Trigger strategy in repeated tests on multiple hypotheses (DRAFT)

Jiangtao Gou
Department of Mathematics and Statistics, Villanova University
March 8, 2020

Abstract. The proposed trigger strategy is a general framework for a multistage statistical design with multiple hypotheses, allowing an adaptive selection of interim analyses. The selection of interim stages can be associated with some prespecified endpoints which serve as the trigger. This selection allows us to refine the critical boundaries in hypotheses testing procedures, and potentially increase the statistical power.

Keywords. Clinical trial; Multiple tests; Multistage designs; Power; Trigger

1 Introduction

It is common for a modern clinical trial design to involve more than one endpoint along with repeated tests in order to achieve multiple goals simultaneously and as early as possible. An issue that arises as a result of the increase of the number of hypotheses of interest and the increase of the times of repeated tests is the reduced statistical power when the overall type I error rate and the maximum sample size have been set. To enhance the statistical power, especially the power of testing secondary endpoints which are possibly underpowered in the trial design, one approach is to reduce some unnecessary tests or interim analyses. A following question is how to determine these unnecessary analyses so that we can spend the type I error that was cost by these excessive analyses on some essential hypothesis tests.

To answer this question, we consider using results from the primary hypothesis test during the study to determine the good timing for testing the secondary hypotheses. We call this approach the trigger strategy where the result of testing the primary endpoint serves as a trigger for testing the secondary endpoints. Here we borrow the term from game theory (Fudenberg and Maskin, 1986), where a player utilizing a certain trigger strategy originally cooperates but penalizes the opponent if a trigger is observed. Under the setting of a clinical trial, we put the secondary hypothesis aside initially but start the test if a specific trigger is perceived.

The trigger strategy is a flexible generalization of hierarchical testing of multiple hypotheses in the group sequential design (Hung et al., 2007; Tamhane et al., 2010; Glimm et al., 2010; Tamhane et al., 2018; Gou and Xi, 2019; Gou and Chén, 2019; Zhang and Gou, 2020) that allows a variety of triggers to activate the hypothesis testing procedures for some endpoints of interest. The article is organized as follows. Section 2 provides the general framework of trigger strategy. Then, in Section 3, a specific trigger strategy using a significance trigger and a time trigger is described. Power analysis of the procedure explored in Section 2 and 3 is discussed in Section 4. Section 5 contains some final discussion and comments. The proofs of all propositions are included in Appendix.

2 Trigger strategy: a general framework

Our application of the trigger strategy begins with choosing one or more endpoints which serve as triggers for activation of statistical testing procedures for other endpoints. The hypothesis testing procedures of these trigger endpoints start automatically at the beginning of a clinical trial. It is most likely that the primary endpoints will serve as the trigger endpoints. Note that the design with two or more trigger endpoints and that with only one trigger endpoint have no essential difference when considering the type I error control. Without loss of generality, we consider the design with one trigger endpoint in this article. We denote the only trigger endpoint by H_0, and the other n endpoints which are triggered by H_0 by $H_1, \ldots, H_n$. The endpoint H_0 is scheduled to be tested at a collection of K_0 calendar time points $T_0^c = \{t_{0,1}^c, \ldots, t_{0,K_0}^c\} \subset (0, t_{\max}^c]$, where $(0, t_{\max}^c]$ is the span of calendar time of a trial study. The critical boundary for testing H_0 is determined by the calendar time T_0^c and the corresponding information time T_0 using the method of repeated tests on accumulating data or group sequential procedures (Armitage et al., 1969; Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983; Jennison and Turnbull, 2000; Wassmer, 2000; Proschan et al., 2006). Similarly, the hypothesis H_i is potentially tested at K_i different calendar time points $T_i^c = \{t_{i1}^c, \ldots, t_{iK_i}^c\} \subset (0, t_{\max}^c]$, and the corresponding information time points or information fractions are $T_i = \{t_{i1}, \ldots, t_{iK_i}\}$. These hypotheses $\{H_i\}_{i=1}^n$ are tested only if some triggers are observed.

The triggers in a clinical trial include (1) significance triggers and (2) time triggers. The significance trigger counts on a total of n sequences of sets $\{S_{0i,k}\}_{k=1}^{K_0}$, $i = 1, \ldots, n$, for the observed interim test statistics $\{z_{0,k}\}_{k=1}^{K_0}$ of the trigger endpoint. The endpoint i will not be tested at calendar time $t_{0,k}^c$ if the test statistics of the trigger endpoint $z_{0,k'}$ does not fall into the corresponding set $S_{0i,k'}$ for all $k' = 1, \ldots, k$. We will start to test H_i after we observe $z_{0,k}$ falls in $S_{0i,k}$. Precisely, the hypothesis H_i will be tested since the $k_i^\star$th interim analysis, where the corresponding calendar time is

$$t_{i,k_i^\star}^c = \min\{t_i^c \in T_i^c : t_i^c \geq t_{0,k_i'}^c\}, \text{ where } k_i' = \min\{k \in \{1, \ldots, K_0\} : z_{0,k} \in S_{0i,k}\}.$$

In addition to the significance trigger, we can set a time trigger so that we can start to test H_i if we have waited long enough. For example, the ith endpoint will be tested since the $k_i^\dagger$th interim analysis, and the calendar time is

$$t_{i,k_i^\dagger}^c = \min\{t_i^c \in T_i^c : t_i^c \geq \tau_i^c\},$$

where τ_i^c serves as the time trigger for testing H_i. A common choice is to let $\tau_i^c = t_{i,K_i}^c$. With this timer trigger, we are allowed to test H_i at the final stage (K_ith stage) if no other trigger is triggered during the interim analyses. Moreover, for simplicity, let $\mathcal{T}_S$ denote the design with significance triggers, $\mathcal{T}_T$ denote the trial with time triggers, and $\mathcal{T}_{ST}$ denote the trigger strategy with both types of triggers.

A trial design with suitable triggers is generally more powerful than the design without triggers. When there is no hierarchy structure among endpoints in a clinical trial, different hypotheses are tested separately. This scheme is usually called the co-equal strategy (Jennison and Turnbull, 1993, 1997; Glimm et al., 2010). Consider a total of $n + 1$ co-equal

endpoints $H_0, H_1, \ldots, H_n$ which are tested individually. At the design stage, the critical boundaries are chosen based on

$$\Pr\left(\cup_{k=1}^{K_i}\{Z_{ik} > c_{ik}\}\right) = w_i\alpha, \text{ for } i = 0, 1, \cdots, n, \text{ and } \sum_{i=0}^{n} w_i = 1, \tag{1}$$

where α is the level of significance, Z_{ik} is the test statistic of testing H_i at kth stage, and c_{ik} is the corresponding critical boundary. Please note that in some closure-principle-based methods, for example, the group sequential trial using graphical approaches and the group sequential Holm procedure (Maurer and Bretz, 2013; Ye et al., 2013), although the critical boundary may be updated in the interim analysis if some endpoints appear to be significant at early stages, the maximum sample size estimation has to follow the worst case where no early rejection happens and uses the critical values calculated from equation (1).

Proposition 1. *In a Bonferroni-based multiple testing procedure with interim analyses, including a trigger strategy allows refinement of critical boundaries for testing non-trigger endpoints.*

The Bonferroni-based multiple testing procedures are commonly used in practice, because it is highly flexible, very simple to compute, and can be applied with flexible dependence assumptions. Many Bonferroni-based methods, for example, Holm (1979)'s procedure, the graphical approach (Bretz et al., 2009; Burman et al., 2009), can be extended and applied to repeated tests with multistages. For these methods, we can further increase the power by including hierarchical structures and triggers. For example, consider a clinical trial with a progression-free survival (PFS) endpoint H_0 and an overall survival (OS) endpoint H_1. Assume that the hazard ratios for PFS and OS are non-negatively correlated (Adunlin et al., 2015). Suppose one analysis for the PFS endpoint H_0 and two analyses for the OS endpoint H_1 are conducted. Let the significance level for one-sided test $\alpha = 0.05$, the type I error rate spent on the PFS endpoint $\alpha_0 = 0.02$, the information fraction for the interim analysis of the OS endpoint $t_1 = 0.6$. Using the logarithm of the hazard ratio as the test statistics, the critical boundary for testing H_0 is $c_0 = 2.054$. For testing H_1, using the Hwang-Shih-DeCani error spending function with parameter $\gamma = 1$ (Hwang et al., 1990), a direct Bonferroni type I error allocation results the critical boundaries for the interim and final analyses are $c_{11} = 2.025$ and $c_{12} = 2.156$. When applying a group sequential Holm procedure (Ye et al., 2013) or a group sequential trial using the graphical approach (Maurer and Bretz, 2013), the critical boundaries for testing H_1 are still $c_{11} = 2.025$ and $c_{12} = 2.156$ if H_0 is failed to be rejected. If we observe statistical significance from testing the PFS endpoint H_0, we update the critical boundaries for testing H_1 and use $c_{11} = 1.803$ $c_{12} = 1.917$. Alternatively, we can apply the trigger strategy and let H_0 be the trigger endpoint. If the trigger hypothesis H_0 is rejected, we test the OS endpoint H_1 at both the interim and final stages, using critical boundaries $c_{11} = 1.839$, $c_{12} = 1.872$. Otherwise, if H_0 is failed to be rejected, we skip the interim stage and directly test H_1 at the final stage with $c_{12} = 1.872$. Note that we only use one set of critical boundaries in this design using the trigger strategy. Next we compare the group sequential design using graphical approaches and the trigger strategy. For the time-to-event endpoint H_1, suppose the hazard ratio is 2.0 and the number of observations is 75. The drift parameter of the test statistic is therefore 3.0. When the trigger endpoint H_0

is failed to be rejected, the statistical power of graphical-approach-based group sequential design for testing H_1 is 80.7%, while the trigger strategy yields a power of 87.0% for testing H_1. If H_0 is rejected, the graphical approach has a power of 86.8% for testing H_1 and the trigger strategy achieves a power of 87.4%. We can observe that adding a suitable trigger into the design can boost the statistical power.

As a general method, trigger strategy can be applied to many multistage designs with multiple hypotheses including but not limited to the Bonferroni-based multiple testing procedures, and provide the potential for increasing the statistical power.

Proposition 2. *Trigger strategy can refine the critical boundaries used in multistage trials with multiple endpoints.*

Trigger strategy helps picking the right timing for some non-primary endpoints in a clinical trial. The beneficial effect can be quite significant when the trigger hypothesis is selected properly, the effect size of the non-primary endpoint is not large, and the sample size is limited.

3 Trigger strategy: a specific $\mathcal{T}_{ST}$ strategy

In this section we consider a type of trigger strategy $\mathcal{T}_{ST}$ with both significance and time triggers. For the sake of simplicity, we consider a one-sample normal hypothesis testing problem to start with. Other types of endpoints, like binomial endpoints, time-to-event endpoints (Lan and Lachin, 1990), can be considered in a similar way. Suppose a bivariate normal response (X, Y) for each patient is observed with unknown mean (θ_x, θ_y), known variance σ_x^2 and σ_y^2, and unknown correlation ρ. The effect sizes (δ_x, δ_y) for (X, Y) are equal to $(\theta_x/\sigma_x, \theta_x/\sigma_x)$. We would like to test two hypotheses $H_x : \theta_x = 0$ and $H_y : \theta_y = 0$ against their corresponding upper one-sided alternatives.

We use a group sequential design with $K + L$ grand stages, where $K \geq 1$ and $L \geq 0$. For the first K grand stages, each grand stage k ($k = 1, \cdots, K$) is separated into I_k stages ($I_k \geq 1$). Hypothesis H_x is planned to be tested at each stage in the first K grand stage, and the planned final stage is the last stage in grand stage K. Hypothesis H_y will be potentially tested at each grand stage. The final stage of testing H_y is the $(K + L)_{\text{th}}$ grand stage.

In this group sequential design, inspections are carried out after groups of observations at the corresponding calendar time t^c (Lan and DeMets, 1989). Denote the calendar time of testing H_x at grand stage k ($1 \leq k \leq K$), stage i ($1 \leq i \leq I_k$) by t^c_{xki}, and denote the calendar time of testing H_y at grand stage k ($1 \leq k \leq K + L$) by t^c_{yk}. Note that $t^c_{xkI_k} = t^c_{yk}$ ($1 \leq k \leq K$), since stage I_k is the last stage in the stages of the k_{th} grand stage. The following table illustrates the schedule of interim evaluations, where the vertical bars separate two consecutive grand stages, and the double vertical bar indicates the end of testing H_x.

Grand stage 1			Grand stage 2				Grand stage K				Grand stage $K+1$		Grand stage $K+L$
t^c_{x11}	$\cdots$	$t^c_{x1I_1}$	t^c_{x21}	$\cdots$	$t^c_{x2I_2}$	$\cdots$	t^c_{xK1}	$\cdots$	$t^c_{xKI_K}$	$\Vert$			
		t^c_{y1}			t^c_{y2}	$\cdots$			t^c_{yK}	$\Vert$	$t^c_{y,K+1}$	$\cdots$	$t^c_{y,K+L}$

Let H_x serve as the trigger hypothesis. We will start testing H_y immediately once H_x is rejected. In addition, we have a time trigger at t^c_{yK}: if H_x is failed to be rejected at its final stage at $t^c_{xKI_K}$, we will begin to test H_y at t^c_{yK} until we reject H_y or reach the final stage of testing H_y at $t^c_{y,K+L}$. This procedure using trigger strategy is:

Step 0. Select a total of $\sum_{k=1}^{K} I_k$ time points $\{t^c_{x11}, \dots t^c_{xKI_K}\}$ for testing H_x, where $t^c_{xki} < t^c_{xk'i'}$ if $k < k'$, and $t^c_{xki} < t^c_{xkj}$ if $i < j$. Set $t^c_{y1} = t^c_{x1I_1}, \dots, t^c_{yK} = t^c_{xKI_K}$ as the first K time points to test H_y potentially. In addition, include another L time points $t^c_{y,K+1}, \dots, t^c_{K+L}$ for testing H_y.

Step 1. Test H_x at each time point. Keep the test of H_y paused. If H_x is rejected at calendar time t^c_{xki}, we start to test H_y at time t^c_{yk}. If we fail to reject H_x at its final stage, we will test H_y starting from time t^c_{yK}.

Step 2. Continue testing H_y until either it is rejected or we reach its final stage at time $t^c_{y,K+L}$.

For example, consider a group sequential design with four grand stages, where $K = 2$ and $L = 2$. Each of the first two grand stages includes two stages ($I_1 = 2$ and $I_2 = 2$). Let calendar time (in month) $t^c_{x11} = 3$, $t^c_{x12} = 6$, $t^c_{x21} = 9$, $t^c_{x22} = 12$, $t^c_{y1} = 6$, $t^c_{y2} = 12$, $t^c_{y3} = 18$, and $t^c_{y4} = 24$. In this design, H_x will be tested at month 3, 6, 9 and 12, while H_y will be potentially tested at month 6, 12, 18 and 24.

While interim assessments occur in calendar time, the interim test statistics Z_x and Z_y depend on the information available. Information time t, which is another time scale in group sequential design, is the proportion of maximum information that has already been observed, where $0 \le t \le 1$. Denote the information time of testing H_x and H_y at a specific interim stage by $t_{xki} = \mathcal{I}_x(t^c_{xki})/\mathcal{I}_x(t^c_{xKI_K})$ and $t_{yk} = \mathcal{I}_y(t^c_{yk})/\mathcal{I}_y(t^c_{y,K+L})$, where $\mathcal{I}_x(t^c)$ and $\mathcal{I}_y(t^c)$ are the Fisher information of X and Y at calendar time t^c respectively. In the setting of one-sample hypothesis test, the information time of X at calendar time t^c_{xki} and that of Y at t^c_{yk} are $t_{xki} = n_{xki}/n_{xKI_K}$ and $t_{yk} = n_{yk}/n_{y,K+L}$, where n_{xki} is the sample size of X at time t^c_{xki} and n_{yk} is the sample size of Y at time t^c_{yk}. The maximum sample sizes of X and Y are n_{xKI_K} and $n_{y,K+L}$. For simplicity, we also denote the maximum sample sizes n_{xKI_K} and $n_{y,K+L}$ by $n_x^{\max}$ and $n_y^{\max}$.

The standardized sample mean test statistic of X at grand stage k, stage i is denoted by Z_{xki}, which is

$$Z_{xki} = \frac{\sum_{l=1}^{n_{xki}} X_l}{\sqrt{n_{xki}\sigma_x^2}}$$

with mean $\sqrt{t_{xki}}\Delta_x$ and variance 1, where the drift parameter $\Delta_x = \delta_x\sqrt{n_x^{\max}}$ in the setting of one-sample hypothesis test. Similarly, the test statistic of Y at the k_{th} grand stage is denoted by Z_{yk} with drift parameter Δ_y. The correlations between test statistics in different stages are shown as follows. For any $t_{xk_1i_1} < t_{xk_2i_2}$ and $t_{yk_1} < t_{yk_2}$,

$$\text{corr}\,(Z_{xk_1i_1}, Z_{xk_2i_2}) = \sqrt{\frac{t_{xk_1i_1}}{t_{xk_2i_2}}}, \quad \text{corr}\,(Z_{yk_1}, Z_{yk_2}) = \sqrt{\frac{t_{yk_1}}{t_{yk_2}}}. \tag{2}$$

For correlations between test statistics for testing H_x and H_y, we consider a general situation in which some pairs of observations are incomplete—in other words, either X value or Y value is missing in some bivariate responses. Consider a sequence of observations $\{X_l\}_{l=1}^{n_{xki}}$ until calendar time t_{xki}^c and another sequence $\{Y_l\}_{l=1}^{n_{yj}}$ until calendar time t_{yj}^c. We denote the number of the complete pairs $\{(X_l, Y_l)\}_{l\leq \min\{n_{xki}, n_{yj}\}}$ by $n_{xki,yj}$. It is clear that $n_{xki,yj} \leq \min\{n_{xki}, n_{yj}\}$ and the equality holds only when all pairs of observations are complete. We can show that $\text{corr}\,(Z_{xki}, Z_{yj}) = \frac{\rho \cdot n_{xki,yj}}{\sqrt{n_{xki}}\sqrt{n_{yj}}}$. By defining an information time of paired data $t_{xki,yj} = \rho \cdot n_{xki,yj}/\sqrt{n_x^{\max} n_y^{\max}}$, the correlation between Z_{xki} and Z_{yj} is shown as follows.

$$\text{corr}\,(Z_{xki}, Z_{yj}) = \frac{\rho \cdot t_{xki,yj}}{\sqrt{t_{xki}}\sqrt{t_{yj}}} \tag{3}$$

Note that when all pairs of observations are complete, it follows that $t_{xki,yj} = \min\{t_{xki}, t_{yj}\}$, and equation (3) can be simplified to $\text{corr}\,(Z_{xki}, Z_{yj}) = \rho \cdot \sqrt{\min\{t_{xki}, t_{yj}\}/\max\{t_{xki}, t_{yj}\}}$.

Besides monitoring the Z-test statistic, a transformed Z-value, which is called the B-value by Lan and Wittes (1988), can also be computed as a discrete Brownian motion process. In this group sequential design, the B-values

$$B_{xki} = Z_{xki}\sqrt{t_{xki}}, \quad B_{yk} = Z_{yk}\sqrt{t_{yk}}$$

have mean $\Delta_x t_{xki}$ and $\Delta_y t_{yk}$, and variance t_{xki} and t_{yk}. The B-values have the same correlation structures with the Z-value, as shown in equation (2) and (3).

Denote the critical boundary for testing H_x by c_{xki} where $i = 1, \ldots, I_k$, $k = 1, \ldots, K$ and that for testing H_y by c_{yk} where $k = 1, \ldots, K+L$. Critical boundary c_{xki} is associated with the test statistic Z_{xki} at calendar time t_{xki}^c, and c_{yk} is associated with Z_{yk} at time t_{yk}^c. These critical boundaries are calculated to satisfy the strong control of the familywise error rate (FWER) for any specified level α (Hochberg and Tamhane, 1987; Gou et al., 2014; Tamhane and Gou, 2018; Gou and Tamhane, 2018). By following the closed testing principle by Marcus et al. (1976), the control of the FWER at level α requires (i) $\Pr\,(\text{reject } H_x \mid H_x) \leq \alpha$, (ii) $\Pr\,(\text{reject } H_y \mid H_y) \leq \alpha$, and (iii) $\Pr\,(\text{reject } H_x \text{ or } H_y \mid H_x \cap H_y) \leq \alpha$. Under this $\mathcal{T}_{ST}$ strategy, these probabilities are

$$\Pr\,(\text{reject } H_x \mid H_x) = 1 - \Pr\left(\cap_{k=1}^{K} \cap_{i=1}^{I_k} \{Z_{xki} \leq c_{xki}\}\right), \tag{4}$$

$$\begin{aligned}\Pr\,(\text{reject } H_y \mid H_y) &= 1 - \Pr\left(\cap_{k=1}^{K+L} \{Z_{yk} \leq c_{yk}\}\right) \\ &\quad - \sum_{j=1}^{K-1} \Pr\left(\left\{\cap_{k=1}^{j} \cap_{i=1}^{I_k} \{Z_{xki} \leq c_{xki}\}\right\} \cap \{Z_{yj} > c_{yj}\} \cap \left\{\cap_{k=j+1}^{K+L} \{Z_{yk} \leq c_{yk}\}\right\}\right),\end{aligned} \tag{5}$$

$$\begin{aligned}&\Pr\,(\text{reject } H_x \text{ or } H_y \mid H_x \cap H_y) \\ &\quad = 1 - \Pr\left(\left\{\cap_{k=1}^{K} \cap_{i=1}^{I_k} \{Z_{xki} \leq c_{xki}\}\right\} \cap \left\{\cap_{k=K}^{K+L} \{Z_{yk} \leq c_{yk}\}\right\}\right).\end{aligned} \tag{6}$$

Denote the probability $\Pr\,(\text{reject } H_x \mid H_x)$ in equation (4) by α_x. We also define the marginal significance level for H_y by α_y in equation (7)

$$\alpha_y = 1 - \Pr\left(\cap_{k=1}^{K+L} \{Z_{yk} \leq c_{yk}\}\right), \tag{7}$$

where the marginal significance level is the probability of rejecting the null hypothesis when each single hypothesis is tested individually in a group sequential design. The following proposition shows the relation between the marginal significance level of testing H_y and the type I error rate under H_y.

Proposition 3. *The type I error rate* $\Pr(\textit{reject } H_y \mid H_y)$ *under the strategy* $\mathcal{T}_{ST}$ *is bounded from above by* α_y*. This upper bound is sharp if and only if either (i) or (ii) holds: (i) for each* $j = 1, \ldots, K-1$*, there exist index* $k_j \leq j$ *and index* $1 \leq i_j \leq I_{k_j}$ *such that* $t_{yj} = t_{xk_ji_j}$*; in addition,* $\rho = 1$ *and* $\Delta_x \geq (c_{xk_ji_j} - c_{yj})/\sqrt{t_{yj}}$*, (ii)* $\Delta_x = +\infty$*.*

A consequence of Proposition 3 is that we may use α_y instead of $\Pr(\text{reject } H_y \mid H_y)$ for the type I error control under H_y when there is no prior information on Δ_x. Next consider the type I error rate under $H_x \cap H_y$.

Proposition 4. *Consider a design following the strategy* $\mathcal{T}_{ST}$ *satisfying* $\alpha_x = \Pr(\textit{reject } H_x \mid H_x) < \alpha$*. A necessary and sufficient condition for* $\Pr(\textit{reject } H_x \textit{ or } H_y \mid H_X \cap H_Y) \leq \alpha$ *for any non-negative correlation* ρ *is* $\Pr\left(\cap_{k=K}^{K+L}\{Y_k \leq c_{yk}\}\right) \geq (1-\alpha)/(1-\alpha_x)$*.*

Based on Proposition 3 and 4, a $\mathcal{T}_{ST}$ strategy design can be constructed by applying the closed testing principle. Before presenting Proposition 5, it is useful to define the partial marginal significance level for testing H_y in equation 8 by α_{y-}

$$\alpha_{y-} = 1 - \Pr\left(\cap_{k=K}^{K+L}\{Y_k \leq c_{yk}\}\right), \tag{8}$$

which is the significance level of testing H_y in a group sequential design with $L+1$ stages by using critical boundary $(c_{yK}, c_{y,K+1}, \ldots, c_{y,K+L})$. If we intentionally skip the first $K-1$ stages in the $(K+L)$-stage design for H_y, the corresponding type I error rate will be α_{y-}.

Proposition 5. *A* $\mathcal{T}_{ST}$ *strategy design controls the familywise error rate (FWER) at level* α *for non-negative correlated* X *and* Y*, if the following three conditions are satisfied: (i)* $\alpha_x \leq \alpha$*, (ii)* $\alpha_y \leq \alpha$*, and (iii)* $\alpha_{y-} \leq (\alpha - \alpha_x)/(1-\alpha_x)$*.*

Proposition 5 provides a simple method to construct a $\mathcal{T}_{ST}$ strategy design. By using Proposition 5, the calculation of the critical boundary for testing H_x and that for H_y are relatively independent and self-sufficient. Only α_x, the significance level for H_x, need to be known when computing the boundary for H_y.

4 Power analysis

In this section we first bring in a proposition about power comparison based on a simple trigger strategy trial, then we conduct numerical calculation to show the power difference. The simple trigger strategy involves one primary hypothesis H_x as the trigger endpoint, and one key secondary hypothesis H_y. The endpoint H_x is only tested once. If H_x is rejected, the endpoint H_y will be tested twice in both the interim and final stages, otherwise H_y will be only tested once at the final analysis. We call this method the simple $\mathcal{T}_{ST}$ strategy. The example with PFS as the primary endpoint and OS as the key secondary endpoint in Section 2 follows the simple $\mathcal{T}_{ST}$ strategy. Proposition 6 shows that the simple $\mathcal{T}_{ST}$ strategy is uniformly more powerful for testing H_y than the Bonferroni-based method when H_x is failed to be rejected.

Proposition 6. *Consider testing H_x using a single-stage test and H_y using a two-stage test at level α. The statistical power for testing H_y conditioning on that H_x is not rejected using the simple $\mathcal{T}_{ST}$ strategy is denoted by $P_{\mathcal{T}}$, and the power under the same condition using the Bonferroni-based method is denoted by P_B. Then $P_{\mathcal{T}} > P_B$ and the ratio $P_{\mathcal{T}}/P_B$ is approximately bounded from below by*

$$\frac{P_{\mathcal{T}}}{P_B} \gtrsim 1 + \frac{\alpha_x \cdot \phi(\Delta_y - z_{\alpha-\alpha_x})}{z_{\alpha-\alpha_x}}, \tag{9}$$

where α_x denotes, Δ_y denotes the drift parameter of test statistic for H_y, $z_{\alpha-\alpha_x}$ denotes $\Phi^{-1}(1-\alpha+\alpha_x)$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the univariate standard probability density function and distribution function.

Next we consider the trigger strategy described in Section 3 with multiple stages for each endpoint. We follow the multivariate normal setup and assume hypotheses are tested at one-sided 5% significance level. For non-normal setup, the power analysis model provided in Zhang and Gou (2016) can be an alternative. The primary endpoint is tested at calendar time (month) $t_x^c = c(3, 6, 9, 12, 18)$ and the key secondary endpoint H_y is tested at calendar time (moth) $t_y^c = (6, 12, 18, 36)$. Following the notations in Section 3, these calendar time points are $t_{x11}^c = 3$, $t_{x12}^c = 6$, $t_{x21}^c = 9$, $t_{x22}^c = 12$, $t_{x31}^c = 18$, $t_{y1}^c = 6$, $t_{y2}^c = 12$, $t_{y3}^c = 18$, and $t_{y4}^c = 24$. The information time points are assumed to be equispaced for both H_x and H_y. The critical boundaries for testing H_x are calculated using the Hwang et al. (1990) error spending function (HSD) with parameter $\gamma = -4$ under level $\alpha_x = 1\%$. These boundaries are $c_x = (3.505, 3.261, 2.993, 2.707, 2.398)$ For the critical boundaries for testing H_y, these values based on the direct Bonferroni method and those based on the trigger strategy are compared in Table 1. The error spending functions (ESFs) include the O'Brien and Fleming (1979)-like Lan and DeMets (1983) (OBF), Pocock (1977)-like Lan and DeMets (1983) (POC), and Hwang et al. (1990) (HSD) with parameters $\gamma = -4$, -2 and 1. The HSD error spending function with $\gamma = -4$ or -5 is similar to OBF and that with $\gamma = 1$ is similar to POC. For the calculation of the critical boundary of trigger strategy, we first calculate the tail half of the critical boundary vector using equation (8), and then calculate the front half using equation (7) given the tail half.

Table 1: Comparison between the Bonferroni-based unrefined boundary and the refined boundary using trigger strategy for testing H_y, where $\alpha = 2.5\%$, $\alpha_x = 1.0\%$, $t_x = (0.2, 0.4, 0.6, 0.8, 1.0)$, $t_y = (0.25, 0.50, 0.75, 1.00)$, $t_x^c = (3, 6, 9, 12, 18)$ and $t_y^c = (6, 12, 18, 36)$

ESF	Unrefined boundary				Refined boundary			
HSD(-4)	3.301	2.982	2.624	2.227	2.916	2.196	2.549	2.214
HSD(-2)	2.963	2.757	2.535	2.304	2.646	2.233	2.397	2.274
HSD(1)	2.559	2.553	2.553	2.566	2.360	2.420	2.237	2.475
OBF	4.726	3.248	2.591	2.213	3.251	2.178	2.573	2.208
POC	2.552	2.563	2.561	2.558	2.362	2.419	2.240	2.467

In Table 2 we compare the statistical power Pr(reject $H_y \mid \overline{H}_y$), using the direct Bonferroni method (Bonf), Ye et al. (2013)'s group sequential Holm procedure (Holm-Ye), and

the trigger strategy $\mathcal{T}_{ST}$ (Trigger). The Holm-Ye method allows using the significance level α instead of $\alpha - \alpha_x$ once H_x is rejected during a group sequential trial. The drift parameter Δ_y ranges from 1 to 4, and the correlation coefficient ρ defined in Section 3 includes 0 (independence) and 0.5 (dependence). We use the HSD error spending function with $\gamma = -4$ and $\gamma = 1$ to calculate the critical boundaries. These numbers are calculated using numerical integration. The methods with the greatest statistical power under these combinations are shown in bold. From Table 2, we can observe that Ye et al. (2013)'s group sequential Holm procedure and the trigger strategy $\mathcal{T}_{ST}$ are more powerful than the direct Bonferroni method. The trigger strategy is more powerful than the Holm-Ye method under positive dependence or using Pocock-like error spending functions, like HSD($\gamma = 1$). Under independence and using O'Brien-Fleming-like error spending functions, like HSD($\gamma = -4$), the Holm-Ye method is more powerful than the trigger strategy, but the power difference is relatively small.

Table 2: Power (%) comparison among the direct Bonferroni method (Bonf), Ye et al. (2013)'s group sequential Holm procedure (Holm-Ye), and the trigger strategy $\mathcal{T}_{ST}$ (Trigger) for testing H_y, where $\alpha = 2.5\%$, $\alpha_x = 1.0\%$, $t_x = (0.2, 0.4, 0.6, 0.8, 1.0)$, $t_x^c = (3, 6, 9, 12, 18)$, $\Delta_x = 1$, $t_y^c = (6, 12, 18, 36)$, and $t_y = (0.25, 0.50, 0.75, 1.00)$

ESF	Δ_y	$\rho = 0.0$			$\rho = 0.5$		
		Bonf	Holm-Ye	Trigger	Bonf	Holm-Ye	Trigger
HSD(−4)	$\Delta_y = 1$	11.74	**12.17**	12.04	11.74	**12.59**	12.17
	$\Delta_y = 2$	42.17	**42.93**	42.77	42.17	42.66	**42.71**
	$\Delta_y = 3$	78.71	**79.22**	79.15	78.71	78.82	**79.11**
	$\Delta_y = 4$	96.35	**96.47**	96.46	96.35	96.32	**96.46**
HSD(1)	$\Delta_y = 1$	9.40	9.73	**10.83**	9.40	10.23	**10.95**
	$\Delta_y = 2$	34.68	35.40	**38.65**	34.68	35.34	**38.65**
	$\Delta_y = 3$	71.29	71.92	**74.99**	71.29	71.58	**74.99**
	$\Delta_y = 4$	93.76	93.96	**95.03**	93.76	93.75	**94.99**

5 Discussion

Relying on the decision of testing the prespecified trigger hypotheses, the method applying the trigger strategy can choose suitable starting stages for testing other hypotheses in an adaptive manner. Because of the possibility that a trial may skip the first few interim stages, the trigger strategy provides us space to refine the critical boundaries. Meanwhile, trigger strategy is a flexible framework, where various triggers can be applied and different types of information can be included. For example, if the correlation information is known, we can further refine the critical boundaries and increase the statistical power. Table 3 shows the further refined critical boundaries under the same setting in Section 4, Table 1 and 2. With knowing the value of the correlation coefficient, the critical boundaries are further refined comparing with the boundaries in Table 1, and the statistical powers are improved comparing with the values calculated in Table 2 when $\rho = 0.5$.

Table 3: Critical boundaries and statistical powers (%) for testing H_y using the trigger strategy $\mathcal{T}_{ST}$ with incorporating the correlation information, where $\alpha = 2.5\%$, $\alpha_x = 1.0\%$, $t_x = (0.2, 0.4, 0.6, 0.8, 1.0)$, $t_x^c = (3, 6, 9, 12, 18)$, $\Delta_x = 1$, $t_y^c = (6, 12, 18, 36)$, $t_y = (0.25, 0.50, 0.75, 1.00)$ and $\rho = 0.5$

Refined boundary	c_{y1}	c_{y2}	c_{y3}	c_{y4}
HSD(-4)	2.953	2.243	2.514	2.173
HSD(1)	2.404	2.466	2.200	2.437
Power (%)	$\Delta_y = 1$	$\Delta_y = 2$	$\Delta_y = 3$	$\Delta_y = 4$
HSD(-4)	13.00	44.36	80.27	96.78
HSD(1)	11.47	40.12	76.16	95.43

Acknowledge

The author thanks Dana Sylvan, Olympia Hadjiliadis, Xiaojia Lu and Sean Ammirati for helpful discussion. Some preliminary results have been present at the ENAR 2019 Spring Meeting, Session 105 Biopharmaceutical research and clinical trials, *Critical boundary refinement for the generalized partially hierarchical test procedure in group sequential trials* on March 27, 2019.

Conflict of Interest

The authors have declared no conflict of interest.

Appendix

Proof of Proposition 1. In the design stage of a Bonferroni-based multiple testing procedure with multiple interim analyses, the critical boundaries are calculated using equation (1) with a choice of error spending functions (Lan and DeMets, 1983; Proschan, 1999; Lan and DeMets, 2009), and are denoted by $\{c_{ik}, k = 1, \ldots, K_i, i = 0, 1, \ldots, n\}$. Using the trigger strategy with a significance trigger, the type I error rate for testing a non-trigger endpoint H_i, $i \in \{1, \ldots, n\}$, is

$$\begin{aligned}&\Pr\left(\cup_{k=1}^{K_i}\left\{\cup_{k'=1}^{K_0}\{Z_{0,k'} \in S_{0i,k'}\}\right\} \cap \{Z_{ik} > c_{ik}\} \cap \left\{t_{0,k'}^c \le t_{i,k}^c\right\}\right)\\ &\le \Pr\left(\cup_{k=1}^{K_i}\{Z_{ik} > c_{ik}\}\right) = w_i\alpha.\end{aligned}$$

Similarly, when applying a time trigger, the type I error rate for testing H_i is

$$\Pr\left(\left\{\cup_{k=1}^{K_i}\{Z_{ik} > c_{ik}\}\right\} \cap \{t_{ik}^c \ge \tau_i^c\}\right) \le \Pr\left(\cup_{k=1}^{K_i}\{Z_{ik} > c_{ik}\}\right) = w_i\alpha.$$

Therefore, we can refine the critical boundaries and have $\{c'_{ik}, k = 1, \ldots, K_i, i = 0, 1, \ldots, n\}$ that satisfies $c'_{ik} \le c_{ik}$. □

Proof of Proposition 2. The proof is similar to the proof of Proposition 1. The familywise error rate control of a multiple testing procedure is based on the control of type I error rate of a family of intersection hypotheses $\cap_{i\in I} H_i$, $I \subset \{0, 1, \ldots, n\}$, which is

$$\Pr\left(\cup_{i\in I} \cup_{k=1}^{K_i} \{Z_{ik} > c_{ik}\}\right) = \alpha_I,$$

where α_I is the significance level of testing the intersection of null hypotheses in I. When applying the trigger strategy, the type I error rate with a significance trigger is

$$\Pr\left(\cup_{i\in I} \left\{\cup_{k=1}^{K_i} \left\{\cup_{k'=1}^{K_0} \{Z_{0,k'} \in S_{0i,k'}\}\right\} \cap \{Z_{ik} > c_{ik}\} \cap \left\{t_{0,k'}^c \le t_{i,k}^c\right\}\right\}\right) \le \alpha_I,$$

and that with a time trigger is

$$\Pr\left(\cup_{i\in I} \left\{\left\{\cup_{k=1}^{K_i} \{Z_{ik} > c_{ik}\}\right\} \cap \{t_{ik}^c \ge \tau_i^c\}\right\}\right) \le \alpha_I.$$

Therefore, the critical boundaries can be refined. □

Proof of Proposition 3. From equation (5), the upper bound follows directly by noting that

$$P_j = \Pr\left(\left\{\cap_{k=1}^{j} \cap_{i=1}^{I_k} \{Z_{xki} \le c_{xki}\}\right\} \cap \{Z_{yj} > c_{yj}\} \cap \left\{\cap_{k=j+1}^{K+L} \{Z_{yk} \le c_{yk}\}\right\}\right) \ge 0$$

for any $j = 1, \ldots, K-1$. Under H_y, we have $\Delta_y = 0$, and the upper bound α_y is achieved when and only when all $P_j = 0$, $j = 1, \ldots, K-1$. If $\rho < 1$, correlation $\text{corr}(Z_{xki}, Z_{yj})$ is strictly less than 1, it follows that $P_j = 0$ only when $\Delta_x = +\infty$. If $\rho = 1$ and Δ_x is finite, $P_j > 0$ unless the distribution of $(Z_{x11}, \ldots, Z_{xjI_j}, Z_{yj})$ is degenerate. In this case, there exist a grand stage k_j and a stage i_j such that $t_{xk_ji_j} = t_{yj}$ and $Z_{xk_ji_j} = Z_{yj} + \sqrt{t_{yj}}\Delta_x$. When $\{Z_{yj} + \sqrt{t_{yj}}\Delta_x \le c_{xk_ji_j}\} \cap \{Z_{yj} > c_{yj}\} = \emptyset$, we have $P_j = 0$. □

Proof of Proposition 4. In a strategy $\mathcal{T}_{ST}$ design, the probability $\Pr(\text{reject } H_x \text{ or } H_y \mid H_X \cap H_Y) \ge \Pr(\text{reject } H_x \mid H_X \cap H_Y) = \Pr(\text{reject } H_x \mid H_X)$. From equation 4 and 6, we compute the difference

$$\begin{aligned}
&\Pr(\text{reject } H_x \text{ or } H_y \mid H_x \cap H_y) - \Pr(\text{reject } H_x \mid H_x) \\
&= \Pr\left(\cap_{k=1}^{K} \cap_{i=1}^{I_k} \{Z_{xki} \le c_{xki}\}\right) - \Pr\left(\left\{\cap_{k=1}^{K} \cap_{i=1}^{I_k} \{Z_{xki} \le c_{xki}\}\right\} \cap \left\{\cap_{k=K}^{K+L} \{Z_{yk} \le c_{yk}\}\right\}\right) \\
&\le \Pr\left(\cap_{k=1}^{K} \cap_{i=1}^{I_k} \{Z_{xki} \le c_{xki}\}\right) \left(1 - \Pr\left(\cap_{k=K}^{K+L} \{Z_{yk} \le c_{yk}\}\right)\right),
\end{aligned}$$

where $\Pr\left(\left\{\cap_{k=1}^{K} \cap_{i=1}^{I_k} \{Z_{xki} \le c_{xki}\}\right\} \cap \left\{\cap_{k=K}^{K+L} \{Z_{yk} \le c_{yk}\}\right\}\right) \ge \Pr\left(\cap_{k=1}^{K} \cap_{i=1}^{I_k} \{Z_{xki} \le c_{xki}\}\right) \cdot \Pr\left(\cap_{k=K}^{K+L} \{Z_{yk} \le c_{yk}\}\right)$ holds for any non-negative ρ by using Slepian (1962)'s inequality. If $\Pr\left(\cap_{k=K}^{K+L} \{Y_k \le c_{yk}\}\right) \ge (1-\alpha)/(1-\alpha_x)$, then $\Pr(\text{reject } H_x \text{ or } H_y \mid H_x \cap H_y) \le \alpha_x + (1-\alpha_x) \cdot (1 - (1-\alpha)/(1-\alpha_x)) = \alpha$. Meanwhile, Slepian (1962)'s inequality guarantees that $\Pr\left(\left\{\cap_{k=1}^{K} \cap_{i=1}^{I_k} \{Z_{xki} \le c_{xki}\}\right\} \cap \left\{\cap_{k=K}^{K+L} \{Z_{yk} \le c_{yk}\}\right\}\right)$ is an increasing function of ρ. It follows that $\alpha_x + (1-\alpha_x) \cdot \left(1 - \Pr\left(\cap_{k=K}^{K+L} \{Z_{yk} \le c_{yk}\}\right)\right)$ is the maximum value of the probability $\Pr(\text{reject } H_x \text{ or } H_y \mid H_x \cap H_y)$ when $\rho \ge 0$. Consequently, the type I error control of testing $H_x \cap H_y$ at level α leads to $\alpha_x + (1-\alpha_x) \cdot \left(1 - \Pr\left(\cap_{k=K}^{K+L} \{Z_{yk} \le c_{yk}\}\right)\right) \le \alpha$, which is $\Pr\left(\cap_{k=K}^{K+L} \{Y_k \le c_{yk}\}\right) \ge (1-\alpha)/(1-\alpha_x)$. □

Proof of Proposition 5. This proposition follows from Proposition 3 and 4 via the closed testing principle. □

Proof of Proposition 6. Given that H_x is failed to be rejected, the Bonferroni-based methods, including the graphical approach and the group sequential Holm procedure, test H_y at level $\alpha - \alpha_x$. For a fixed drift parameter Δ_y, the statistical power of testing H_y is bounded from above by $P_B \leq \Phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right) + \Delta_y\right)$ (Jennison and Turnbull, 2000). By Proposition 5, the power of testing H_y using the simple $\mathcal{T}_{ST}$ strategy is

$$P_{\mathcal{T}} = \Phi\left(\Phi^{-1}\left(\frac{\alpha - \alpha_x}{1 - \alpha_x}\right) + \Delta_y\right) \approx \Phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right) + \frac{\alpha_x\left(\alpha - \alpha_x\right)}{\phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right)\right)} + \Delta_y\right)$$
$$\approx \Phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right) + \Delta_y\right) + \frac{\phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right) + \Delta_x\right)}{\phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right)\right)} \cdot \alpha_x\left(\alpha - \alpha_x\right).$$

Thus

$$\frac{P_{\mathcal{T}}}{P_B} \gtrsim 1 + \frac{\phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right) + \Delta_x\right)}{\Phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right) + \Delta_x\right)} \cdot \frac{\alpha_x\left(\alpha - \alpha_x\right)}{\phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right)\right)} \approx 1 + \frac{\alpha_x \cdot \phi\left(\Phi^{-1}\left(\alpha - \alpha_x\right) + \Delta_x\right)}{\Phi^{-1}\left(1 - \alpha + \alpha_x\right)},$$

where $\phi(-c)/\Phi(-c) \approx c$ for $c > 0$ using the Mills ratio (Mills, 1926). Therefore, $P_{\mathcal{T}} > P_B$ and the result in equation (9) is obtained. □

References

Adunlin, G., Cyrus, J. W. W., and Dranitsaris, G. (2015). Correlation between progression-free survival and overall survival in metastatic breast cancer patients receiving anthracyclines, taxanes, or targeted therapies: a trial-level meta-analysis. *Breast Cancer Research and Treatment* **154,** 591—-608.

Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)* **132,** 235–244.

Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* **28,** 586–604.

Burman, C.-F., Sonesson, C., and Guilbaud, O. (2009). A recycling framework for the construction of bonferroni-based multiple tests. *Statistics in Medicine* **28,** 739–761.

Fudenberg, D. and Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54,** 533–554.

Glimm, E., Maurer, W., and Bretz, F. (2010). Hierarchical testing of multiple endpoints in group-sequential trials. *Statistics in Medicine* **29,** 219–228.

Gou, J. and Chén, O. Y. (2019). Critical boundary refinement in a group sequential trial when the primary endpoint data accumulate faster than the secondary endpoint. In Zhang,

L., Chen, D.-G. D., Jiang, H., Li, G., and Quan, H., editors, *Contemporary Biostatistics with Biopharmaceutical Applications*, pages 205–224. Springer International Publishing, Cham.

Gou, J. and Tamhane, A. C. (2018). Hochberg procedure under negative dependence. *Statistica Sinica* **28,** 339–362.

Gou, J., Tamhane, A. C., Xi, D., and Rom, D. (2014). A class of improved hybrid Hochberg-Hommel type step-up multiple test procedures. *Biometrika* **101,** 899–911.

Gou, J. and Xi, D. (2019). Hierarchical testing of a primary and a secondary endpoint in a group sequential design with different information times. *Statistics in Biopharmaceutical Research* **11,** 398–406.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley and Sons, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6,** 65–70.

Hung, H. M. J., Wang, S.-J., and O'Neill, R. (2007). Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *Journal of Biopharmaceutical Statistics* **17,** 1201–1210.

Hwang, I. K., Shih, W. J., and DeCani, J. S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine* **9,** 1439–1445.

Jennison, C. and Turnbull, B. W. (1993). Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49,** 741–752.

Jennison, C. and Turnbull, B. W. (1997). Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92,** 1330–1341.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, New York.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70,** 659–663.

Lan, K. K. G. and DeMets, D. L. (1989). Group sequential procedures: calendar versus information time. *Statistics in Medicine* **8,** 1191–1198.

Lan, K. K. G. and DeMets, D. L. (2009). Further comments on the alpha spending function. *Statistics in Biosciences* **1,** 95–111.

Lan, K. K. G. and Lachin, J. M. (1990). Implementation of group sequential logrank tests in a maximum duration trial. *Biometrics* **46,** 759–770.

Lan, K. K. G. and Wittes, J. (1988). The B-value: A tool for monitoring data. *Biometrics* **44,** 579–585.

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.

Maurer, W. and Bretz, F. (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research* **5,** 311–320.

Mills, J. P. (1926). Table of the ratio: Area to bounding ordinate, for any portion of normal curve. *Biometrika* **18,** 395–400.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35,** 549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191–199.

Proschan, M. A. (1999). Properties of spending function boundaries. *Biometrika* **86,** 466–473.

Proschan, M. A., Lan, K. K. G., and Wittes, J. T. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer, New York.

Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal* **41,** 463–501.

Tamhane, A. C. and Gou, J. (2018). Advances in p-value based multiple test procedures. *Journal of Biopharmaceutical Statistics* **28,** 10–27.

Tamhane, A. C., Gou, J., Jennison, C., Mehta, C. R., and Curto, T. (2018). A gatekeeping procedure to test a primary and a secondary endpoint in a group sequential design with multiple interim looks. *Biometrics* **74,** 40–48.

Tamhane, A. C., Mehta, C. R., and Liu, L. (2010). Testing a primary and a secondary endpoint in a group sequential design. *Biometrics* **66,** 1174–1184.

Wassmer, G. (2000). Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers* **41,** 253–279.

Ye, Y., Li, A., Liu, L., and Yao, B. (2013). A group sequential Holm procedure with multiple primary endpoints. *Statistics in Medicine* **32,** 1112–1124.

Zhang, F. and Gou, J. (2016). A p-value model for theoretical power analysis and its applications in multiple testing procedures. *BMC Medical Research Methodology* **16,** 135.

Zhang, F. and Gou, J. (2020). Refined critical boundary with enhanced statistical power for non-directional two-sided tests in group sequential designs with multiple endpoints. *Statistical Papers* , to appear, `https://doi.org/10.1007/s00362-019-01134-7`.