Gatekeeping Testing via Adaptive Alpha Allocation

Jianjun(David) Li[*, 1] and **Devan V. Mehrotra**[2]

[1] Wyeth Research Devision, 500 Arcola Road, Collegeville, PA 19425, USA
[2] Merck Research Laboratories, UG1CD-44, 351 N. Sumneytown Pike, North Wales, PA 19454, USA

Received 7 November 2007, revised 14 March 2008, accepted 23 June 2008

Summary

In a typical clinical trial, there are one or two primary endpoints, and a few secondary endpoints. When at least one primary endpoint achieves statistical significance, there is considerable interest in using results for the secondary endpoints to enhance characterization of the treatment effect. Because multiple endpoints are involved, regulators may require that the familywise type I error rate be controlled at a pre-set level. This requirement can be achieved by using "gatekeeping" methods. However, existing methods suffer from logical oddities such as allowing results for secondary endpoint(s) to impact the likelihood of success for the primary endpoint(s). We propose a novel and easy-to-implement gatekeeping procedure that is devoid of such deficiencies. A real data example and simulation results are used to illustrate efficiency gains of our method relative to existing methods.

Key words: Clinical trials; Familywise error rate; Multiple tests: Gatekeeping; *p*-value.

1 Introduction

In a clinical trial, multiple endpoints are commonly classified as primary and secondary per regulatory guidelines and/or certain conventions (ICH guidline; CPMP-EU guidline; Capizzi and Zhang, 1996; Chi, 1998). A key factor to determine an endpoint as primary or secondary is its clinical relevance. A primary endpoint is the variable that is capable of providing the most clinically relevant evidence for the study treatment benefit. Significant results on the primary endpoints indicate the treatment benefits and would lead to regulatory claims. However, a secondary endpoint usually just provides additional or supportive information on the treatment. Significant findings on the secondary endpoints alone are not sufficient for any claims, but they can be used to enhance characterization of the treatment effect. Per the main objective of the clinical trial, the primary endpoints are adequately powered by the design, while the secondary may not be.

In some settings, a clinically pivotal endpoint may be demoted to key secondary due to lack of power. This type of key secondary endpoint is different from what has been called "secondary" conventionally, and analysis of the former type of secondary endpoint has been controversial (D'Agostino, 2000; Moyé, 2000; Koch, 2000; O'Neill, 2000). We have proposed an approach to tackle this problem elsewhere (Li and Mehrotra, 2008, in press). The purpose of this paper is to address analysis for conventional secondary endpoints, which only play a supportive role in the trial, and are not deemed to be clinically pivotal.

In a trial with primary endpoints and supportive secondary endpoints, the gatekeeping testing framework works well. In the framework, testing of the hypotheses associated with the primary endpoints (or primary hypotheses) serves as a gatekeeper for testing the hypotheses associated with the secondary endpoints (or secondary hypotheses). The secondary hypotheses are tested only if the

[*] Corresponding author: e-mail: lij20@wyeth.com, Phone: +001 484 865 2970, Fax: +001 484 865 0061

gate-keeper is passed, i.e., at least one or all primary hypotheses have been rejected. There are two types of gatekeeping procedures: serial and parallel. A serial gatekeeping procedure allows one to test the secondary hypotheses only if all primary hypotheses are rejected, while a parallel procedure allows one to test the secondary hypotheses if at least one primary hypothesis is rejected. Two procedural requirements for a parallel gatekeeping procedure are: (1) significance of primary hypotheses does not depend on the significance of secondary hypotheses, and (2) secondary hypotheses can be tested only if at least one primary hypothesis is significant. These two properties have been referred as 'parallel gatekeeping criteria' in the literature (Dmitrienko, Offen, and Westfall, 2003).

At present, the parallel gatekeeping testing procedures that can be used in general are from Dmitrienko, Offen, and Westfall (2003) and Chen, Luo, and Capizzi (2005). Dmitrienko et al. (2003) developed two parallel gatekeeping procedures based on the closure principle (Marcus, Peritz, and Gabriel, 1976) using weighted tests: weighted Bonferroni procedure (WB) and weighted Simes procedure (WS). Chen et al. (2005) proposed a modified version of weighted Simes procedure (MWS) where the weights for the secondary hypotheses are capped, and the matching relationship between the individual primary and secondary hypotheses, if such a relationship exists, can be incorporated.

Being a valid parallel gatekeeping procedure, WB indeed satisfies the two parallel gatekeeping criteria. However, it is not very powerful compared to the weighted Simes procedures (Dmitrienko et al. 2003; Chen et al., 2005); for example, a power drop from 82% to 78% for a primary hypothesis is possible. Meanwhile, it has been observed that WS and MWS, though powerful, may not satisfy the parallel gatekeeping criteria. The conclusions for the primary hypotheses can be affected by the significance levels of the secondary hypotheses if these procedures are used. Table 1 below lists two scenarios to show the degree of the impact. In Scenario 1, there are two primary hypotheses and each of them has a raw p-value of 0.035. With a p-value of 0.051 from each secondary hypothesis, neither of primary hypotheses is significant when the familywise type I error rate is controlled at 0.05. The second scenario includes 4 primary hypotheses and 4 secondary hypotheses. P-values as small as

Table 1 Examples on how significance of the primary hypotheses can be affected by non-significance of the secondary hypotheses.

Hypothesis	Weight	Raw p-value	Adjusted p-value	
			WS	MWS
Scenario 1				
Primary 1	0.50	0.035	0.051	0.051
Primary 2	0.50	0.035	0.051	0.051
Secondary 1	0.50	0.051	0.051	0.051
Secondary 2	0.50	0.051	0.051	0.051
Scenario 2				
Primary 1	0.25	0.035	0.051	0.051
Primary 2	0.25	0.018	0.051	0.051
Primary 3	0.25	0.018	0.051	0.051
Primary 4	0.25	0.018	0.051	0.051
Secondary 1	0.25	0.051	0.051	0.051
Secondary 2	0.25	0.051	0.051	0.051
Secondary 3	0.25	0.051	0.051	0.051
Secondary 4	0.25	0.051	0.051	0.051

A threshold $\gamma = 2/3$ is used in MWS.

0.018 for the three primary hypotheses and 0.035 for the fourth one will not lead to any statistical significance, if all the secondary hypotheses have a p-value of 0.051. In both scenarios, all primary hypotheses will be significant using Hochberg's procedure (Hochberg, 1988) if only the primary hypotheses are tested. A study with either scenario will unfortunately be concluded as negative as a consequence of including the supportive secondary hypotheses in the testing plan, if one of the procedures (WS or MWS) is used.

In the gatekeeping testing framework, the power of the secondary hypotheses may be quite low due to their supportive role. Applying WS and MWS could pose a great risk to the success of trial which is typically determined by the results of the primary hypotheses. This unpleasant property raises concerns for applying both procedures in practice. Moreover, both procedures are relatively difficult to implement and interpret. Work to search for short-cut versions of the parallel gatekeeping procedures has been done recently (Dmitrienko et al., 2006; Hommel, Bretz, and Maurer, 2007), but some of the aforementioned concerns remain unresolved.

Our goal in this paper is to present a simple, powerful parallel gatekeeping procedure that can be used in most late phase clinical trials. The proposed procedure has the following nice properties: (1) it controls familywise type I error rate, (2) it maintains the power of primary hypotheses, and power loss due to inclusion of secondary hypotheses in the analysis is generally negligible, (3) it satisfies the two parallel gatekeeping criteria, (4) it is simple in implementation and easy in interpretation, and (5) it is comfortably straightforward and the conclusion for each hypothesis can be drawn by directly reviewing the respective raw p-value in general.

The paper is organized as follows. We introduce the proposed procedure in Section 2. A real clinical trial example is used in Section 3 to show how the proposed procedure works, as contrasted to others. Section 4 includes simulation-based evaluations of power performance of the proposed procedure. Final remarks are in Section 5.

2 Proposed Procedure

2.1 Introduction of the procedure

For introduction, let $H_{11}, \cdots, H_{1m}$ denote the primary (null) hypotheses and $H_{21}, \cdots, H_{2n}$ denote the supportive secondary (null) hypotheses in a trial. If there is only one primary hypothesis, the parallel gatekeeping procedures degenerate to serial procedures, so we assume $m \geq 2$ to exclude this trivial case. According to the hierarchical structure of the hypotheses, $H_{11}, \cdots, H_{1m}$ serve as the gatekeeper, and $H_{21}, \cdots, H_{2n}$ are tested only if at least one of $H_{11}, \cdots, H_{1m}$ has been rejected. For convenience, we say $H_{11}, \cdots, H_{1m}$ are hypotheses in the primary or gatekeeper family and $H_{21}, \cdots, H_{2n}$ are hypotheses in the secondary family.

One commonly used multiple comparison procedure for multiple hypotheses in clinical trials is Hochberg's procedure (Hochberg, 1998). Sarkar (1998) proved mathematically that the procedure is valid, i.e., controls the familywise type I error rate at a pre-set target level, if the distribution of the test statistics has the MTP_2 property. Karlin and Rinott (1981) presented necessary and sufficient conditions for MTP_2 when the distribution is multivariate normal or absolute-valued multivariate normal. For example, an absolute-valued multivariate normal distribution possesses MTP_2 if $\text{cov}(X_i, X_j) = a_i a_j$ for $i \neq j$. Of note a multivariate normal may not be MTP_2 even if the correlations are all non-negatives. Throughout this paper, we simply assume Hochberg's procedure can be validly used to test the hypotheses within the primary family and the secondary family.

Let α be the pre-specified familywise type I error rate for the trial. That is, the type I error rate for testing all the hypotheses in the trial, including primary hypotheses $H_{11}, \cdots, H_{1m}$ and secondary hypotheses $H_{21}, \cdots, H_{2n}$, is to be controlled at the overall α level. Our proposed procedure includes the following 3 simple steps.

Step 1. Choose $\alpha_p (< \alpha)$ as the total error rate for testing the primary hypotheses $H_{11}, \cdots, H_{1m}$, and test the primary hypotheses using Hochberg's procedure.

Step 2. Determine the error rate for testing the secondary hypotheses $H_{21}, \cdots, H_{2n}$, if needed, as

$$\alpha_s = \alpha I_{[p_{(m)} \le \alpha_p]} + \min\left(\frac{\lambda \alpha_t}{P_{(m)}^2}, \alpha_p\right) I_{[P_{(m)} > \alpha_p]} \tag{1}$$

where $P_{(m)}$ is the maximum of the p-values from the primary hypotheses $H_{11}, \cdots, H_{1m}$, $I_{[P_{(m)} \le \alpha_p]}$ (or $I_{[P_{(m)} > \alpha_p]}$) is the indicator function equal to 1 if $P_{(m)} \le \alpha$ (or $P_{(m)} > \alpha$) and 0 otherwise, and α_t is determined by

$$\alpha_t = \begin{cases} \alpha_p \left(1 - \sqrt{\dfrac{2\alpha_p - \alpha - \alpha_p^2/(m-1)}{\alpha_p}}\right)^2 & \text{if} \quad \alpha_p + \dfrac{\alpha_p^2}{m-1} - \dfrac{\alpha_p^3}{(m-1)^2} \le \alpha \\ \alpha_p \dfrac{\alpha - \alpha_p}{(m-1) - \alpha_p} & \text{if} \quad \alpha_p + \dfrac{\alpha_p^2}{m-1} - \dfrac{\alpha_p^3}{(m-1)^2} > \alpha\,. \end{cases} \tag{2}$$

In (1), λ is the (preferably largest) constant such that

$$P\left\{p_A > \frac{\alpha_p}{m-1}, p_B \le \min\left(\frac{\lambda \alpha_t}{P_A^2}, \alpha_p\right)\right\} \le \alpha - \alpha_p \tag{3}$$

when the null hypothesis for both A and B is true, for any $A \in \{H_{11}, \cdots, H_{1m}\}$ and $B \in \{H_{21}, \cdots, H_{2n}\}$.

Step 3. Test secondary hypotheses $H_{21}, \cdots, H_{2n}$ using Hochberg's procedure with the total error rate α_s, if at least one primary hypothesis is significant.

2.2 Implementation of the procedure

In the proposed procedure, the error rate α_p is pre-specified. Since preserving the power of the primary hypotheses is critical in the gatekeeping framework where the secondary hypotheses are only supportive, α_p should be chosen close to α. For common familywise type I error rate $\alpha = 0.05$, $\alpha_p = 0.048$ is recommended, though in general any α_p no less than 0.045 can be used. As one can show, using $\alpha_p = 0.045$ rather than $\alpha_p = 0.05$ has little impact on the power for the primary hypotheses. If the familywise type I error rate is $\alpha = 0.025$, $\alpha_p = 0.022$ or larger can be used.

The error rate α_s in the proposed procedure is calculated via (1), a pre-specified function of m, α_p, λ and $P_{(m)}$. When all primary hypotheses are rejected by Hochberg's procedure ($P_{(m)} \le \alpha_p$), α can be used as the error rate for testing secondary hypotheses, which is similar to the fallback approach (Wiens and Dmitrienko, 2005). When only some, not all, primary hypotheses are rejected ($P_{(m)} > \alpha_p$), the error rate that can be used for testing secondary hypotheses is essentially $\alpha - \alpha_p$. An ordinary approach will determine α_s equal to $\alpha - \alpha_p$ no matter what $P_{(m)}$ is, which leads to a quite small error rate for the secondary hypotheses. In (1), α_s is determined adaptively (to $P_{(m)}$) so that α_s can be relatively large if $P_{(m)}$ is small and small if $P_{(m)}$ is large. As detailed below, this adaptation takes advantage of a unique characteristic of late phase clinical trials – all primary hypotheses are typically adequately powered by design, and hence $P_{(m)}$ will be "small" most likely.

For calculating α_s via (1), λ is to be obtained. It can be shown that $\lambda = 1$ if the test statistics for $H_{11}, \cdots, H_{1m}$ and $H_{21}, \cdots, H_{2n}$ are independent (see Appendix). When the test statistics are correlated, using $\lambda = 1$ will not gurantee the control of the familywise type I error rate even if the test statistics are positively correlated. Computing λ via (3) is necessary and can be done if the distribution of the test statistics under the nulls is identified.

For most clinical trial applications it is reasonable to assume that the joint distribution of the test statistics for $H_{11}, \cdots, H_{1m}$ and $H_{21}, \cdots, H_{2n}$ is a multivariate normal with non-negative correlations, and λ can be chosen as the constant such that

$$\max_{0 \le \varrho \le 1} P\left\{p_A > \frac{\alpha_p}{m-1},\ p_B \le \min\left(\frac{\lambda \alpha_t}{P_A^2},\ \alpha_p\right)\right\} = \alpha - \alpha_p \tag{4}$$

Table 2 Values of λ for different $\alpha_p(<\alpha)$ and m for 2-sided tests with $\alpha = 0.05$.

$\alpha_p =$	0.045	0.046	0.047	0.048	0.049
$m = 2$	0.3129	0.3630	0.4115	0.4411	0.4589
3	0.1021	0.1295	0.1659	0.2099	0.2432
4	0.0508	0.0645	0.0854	0.1188	0.1609

Table 3 Values of λ for different $\alpha_p(<\alpha)$ and m for 1-sided tests with $\alpha = 0.025$.

$\alpha_p =$	0.020	0.021	0.022	0.023	0.024
$m = 2$	0.0595	0.0685	0.0896	0.1344	0.2078
3	0.0139	0.0171	0.0241	0.0406	0.0791
4	0.0069	0.0084	0.0115	0.0189	0.0402

where ϱ denotes the correlation coefficient between the test statistics for some $A \in \{H_{11}, \cdots, H_{1m}\}$ and $B \in \{H_{21}, \cdots, H_{2n}\}$. A simple R or SAS program can be written to compute λ. For illustrative purpose, Table 2 and Table 3 above show the values of λ computed via an R program for different values of $\alpha_p(<\alpha)$ and m, assuming the test for $H_{11}, \cdots, H_{1m}$ and $H_{21}, \cdots, H_{2n}$ is 2-sided ($\alpha = 0.05$) or 1-sided ($\alpha = 0.025$), respectively.

2.3 Properties of the procedure

The proposed procedure is valid, i.e., it controls the familywise type I error rate when the test statistics for the primary and secondary hypotheses are independent; a mathematical proof is provided in Appendix. For the correlated case where the test statistics jointly follow a multivariate normal distribution with positive correlations, we conjecture that it is valid based on extensive simulations. The simulation results are also provided in Appendix.

It is obvious that the proposed procedure satisfies the two parallel gatekeeping criteria. In the proposed procedure, the primary hypotheses are tested at the α_p level and it is pre-specified independent of any p-value from the secondary hypotheses. Significance of the primary hypotheses will not be impacted by the observed significance levels of the secondary hypotheses.

The proposed procedure also takes into account two facts of the considered clinical trial scenario: (1) the secondary hypotheses are supportive and rejection of at least one primary hypothesis to achieve the trial success is critical, and (2) the primary hypotheses are adequately powered by design. By allocating a large portion of the familywise alpha to the primary hypotheses, the power of the primary hypotheses is essentially maintained, i.e., the power of the primary hypotheses is not materially impacted by including testing of supportive secondary hypotheses. Given that the primary hypotheses are well powered in the trial, the p-values from the primary hypotheses, if not smaller than α_p, are likely to be close to α_p. For example, if the power for a hypothesis is 85%, the median of the p-value distribution is 0.004, and the 90$^{\text{th}}$ percentile is 0.15 (Hung et al., 1997). Hence, per (1), α_s can be quite large in the scenarios we encounter frequently in practice. For example, if $m = 2$ and $\alpha_p = 0.048$, then $\alpha_s = 0.0124$ if $P_{(m)} = 0.06$, $\alpha_s = 0.0044$ if $P_{(m)} = 0.1$, and $\alpha_s = 0.0020$ if $P_{(m)} = 0.15$, if the test statistics are positively correlated.

Table 4 Study results and conclusions in a depression trial.

Hypothesis	Raw p-value	Significant or not ? ($\alpha = 0.05$)		
		WS	MWS	Proposed
Primary				
HAMD-17	0.043	No (0.086)	No (0.065)	Yes
CGI-I	0.015	Yes (0.030)	Yes (0.030)	Yes
Secondary				
HAMD-1	0.007	Yes (0.030)	No (0.065)	Yes
HAMA	0.128	No (0.128)	No (0.128)	No

Number in the parenthesis is adjusted p-value.

3 A Clinical Trial Example

A late phase clinical trial was conducted to investigate a new drug for treatment of depression as compared with placebo. The trial was a randomized double-blinded multicenter trial with 3 parallel groups: new drug, active control, and placebo. The primary endpoint was the Hamilton Depression Scale 17-item total score (HAMD-17). The secondary endpoints included: Clinical Global Impression Improvement score (CGI-I), Item No.1 score of the Hamilton Depression Scale (HAMD-1), and Hamilton Rating Scale for Anxiety total score (HAMA). For illustrative purpose, we consider the results of comparing active control to placebo from the trial and assume that the comparisons related to both HAMD-17 and CGI-I are primary hypotheses and the comparisons related to HAMD-1 and HAMA are secondary. The familywise error rate is set at $\alpha = 0.05$.

The study results and conclusions are summarized in Table 4. For WS and MWS, equal weight is assigned to each hypothesis within the primary family and within the secondary family. In the application of MWS, we hypothetically assume matching relationship between HAMD-17 and HAMD-1, and between CGI-I and HAMA. Note that for WS and the proposed procedure, such matching relationship is not required. Per recommendation by Chen et al. (2005), a threshold $\gamma = 2/3$ is used in MWS. As presented in Table 4, WS concludes that the primary hypothesis related to CGI-I is significant and the secondary hypothesis related to HAMD-1 is also significant. Only the primary hypothesis related to CGI-I is declared significant by MWS.

In the proposed procedure, $\alpha_p = 0.048$ is used. Since all the p-values from two primary hypotheses are less than 0.048, $\alpha_s = 0.05$ per (1). By applying Hochberg's procedure to the hypotheses within the primary family and within the secondary family separately, the p-values for both primary hypotheses and the secondary hypothesis related to HAMD-1 are declared significant.

4 Power Comparisons

A simulation study was conducted to compare the performance of WS, MWS, the proposed procedure, and a serial gatekeeping procedure (SGP), in the setting where there are two primary hypotheses H_{11}, H_{12} and two secondary hypotheses H_{21}, H_{22}. Each hypothesis is of the form $H_{ij} : d_{ij} = 0$, where $d_{11},\ d_{12},\ d_{21},\ d_{22}$ represent the means of four normally distributed variables $X_{11},\ X_{12},\ X_{21},\ X_{22}$ with standard deviation 1 and a common correlation coefficient ϱ. The 2-sided marginal p-value for testing H_{ij} is given by $2 * [1 - \Phi(|X_{ij}|)]$, where $\Phi(x)$ is the distribution function of the standard normal distribution. For a given set of marginal powers $(q_{11},\ q_{12},\ q_{21},\ q_{22})$ or equivalently a set of means $(d_{11},\ d_{12},\ d_{21},\ d_{22})$, 1,000,000 sets of normal variables $(X_{11},\ X_{12},\ X_{21},\ X_{22})$ were generated. The power of each procedure was then computed as the average of rejections over 1,000,000 sets by that procedure. For WS and MWS, all hypotheses were equally weighted within each family. For MWS,

Table 5. Power (in percent) of a serial gatekeeping procedure and different parallel gatekeeping testing procedures ($\alpha = 0.05$).

Marginal Power	H_{11}				H_{21}			
	WS	MWS	Proposed	SGP	WS	MWS	Proposed	SGP
$\varrho = 0.2$								
(90, 90, 90, 90)	88.0	88.8	88.9	89.2	85.3	79.4	81.2	74.3
(90, 90, 70, 70)	86.3	88.2	88.9	89.2	63.2	59.6	59.9	57.0
(90, 90, 50, 50)	85.2	87.6	88.9	89.2	42.2	40.3	40.2	39.0
(90, 90, 10, 10)	84.2	86.6	88.9	89.2	6.1	6.0	6.0	5.9
(90, 70, 90, 90)	87.0	87.6	87.4	87.8	81.0	76.7	72.0	59.1
(90, 70, 70, 70)	85.8	87.1	87.4	87.8	59.5	56.9	52.0	46.3
(90, 70, 50, 50)	85.0	86.7	87.4	87.8	39.4	38.2	34.7	32.2
(90, 70, 10, 10)	84.2	86.0	87.4	87.8	5.6	5.7	5.2	5.1
(70, 90, 90, 90)	67.1	68.3	68.2	68.9	80.9	62.0	72.0	59.1
(70, 90, 70, 70)	64.1	67.1	68.2	68.9	59.4	47.4	52.0	46.3
(70, 90, 50, 50)	61.9	66.0	68.2	68.9	39.4	32.5	34.7	32.2
(70, 90, 10, 10)	**59.7**	**63.9**	**68.2**	68.9	5.6	5.0	5.2	5.1
(70, 70, 90, 90)	65.4	66.2	66.0	66.7	72.1	59.2	62.3	47.7
(70, 70, 70, 70)	63.2	65.4	66.0	66.7	53.4	44.9	44.6	38.0
(70, 70, 50, 50)	61.5	64.5	66.0	66.7	35.6	30.7	30.0	27.0
(70, 70, 10, 10)	59.6	62.9	66.0	66.7	5.1	4.7	4.6	4.4
(50, 50, 90, 90)	43.8	**44.3**	**43.8**	**44.6**	**54.1**	39.7	**42.5**	**26.5**
(50, 50, 70, 70)	42.2	43.7	43.8	44.6	40.4	30.3	29.4	21.8
(50, 50, 50, 50)	40.8	43.0	43.8	44.6	27.1	21.0	19.6	16.1
(50, 50, 10, 10)	38.9	41.5	43.8	44.6	3.9	3.4	3.1	2.9
$\varrho = 0.6$								
(90, 90, 90, 90)	87.3	88.2	88.4	88.7	84.4	81.3	81.7	79.0
(90, 90, 70, 70)	85.4	87.3	88.4	88.7	64.2	63.2	63.2	62.6
(90, 90, 50, 50)	84.6	86.7	88.4	88.7	44.1	43.8	43.9	43.8
(90, 90, 10, 10)	84.2	86.2	88.4	88.7	6.8	6.8	6.9	6.9
(90, 70, 90, 90)	86.1	86.5	86.3	86.7	79.9	78.1	72.7	65.1
(90, 70, 70, 70)	85.1	86.1	86.3	86.7	61.0	60.4	56.9	54.5
(90, 70, 50, 50)	84.5	85.7	86.3	86.7	42.2	42.3	40.5	39.8
(90, 70, 10, 10)	84.2	85.3	86.3	86.7	6.7	6.8	6.7	6.7
(70, 90, 90, 90)	67.7	68.6	68.4	69.0	**79.9**	65.8	**72.7**	65.0
(70, 90, 70, 70)	64.3	67.1	68.4	69.0	61.0	54.3	56.9	54.5
(70, 90, 50, 50)	61.7	65.7	68.4	69.0	42.2	39.5	40.5	39.9
(70, 90, 10, 10)	**59.6**	**63.7**	**68.4**	69.0	6.7	6.7	6.7	6.7
(70, 70, 90, 90)	65.6	66.1	65.7	66.3	70.5	62.8	64.4	56.0
(70, 70, 70, 70)	63.4	65.2	65.7	66.3	56.3	52	51.7	48.7
(70, 70, 50, 50)	61.4	64.2	65.7	66.3	40	38.2	38.0	37.1
(70, 70, 10, 10)	59.6	62.6	65.7	66.3	6.6	6.6	6.6	6.6
(50, 50, 90, 90)	44.6	**44.8**	**44.2**	44.9	52.5	43.4	46.3	34.9
(50, 50, 70, 70)	43.4	44.3	44.2	44.9	44.3	37.7	37.7	32.3
(50, 50, 50, 50)	41.7	43.6	44.2	44.9	33.1	29.5	29.0	26.8
(50, 50, 10, 10)	39.0	41.7	44.2	44.9	6.2	6.0	6.0	6.0

the minimum weight on the primary hypotheses was set as $\gamma = 2/3$, and matching relationships between H_{11} and H_{21}, and between H_{12} and H_{22} were incorporated. For the proposed procedure, $\alpha_p = 0.048$ was used and Hochberg's procedure was applied within each family. For the serial gatekeeping procedure SGP, the primary hypotheses were tested at $\alpha = 0.05$ and the secondary hypotheses were tested at $\alpha = 0.05$ (via Hochberg's procedure) only if both primary hypotheses were rejected. Note that SGP is a special case of the proposed procedure (by taking $\alpha_p = \alpha = 0.05$).

To mimic the common clinical trial settings, we set the power for the primary hypotheses as (90%, 90%) or (90%, 70%) while the power for the secondary hypotheses ranged from 90% down to 10%. To check the robustness of the proposed procedure, we also considered two unusual scenarios where the power for both the primary hypotheses was 70% or 50%. We computed the results for two correlation scenarios $\varrho = 0.2$ and $\varrho = 0.6$, representing low-level and high-level correlations (Sankoh, D'Agostino, and Huque, 2003).

The simulation results are summarized in Table 5. The powers for H_{11} and H_{21} are presented. The powers for H_{12} and H_{22} are not presented since they can be obtained from those for H_{11} and H_{21} by symmetry.

The simulations show that the proposed procedure has a higher power than WS for the primary hypotheses in all cases expect one where both the primary hypotheses have 50% power and both the secondary hypotheses have 90% power. A maximum increase in power of 8.5% is observed if the common correlation coefficient $\varrho = 0.2$ and 8.8% if $\varrho = 0.6$. For the secondary hypotheses, the proposed procedure has a lower power than WS. The maximum power loss is 11.6% if $\varrho = 0.2$ and is reduced to 7.2% if $\varrho = 0.6$. The power loss diminishes in general as the marginal power for the secondary hypotheses becomes smaller.

Compared to MWS, the proposed procedure usually has a higher power (maximum gain = 4.3% for $\varrho = 0.2$, and 4.7% for $\varrho = 0.6$) for the primary hypotheses except in the unusual situations where a primary hypothesis actually has a smaller marginal power compared to its matching secondary. In the latter case, the power loss of using the proposed procedure rather than WMS is negligible (maximum loss = 0.5% for $\varrho = 0.2$, and 0.6% for $\varrho = 0.6$). For the power of the secondary hypotheses, the proposed procedure preforms similarly to MWS if both primary hypotheses have the same marginal power. If the marginal power of two primary hypotheses is different, the power of a secondary hypothesis by MWS can be higher or lower depending on the marginal power of its matching primary, and in this case the power by the proposed procedure is in-between.

As expected, SGP has a higher power than the proposed procedure for the primary hypotheses and lower power for the secondary hypotheses. The differences between two procedures in power are negligible for the primary hypotheses (at most 0.8%), but noteworthy for the secondary hypotheses in some scenarios (as large as 16%).

The simulations also confirm that the power of the primary hypotheses by the proposed procedure is not influenced by the marginal power of the secondary hypotheses; this is not true for WS and MWS.

5 Conclusions and Discussions

Parallel gatekeeping testing approaches are very useful in the setting where secondary hypotheses cannot be tested unless at least one primary hypothesis is significant. Current available powerful approaches such as WS and MWS are based on the closure principle and are deemed unsatisfactory primarily because they do not satisfy the parallel gatekeeping criteria. As shown in this paper, the power of primary hypotheses can be severely affected by the power of secondary hypotheses when using WS or MWS. This property is critical because in the gatekeeping testing framework, secondary hypotheses play a supportive role and may not be well powered by the study design. The trial success could be jeopardized due to an aggressive plan of testing a (less important) secondary hypothesis.

The proposed procedure pre-specifies the error rate for the primary hypotheses and adaptively determines the error rate for the secondary hypotheses. By reserving a large portion of the total familywise alpha, the new procedure preserves the power of the primary hypotheses. Adaptive way of determining

the alpha level for the secondary hypotheses usually leads to a big error rate allowance for them in practice. So the proposed procedure essentially preserves the power of the primary hypotheses (and therefore the probability of trial success) and has decent power for the supportive secondary hypotheses.

As an additional advantage, the proposed procedure is simple in implementation and straightforward in drawing conclusions directly based on raw-p-values. This feature makes the procedure easy to understand and interpret.

Appendix

a. Proof that the proposed procedure controls the familywise type I error rate at α when the test statistics are (continuous and) independent.

Note that we say $H_{11}, \cdots, H_{1m}$ are hypotheses in the primary or gatekeeper family and $H_{21}, \cdots, H_{2n}$ are hypotheses in the secondary family. Let $\lambda = 1$ in Eq. (1).

Case 1: At least one null hypothesis in the primary family is true. The probability of making a type I error in this case is:

$$\text{FWER} = P(E_1) + P(F_1)$$

where

$E_1 = \{$rejection of at least one true null hypothesis in the primary family$\}$

$F_1 = \{$acceptance of all true null hypotheses in the primary family and rejection of at least one false null hypothesis in the primary family and rejection of at least one true null hypothesis in the secondary family$\}$.

By using Hochberg's procedure for the primary family, $P(E_1) \le \alpha_p$. If all true null hypotheses in the primary family are accepted, then $p_{(m)} > \alpha_p$. Moreover, $p_{(m)} \ge p_{m_i} \ge \alpha_p/(m-1)$, where p_{m_i} is the p-value associated with (any) one of the true null hypotheses. Let

$F_{12} = \{$rejection of at least one true null hypothesis in the secondary family$\}$.

Next we need to show that

$$P(F_1) \le P(F_{12}[p_{(m)} > \alpha_p, p_{m_i} > \alpha_p/(m-1)]) \le \alpha - \alpha_p$$

where $[p_{(m)} > \alpha_p, p_{m_i} > \alpha_p/(m-1)]$ defines the event of occurrence of $p_{(m)} > \alpha_p$ and $p_{m_i} > \alpha_p/(m-1)$. When all the test statistics are independent,

$$P(F_1) \le P(F_{12}[p_{(m)} > \alpha_p, p_{m_i} > \alpha_p/(m-1)]) = E(P(F_{12} \mid p_{(m)}, p_{m_i})\, I_{[p_{(m)} \ge p_{m_i} > \alpha_p/(m-1)]})$$
$$\le E\left(\min\left(\frac{\alpha_t}{p_{(m)}^2}, \alpha_p\right) I_{[p_{(m)} \ge p_{m_i} > \alpha_p/(m-1)]}\right) \le E\left(\min\left(\frac{\alpha_t}{p_{m_i}^2}, \alpha_p\right) I_{[p_{m_i} > \alpha_p/(m-1)]}\right).$$

Note that since p_{m_i} is the p-value under a null hypothesis, it is a uniform (0, 1) random variable. If $\alpha_p + \frac{\alpha_p^2}{m-1} - \frac{\alpha_p^3}{(m-1)^2} \le \alpha$ as in (2), then $2\alpha_p - \alpha - \frac{\alpha_p^2}{m-1} \le \alpha_p\left(1 - \frac{\alpha_p}{m-1}\right)^2$. As an immediate result, $\alpha_t \ge \frac{\alpha_p^2}{(m-1)^2}$. This implies

$$P(F_1) \le E\left(\min\left(\frac{\alpha_t}{p_{m_i}^2}, \alpha_p\right) I_{[p_{m_i} > \alpha_p/(m-1)]}\right) = E(\alpha_p I_{[\alpha_p/(m-1) < p_{m_i} \le \sqrt{\alpha_t/\alpha_p}]}) + E\left(\frac{\alpha_t}{p_{m_i}^2} I_{[p_{m_i} > \sqrt{\alpha_t/\alpha_p}]}\right)$$
$$= \alpha_p\left[\sqrt{\frac{\alpha_t}{\alpha_p}} - \frac{\alpha_p}{m-1}\right] + \alpha_t \int_{\sqrt{\alpha_t/\alpha_p}}^{1} \frac{dx}{x^2} = \sqrt{\alpha_t\alpha_p} - \frac{\alpha_p^2}{m-1} + \sqrt{\alpha_t\alpha_p} - \alpha_t$$
$$= 2\sqrt{\alpha_t\alpha_p} - \alpha_t - \frac{\alpha_p^2}{m-1} = \alpha - \alpha_p\,.$$

Conversely, if $\alpha_p + \dfrac{\alpha_p^2}{m-1} - \dfrac{\alpha_p^3}{(m-1)^2} > \alpha$, then $\alpha_t < \dfrac{\alpha_p^3}{(m-1)^2}$ and

$$P(F_1) \le E(\min\left(\frac{\alpha_t}{p_{m_i}^2}, \alpha_p\right) I_{[p_{m_i} > \alpha_p/(m-1)]}) = E\left(\frac{\alpha_t}{p_{m_i}^2} I_{[p_{m_i} > \alpha_p/(m-1)]}\right) = \alpha_t\left(\frac{m-1}{\alpha_p} - 1\right)$$
$$= \alpha - \alpha_p\,.$$

Therefore, FWER $= P(E_1) + P(F_1) \le \alpha_p + (\alpha - \alpha_p) = \alpha$.

Case 2: All null hypotheses in the primary family are false. The probability of making a type I error in this case is:

$$\text{FWER} = P(F_2)$$

where

$F_2 =$ {rejection of at least one null hypothesis in the primary family and rejection of at least one true null hypothesis in the secondary family}.

Then

$$\text{FWER} = P(F_2) \le E\left(\alpha I_{[p_{(m)} \le \alpha_p]} + \min\left(\frac{\alpha_t}{p_{(m)}^2}, \alpha_p\right) I_{[p_{(m)} > \alpha_p]}\right) \le \alpha\,.$$

Since the familywise type I error rate is controlled at level α in both Case 1 and Case 2 (which cover all possibilities), the proof is complete.

b. Demonstration that the proposed gatekeeping testing approach controls the familywise type I error rate at α when the test statistics jointly follow a multivariate normal with positive correlations.

Following the proof for the independent scenario, we just need to show that

$$P(F_1) \le P(F_{12}[p_{(m)} > \alpha_p,\, p_{m_i} > \alpha_p/(m-1)]) \le \alpha - \alpha_p\,.$$

Note that Hochberg's procedure is used to test the hypotheses in the secondary family. By the closure principle, we only need to show that for any $1 \le k \le n$

$$P\left(p_{(k)}^S \le \alpha_s \text{ or } p_{(k-1)}^S \le \frac{\alpha_s}{2} \text{ or } \cdots \text{ or } p_{(1)}^S \le \frac{\alpha_s}{k},\ p_{(m)} \ge \alpha_p,\ p_{m_i} > \frac{\alpha_p}{m-1}\right) \le \alpha - \alpha_p \qquad \text{(a1)}$$

where $p_{(k)}^S \ge p_{(k-1)}^S \ge \cdots \ge p_{(1)}^S$ are ordered null p-values $\{p_1^S, p_2^S, \cdots, p_k^S\} \subset \{p_{H_{21}}, p_{H_{22}}, \cdots, p_{H_{2n}}\}$, which are computed under the nulls. Or we need to show that for any $1 \le k \le n$

$$\Delta \equiv P\left(p_{B(k)} \le \min\left(\frac{\lambda\alpha_t}{p_A^2}, \alpha_p\right) \text{ or } \cdots \text{ or } p_{B(1)} \le \frac{1}{k}\min\left(\frac{\lambda\alpha_t}{p_A^2}, \alpha_p\right),\ p_A > \frac{\alpha_p}{m-1}\right) + \alpha_p \le \alpha \qquad \text{(a2)}$$

where the null p-value p_A and unordered null p-values $p_{B1} \cdots p_{Bk}$ are from a $(k+1)$-dimensional multivariate normal with positive correlations.

By determination of λ (Eq. (4)), we know that (a2) holds for $k = 1$. We verify (a2) via simulation for $k > 1$.

To provide a comprehensive evaluation of Δ, we simulated the correlation coefficient matrix for the $(k+1)$-dimensional multivariate normal with each component generated independently from a *U(0,1)* distribution. We considered 1000 sets of such matrix. For each set, 1,000,000 multivariate normal variables were simulated to compute Δ in (a2). We summarize simulation results with $m \le 4$ and $k \le 4$ for 2-sided tests, $\alpha_p = 0.045$ and $\alpha_p = 0.048$ in Table 6, and for 1-sided tests, $\alpha_p = 0.022$ and $\alpha_p = 0.024$ in Table 7. The results for other scenarios simulated ($m \le 4$ and $5 \le k \le 8$) are similar, and not presented.

Table 6 Summary of simulation results ($\alpha = 0.05$).

(m, k)	Δ [as defined in (a2)]			
	$\alpha_p = 0.048$		$\alpha_p = 0.045$	
	Average	Maximum	Average	Maximum
(2, 2)	0.0494	0.0500	0.0483	0.0499
(3, 2)	0.0492	0.0499	0.0476	0.0496
(4, 2)	0.0490	0.0499	0.0473	0.0499
(2, 3)	0.0482	0.0488	0.0456	0.0470
(3, 3)	0.0482	0.0488	0.0455	0.0468
(4, 3)	0.0482	0.0487	0.0454	0.0469
(2, 4)	0.0493	0.0498	0.0480	0.0494
(3, 4)	0.0491	0.0499	0.0473	0.0490
(4, 4)	0.0489	0.0497	0.0470	0.0488

Table 7 Summary of simulation results ($\alpha = 0.025$).

(m, k)	Δ [as defined in (a2)]			
	$\alpha_p = 0.024$		$\alpha_p = 0.022$	
	Average	Maximum	Average	Maximum
(2, 2)	0.0246	0.0250	0.0239	0.0249
(3, 2)	0.0246	0.0250	0.0235	0.0248
(4, 2)	0.0245	0.0250	0.0233	0.0250
(2, 3)	0.0241	0.0244	0.0224	0.0232
(3, 3)	0.0241	0.0244	0.0223	0.0231
(4, 3)	0.0241	0.0244	0.0223	0.0232
(2, 4)	0.0246	0.0249	0.0237	0.0246
(3, 4)	0.0245	0.0249	0.0233	0.0243
(4, 4)	0.0245	0.0249	0.0231	0.0242

In Tables 6 and 7, the average and maximum of Δ over 1000 sets are presented. For all scenarios, the maximum of Δ over 1000 sets is no more than α. We can conclude that the proposed procedure controls the familywise type I error well no matter what the (positive) correlations among the test statistics are.

Acknowledgements *We are grateful to a referee and one of the Guest Editors for their constructive comments, which led to a more precise presentation of the paper.*

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

Capizzi, T. and Zhang, J. (1996). Testing the hypothesis that matters for multiple primary endpoints. *Drug Information Journal*. **30**, 949–956.

Chen X., Luo X., and Capizzi T. (2005). The application of enhanced parallel gatekeeping strategies. *Statistics in Medicine* **24**, 1385–1397.

Chi, GHY. (1998). Multiple testings: multiple comparisons and multiple endpoints. *Drug Information Journal* **32**, 1347S–1362S.

CPMP-EU guidline 'Points to consider on multiplicity issues in clinical trials'.

D'Agostino, Sr., R. B. (2000). Controlling alpha in a clinical trial: the case for secondary endpoints. *Statistics in Medicine* **19**, 763–766.

Dmitrienko A., Offen W., and Westfall P. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* **22**, 2387–2400.

Dmitrienko A., Tamhane A., Wang X., and Chen X. (2006) Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal* **48**, 984–991.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–803.

Hommel, G., Bretz, F., and Maurer W. (2007). Powerful short-cuts of multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine*. Accepted manuscript.

Hung, H. M., O'Neill R., Bauer, P., and Kohne, K. (1997). The behavior of the *p*-value when the alternative hypothesis is true. *Biometrics* **53**, 11–22.

Karlin, S. and Rinott Y. (1981). Total positive properties of absolute value multinormal variables with applications to confidence interval estimates and related probabilistic inequalities. *Annals of Statistics* **9**, 1035–1049.

Koch, G. G. (2000). Discussion for 'Alpha calculation in clinical trials: consideration and commentary for the new millennium'. *Statistics in Medicine* **19**, 781–784.

International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use: ICH Tripartite Guideline- E9 Document: Statistical Principles for Clinical Trials.

Li, J. and Mehrotra, D. (2008). An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine*, in press.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Moyé, L. A. (2000). Alpha calculus in clinical trials: Considerations and commentary for the new millennium. *Statistics in Medicine* **19**, 767–779.

Moyé, L. A. (2000). Response to commentaries on 'Alpha calculus in clinical trials: Considerations'. *Statistics in Medicine* **19**, 795–799.

O'Neill, R. T. (2000). Commentary on 'Alpha calculation in clinical trials: consideration and commentary for the new millennium'. *Statistics in Medicine* **19**, 785–793.

Sankoh, A. J., D'Agostino, R. B., and Huque, M. F. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine* **22**, 3133–3150.

Sarkar, S. (1998). Some probability inequalities for ordered MTP2 random varaibels: a proof of Simes' conjecture. *Ann. Statist.* **26**, 494–504.

Wiens, B. and Dmitrienko, A. (2005). The fallback procedure for evaluating a single family of hypotheses. *Journal of Pharmaceutical Statistics* **15**, 1–14.

