Allocating recycled significance levels in group sequential procedures for multiple endpoints

Dong Xi*[1] and **Ajit C. Tamhane**[2]

[1] IIS Statistical Methodology, Novartis Pharmaceuticals Corporation, One Health Plaza, East Hanover, NJ 07936, USA

[2] Department of Industrial Engineering and Management Sciences, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA

Received 3 August 2013; revised 15 July 2014; accepted 11 September 2014

Graphical approaches have been proposed in the literature for testing hypotheses on multiple endpoints by recycling significance levels from rejected hypotheses to unrejected ones. Recently, they have been extended to group sequential procedures (GSPs). Our focus in this paper is on the allocation of recycled significance levels from rejected hypotheses to the stages of the GSPs for unrejected hypotheses. We propose a delayed recycling method that allocates the recycled significance level from Stage r onward, where r is prespecified. We show that r cannot be chosen adaptively to coincide with the random stage at which the hypothesis from which the significance level is recycled is rejected. Such an adaptive GSP does not always control the FWER. One can choose r to minimize the expected sample size for a given power requirement. We illustrate how a simulation approach can be used for this purpose. Several examples, including a clinical trial example, are given to illustrate the proposed procedure.

Keywords: Error spending function; Familywise error rate; Graphical approach; O'Brien-Fleming boundary; Pocock boundary; Recycling.

1 Introduction

Bretz et al. (2009) and Burman et al. (2009) proposed graphical approaches for testing multiple hypotheses using weighted Bonferroni tests, which recycle significance levels from rejected hypotheses to unrejected ones. Maurer and Bretz (2013) extended these approaches to group sequential procedures (GSPs). They did not explicitly address the problem of the timing of the allocation of recycled significance level but Maurer in a personal communication to us pointed out that by choosing an appropriate family of error spending functions in their method any desired allocation can be achieved.

Ye et al. (2013) considered the significance level allocation problem. For ease of explanation, consider testing two null hypotheses, H_1 and H_2, corresponding to two primary endpoints using two inter-linked GSPs with a Bonferroni split, α_1 and α_2, of the overall significance level α. If H_1 is rejected before H_2 then α_1 is recycled to H_2 and it is tested at the full $\alpha_1 + \alpha_2 = \alpha$ level with a corresponding modified boundary. Ye et al. (2013) proposed two methods for modifying the boundary. The first method, which they called the *group sequential Holm variable (GSHv) procedure*, allocates the recycled significance level to all stages, thus modifying the entire boundary of the GSP. The second method, which they called the *group sequential Holm fixed (GSHf) procedure*, allocates the recycled significance level only to the final stage, thus modifying only the final critical constant.

GSHv allocates a portion of the recycled significance level to the stages prior to recycling (unless recycling takes place at the first stage) and these stages cannot be revisited. The reason is that if one

*Corresponding author: e-mail: dong.xi@novartis.com, Phone: +1-862-778-7498, Fax: +1-973-781-8891

modifies the critical boundaries of previous stages it may turn out at a later stage that the hypotheses could have been rejected at an earlier stage. This can lead to contradiction since the decision should be based on the cumulative data until the current stage. Therefore GSHv does not exploit the full potential of power gains due to recycling of the significance level. On the other hand, the power gain associated with the GSHf procedure is realized only if the trial continues to the final stage.

In a sense, GSHv assumes that recycling occurs at the first stage while GSHf assumes that recycling occurs at the final stage. We propose to strike a balance between these two extremes by assuming that recycling occurs at some specified stage r ($1 \leq r \leq m$) where m is the total number of stages. We refer to r as the common *planned recycling stage* for both hypotheses. Just as GSHv modifies the boundary according to $r = 1$ and GSHf modifies the boundary according to $r = m$ regardless of when the actual recycling occurs, the same is true for the proposed procedure. We will denote the GSP with planned recycling stage r by GSP(r). Using this notation, the GSHv procedure will be denoted by GSP(1) and the GSHf procedure by GSP(m).

Let s denote the observed stage at which either hypothesis is rejected and the significance level assigned to it is recycled to the other hypothesis. If neither hypothesis is rejected in the trial then we can set $s > m$. We refer to s as the *actual recycling stage*. The critical boundary of the unrejected hypothesis is modified at stage $u = \max(r, s)$, which we refer to as the *effective recycling stage* for that hypothesis. Note that for the GSHv procedure $u = s$ and for the GSHf procedure $u = m$ if $s \leq m$.

The error spending function (e.s.f.) approach of Lan and DeMets (1983) does not require specification of the number or the times of the interim analyses (stages). In that case, we will replace r by the corresponding time as discussed in Section 3.2. Strictly speaking, the notation GSP(r) is applicable only when the number and times of stages are specified but we will also use it when the e.s.f. approach is used to calculate the modified boundaries of the GSP upon recycling.

One may ask: why not choose $r = s$ so that the resulting GSP modifies the boundary exactly at the actual recycling stage and thus fully utilizes the recycled significance level? We will show in Section 4 that such an *adaptive* procedure does not always satisfy the strong familywise error rate (FWER) control requirement (Hochberg and Tamhane 1987):

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

for any combination of the true and false H_i's where α is specified global significance level.

The paper is organized as follows. Section 2 gives a brief review of the GSPs for a single endpoint. Two methods for allocating the recycled significance level, the boundary method and the e.s.f method, are presented in Section 3. It is shown in Section 4 that the adaptive version of GSP(r) does not always control the FWER. Section 5 gives sample size comparisons for different choices of r and GSP boundaries. Section 6 incorporates the allocation methods in the graphical approach of Bretz et al. (2009) for testing multiple hypotheses and discusses the choice of r via simulation. The paper concludes with summary remarks in Section 7. Proofs of the results are given in the appendix. `R` programs for various computations and simulations are given in supplementary materials.

2 Group sequential procedures for a single hypothesis

GSPs have been studied in the literature for nearly 40 years starting with Armitage (1975). Pocock (1977) and O'Brien and Fleming (1979) proposed two popular GSPs that we will denote by POC and OBF, respectively. The books by Jennison and Turnbull (2000) and Whitehead (1997) provide comprehensive overviews of this area. Here, we give a brief review for the single hypothesis case mainly to set up the background and notation for the present paper.

Consider testing a null hypothesis $H_0 : \theta = 0$ against an upper one-sided alternative using a GSP with $m \geq 2$ stages. Denote by $\mathcal{I}_k$ the cumulative statistical information available up to Stage k. Let Z_k denote the test statistic and $t_k = \mathcal{I}_k/\mathcal{I}_m$ denote the information time or fraction at Stage k ($1 \leq k \leq m$)

and let $t_0 = 0$; thus $0 = t_0 < t_1 < \cdots < t_m = 1$. We assume that the test statistics $Z_1, \ldots, Z_m$ follow an m-variate normal distribution with $E(Z_k) = \theta\sqrt{\mathcal{I}_k}$, $\mathrm{Var}(Z_k) = 1$ and $\mathrm{Corr}(Z_k, Z_\ell) = \sqrt{t_k/t_\ell}$ for $1 \le k < \ell \le m$. In this normal theory setup, $\mathcal{I}_k$ is proportional to the cumulative sample size up to Stage k.

Consider a GSP with a significance level γ that rejects H_0 at Stage k if $Z_j \le c_j(\gamma)$ $(1 \le j \le k-1)$ and $Z_k > c_k(\gamma)$. (We use the notation γ for a significance level instead of the usual α since α is used to denote the overall significance level at which the FWER is controlled when testing multiple hypotheses.) The critical constants $c_k(\gamma)$ are chosen to satisfy the following equation:

$$P_{H_0}\left\{Z_1 \le c_1(\gamma), \ldots, Z_m \le c_m(\gamma)\right\} = 1 - \gamma. \tag{1}$$

For the POC boundary, $c_1(\gamma) = \cdots = c_m(\gamma) = c_{\mathrm{POC}}(\gamma)$ while for the OBF boundary, $c_k(\gamma) = c_{\mathrm{OBF}}(\gamma)\sqrt{1/t_k}$ $(1 \le k \le m)$ where the constants $c_{\mathrm{POC}}(\gamma)$ and $c_{\mathrm{OBF}}(\gamma)$ are chosen to satisfy (1).

Lan and DeMets (1983) proposed a flexible approach to constructing GSPs based on e.s.f.'s. We denote the e.s.f. by $\varepsilon(\gamma, t)$, which is a monotone nondecreasing function of time $t \in [0, 1]$ with $\varepsilon(\gamma, 0) = 0$ and $\varepsilon(\gamma, 1) = \gamma$. If interim analyses have been performed previously at times $t_1, \ldots, t_{k-1}$ then the critical constant $c_k(\gamma)$ for the kth analysis can be computed from the e.s.f. by first calculating the so-called spent level:

$$\alpha_k(\gamma) = \varepsilon(\gamma, t_k) - \varepsilon(\gamma, t_{k-1}) \tag{2}$$

and then solving for $c_k(\gamma)$ recursively from the following set of equations:

$$\alpha_k(\gamma) = P_{H_0}\left[\bigcap_{j=1}^{k-1}\{Z_j \le c_j(\gamma)\}\bigcap\{Z_k > c_k(\gamma)\}\right] \quad (1 \le k \le m). \tag{3}$$

If there are m total analyses then $\sum_{k=1}^{m}\alpha_k(\gamma) = \gamma$.

In the sequel we will require the critical constants to satisfy the monotonicity condition (Liu and Anderson, 2008):

$$\gamma' > \gamma \implies c_k(\gamma') \le c_k(\gamma) \text{ for all } k, \tag{4}$$

so that if the significance level increases then the group sequential boundary shrinks enabling easier rejection of H_0. The following sufficient condition on the underlying e.s.f to ensure (4) was given by Maurer and Bretz (2013):

$$\gamma' > \gamma \implies \alpha_k(\gamma') \ge \alpha_k(\gamma) \text{ for all } k. \tag{5}$$

The e.s.f.'s of the POC and OBF boundaries are approximately given by

$$\varepsilon_{\mathrm{POC}}(\gamma, t) \approx \gamma \ln\{1 + (e-1)t\} \quad \text{and} \quad \varepsilon_{\mathrm{OBF}}(\gamma, t) \approx 2\Phi(-z_{\gamma/2}/\sqrt{t}),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and $\Phi(-z_\gamma) = \gamma$. We use the e.s.f. of the POC boundary in the examples later. It can be shown that it satisfies the monotonicity condition (5).

3 Methods for allocating recycled significance levels

We next discuss two methods, the boundary method and the e.s.f. method, for allocating the significance level transferred from a rejected hypothesis to the stages of the GSP of an unrejected hypothesis. In the boundary method, a desired parametric form can be specified for boundaries while the e.s.f. method is more flexible as noted before. To make the essential ideas clear, we will focus attention on a single unrejected hypothesis $H_0 : \theta = 0$, which is tested using an m-stage GSP(r) initially at a significance

level γ. Due to recycling the significance level from another rejected hypothesis, the level γ is increased to γ'. The additional significance level $\gamma' - \gamma$ is allocated only to stages $k \geq r$. It should be noted that the recycling method must be independent of the recycling sequence and therefore can be implicitly interpreted as a consonant weighted Bonferroni test as in Bretz et al. (2009) and Burman et al. (2009).

3.1 Boundary method

Let $(c_1(\gamma), \ldots, c_m(\gamma))$ denote the initial γ-level group sequential boundary for testing H_0 that satisfies (1). Then the *delayed recycling boundary* $(c_1(\gamma), \ldots, c_{r-1}(\gamma), c'_r(\gamma'), \ldots, c'_m(\gamma'))$ is calculated from the following equation:

$$P_{H_0}\{Z_1 \leq c_1(\gamma), \ldots, Z_{r-1} \leq c_{r-1}(\gamma), Z_r \leq c'_r(\gamma'), \ldots, Z_m \leq c'_m(\gamma')\} = 1 - \gamma', \tag{6}$$

where $c'_k(\gamma') \leq c_k(\gamma)$ for $k = r, \ldots, m$. Note that the critical constants $c'_k(\gamma')$ also depend on γ but we have suppressed this dependence for notational simplicity.

We may choose $c'_k(\gamma')$ to have the same form as the initial boundary. Thus, if the initial boundary is OBF then we can set $c'_k(\gamma') = c'_{\mathrm{OBF}}(\gamma')\sqrt{1/t_k}$ and if the initial boundary is POC then we can set $c'_k(\gamma') = c'_{\mathrm{POC}}(\gamma')$ for $k = r, \ldots, m$. In both cases, (6) involves a single unknown constant, $c'_{\mathrm{OBF}}(\gamma')$ or $c'_{\mathrm{POC}}(\gamma')$, for which it can be solved.

The boundaries for two choices of the planned recycling stage, $r < r'$, can be compared if both boundaries have the same parametric form. For example, we may require the POC type delayed recycling boundaries for both cases. Denote by

$$(c_1(\gamma), \ldots, c_{r-1}(\gamma), c'_r(\gamma'), \ldots, c'_m(\gamma')) \text{ and } (c_1(\gamma), \ldots, c_{r'-1}(\gamma), c''_{r'}(\gamma'), \ldots, c''_m(\gamma'))$$

the delayed recycling boundaries calculated using (6) for r and r', respectively. Then the critical constants of the two boundaries are equal for $k = 1, \ldots, r-1$ and it is easy to see that $c'_k(\gamma') \leq c_k(\gamma)$ for $k = r, \ldots, r' - 1$ and $c'_k(\gamma') \geq c''_k(\gamma')$ for $k = r', \ldots, m$.

As an example of the above, consider GSP(r) boundaries for $r = 1$ and $r = m$. The GSP(1) boundary equals $(c_1(\gamma'), \ldots, c_m(\gamma'))$, which is simply the initial boundary but with γ' level. The GSP(m) boundary equals $(c_1(\gamma), \ldots, c_{m-1}(\gamma), c'_m(\gamma'))$, which is given by (6) with $r = m$. Then $c_k(\gamma') \leq c_k(\gamma)$ for $1 \leq k \leq m - 1$ and $c_m(\gamma') \geq c'_m(\gamma')$.

3.2 Error spending function method

In the e.s.f. method instead of specifying the planned recycling stage r, we specify the *planned recycling time* $t^* \in (0, 1)$ such that recycling can take place only at any time $t > t^*$. This is analogous to specifying r, which implies that recycling can take place only after stage $r - 1$. Thus the relationship between the two methods is $t_{r-1} = t^*$.

In this method, we seek an e.s.f. $\varepsilon'(\gamma, \gamma', t|t^*)$, which satisfies the following conditions for a given initial e.s.f. $\varepsilon(\gamma, t)$, $\gamma' > \gamma$ and t^*:

(i) $\varepsilon'(\gamma, \gamma', t|t^*)$ satisfies the monotonicity condition (5).
(ii) $\varepsilon'(\gamma, \gamma', t|t^*) = \varepsilon(\gamma, t)$ for $0 \leq t \leq t^*$.
(iii) $\varepsilon'(\gamma, \gamma', 1|t^*) = \gamma' > \gamma = \varepsilon(\gamma, 1)$.

We refer to such an e.s.f. as a delayed recycling e.s.f.

Many choices are possible for $\varepsilon'(\gamma, \gamma', t|t^*)$. We choose the following:

$$\varepsilon'(\gamma, \gamma', t|t^*) = \begin{cases} \varepsilon(\gamma, t) & \text{for} \quad 0 \leq t \leq t^* \\ \varepsilon(\gamma, t^*) + \eta(\gamma, \gamma', t|t^*) & \text{for} \quad t^* < t \leq 1, \end{cases} \tag{7}$$

where $\eta(\gamma, \gamma', t|t^*)$ is a nondecreasing function of t such that $\eta(\gamma, \gamma', t^*|t^*) = 0$ and $\eta(\gamma, \gamma', 1|t^*) = \gamma' - \varepsilon(\gamma, t^*)$.

A possible choice for $\eta(\gamma, \gamma', t|t^*)$ is

$$\eta(\gamma, \gamma', t|t^*) = \varepsilon^*(\gamma^*, t) - \varepsilon^*(\gamma^*, t^*) \tag{8}$$

where $\varepsilon^*(\gamma^*, t)$ is an arbitrary e.s.f. and γ^* is the solution to the equation

$$\gamma^* - \varepsilon^*(\gamma^*, t^*) = \gamma' - \varepsilon(\gamma, t^*). \tag{9}$$

If we choose $\varepsilon^*(\gamma^*, t)$ to be the initial e.s.f. $\varepsilon(\gamma^*, t)$ then it can be shown that $\varepsilon'(\gamma, \gamma', t|t^*)$ satisfies the monotonicity condition (5).

From (8) we see that $\varepsilon'(\gamma, \gamma', t^*|t^*) = \varepsilon(\gamma, t^*)$ and from (9) we see that

$$\begin{aligned}
\varepsilon'(\gamma, \gamma', 1|t^*) &= \varepsilon(\gamma, t^*) + \eta(\gamma, \gamma', 1|t^*) \\
&= \varepsilon(\gamma, t^*) + \varepsilon^*(\gamma^*, 1) - \varepsilon^*(\gamma^*, t^*) \\
&= \varepsilon(\gamma, t^*) + \gamma^* - \varepsilon^*(\gamma^*, t^*) \\
&= \gamma'.
\end{aligned}$$

Another example of $\varepsilon^*(\gamma^*, t)$ is the linear e.s.f. $\varepsilon^*(\gamma^*, t) = \gamma^* t$. It is easy to see that for this choice, $\gamma^* = (\gamma' - \varepsilon(\gamma, t^*))/(1 - t^*)$.

Remark: A referee suggested an alternative form for the delayed recycling e.s.f. given by

$$\varepsilon'(\gamma, \gamma', t|t^*) = \begin{cases} \varepsilon(\gamma, t) & \text{for} \quad 0 \le t \le t^* \\ \varepsilon(\gamma, t)\eta(\gamma, \gamma', t|t^*) & \text{for} \quad t^* < t \le 1, \end{cases}$$

where

$$\eta(\gamma, \gamma', t|t^*) = \frac{\gamma(1 - t) + \gamma'(t - t^*)}{\gamma(1 - t^*)}. \tag{10}$$

Note that this e.s.f. uses a multiplicative factor rather than an additive factor used in (7). This factor scales up $\varepsilon(\gamma, t)$ linearly in t for $t > t^*$.

Although our proposal (7) requires the computation of γ^*, it is more flexible in that one can choose any e.s.f. (subject to the monotonicity condition (5)) for $\varepsilon^*(\gamma^*, t)$ in (8). Therefore we will follow our proposal.

For a given analysis k at time t_k $(k = 1, \ldots, m)$, the delayed recycling boundary $(c'_1(\gamma'), \ldots, c'_m(\gamma'))$ can be calculated by using

$$\alpha'_k(\gamma') = \varepsilon'(\gamma, \gamma', t_k|t^*) - \varepsilon'(\gamma, \gamma', t_{k-1}|t^*) \ (1 \le k \le m). \tag{11}$$

in (3) in place of $\alpha_k(\gamma)$. Let t^* be the time of the $(r-1)$th analysis and s be the actual recycling stage. Then the effective boundary is given by $(c_1(\gamma), \ldots, c_{u-1}(\gamma), c'_u(\gamma'), \ldots, c'_m(\gamma'))$ where $u = \max(r, s)$ is the effective recycling stage. If $r \ge s$ then the effective boundary is the same as the delayed recycling boundary and there is no loss of significance level although it is not always recycled at the actual recycling stage. This is always the case when $r = m$. On the other hand, if $r < s$ then the significance level recycled to stages $r, r + 1, \ldots, s - 1$ is wasted since those stages cannot be revisited. This is the case when $r = 1$ and $s > 1$.

Figure 1 shows the e.s.f.'s of GSP(r) for $r = 1, 2, 3$ and $m = 3$, $\gamma = 0.025$, $\gamma' = 0.05$ when the POC boundary is used both before and after the rth stage. It is seen that GSP(r) allocates the largest increase in the e.s.f. at the rth stage.

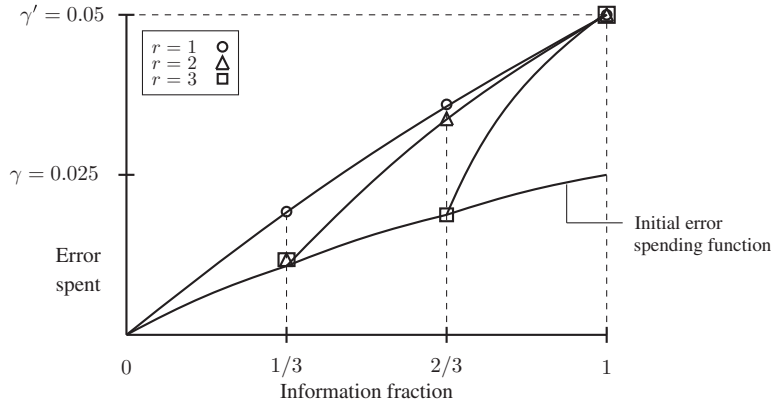

Figure 1 Error spending functions for GSP(1), GSP(2), and GSP(3) using the POC boundary ($m = 3$, $\gamma = 0.025$, $\gamma' = 0.05$).

3.3 Example

Suppose we want to test null hypotheses H_1 and H_2 on two primary endpoints against upper one-sided alternatives. Assume that the nominal level $\alpha = 0.05$ is split equally between them; thus if H_1 is rejected at the 0.025 level then H_2 is tested at the full 0.05 level and vice versa. A GSP with three stages (two interim analyses and one final analysis) is used for each hypothesis with the interim analyses carried out after 1/3 and 2/3 of the total number of patients are observed.

Suppose that H_1 is rejected at Stage 2 but H_2 is not yet rejected, and the POC boundary is used to test H_2. For $\alpha = 0.025$, this boundary is (2.289, 2.289, 2.289). We now show how to calculate the delayed recycling and effective boundaries for H_2 for $r = 1, 2, 3$ using the boundary and the e.s.f. methods. Multivariate normal probabilities needed in these calculations were evaluated using the `R` package `mvtnorm` by Genz et al. (2014) that is based on the works of Miwa et al. (2003) and Genz and Bretz (2009).

Boundary Method: For $r = 1$, the delayed recycling boundary is simply the 0.05-level POC boundary, which is (1.992, 1.992, 1.992). Since $s = 2$, the effective boundary for GSP(1) is (2.289, 1.992, 1.992).

For $r = 2$, we solve for $c_2'(0.05) = c_3'(0.05) = c'(0.05)$ from $P\{Z_1 \le 2.289, Z_2 \le c'(0.05), Z_3 \le c'(0.05)\} = 0.95$. The solution is $c'(0.05) = 1.889$. So the delayed recycling and effective boundaries for GSP(2) are both (2.289, 1.889, 1.889).

For $r = 3$, we solve for $c_3'(0.05)$ from $P\{Z_1 \le 2.289, Z_2 \le 2.289, Z_3 \le c_3'(0.05)\} = 0.95$. The solution is $c_3'(0.05) = 1.737$. So the delayed recycling and effective boundaries for GSP(3) are both (2.289, 2.289, 1.737).

Error Spending Function Method: Note that for $t \le t^* = (r-1)/3$ $(r = 1, 2, 3)$, the delayed recycling e.s.f. is the same as the initial POC e.s.f., which is $0.025\ln\{1 + (e-1)t\}$. For $t > t^* = (r-1)/3$, we compute the delayed recycling e.s.f. using (7) where $\eta(\gamma, \gamma', t|t^*)$ is given by (8). For $\varepsilon^*(\gamma^*, t)$ we use the POC e.s.f. $\gamma^* \ln\{1 + (e-1)t\}$.

For $r = 1$, since $t^* = 0$ we have $\varepsilon(\gamma, t^*) = 0$ and so $\eta(0.025, 0.05, t|0) = \gamma^* \ln\{1 + (e-1)t\}$ where $\gamma^* = \gamma' = 0.05$ from (9). Thus, $\varepsilon'(0.025, 0.05, t|0) = 0.05\ln\{1 + (e-1)t\}$ for $t > 0$ that is the 0.05-level POC e.s.f. Therefore the delayed recycling boundary is the 0.05-level POC boundary (1.992, 1.992, 1.992), the same as that obtained using the boundary method. Since $s = 2$, the effective boundary for GSP(1) is (2.289, 1.992, 1.992). If we use the method proposed by the referee then from (10) we get $\eta(0.025, 0.05, t|0) = 1 + t$. Therefore $\varepsilon'(0.025, 0.05, t|0) = 0.025(1 + t)\ln\{1 + (e-1)t\}$,

which is not the POC e.s.f. but some nonstandard e.s.f. So the delayed recycling boundary will not be the simple 0.05-level POC boundary.

For $r = 2$, we first calculate $\varepsilon(0.025, 1/3) = 0.025\ln\{1 + (e-1)(1/3)\} = 0.0113$. To evaluate the delayed recycling e.s.f. (7) using (8), we need to solve the following equation obtained from (9) for γ^*:

$$\gamma^* - \gamma^* \ln\{1 + (e-1)(1/3)\} = 0.05 - 0.025\ln\{1 + (e-1)(1/3)\}.$$

The solution can be checked to be $\gamma^* = 0.0707$. Hence

$$\eta(0.025, 0.05, 2/3|1/3) = 0.0707\ln\{1 + (e-1)(2/3)\} - 0.0707\ln\{1 + (e-1)(1/3)\} = 0.0220.$$

Therefore the spent levels are $\alpha_2'(0.05) = 0.0220$ and $\alpha_3'(0.05) = 0.05 - 0.0220 - 0.0113 = 0.0167$. So $c_2'(0.05)$ and $c_3'(0.05)$ can be determined recursively from the following two equations:

$$P\{Z_1 \le 2.289, Z_2 > c_2'(0.05)\} = 0.0220,$$

the solution to which is $c_2'(0.05) = 1.925$ and

$$P\{Z_1 \le 2.289, Z_2 \le 1.925, Z_3 > c_3'(0.05)\} = 0.0167,$$

the solution to which is $c_3'(0.05) = 1.865$. So the delayed recycling and effective boundaries for GSP(2) are both (2.289, 1.925, 1.865). Note that although we used the POC e.s.f. to obtain the delayed recycling e.s.f., we did not get $c_2'(0.05) = c_3'(0.05)$.

Finally for $r = 3$, the GSP(3) boundary is the same as that obtained using the boundary method that is (2.289, 2.289, 1.737).

Note that for $r = 1$ and $r = m$ the e.s.f. method always gives the same delayed recycling boundary as the boundary method gives regardless of the choice of the delayed e.s.f. We will focus on the boundary method in the rest of the paper.

4 Adaptive choice of planned recycling stage *r*

It is tempting to make GSP(r) adaptive by setting $r = s$ instead of prespecifying it. This leads to a random adaptive boundary for H_2 depending on when H_1 is rejected. Although GSP(s) fully utilizes the recycled significance level, it does not control the FWER in general. This result could be viewed as a generalization of the result observed by Hung et al. (2007) and proved by Tamhane et al. (2010) who showed that in a hierarchical test of a primary and a secondary endpoint, the FWER is not always controlled if the primary hypothesis is rejected at an interim analysis and the secondary hypothesis is tested at the full level α at the same stage.

For convenience of notation and explanation, the following propositions are given only for two null hypotheses, $H_1 : \theta_1 = 0$ and $H_2 : \theta_2 = 0$, which are assumed to be tested with a Bonferroni split of the significance level α. Let GSP(s) denote the adaptive version of the GSP(r) procedure, which sets $r = s$ where s is the actual recycling stage. First note that GSP(s) weakly controls the FWER under $H_1 \bigcap H_2$ because of the Bonferroni split of α. Therefore in the following propositions we consider the partial null hypothesis: H_1 is false and H_2 is true.

Proposition 4.1. *GSP(s) controls the FWER at level α if the test statistics for the two hypotheses are independent.*

Proposition 4.2. *GSP(s) does not control the FWER at level α if the test statistics are positively correlated.*

Table 1 Maximum FWER as a function of ρ for testing H_1 and H_2 using a two-stage GSP when H_1 is false and H_2 is true ($\gamma = 0.025,\ \gamma' = 0.05$).

ρ	0.0	0.2	0.4	0.6	0.8	1.0
max FWER	0.0500	0.0506	0.0515	0.0527	0.0545	0.0603

The analytical proof in the dependence case is given for the limiting case of correlation $\rho = 1$ between the test statistics for H_1 and H_2. In a two-stage GSP, no matter which boundaries are used for H_1 and H_2, the FWER $> \alpha$ when $\rho = 1$ if $\delta_1 = c_{11}/\sqrt{t_1}$, where $\delta_1 = \theta_1\sqrt{\mathcal{I}_m}$ is the drift parameter for H_1 and c_{11} is the critical constant for the first stage for H_1 before recycling occurs. But in fact numerical results show that for every $\rho > 0$, max FWER $> \alpha$ where the maximum is taken over δ_1. This numerical study is described in the following section.

4.1 Numerical study of FWER of adaptive GSP

Consider a two-stage GSP in which the OBF boundary is used for H_1 and the POC boundary is used for H_2 with $\gamma = 0.025$ and $\gamma' = 0.05$. Denote by GSP_i the initial boundary (for $\gamma = 0.025$) and by $\text{GSP}_i(r)$ the delayed recycling boundary (for $\gamma' = 0.05$) for hypothesis H_i ($i = 1, 2$) using the planned recycling stage $r = 1, 2$. These boundaries can be calculated using the boundary method as

$$\text{GSP}_1 = (2.7965, 1.9774), \text{GSP}_1(1) = (2.3729, 1.6779), \text{GSP}_1(2) = (2.7965, 1.6507)$$

and

$$\text{GSP}_2 = (2.1782, 2.1782), \text{GSP}_2(1) = (1.8754, 1.8754), \text{GSP}_2(2) = (2.1782, 1.7145).$$

The adaptive GSP operates as follows. Suppose that H_1 is rejected first using GSP_1 at stage s then H_2 is tested using $\text{GSP}_2(s)$ ($s = 1, 2$). Similarly, if H_2 is rejected first using GSP_2 at stage s then H_1 is tested using $\text{GSP}_1(s)$ ($s = 1, 2$). If neither hypothesis is rejected then no recycling takes place and both hypotheses are accepted.

We numerically evaluated the FWER of this procedure for $\rho = 0.0(0.1)1.0$ and $\delta_1 = 0.00(0.01)5.00$. These numerical results are graphed in Fig. 2 and the max FWER values for $\rho = 0.0(0.2)1.0$ are given in Table 1. Observe that max FWER > 0.05 for all $\rho > 0$ and could be as high as 0.0603 for $\rho = 1$. Furthermore, FWER $\to 0.05$ as $\delta_1 \to \infty$ for all $\rho \geq 0$.

It is curious to note that the FWER curve for $\rho = 1$ has a plateau for a small interval of δ_1 with FWER $= 0.05$. In the following proposition we give a formula for this interval.

Proposition 4.3. *When $\rho = 1$, the FWER of the adaptive two-stage GSP equals α for $\delta_1 \in [L, U]$, where L and U are given by*

$$L = \max\{0, c_2(GSP_1) - c_2(GSP_2(2))\} \quad \textit{and} \quad U = \max\left\{0, t_1^{-1/2}[c_1(GSP_1) - c_1(GSP_2)]\right\}.$$

In fact, when $\delta_1 \in [L, U]$, the FWER of the adaptive $GSP_2(s)$ is the same as that of $GSP_2(2)$.

Substituting the values of the necessary critical constants in the above equation for $\alpha = 0.05$ we get

$$L = \max\{0, 1.9774 - 1.7145\} = 0.2629 \quad \text{and} \quad U = \max\{0, \sqrt{2}(2.7965 - 2.1782)\} = 0.8744.$$

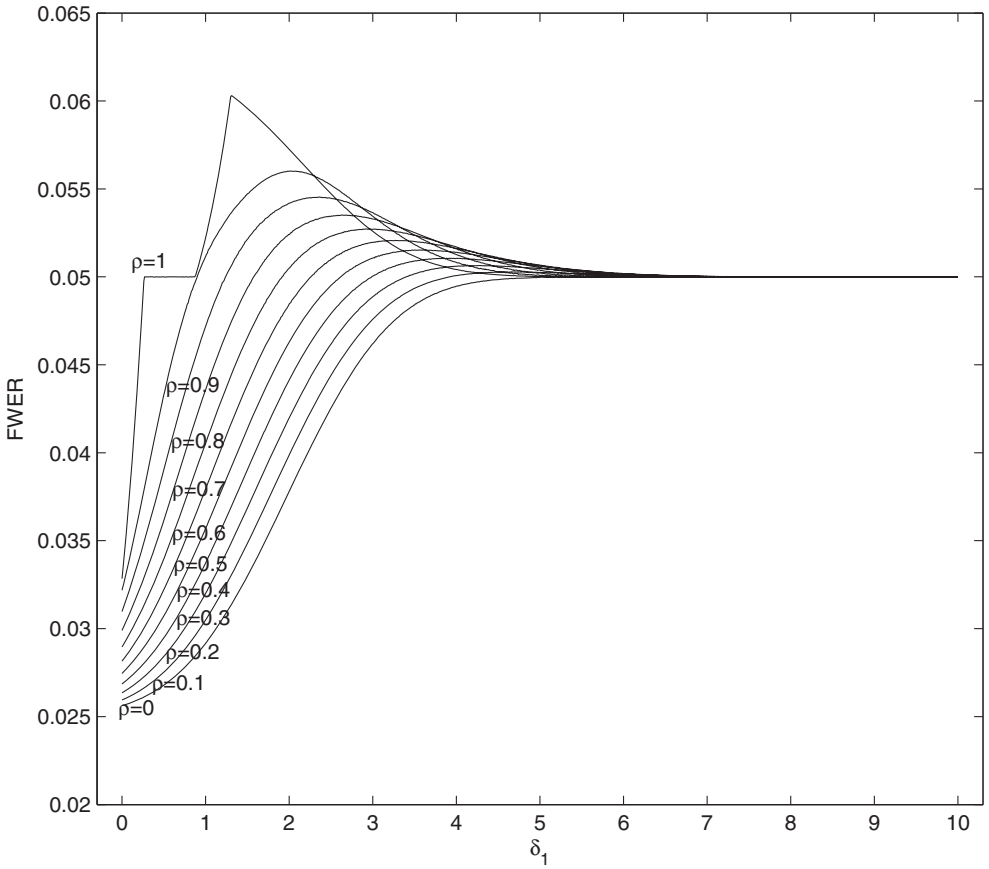

Figure 2 FWER as a function of ρ and the drift parameter δ_1 for H_1 when testing two hypotheses using a two-stage adaptive GSP ($\gamma = 0.025,\ \gamma' = 0.05$).

5 Performance comparisons for a single hypothesis

In order to gain a better understanding of the differences between the performances of GSP(r) for different values of r, we will consider the problem of testing a single hypothesis $H_0 : \theta = 0$ using an m-stage GSP. Assume that the hypothesis is initially tested at level $\gamma = 0.025$ and an additional significance level 0.025 is propagated to it at some stage s during the course of the trial, raising the level to $\gamma' = 0.05$ for the remainder of the trial. This setup simplifies the comparison since the factors such as the choices of the e.s.f.'s for other hypotheses, their drift parameters, etc. do not complicate the comparison. As measures of performance we will study the expected and maximum sample sizes. The power to reject H_0 is an increasing function of the maximum sample size M and so we do not study it here.

Denote the maximum or planned total sample size by M and the expected sample size by $E(N)$. Assume that the test statistics $Z_1, \ldots, Z_m$ have an m-variate normal distribution with $E(Z_k) = \theta\sqrt{Mt_k}$, $\mathrm{Var}(Z_k) = 1$ and $\mathrm{Corr}(Z_k, Z_\ell) = \sqrt{t_k/t_\ell}$ for $1 \le k < \ell \le m$. For given m and the group sequential boundary $(c_1, \ldots, c_m)$, we can determine M under a specified power requirement that the probability of rejecting $H_0 : \theta = 0$ is at least $1 - \beta$ when the true parameter equals $\theta > 0$. Thus, M is the smallest

Table 2 Maximum and expected sample sizes (expressed as percentages of the fixed sample size) for GSP(r) for all possible outcomes s ($m = 3$, $\gamma = 0.025$, $\gamma' = 0.05$, Power $1 - \beta = 0.80$ at $\theta = 1$).

Initial boundary	r	$s = 1$		$s = 2$		$s = 3$	
		M	$E(N)$	M	$E(N)$	M	$E(N)$
OBF	1	102.7	83.7	102.8	85.4	104.6	91.5
	2	102.5	85.1	102.5	85.1	104.4	91.3
	3	100.4	88.4	100.4	88.4	100.4	88.4
POC	1	118.4	80.7	120.5	86.5	124.6	92.7
	2	111.6	80.6	111.6	80.6	116.6	88.5
	3	104.8	81.9	104.8	81.9	104.8	81.9

total sample size that satisfies

$$P(Z_1 \le c_1, \ldots, Z_m \le c_m|\theta) = \beta$$

for given r and s. In practice, M is rounded up so that the group sample sizes are integers, but in the calculations presented here we have left M to be the exact fractional solution to the above equation. Then $E(N)$ is given by

$$\begin{aligned} E(N) &= M \sum_{k=1}^{m-1} t_k P(\text{GSP stops and rejects } H_0 \text{ at Stage } k|\theta) \\ &\quad + M \times P(\text{GSP stops at Stage } m|\theta) \\ &= M \sum_{k=1}^{m-1} t_k P(Z_1 \le c_1, \ldots, Z_{k-1} \le c_{k-1}, Z_k > c_k|\theta) \\ &\quad + M \times P(Z_1 \le c_1, \ldots, Z_{m-1} \le c_{m-1}|\theta). \end{aligned}$$

We calculated M and $E(N)$ with the OBF and POC group sequential boundaries with equal group sizes for $m = 2, 3, 4$ and $\theta = 0.5, 1, 1.5$. In each case, we considered the GSP(r) procedure for $r = 1, \ldots, m$ and for all possible outcomes $s = 1, \ldots, m$. The boundary is changed at the effective recycling stage $u = \max(r, s)$ from level $\gamma = 0.025$ to level $\gamma' = 0.05$. For lack of space, only the results for $m = 3$ and $\theta = 1$ are presented in Table 2.

We can draw the following conclusions from these results.

1. To guarantee the power of $1 - \beta$, the maximum sample size M decreases with r (for any given s). From this we can infer that GSP(1) is the least powerful and GSP(m) is the most powerful with a common maximum sample size.
2. Although $r = m$ maximizes the power, it also increases the expected sample size since the power for rejecting H_0 in an earlier stage is not increased as no significance level is propagated to those stages. This situation is similar to comparing a group sequential trial with a fixed maximum sample size to a nonsequential trial. The latter has more power but a larger expected sample size.
3. For the OBF boundary, the minimum of $E(N)$ is achieved for given s when $r = s$, i.e. when the planned recycling stage coincides with the actual recycling stage. This is also true for the POC boundary except for $s = 1$ when the minimum is achieved at $r = 2$.

4. For the OBF boundary, $r = m$ yields the largest $E(N)$ over different choices of r if $s < m$ that is a drawback of choosing $r = m$ although it yields the smallest M. On the other hand, for the POC boundary, GSP(1) has the largest $E(N)$ if $s > 1$.
5. The expected sample size $E(N)$ of GSP(r) for each r is constant for $s \leq r$ and then increases for $s > r$.
6. If we apply the minimax criterion to choose r to minimize the maximum $E(N)$ over all outcomes s for each r, $r = 3$ is the optimum choice for both the OBF and POC boundaries.

6 Multiple hypotheses

Next, we consider the problem of testing a family of $n \geq 2$ null hypotheses, $H_1, \ldots, H_n$, using a GSP with $m \geq 2$ stages. The hypotheses could be unordered, e.g. hypotheses concerning primary or coprimary endpoints or they could be hierarchically ordered as primary and secondary. Previous works on GSPs for multiple unordered primary endpoints include Pocock et al. (1987), Tang and Geller (1999), and Liu and Anderson (2008). GSPs for primary and secondary endpoints have been studied by Glimm et al. (2010) and Tamhane et al. (2010).

6.1 Algorithm for graphical implementation

We now give an algorithm for the closed procedure using the graphical approach with recycling proposed by Maurer and Bretz (2013). We provide additional steps necessary for calculation of the delayed recycling boundary since we use test statistics Z_{ik} instead of their p-values. The algorithm involves choosing initial group sequential boundaries or their associated e.s.f.'s $\varepsilon_i(\gamma, t)$ and assigning a local weight $w_i(I)$ to each $H_i,\ i \in I \subseteq \{1, \ldots, n\}$ where $0 \leq w_i(I) \leq 1$ and $\sum_{i \in I} w_i(I) \leq 1$. The local test of the intersection hypothesis $H(I)$ is the weighted Bonferroni test that rejects $H(I)$ if at least one $H_i,\ i \in I$ is rejected at level $w_i(I)\alpha$. In the GSP this test is applied at each analysis after recalculating the set I, the weights $w_i(I)$, the test statistics Z_{ik} and the critical constants c_{ik} according to the algorithm below. The weights are assumed to satisfy the following monotonicity condition due to Hommel et al. (2007):

$$w_i(I) \leq w_i(J) \text{ for all } i \in J \subseteq I \subseteq \{1, \ldots, n\}.$$

Maurer and Bretz (2013) have shown that if the critical constants of the closed procedure with group sequential boundaries satisfy the monotonicity condition (4) of Liu and Anderson (2008) then it is consonant and hence has a stepwise shortcut at each stage.

In the graphical approach the hypotheses are represented as nodes in a directed graph with transition weight $g_{ij} \geq 0$ on the arc connecting H_i to H_j for each pair $i \neq j$ subject to $\sum_j g_{ij} \leq 1$ for each H_i where g_{ij} is the fraction of the significance level assigned to H_i that is recycled to H_j if H_i is rejected. The hypothesis H_i is removed from the graph and the g_{ij} are updated.

Algorithm

Step 0: Set $I = \{1, \ldots, n\}$. Assign weights $w_i = w_i(I)$ to all hypotheses $H_i,\ i \in I$ such that $w_i \geq 0$ and $\sum_{i \in I} w_i \leq 1$. Also assign transition weights $g_{ij} = g_{ij}(I) \geq 0$ to directed edges from H_i to H_j $(i, j \in I, i \neq j)$ such that $\sum_{j \neq i} g_{ij} \leq 1$ for all $i \in I$. Assume that the planned recycling stages r_i or planned recycling times t_i^* are specified for all H_i. Calculate the critical boundary $(c_{i1}, \ldots, c_{im})$ at level $\gamma_i^0 = w_i\alpha$ for each $H_i,\ i \in I$. Set $\gamma_i = \gamma_i^0$ and $k = 1$.

Stage k:

Step 1: Compute the test statistics Z_{ik} for $i \in I$.

Step 2: If there exists an $i \in I$ such that $Z_{ik} > c_{ik}$ then reject H_i and proceed to Step 3; otherwise go to Step 6.

Step 3: Update the graph:

$$
\begin{aligned}
I &\to I \setminus \{i\},\\
w_j &\to \begin{cases} w_j + w_i g_{ij} & j \in I,\\ 0 & \text{otherwise,} \end{cases}\\
g_{jh} &\to \begin{cases} (g_{jh} + g_{ji}g_{ih})/(1 - g_{ji}g_{ij}) & j, h \in I,\ j \neq h,\ g_{ji}g_{ij} < 1,\\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

Step 4: Calculate the new critical constants c'_{ij} at level $\gamma'_i = w_i\alpha$ for all $i \in I$ using either the boundary method (Eq. (6)) or the e.s.f. method (Eqs. (11) and (3)).

Step 5: Calculate the effective boundary by setting $c_{ij} \to c'_{ij}$ where $j = \max(r_i, k), \ldots, m$. Go to Step 2.

Step 6: If $|I| \geq 1$ and $k < m$ then set $k \to k + 1$ and go to Step 1 of Stage k; otherwise stop.

The following points should be noted about this algorithm.

1. Since the GSP(r_i) is used for H_i, the critical constants in stages $1, \ldots, r_i - 1$ are unchanged in Step 4 from those in Step 0.
2. The algorithm allows for more than one hypothesis to be tested and rejected at any stage, each time recycling the corresponding significance level and modifying the critical boundaries of the remaining unrejected hypotheses. In this case, the significance level may increase from the initial level γ to γ' and then to γ'' and so on. It is clear from (7) that the delayed recycling boundary at level γ'', $(c_1(\gamma), \ldots, c_{r-1}(\gamma), c''_r(\gamma''), \ldots, c''_m(\gamma''))$, does not depend on γ' but only upon γ and γ''. Therefore the same is true for the effective boundary. As a result, we only need to save in memory the last set of the critical constants based on the last increased level when proceeding from one stage to the next.
3. Another possibility is that the significance level for some hypothesis could increase from γ to $\gamma' > \gamma$ at Stage k' and then to $\gamma'' > \gamma'$ at Stage $k'' > k'$ due to rejections of other hypotheses at each of the stages. It is clear from the above point that the delayed recycling boundary at γ'' does not depend on γ'. But the effective boundary at γ'' uses the critical constants at γ' for Stage k' onwards to $k'' - 1$ since in Step 5 only the constants at the current and future stages are updated without revisiting previous stages.
4. In practice, it is usually the Data Monitoring Committee (DMC) that makes the decision on stopping or continuing the trial given the evidence on efficacy (or futility) and safety. This option could be incorporated in Step 6.

6.2 Example

To illustrate the application of the proposed GSP(r) procedure to multiple hypotheses we use the diabetes trial example of Maurer and Bretz (2013) in which they used GSP(1). This is a three-stage trial with equal group sizes; thus $(t_1, t_2, t_3) = (1/3, 2/3, 1)$. It compares a high and a low dose of a drug against placebo (on top of the standard-of-care) on two endpoints: HbA1c and body weight. HbA1c is the primary endpoint while body weight is the secondary endpoint. Thus there are four null hypotheses. Within each dose the secondary null hypothesis cannot be tested unless the primary null hypothesis is rejected. Denote the null hypotheses on HbA1c for low and high doses by H_1 and H_2 and those on body weight by H_3 and H_4, respectively. Thus H_3 cannot be tested unless H_1 is rejected. Similarly, H_4 cannot be tested unless H_2 is rejected.

The initial graph for this testing problem is shown in Fig. 3A. Suppose that the nominal level $\alpha = 0.025$ is split equally, $\alpha_1 = \alpha_2 = 0.0125$, between H_1 and H_2. If a primary hypothesis is rejected then its associated significance level is split equally between the other primary hypothesis and the descendent secondary hypothesis of the rejected primary hypothesis. If both hypotheses under one

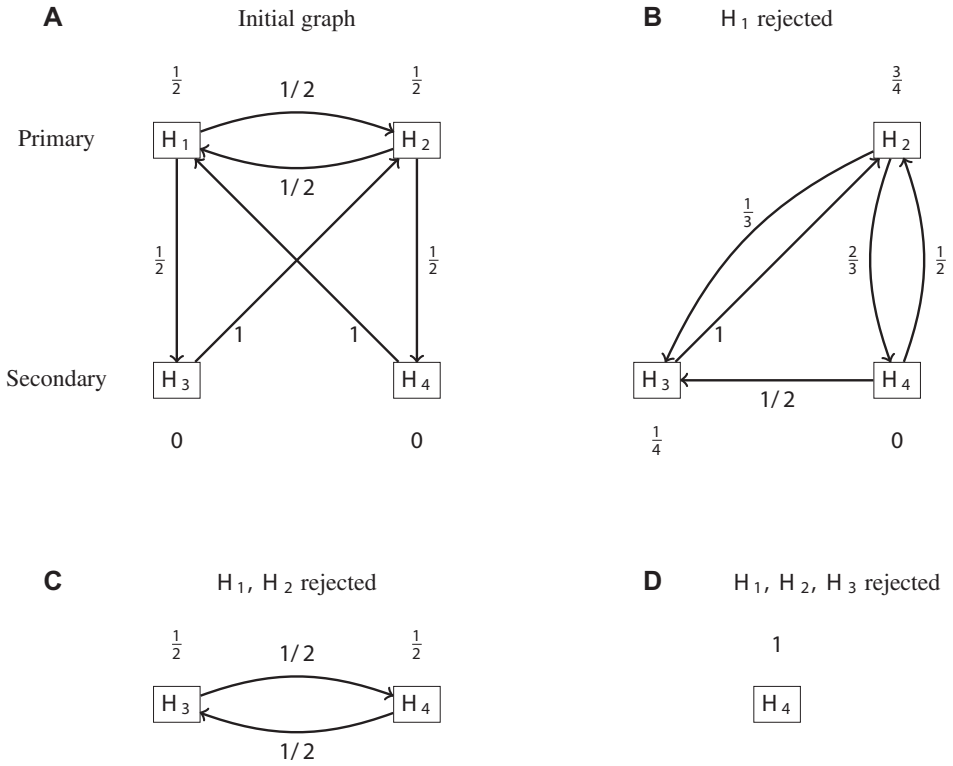

Figure 3 A graphical representation of GSP(2) for the diabetes trial example.

dose level are rejected then the hypotheses under the other dose level can be tested at the full $\alpha = 0.025$ level. We will use the POC boundaries for all four hypotheses instead of the OBF boundaries used in the original example in order to illustrate the difference in the decisions reached with $r = 1$ and $r = 2$. For hypothesis H_i, denote the test statistic at Stage k by Z_{ik} and the corresponding critical constant by c_{ik} ($1 \le i \le 4, 1 \le k \le 3$). Assume that $r = 2$ is specified for all four hypotheses. The delayed recycling boundaries are calculated using the boundary method. We will now walk through the example stage by stage.

Stage 1: We have $I_1 = \{1, 2, 3, 4\}$ and $(w_1(I_1), w_2(I_1), w_3(I_1), w_4(I_1)) = (0.5, 0.5, 0, 0)$. Suppose that the test statistics are $Z_{11} = 2.50, Z_{21} = 2.12, Z_{31} = 2.37, Z_{41} = 1.13$. The POC boundary for $\alpha_1 = \alpha_2 = 0.0125$ is (2.555, 2.555, 2.555). Thus neither H_1 nor H_2 can be rejected since both Z_{11} and Z_{21} are < 2.555. Hence H_3 and H_4 cannot be tested.

Stage 2: Suppose that the test statistics are $Z_{12} = 2.84, Z_{22} = 2.39, Z_{32} = 2.61, Z_{42} = 1.55$. Since we are using the POC boundary and no hypotheses were rejected at Stage 1, the critical constants at Stage 2 are unchanged: $c_{12} = c_{22} = 2.555$. Since $Z_{12} > 2.555$ we reject H_1. Thus $I_2 = \{2, 3, 4\}$ and $(w_2(I_2), w_3(I_2), w_4(I_2)) = (0.75, 0.25, 0)$ using Algorithm 1. The transition parameters also change as shown in Fig. 3B. The new significance levels are $\alpha_2 = 0.01875$ and $\alpha_3 = 0.00625$, respectively. Since no significance level is transferred to H_4, $\alpha_4 = 0$. The delayed recycling POC boundary using GSP(2) can be calculated as (2.555, 2.339, 2.339) for H_2 and $(\infty, 2.671, 2.671)$ for H_3 where the first critical constant is set equal to ∞ because initially $\alpha_3 = 0$. Since $Z_{22} > 2.339$ we reject H_2. Thus one can test H_3 and H_4 at Stage 2 at the full $\alpha = 0.025$ level.

Upon rejection of H_2, I_2 is updated to $I_2 = \{3, 4\}$ and $(w_3(I_2), w_4(I_2)) = (0.5, 0.5)$. The resulting graph along with its transition parameters is shown in Fig. 3C. From the symmetry of this graph it follows that $\alpha_3 = \alpha_4 = 0.0125$. The delayed recycling POC boundary using GSP(2) can be calculated as $(\infty, 2.421, 2.421)$ for both H_3 and H_4. Since $Z_{32} > 2.421$ we reject H_3 and transfer its level to H_4 that is then tested at $\alpha_4 = 0.025$. The resulting graph with a single node H_4 is shown in Fig. 3D. The new delayed recycling POC boundary can be calculated as $(\infty, 2.146, 2.146)$. Since $Z_{42} < 2.146$ we cannot reject H_4 and so the trial proceeds to Stage 3. At this point the DMC may decide to terminate the trial to save resources since both primary hypotheses and one secondary hypothesis have been rejected.

Note that if we had used $r = 1$, the delayed recycling POC boundaries at Stage 2 after rejecting H_1 would have been $(2.556, 2.403, 2.403)$ for H_2 and $(\infty, 2.798, 2.798)$ for H_3. So neither H_2 nor H_3 could have been rejected at Stage 2. On the other hand, if we could have rejected H_1 and/or H_2 at Stage 1 then we could have tested their descendant secondary hypotheses H_3 and/or H_4 at that stage itself instead of waiting until the second stage when $r = 2$ as in the present example.

6.3 Choice of *r* for multiple hypotheses

In this section, we illustrate how to choose r using simulation in a multiple hypotheses setting. The example is too small to draw any general conclusions. Its purpose is merely to demonstrate the use of simulation in selection of r.

To keep the discussion simple, we return to the example of two hypotheses with three-stage GSPs from Section 3.3. Consider testing H_1 and H_2 each at level 0.025 initially. If H_1 (H_2) is rejected, H_2 (H_1) can be tested at level 0.05 using GSP(r). We assume the trial is stopped early for superiority only if both hypotheses are rejected.

In practice, one could choose a different r for each hypothesis depending on when the other hypothesis is likely to be rejected. For the sake of simplicity, we will choose a common r for both hypotheses. Assume that we require 80% power for H_1 (H_2) under some true parameter θ_1 (θ_2). We studied various combinations of θ_1 and θ_2 for different choices of the correlation ρ between the two endpoints, but here we report the results only for $\rho = 0.5$ since the best choice of r was found to be relatively insensitive to ρ although $E(N)$ and M both depend on ρ. Tamhane et al. (2010) and Glimm et al. (2010) have shown that different combinations of boundaries also affect the performance of a GSP. We considered two combinations of boundaries for H_1 and H_2: OBF-OBF and OBF-POC. Under our assumptions, we focused on the larger of the two $E(N)$ required to reject H_1 and H_2 with 80% marginal power for each. The choice of r was based on first minimizing $E(N)$ and then on minimizing M. The sample sizes M and $E(N)$ for different combinations of (θ_1, θ_2) and boundaries are given in Table 3. The minimum $E(N)$ for each case is shown in bold.

7 Concluding remarks

We have focused on minimizing $E(N)$ in this paper, but it may be noted that the differences in $E(N)$ are not large at least in this small example, and that other considerations could also be important. In addition, the assumptions that lead to differences in the optimal r are quite uncertain. Choosing $r = 1$ has certain practical benefits although it may not be necessarily an optimal choice in terms of $E(N)$. For example, one can directly use standard GSPs (including POC and OBF) that are offered in software packages. On the other hand, the EMA (2007) guideline states that "Often it may not be acceptable to stop a trial very early, despite convincing efficacy results, because insufficient data on safety, or on secondary endpoints may be available ... ". For this reason, it may be advisable to choose an r greater than the one that minimizes $E(N)$. The lesson here is that many different considerations dictate the choice of r.

Table 3 Maximum and expected sample sizes for 80% power to reject both H_1 and H_2 for different scenarios and boundary combinations ($m = 3, \alpha = 0.05, \rho = 0.5$).

Case	θ_1	θ_2	Boundary	$r = 1$		$r = 2$		$r = 3$	
#				M	$E(N)$	M	$E(N)$	M	$E(N)$
1	0.1	0.1	OBF-POC	803	646	772	635	742	**634**
2	0.1	0.1	OBF-OBF	703	**592**	702	593	694	600
3	0.1	0.2	OBF-POC	636	**520**	634	527	622	548
4	0.1	0.2	OBF-OBF	636	**523**	634	528	621	547
5	0.2	0.1	OBF-POC	740	519	691	**500**	648	507
6	0.2	0.1	OBF-OBF	636	**523**	634	528	621	547
7	0.2	0.2	OBF-POC	201	162	193	159	186	**159**
8	0.2	0.2	OBF-OBF	176	**149**	176	149	174	151
9	0.3	0.2	OBF-POC	188	135	176	**128**	165	129
10	0.3	0.2	OBF-OBF	162	135	161	**135**	158	139

The minimum $E(N)$ over different choices of r is shown in bold for each case.

From Fig. 2 we see that the inflation in FWER of the adaptive GSP is rather small for practically encountered ranges of ρ, e.g. $0.2 \leq \rho \leq 0.8$, and occurs only over a small interval of δ_1-values. Thus it may be possible to use the adaptive GSP with its attendant power gain if we can statistically rule out those particular combinations of ρ and δ_1 by constructing a joint confidence set for these parameters. An approach similar to the one adopted in Tamhane et al. (2012) may be followed to address this problem.

Acknowledgments The authors would like to thank Dr. Ekkehard Glimm for pointing out that the adaptive version of GSP(r) does not control FWER. We are also grateful to Dr. Willi Maurer for helpful suggestions. We acknowledge the constructive comments of the editor, the associate editor, and three referees, which led to an improved article. This paper forms a part of the first author's doctoral dissertation under the second author's supervision in the Department of Statistics at Northwestern University.

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

Appendix

Proof of Proposition 4.1. Let $(Z_{i1}, \ldots, Z_{im})$ be the test statistics for testing hypothesis $H_i : \theta_i = 0$ versus an upper one-sided alternative and let $(c_{i1}, \ldots, c_{im})$ be the corresponding critical boundary at level α_i $(i = 1, 2)$ where $\alpha_1 + \alpha_2 = \alpha$. Assume that H_1 is false and H_2 is true. Then the FWER of GSP(s) is given by

$$\begin{aligned}\text{FWER} = P(\text{Reject } H_2) &= \sum_{s=1}^{m} P(\text{Reject } H_1 \text{ @ stage } s, \text{ reject } H_2 \text{ @ any stage}) \\ &+ \sum_{s=1}^{m} P(\text{Do not reject } H_1, \text{ reject } H_2 \text{ @ stage } s).\end{aligned}$$

We will show that this FWER equals the FWER of GSP(m) procedure that is α.

Note that the second term of the above expression is the same for GSP(m) since in that case H_1 is not rejected and so no significance level is recycled to H_2. So we need to show that the first term of the above expression is the same for both GSP(s) and GSP(m). When the Z_{1j} are mutually independent of the Z_{2j} for $1 \leq j \leq m$, we can write each summand in the first term as

$$\begin{aligned}
&P(\text{Reject } H_1 \text{ @ Stage } s, \text{ reject } H_2 \text{ @ any stage}) \\
&\quad = P(\text{Reject } H_1 \text{ @ stage } s) \times P(\text{Reject } H_2 \text{ using GSP}(s)) \\
&\quad = P(\text{Reject } H_1 \text{ @ stage } s) \times \alpha \\
&\quad = P(\text{Reject } H_1 \text{ @ stage } s) \times P(\text{Reject } H_2 \text{ using GSP}(m)) \\
&\quad = P(\text{Reject } H_1 \text{ @ stage } s, \text{ reject } H_2 \text{ using GSP}(m)).
\end{aligned}$$

The third step follows because GSP(s) fully utilizes the recycled significance level and the fourth step follows because GSP(m) also fully utilizes the recycled significance level. This concludes the proof for the independence case. □

Proof of Proposition 4.2. Here we will assume $m = 2$, i.e. a two-stage GSP, to keep the notation and proof simple. Note that, when H_i is true, (c_{i1}, c_{i2}) satisfy

$$P(Z_{i1} > c_{i1}) + P(Z_{i1} \leq c_{i1}, Z_{i2} > c_{i2}) = \alpha_i \ (i = 1, 2).$$

Denote the delayed recycling critical boundary for H_i using GSP(1) by (c'_{i1}, c'_{i2}), which is simply an α-level boundary, i.e.

$$P(Z_{i1} > c'_{i1}) + P(Z_{i1} \leq c'_{i1}, Z_{i2} > c'_{i2}) = \alpha \ (i = 1, 2)$$

and denote the delayed recycling critical boundary for H_i using GSP(2) by (c_{i1}, c''_{i2}) where c''_{i2} is the solution to the equation

$$P(Z_{i1} > c_{i1}) + P(Z_{i1} \leq c_{i1}, Z_{i2} > c''_{i2}) = \alpha \ (i = 1, 2).$$

Furthermore, let n_j be the sample size at Stage j ($j = 1, 2$). Assume that H_1 is false and H_2 is true. Then $Z_{11} \sim N(\delta_1\sqrt{t_1}, 1)$, $Z_{12} \sim N(\delta_1, 1)$ and $Z_{2j} \sim N(0, 1)$ where $\delta_1 = \theta_1\sqrt{n_1 + n_2}$ is the drift parameter for H_1. Note that $\text{Corr}(Z_{i1}, Z_{i2}) = \sqrt{t_1}$ for $i = 1, 2$. Denote by ρ the correlation between the two endpoints so that $\text{Corr}(Z_{11}, Z_{21}) = \text{Corr}(Z_{12}, Z_{22}) = \rho$.

To calculate the type I error, we divide the event $E = (\text{Reject } H_2)$ into the following mutually exclusive events:

$$\begin{aligned}
E_1 &= (\text{Reject } H_1 \text{ and } H_2 \text{ @ Stage 1}), \\
E_2 &= (\text{Reject } H_1 \text{ @ Stage 1 and } H_2 \text{ @ Stage 2}), \\
E_3 &= (\text{Reject } H_1 \text{ and } H_2 \text{ @ Stage 2}), \\
E_4 &= (\text{Not reject } H_1, \text{ reject } H_2 \text{ @ Stage 1}), \\
E_5 &= (\text{Not reject } H_1, \text{ reject } H_2 \text{ @ Stage 2}).
\end{aligned}$$

Note that events E_4 and E_5 do not involve recycling.

Denote by $P_i = P(E_i)$ when H_1 is false and H_2 is true, so that FWER $= P_1 + P_2 + P_3 + P_4 + P_5$. We can write the following expressions.

$$
\begin{aligned}
P_1 &= P(Z_{11} > c_{11}, Z_{21} > c'_{21}),\\
P_2 &= P(Z_{11} > c_{11}, Z_{21} \le c'_{21}, Z_{22} > c'_{22}),\\
P_3 &= P(Z_{11} \le c_{11}, Z_{21} \le c_{21}, Z_{12} > c_{12}, Z_{22} > c''_{22}),\\
P_4 &= P(Z_{11} \le c_{11}, Z_{21} > c_{21}),\\
P_5 &= P(Z_{11} \le c_{11}, Z_{21} \le c_{21}, Z_{12} \le c_{12}, Z_{22} > c_{22}).
\end{aligned}
$$

Assume that $\rho = 1$. Then $Z_{1j} = Z_{2j} + \delta_1\sqrt{t_j}$ for $j = 1, 2$. For simplicity of notation, denote $X = Z_{21}$ and $Y = Z_{22}$ where X and Y are $N(0, 1)$ with $\mathrm{Corr}(X, Y) = \sqrt{t_1}$. To find a counterexample we consider a situation where H_2 is rejected at the same stage when H_1 is rejected. Therefore, we try to maximize the probability of E_1, E_2, and E_3. We will show that FWER $> \alpha$ if $\delta_1 = c_{11}/\sqrt{t_1}$. Since typically $c_{11} \ge c_{12}$ (e.g. the OBF or the POC boundaries), we will have $\delta_1 > c_{12}$.

$$
\begin{aligned}
P_1 &= P(X > c_{11} - \delta_1\sqrt{t_1}, X > c'_{21}) = P(X > \max(0, c'_{21})) = P(X > c'_{21}),\\
P_2 &= P(X > c_{11} - \delta_1\sqrt{t_1}, X \le c'_{21}, Y > c'_{22}) = P(0 < X \le c'_{21}, Y > c'_{22}),\\
P_3 &= P(X \le c_{11} - \delta_1\sqrt{t_1}, X \le c_{21}, Y > c_{12} - \delta_1, Y > c''_{22})\\
&= P(X \le \min(0, c_{21}), Y > \max(c_{12} - \delta_1, c''_{22}))\\
&= P(X \le 0, Y > c''_{22}) \ \text{ since } c_{12} - \delta_1 < 0,\\
P_4 &= P(X \le c_{11} - \delta_1\sqrt{t_1}, X > c_{21}) = P(c_{21} < X \le 0) = 0,
\end{aligned}
$$

$$
\begin{aligned}
P_5 &= P(X \le c_{11} - \delta_1\sqrt{t_1}, X \le c_{21}, Y \le c_{12} - \delta_1, Y > c_{22})\\
&= P(X \le 0, c_{22} < Y \le c_{12} - \delta_1) = 0.
\end{aligned}
$$

Hence

$$
\begin{aligned}
\text{FWER} &= P_1 + P_2 + P_3\\
&= P(X > c'_{21}) + P(0 < X \le c'_{21}, Y > c'_{22}) + P(X \le 0, Y > c''_{22})\\
&= P(X > c'_{21}) + P(X \le c'_{21}, Y > c'_{22}) + [P(X \le 0, Y > c''_{22}) - P(X \le 0, Y > c'_{22})]\\
&> \alpha
\end{aligned}
$$

since (c'_{21}, c'_{22}) is an α-level boundary, so the first two terms sum to α and the term in square brackets is > 0 because $c''_{22} < c'_{22}$. Hence FWER is inflated when $\rho = 1$ and $\delta_1 = c_{11}/\sqrt{t_1}$, and by continuity when $\rho < 1$ and δ_1 is chosen appropriately. □

Proof of Proposition 4.3. Using the same notations, we then show that if $L \le \delta_1 \le U$, FWER $= \alpha$ for $\rho = 1$, where $L = c_{12} - c''_{22} \ge 0$ and $U = (c_{11} - c_{21})/\sqrt{t_1} \ge 0$. If $L \le \delta_1 \le U$ then we have

$$
\begin{aligned}
P_1 &= P(X > c_{11} - \delta_1\sqrt{t_1}, X > c'_{21}) = P(X > c_{11} - \delta_1\sqrt{t_1}) \text{ since } c_{21} \ge c'_{21}.\\
P_2 &= P(X > c_{11} - \delta_1\sqrt{t_1}, X \le c'_{21}, Y > c'_{22}) = P(X > c_{11} - \delta_1\sqrt{t_1}, X \le c'_{21}, Y > c'_{22}) = 0.\\
P_3 &= P(X \le c_{11} - \delta_1\sqrt{t_1}, X \le c_{21}, Y > c_{12} - \delta_1, Y > c''_{22}) = P(X \le c_{21}, Y > c''_{22}).
\end{aligned}
$$

$$P_4 = P(c_{21} < X \leq c_{11} - \delta_1\sqrt{t_1}),$$
$$P_5 = P(X \leq c_{11} - \delta_1\sqrt{t_1}, X \leq c_{21}, Y \leq c_{12} - \delta_1, Y > c_{22}) = 0.$$

Thus

$$P_1 + P_3 + P_4 = P(X > c_{21}) + P(X \leq c_{21}, Y > c''_{22}) = \alpha. \quad (1)$$

Therefore, if $L \leq \delta_1 \leq U$, then FWER $= \alpha$ for $\rho = 1$. □

References

Armitage, P. (1975). *Sequential Medical Trials, 2nd edn*, Blackwell Scientific Publication, Oxford, UK.

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.

European Medicines Agency (EMA). (2007). Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. EMA, London, UK.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics (Vol. 195). Springer-Verlage, Heidelberg, DE.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2014). mvtnorm: Multivariate Normal and t Distributions. URL http://CRAN.R-project.org/package=mvtnorm. R package version 1.0-0.

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

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

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

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Procedures with Applications to Clinical Trials*. Chapman and Hall/CRC Press, Boca Raton, FL.

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

Liu, Q. and Anderson, K. M. (2008). On adaptive extensions of group sequential trials for clinical investigations. *Journal of the American Statistical Association* **103**, 1621–1630.

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

Miwa, T., Hayter, A. J. and Kuriki, S. (2003). The evaluation of general non-centred orthant probabilities. *Statistics in Medicine* **65**, 223–234.

O'Brien, P. C. and Flemming, 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.

Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1977). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498.

Tang, D. I. and Geller, N. L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* **55**, 1188–1192.

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.

Tamhane, A. C., Wu, Y. and Mehta, C. R. (2012). Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (I): unknown correlation between the endpoints. *Statistics in Medicine* **31**, 2027–2040.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Wiley, Chinchester, UK.

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.

