $K = 2, 3, 4, 5$ stages.

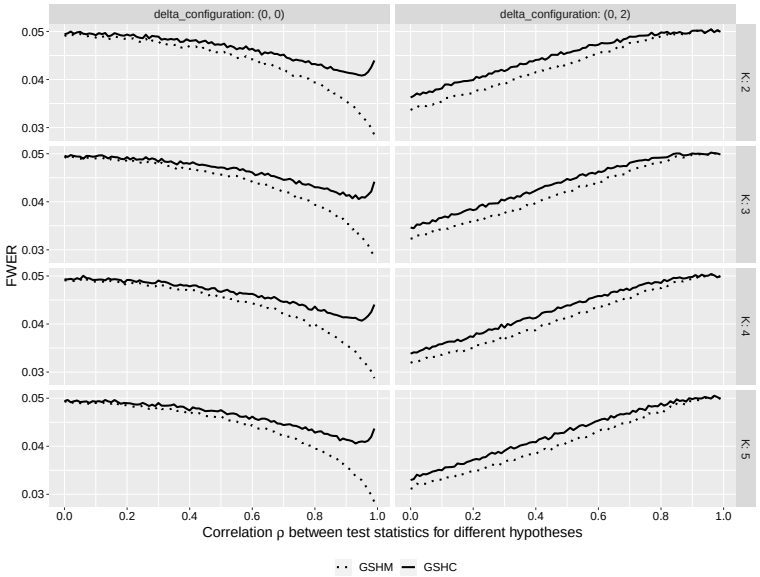

Figure 3: FWER for GSHM (Algorithm 5) and GSHC (Algorithm 6) using POC with $n = 2$ hypotheses and $K = 2, 3, 4, 5$ stages.

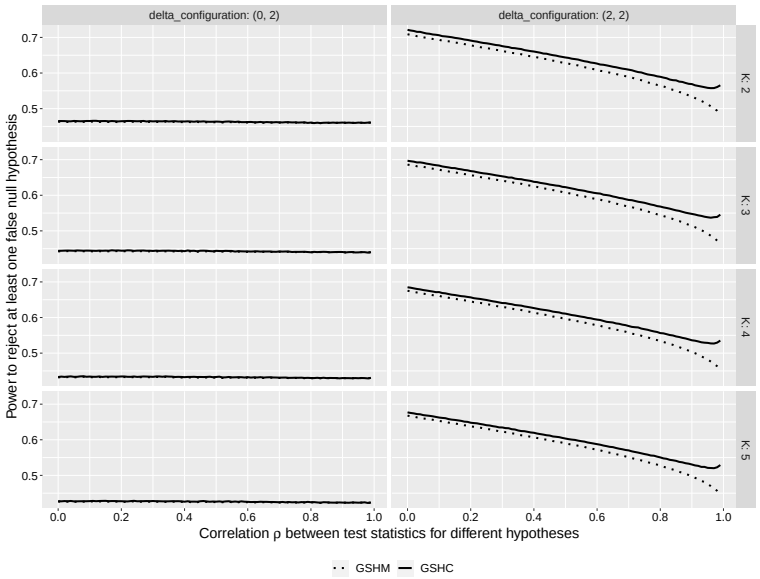

Figure 4: Power for GSHM (Algorithm 5) and GSHC (Algorithm 6) using POC with $n = 2$ hypotheses and $K = 2, 3, 4, 5$ stages.

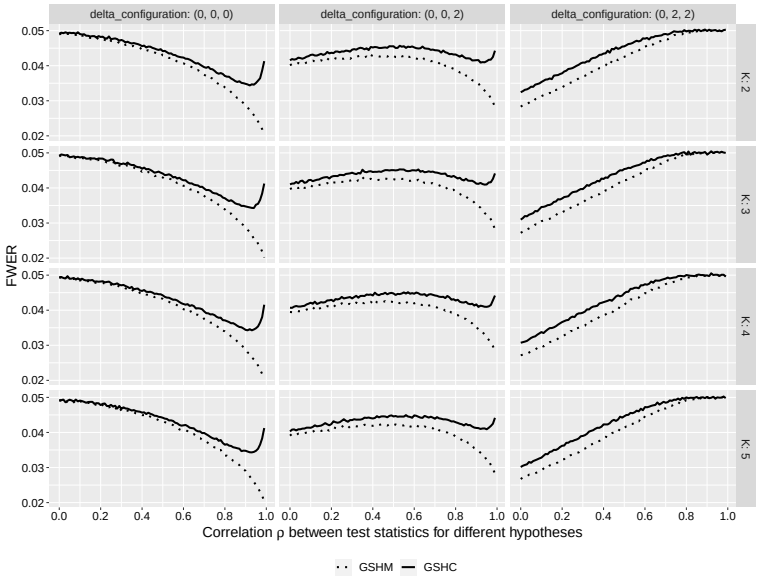

Figure 5: FWER for GSHM (Algorithm 5) and GSHC (Algorithm 6) using OBF with $n = 3$ hypotheses and $K = 2, 3, 4, 5$ stages.

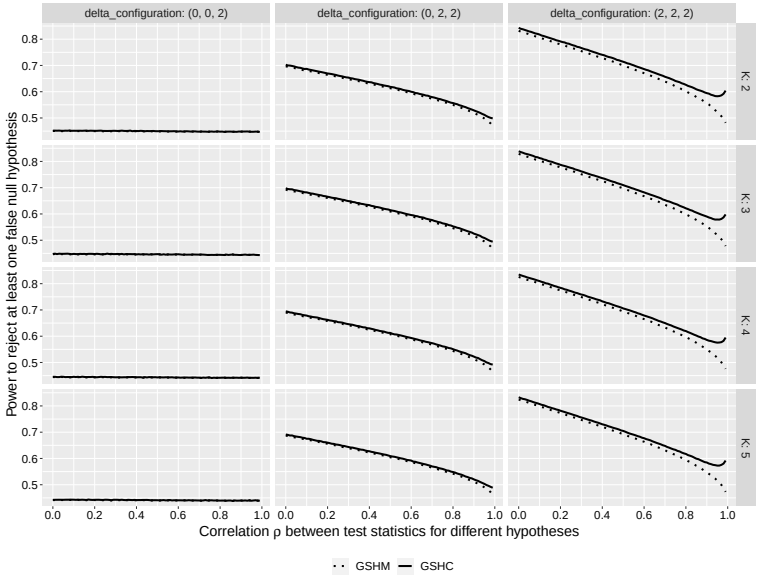

Figure 6: Power for GSHM (Algorithm 5) and GSHC (Algorithm 6) using OBF with $n = 3$ hypotheses and $K = 2, 3, 4, 5$ stages.

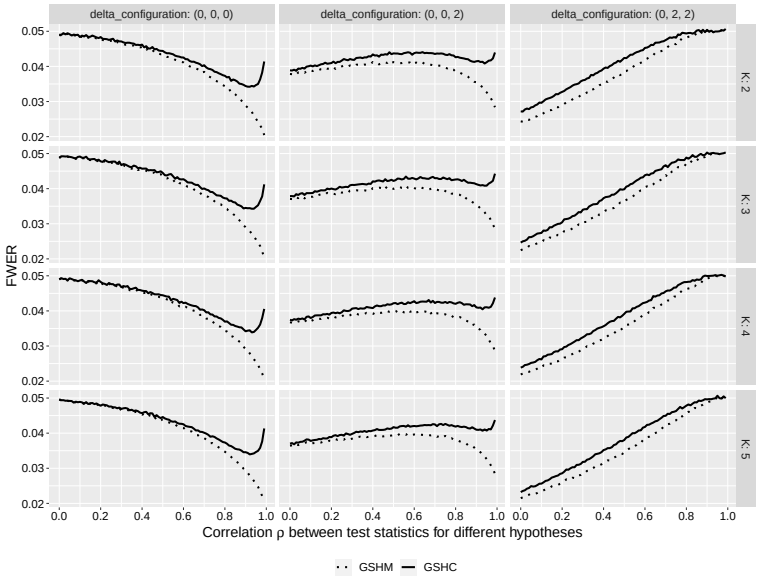

Figure 7: FWER for GSHM (Algorithm 5) and GSHC (Algorithm 6) using POC with $n = 3$ hypotheses and $K = 2, 3, 4, 5$ stages.

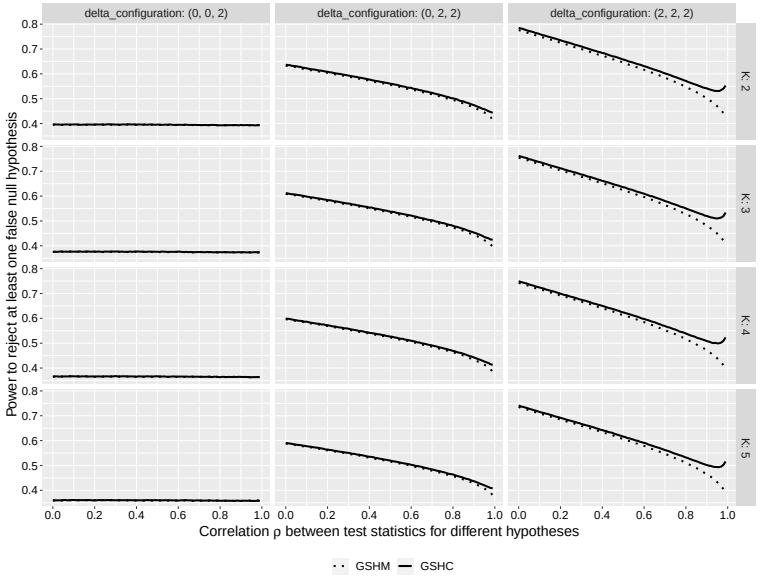

Figure 8: Power for GSHM (Algorithm 5) and GSHC (Algorithm 6) using POC with $n = 3$ hypotheses and $K = 2, 3, 4, 5$ stages.

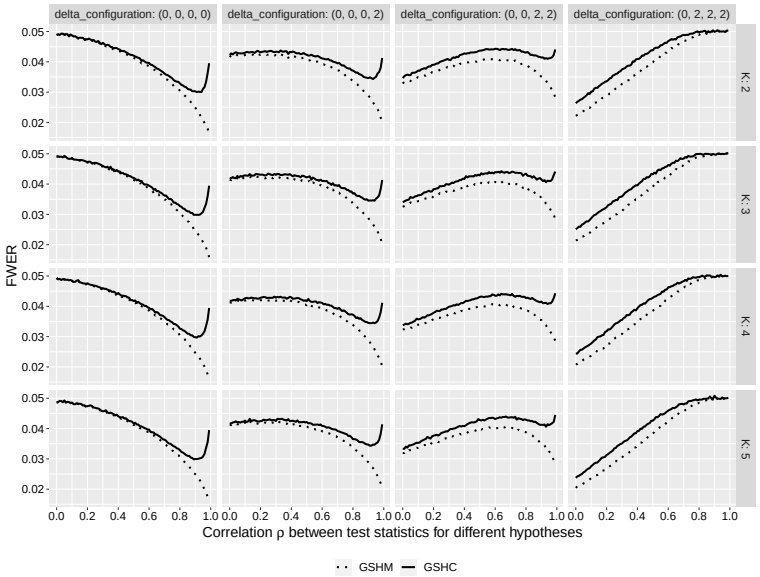

Figure 9: FWER for GSHM (Algorithm 5) and GSHC (Algorithm 6) using OBF with $n = 4$ hypotheses and $K = 2, 3, 4, 5$ stages.

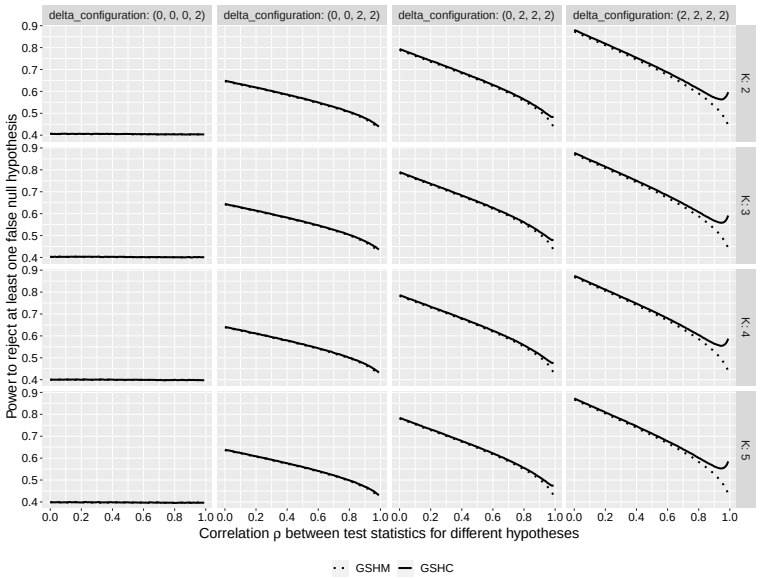

Figure 10: Power for GSHM (Algorithm 5) and GSHC (Algorithm 6) using OBF with $n = 4$ hypotheses and $K = 2, 3, 4, 5$ stages.

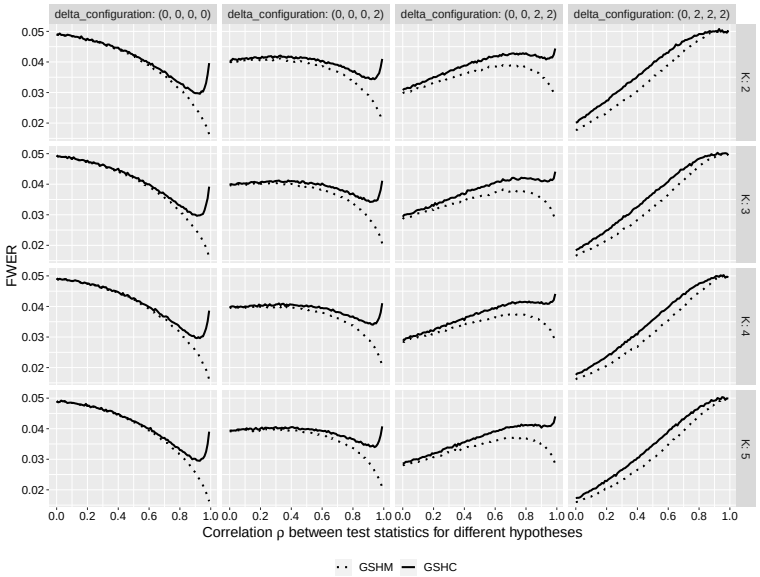

Figure 11: FWER for GSHM (Algorithm 5) and GSHC (Algorithm 6) using POC with $n = 4$ hypotheses and $K = 2, 3, 4, 5$ stages.

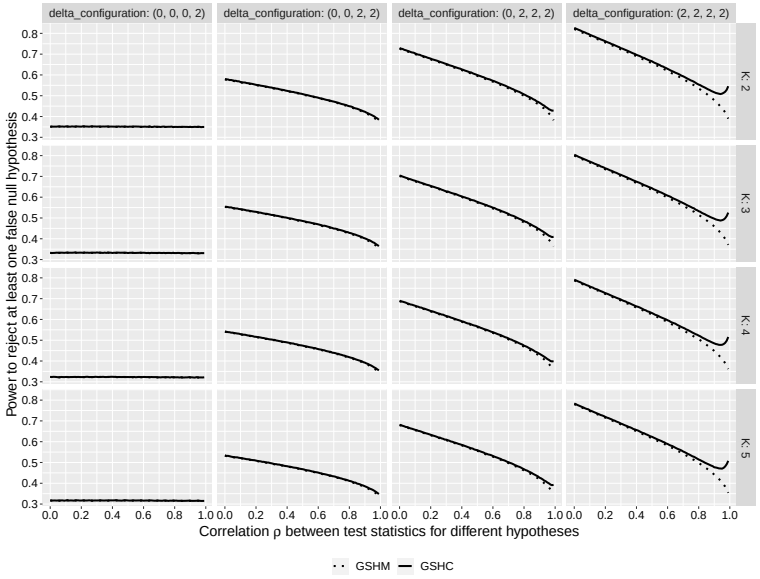

Figure 12: Power for GSHM (Algorithm 5) and GSHC (Algorithm 6) using POC with $n = 4$ hypotheses and $K = 2, 3, 4, 5$ stages.

2 R Programs

2.1 Functions

```r
# Package for the multivariate normal probability
library(mvtnorm)
```

```r
# Function for correlation matrix across stages
cr_function <- function(t){
  # Number of looks
  K <- length(t)
  cr <- diag(K)
  for (i in 1:K){
    for (j in i:K){
      cr[i, j] <- sqrt(t[i] / t[j])
    }
  }
  cr <- cr + t(cr) - diag(K)
  return(cr)
}
```

```r
# Pocock Lan-DeMets spending function
esf_POC_function <- function(alpha, t) {
  alpha * log(1 + (exp(1) - 1) * t)
}

# O'Brien-Fleming Lan-DeMets spending function
esf_OBF_function <- function(alpha, t) {
  2 - 2 * pnorm(qnorm(1 - alpha / 2) / sqrt(t))
}
```

```r
# Function for stopping boundary using a error spending function
solver_boundary_esf_function <- function(t, cumulative){
  K <- length(t)
  cr <- cr_function(t)
  solver <- function(x, cumu, cr = cr, c_past){
    z <- c(c_past, x)
    return(1 - cumu - pmvnorm(upper = z,
                              corr = cr[1:length(z), 1:length(z)],
                              algorithm = Miwa(steps = 128))
    )
  }
  c_boundary <- rep(0,K)
```

```
  c_boundary[1] <- min(qnorm(1 - cumulative[1]), 10)
  if (K > 1){
    for (k in 2:K){
      a <- uniroot(solver, interval = c(0.001, 10),
                   cumu = cumulative[k],
                   cr = cr,
                   c_past = c_boundary[1:(k-1)])$root
      c_boundary[k] <- min(a, 10)
    }
  }
  return(pnorm(c_boundary, lower.tail = F))
}
```

```
# Function for GSHM (Algorithm 5) and GSHC (Algorithm 6)
GSP_function <- function(pvalue, bound, method = c("GSHM", "GSHC")) {
  n <- dim(pvalue)[1] # Number of hypotheses
  K <- dim(pvalue)[2] # Number of stages
  dec <- array(0, dim = c(n, K)) # Decision by stage
  if (method == "GSHM") { # GSHM using Algorithm 5
    for (k in 1:K){ # Stage-by-stage
      # Identify unrejected hypotheses at stage k
      temp <- dec[, k] == 0
      if (sum(temp) > 0) { # At least one unrejected hypothesis
        # Order of pvalues
        order <- rank(pvalue[temp, k], ties.method = "first")
        # Reject the hypothesis with the smallest pvalue
        dec[temp, k][order == 1] <- pvalue[temp, k][order == 1] <
          rev(bound[1:sum(temp), k])[1]
        # If the hypothesis with the smallest pvalue is rejected
        if (dec[temp, k][order == 1]) {
          j <- 1
          # Further rejections by Holm
          while (j < sum(temp) & dec[temp, k][order == j]) {
            dec[temp, k][order == (j + 1)] <- pvalue[temp, k][order == (j + 1)] <
              rev(bound[1:(sum(temp) - j), k])[1]
            j <- j + 1
          }
        }
        for (i in 1:n) {
          if (dec[i, k]) {
            # Once a hypothesis is rejected at one stage,
            # it is rejected for future stages
            dec[i, k:K] <- dec[i, k]
          }
        }
```

```
      }
    }
  } else if (method == "GSHC") { # GSHC using Algorithm 6
    for (k in 1:K){ # Stage-by-stage
      # Identify unrejected hypotheses at stage k
      temp <- dec[, k] == 0
      if (sum(temp) > 0) { # At least one unrejected hypothesis
        # Identify rejected ordered hypotheses
        id <- sort(pvalue[temp, k], decreasing = T) <
          bound[1:sum(temp), k]
        if (any(id)) {
          # Order of pvalues
          order <- rank(pvalue[temp, k], ties.method = "first")
          # Hypotheses with pvalues smaller than ID's are rejected
          dec[temp, k][order %in% c((sum(temp) - which(id)[1] + 1):1)] <- T
        }
        for (i in 1:n) {
          if (dec[i, k]) {
            # Once a hypothesis is rejected at one stage,
            # it is rejected for future stages
            dec[i, k:K] <- dec[i, k]
          }
        }
      }
    }
  } else {
    stop("method has to be 'GSHM' or 'GSHC'")
  }
  rownames(dec) <- paste0("H", 1:n)
  colnames(dec) <- paste0("K=", 1:K)
  return(dec)
}
```

2.2 Case Study: CANTOS Clinical Trial

```
alpha <- 0.025
n <- 3
K <- 3
t <- c(0.5, 0.75, 1)

pvalue <- rbind(c(0.0100, 0.0150, 0.1500),
                c(0.00025, 0.0020, 0.0104),
                c(0.0003, 0.0040, 0.0157))
```

```r
bound <- array(NA, dim = c(n, K))
for (i in 1:n) {
  temp <- alpha / i
  cumulative <- esf_OBF_function(temp, t) # O'Brien-Fleming
  bound[i, ] <- solver_boundary_esf_function(t, cumulative)
}

# Table 4
round(bound, 4)

##        [,1]   [,2]   [,3]
## [1,] 0.0015 0.0092 0.0220
## [2,] 0.0004 0.0038 0.0113
## [3,] 0.0002 0.0023 0.0076

## GSHM using Algorithm 5
# H2 rejected at stage 2
GSP_function(pvalue, bound, method = "GSHM")

##    K=1 K=2 K=3
## H1   0   0   0
## H2   0   1   1
## H3   0   0   0

## GSHC using Algorithm 6
# H2 and H2 rejected at stage 1
GSP_function(pvalue, bound, method = "GSHC")

##    K=1 K=2 K=3
## H1   0   0   0
## H2   1   1   1
## H3   1   1   1
```

3 Proof of FWER Control of GSHC When $n = 2$ and $K \geq 2$

Proposition 1. *Consider $n = 2$ hypotheses, H_1 and H_2 tested at $K \geq 2$ stages. Assume equal weights $1/2$ on each hypothesis and that the test statistics for H_1 and H_2 are independent at all stages. Then the GSHC procedure given in Algorithm 6 controls the FWER (Hochberg and Tamhane, 1987) at level α.*

Proof of Proposition 1. Denote the p-value of H_i at stage k by p_{ik}. The critical values α_{ik} can be written in simplified form since the hypotheses are equally weighted, so the critical values depend on J only through $|J|$, which equals 1 or 2. Therefore we get

$$1 - P\left(p_{i1} > \alpha_{11}, p_{i2} > \alpha_{12}, \cdots, p_{iK} > \alpha_{1K}\right) = \alpha,$$

and

$$1 - P\left(p_{i1} > \alpha_{21}, p_{i2} > \alpha_{22}, \cdots, p_{iK} > \alpha_{2K}\right) = \alpha/2$$

using a pre-specified error spending function $f(t,\alpha)$, where $i \in \{1,2\}$. In order to apply the closure principle (Marcus et al., 1976) to prove the FWER control of the group sequential Hochberg procedure with two hypotheses, we only need to show the following inequality:

$$1 - P\left(p_{(1),1} > \alpha_{21}, p_{(2),1} > \alpha_{11}, \cdots, p_{(1),K} > \alpha_{2K}, p_{(2),K} > \alpha_{1K}\right) < \alpha. \tag{1}$$

Note that this inequality is strict. The idea of this proof is a decomposition and combination process of the rejection region.

$$\begin{aligned}
&P\left(\cap_{j=1}^{K}\left\{p_{(1),j} > \alpha_{2j}, p_{(2),j} > \alpha_{1j}\right\}\right)\\
&= P\left(\cap_{j=1}^{K}\left\{\left\{p_{1,j} > \alpha_{2j}, p_{2,j} > \alpha_{1j}\right\} \cup \left\{p_{1,j} > \alpha_{1j}, \alpha_{2j} < p_{2,j} \le \alpha_{1j}\right\}\right\}\right)\\
&= P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{2j}, p_{2,j} > \alpha_{1j}\right\}\right)\\
&\quad + \sum_{k=1}^{K} P\left(\left\{\cap_{j\neq k}\left\{p_{1,j} > \alpha_{2j}, p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{p_{1,k} > \alpha_{1k}, \alpha_{2k} < p_{2,k} \le \alpha_{1k}\right\}\right)\\
&\quad + \sum_{k_1\neq k_2} P\left(\left\{\cap_{j\neq k_1, j\neq k_2}\left\{p_{1,j} > \alpha_{2j}, p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{p_{1,k_1} > \alpha_{1k_1}, \alpha_{2k_1} < p_{2,k_1} \le \alpha_{1k_1}\right\}\right.\\
&\qquad \left. \cap \left\{p_{1,k_2} > \alpha_{1k_2}, \alpha_{2k_2} < p_{2,k_2} \le \alpha_{1k_2}\right\}\right)\\
&\quad + \cdots\cdots\\
&\quad + \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{1,j} > \alpha_{2j}, p_{2,j} > \alpha_{1j}\right\}\right\}\right.\\
&\qquad \left. \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}} > \alpha_{1j^{Pime}}, \alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right)\\
&\quad + \cdots\cdots\\
&\quad + P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{1j}, \alpha_{2j} < p_{2,j} \le \alpha_{1j}\right\}\right)\\
&= \sum_{r=0}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{1,j} > \alpha_{2j}, p_{2,j} > \alpha_{1j}\right\}\right\}\right.\\
&\qquad \left. \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}} > \alpha_{1j^{Pime}}, \alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right)\\
&= \sum_{r=0}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{1,j} > \alpha_{2j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}} > \alpha_{1j^{Pime}}\right\}\right\}\right)\\
&\qquad \cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right)\\
&= P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{2j}\right\}\right) \cdot P\left(\cap_{j=1}^{K}\left\{p_{2,j} > \alpha_{1j}\right\}\right)\\
&\quad + \sum_{r=1}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{1,j} > \alpha_{2j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}} > \alpha_{1j^{Pime}}\right\}\right\}\right)\\
&\qquad \cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right)\\
&= A.
\end{aligned}$$

$$
\begin{aligned}
A &= P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{2j}\right\}\right) \cdot P\left(\cap_{j=1}^{K}\left\{p_{2,j} > \alpha_{1j}\right\}\right) \\
&\quad + \sum_{r=1}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{p_{1,j} > \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}} > \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\qquad \cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\quad + \sum_{r=1}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{\alpha_{2j} < p_{1,j} \le \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}} > \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\qquad \cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&= P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{2j}\right\}\right) \cdot P\left(\cap_{j=1}^{K}\left\{p_{2,j} > \alpha_{1j}\right\}\right) \\
&\quad + \sum_{r=1}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{1j}\right\}\right) \\
&\qquad \cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\quad + \sum_{r=1}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{\alpha_{2j} < p_{1,j} \le \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}} > \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\qquad \cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&= P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{2j}\right\}\right) \cdot P\left(\cap_{j=1}^{K}\left\{p_{2,j} > \alpha_{1j}\right\}\right) \\
&\quad + P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{1j}\right\}\right) \cdot \sum_{r=1}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{p_{2,j} > \alpha_{1j}\right\}\right\}\right. \\
&\qquad \left. \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\quad + \sum_{r=1}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{\alpha_{2j} < p_{1,j} \le \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}} > \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\qquad \cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&= P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{2j}\right\}\right) \cdot P\left(\cap_{j=1}^{K}\left\{p_{2,j} > \alpha_{1j}\right\}\right) \\
&\quad + P\left(\cap_{j=1}^{K}\left\{p_{1,j} > \alpha_{1j}\right\}\right) \cdot \left[P\left(\cap_{j=1}^{K}\left\{p_{2,j} > \alpha_{2j}\right\}\right) - P\left(\cap_{j=1}^{K}\left\{p_{2,j} > \alpha_{1j}\right\}\right)\right] \\
&\quad + \sum_{r=1}^{K} \sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{\alpha_{2j} < p_{1,j} \le \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}} > \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\qquad \cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\backslash\{k_1,\cdots,k_r\}}\left\{p_{2,j} > \alpha_{1j}\right\}\right\} \cap \left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}} < p_{2,j^{Pime}} \le \alpha_{1j^{Pime}}\right\}\right\}\right) \\
&= B.
\end{aligned}
$$

$$
\begin{aligned}
B &= \left(1-\frac{\alpha}{2}\right)\cdot(1-\alpha)+(1-\alpha)\cdot\left[\left(1-\frac{\alpha}{2}\right)-(1-\alpha)\right] \\
&\quad+\sum_{r=1}^{K}\sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{\alpha_{2j}<p_{1,j}\leq\alpha_{1j}\right\}\right\}\cap\left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}}>\alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\quad\cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{2,j}>\alpha_{1j}\right\}\right\}\cap\left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}}<p_{2,j^{Pime}}\leq\alpha_{1j^{Pime}}\right\}\right\}\right) \\
&= 1-\alpha \\
&\quad+\sum_{r=1}^{K}\sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{\alpha_{2j}<p_{1,j}\leq\alpha_{1j}\right\}\right\}\cap\left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}}>\alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\quad\cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{2,j}>\alpha_{1j}\right\}\right\}\cap\left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}}<p_{2,j^{Pime}}\leq\alpha_{1j^{Pime}}\right\}\right\}\right) \\
&> 1-\alpha.
\end{aligned}
$$

We conclude that

$$1-P\left(\cap_{j=1}^{K}\left\{p_{(1),j}>\alpha_{2j},p_{(2),j}>\alpha_{1j}\right\}\right)<\alpha.$$

The difference between the right-hand side and left-hand side is

$$
\begin{aligned}
&\alpha-\left(1-P\left(\cap_{j=1}^{K}\left\{p_{(1),j}>\alpha_{2j},p_{(2),j}>\alpha_{1j}\right\}\right)\right) \\
&=\sum_{r=1}^{K}\sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{\alpha_{2j}<p_{1,j}\leq\alpha_{1j}\right\}\right\}\cap\left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{1,j^{Pime}}>\alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\quad\cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{2,j}>\alpha_{1j}\right\}\right\}\cap\left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}}<p_{2,j^{Pime}}\leq\alpha_{1j^{Pime}}\right\}\right\}\right).
\end{aligned}
$$

Since two endpoints have the same information fractions, $(p_{1,1},\cdots,p_{1,K})$ and $(p_{2,1},\cdots,p_{2,K})$ follow the same K-dimensional multivariate distribution. Therefore, we can slightly simplify the expression of difference between $1-P\left(\cap_{j=1}^{K}\left\{p_{(1),j}>\alpha_{2j},p_{(2),j}>\alpha_{1j}\right\}\right)$ and α, which is

$$
\begin{aligned}
&\alpha-\left(1-P\left(\cap_{j=1}^{K}\left\{p_{(1),j}>\alpha_{2j},p_{(2),j}>\alpha_{1j}\right\}\right)\right) \\
&=\sum_{r=1}^{K}\sum_{\{k_1,\cdots,k_r\}\in\{1,\cdots,K\}} P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{\alpha_{2j}<p_{j}\leq\alpha_{1j}\right\}\right\}\cap\left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{p_{j^{Pime}}>\alpha_{1j^{Pime}}\right\}\right\}\right) \\
&\quad\cdot P\left(\left\{\cap_{j\in\{1,\cdots,K\}\setminus\{k_1,\cdots,k_r\}}\left\{p_{j}>\alpha_{1j}\right\}\right\}\cap\left\{\cap_{j^{Pime}\in\{k_1,\cdots,k_r\}}\left\{\alpha_{2j^{Pime}}<p_{j^{Pime}}\leq\alpha_{1j^{Pime}}\right\}\right\}\right),
\end{aligned}
$$

where $(p_1,\cdots,p_K)$ follow the same multivariate distribution of $(p_{1,1},\cdots,p_{1,K})$ or $(p_{2,1},\cdots,p_{2,K})$. □

Remark 1. Note that the Simes inequality under independence is an equality. However, the inequality (1) is strict for all $K\geq 2$. For example, for $K=2$ and $\alpha=0.05$, numerical computation shows that the value of the left-hand side in (1) is 0.049936, strictly less than 0.05.

References

Yosef Hochberg and Ajit C. Tamhane. *Multiple Comparison Procedures*. John Wiley and Sons, New York, New York, 1987.

Ruth Marcus, Eric Peritz, and K. Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660, December 1976.