Biometrics 66, 1174–1184
December 2010
DOI: 10.1111/j.1541-0420.2010.01402.x

Testing a Primary and a Secondary Endpoint in a Group Sequential Design

Ajit C. Tamhane,[1,*] Cyrus R. Mehta[2,] and Lingyun Liu[3,***]**

[1]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illnois 60208, U.S.A.
[2]Cytel Corporation and Department of Biostatistics, Harvard School of Public Health, Cambridge, Massachusetts 02139, U.S.A.
[3]Cytel Corporation, Cambridge, Massachusetts 02139, U.S.A.
[*]*email:* atamhane@northwestern.edu
[**]*email:* mehta@cytel.com
[***]*email:* lingyun.liu@cytel.com

Summary. We consider a clinical trial with a primary and a secondary endpoint where the secondary endpoint is tested only if the primary endpoint is significant. The trial uses a group sequential procedure with two stages. The familywise error rate (FWER) of falsely concluding significance on either endpoint is to be controlled at a nominal level α. The type I error rate for the primary endpoint is controlled by choosing any α-level stopping boundary, e.g., the standard O'Brien–Fleming or the Pocock boundary. Given any particular α-level boundary for the primary endpoint, we study the problem of determining the boundary for the secondary endpoint to control the FWER. We study this FWER analytically and numerically and find that it is maximized when the correlation coefficient ρ between the two endpoints equals 1. For the four combinations consisting of O'Brien–Fleming and Pocock boundaries for the primary and secondary endpoints, the critical constants required to control the FWER are computed for different values of ρ. An ad hoc boundary is proposed for the secondary endpoint to address a practical concern that may be at issue in some applications. Numerical studies indicate that the O'Brien–Fleming boundary for the primary endpoint and the Pocock boundary for the secondary endpoint generally gives the best primary as well as secondary power performance. The Pocock boundary may be replaced by the ad hoc boundary for the secondary endpoint with a very little loss of secondary power if the practical concern is at issue. A clinical trial example is given to illustrate the methods.

Key words: Familywise error rate; Gatekeeping procedures; Multiple comparisons; Multiple endpoints; O'Brien–Fleming boundary; Pocock boundary; Primary power; Secondary power.

1. Introduction

Since the pioneering works of Pocock (1977) and O'Brien and Fleming (1979), group sequential designs have been widely studied and are now commonly employed in clinical trials. Jennison and Turnbull (2000) is a comprehensive reference on this topic. Much of the work on group sequential designs deals with a single endpoint. Jennison and Turnbull (1993) discussed a group sequential design for bivariate endpoints (efficacy and safety). They determined two-sided decision boundaries for the test statistics computed for both endpoints so that if the new treatment is shown to be superior on both endpoints then it is accepted; if it is shown to be inferior on one of the endpoints then it is rejected; otherwise sampling is continued to the next stage. We are interested in the problem of testing null hypotheses on hierarchically ordered endpoints, e.g., a primary and a secondary endpoint. For such endpoints, it is generally required that statistical significance be shown on the primary endpoint before the secondary endpoint can be validly tested (O'Neill, 1997). Gatekeeping procedures (Dmitrienko and Tamhane, 2007, 2010) deal with such endpoints; however, these procedures have been studied only for nonsequential designs. This article studies the problem of testing a primary and a secondary endpoint in a two-stage group sequential setting with the former acting as a gatekeeper for the latter. This problem was originally studied by Hung, Wang, and O'Neill (2007). While this article was being revised, we received a manuscript by Glimm, Maurer, and Bretz (2010) which addresses the same problem and reaches similar conclusions but does not give explicit analytical results concerning the familywise error rate (FWER) and the primary and secondary critical boundaries as we derive in this article. These two works are completely independent and complementary to each other.

The outline of the article is as follows: Section 2 describes the group sequential procedure (GSP) for the stated problem. Section 3 discusses control of the FWER for this GSP. This section also gives a table of the critical constants using either the O'Brien–Fleming (OF) or the Pocock (PO) boundaries for the primary and secondary endpoints resulting in four different combinations. Two additional combinations that use the OF or the PO boundary for the primary and an ad hoc (AH) boundary for the secondary endpoint are also studied.

This AH boundary is introduced to address a practical concern that is explained in Section 3.2. The powers of these six different boundary combinations are compared in Section 4. Section 5 gives a clinical trial example. Section 6 outlines extensions for future research to deal with unknown correlation coefficient, multiple stages and multiple endpoints. The proofs of all theoretical results are included in Supplementary Materials.

2. Group Sequential Procedure

Consider a GSP with two stages and a primary and a secondary endpoint. Let n_1 and n_2 be the sample sizes for the two stages. Suppose that the observations on the primary endpoint are i.i.d. $N(\mu_1, \sigma_1^2)$ and those on the secondary endpoint are i.i.d. $N(\mu_2, \sigma_2^2)$. Further suppose that the two endpoints are jointly distributed as bivariate normal with correlation coefficient $\rho \geqslant 0$. Hung et al. (2007) considered the problem of testing the null hypotheses $H_1 : \mu_1 = 0$ and $H_2 : \mu_2 = 0$ against one-sided alternatives with H_2 being tested if H_1 is rejected either at the first stage or at the second stage. Thus, the primary endpoint acts as a gatekeeper for the secondary endpoint. The FWER requirement is

$$\text{FWER} = P\{\text{Reject at least one true } H_i \ (i = 1, 2)\} \leqslant \alpha \tag{1}$$

for specified α.

The standardized cumulative sample means at the two stages are used as the test statistics for both endpoints. Denote by (X_1, X_2) the cumulative test statistics at the two stages for the primary endpoint and (Y_1, Y_2) the corresponding test statistics for the secondary endpoint. Let (c_1, c_2) and (d_1, d_2) denote the corresponding stopping boundaries. We will use the following GSP.

Stage 1: Take n_1 observations and compute (X_1, Y_1). If $X_1 \leqslant c_1$ continue to stage 2. If $X_1 > c_1$, reject H_1 and test H_2. If $Y_1 > d_1$, reject H_2; otherwise accept H_2. In either case terminate the trial.

Stage 2: Take n_2 observations and compute (X_2, Y_2). If $X_2 \leqslant c_2$, accept H_1 and stop testing; otherwise reject H_1 and test H_2. If $Y_2 > d_2$, reject H_2; otherwise accept H_2.

In general, $X_j \sim N(\Delta_{1j}, 1)$ and $Y_j \sim N(\Delta_{2j}, 1)$ for $j = 1, 2$, where

$$\Delta_{i1} = \Delta_i \sqrt{n_1}, \Delta_{i2} = \delta_i \sqrt{n_1 + n_2} \quad \text{and} \quad \delta_i = \frac{\mu_i}{\sigma_i}; \tag{2}$$

under $H_i, \Delta_{ij} = 0$ $(i, j = 1, 2)$. Frequently, we will denote the noncentrality parameters by $\Delta_i = \Delta_{i1} = \gamma \Delta_{i2}$ $(i = 1, 2)$, where

$$\gamma = \sqrt{\frac{n_1}{n_1 + n_2}}.$$

The correlations among (X_1, X_2, Y_1, Y_2) are given by

$$\begin{aligned} \text{corr}(X_1, X_2) &= \text{corr}(Y_1, Y_2) = \gamma, \\ \text{corr}(X_1, Y_1) &= \text{corr}(X_2, Y_2) = \rho, \\ \text{corr}(X_1, Y_2) &= \text{corr}(X_2, Y_1) = \rho\gamma. \end{aligned} \tag{3}$$

3. Familywise Error Rate Control

3.1 *Choice of the Primary Boundary*

To determine the critical boundaries, (c_1, c_2) and (d_1, d_2), in order to satisfy the FWER control requirement (1), we need to consider three configurations, namely, $H_1 \cap H_2, H_1 \cap \bar{H}_2$, and $\bar{H}_1 \cap H_2$, where $\bar{H}_i$ denotes that H_i is false $(i = 1, 2)$. If H_1 is true (the first two cases) then the FWER is controlled at level α if the type I error probability for H_1 is controlled at level α. This is obvious under $H_1 \cap \bar{H}_2$ because there is no type I error for rejecting H_2 in that case. It is also true under $H_1 \cap H_2$ because the event $R_2 = \{\text{Reject } H_2\}$ is a subset of the event $R_1 = \{\text{Reject } H_1\}$ and the probability of R_1 does not depend on the truth or falsity of H_2. Hence, to control $P_{H_1 \cap H_2}(R_1 \cup R_2) = P_{H_1}(R_1) \leqslant \alpha$, one must choose (c_1, c_2) to satisfy

$$P_{H_1}(X_1 > c_1) + P_{H_1}(X_1 \leqslant c_1, X_2 > c_2) \leqslant \alpha. \tag{4}$$

Any pair of constants, (c_1, c_2), that satisfy this inequality will be referred to as an α-level boundary. For the OF boundary, $c_2 = \gamma c_1$, whereas for the PO boundary, $c_2 = c_1$.

More generally, we could use the error spending function approach of Lan and DeMets (1983) to calculate the primary boundary (c_1, c_2). Let $\alpha(t)$ be a nondecreasing function defined over $t \in [0, 1]$ such that $\alpha(0) = 0$ and $\alpha(1) = \alpha$; here t is called the information fraction. Let $t_1 = n_1/(n_1 + n_2)$ and $t_2 = 1$. Then (c_1, c_2) can be calculated from the following two equations:

$$\begin{aligned} P_{H_1}(X_1 > c_1) = \alpha(t_1) \quad \text{and} \quad & P_{H_1}(X_1 > c_1) \\ & + P(X_1 \leqslant c_1, X_2 > c_2) = \alpha(t_2) = \alpha. \end{aligned}$$

For a K-stage GSP $(K \geqslant 2)$ with equal information in each stage, i.e., $t_i = i/K$ $(1 \leqslant i \leqslant K)$, an error spending function that approximates the OF boundary is given by (Lan and DeMets, 1983):

$$\alpha_1(t) = 2\Phi\left(-\frac{z_{\alpha/2}}{\sqrt{t}}\right), \tag{5}$$

where $\Phi(\cdot)$ is the standard normal cdf and $z_{\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the standard normal distribution. Similarly, an error spending function that approximates the PO boundary is given by (Lan and DeMets, 1983):

$$\alpha_2(t) = \alpha \ln\{1 + (e - 1)t\}. \tag{6}$$

These spending functions are used in the example of Section 5.

In conclusion, the FWER control problem is: Given a choice of (c_1, c_2) satisfying equation (4), determine (d_1, d_2) so that equation (1) is satisfied under H_2. We will assume this configuration, which implies that $\Delta_1 \geqslant 0$ and $\Delta_2 = 0$, throughout the next section.

3.2 *Choice of the Secondary Boundary*

To test multiple hypotheses in a group sequential setup, Tang and Geller (1999) showed that if there exists a GSP to test every intersection hypothesis at level α then application of the closure principle of Marcus, Peritz, and Gabriel (1976) leads to a GSP that controls the FWER at level α. With H_1 and H_2 there are only three intersections: $H_1 \cap H_2, H_1$, and H_2.

Table 1
Values of critical constant d for secondary boundary using combinations of the O'Brien–Fleming (OF)[a], Pocock (PO)[b], and ad hoc (AH)[c] boundaries for $n_1 = n_2$ and $\alpha = 0.05$

Primary boundary	Secondary boundary	ρ 0.0	0.2	0.4	0.6	0.8	1.0
OF1	OF2	1.407 (2.041)[d]	1.428 (1.773)	1.476 (1.478)	1.495 (1.150)	1.551 (0.767)	1.678 (0)
OF1	PO2	1.645 (∞)	1.663 (2.638)	1.686 (2.134)	1.717 (1.663)	1.760 (1.184)	1.876 (0.497)
OF1	AH2	1.645 (∞)	1.671 (2.505)	1.714 (0.949)	1.786 (1.430)	1.926 (0.889)	2.515 (0.023)
PO1	PO2	1.645 (∞)	1.655 (2.679)	1.672 (2.080)	1.698 (1.528)	1.742 (0.952)	1.876 (0)
PO1	OF2	1.280 (2.169)	1.304 (1.874)	1.333 (1.553)	1.372 (1.202)	1.429 (0.809)	1.570 (0.216)
PO1	AH2	1.645 (∞)	1.656 (2.635)	1.679 (2.010)	1.720 (1.444)	1.801 (0.876)	2.092 (0.163)

[a]OF primary boundary: $c_1 = 1.678\sqrt{2}, c_2 = 1.678$, OF secondary boundary: $d_1 = d\sqrt{2}, d_2 = d$.
[b]PO primary boundary: $c_1 = c_2 = 1.876$, PO secondary boundary: $d_1 = d_2 = d$.
[c]AH secondary boundary: tabled value is $d_1 \geq d_2 = z_{.05} = 1.645$.
[d]The value in the parentheses is the value of Δ_1 that maximizes FWER.

As seen earlier, an α-level GSP of $H_1 \cap H_2$ is the same as the GSP of H_1 because of the hierarchical testing of H_1 and H_2. Therefore, if we use any α-level secondary boundary for testing H_2 conditional on rejecting H_1, then the overall FWER will be controlled. However, it is possible to use a more liberal secondary boundary by exploiting the dependence of the FWER on the correlation coefficient ρ as calculations in Table 1 shows. The calculations were made using the following expression for FWER under H_2:

$$\begin{aligned}\text{FWER} &= P_{H_2}(X_1 > c_1, Y_1 > d_1) \\ &\quad + P_{H_2}(X_1 \leqslant c_1, X_2 > c_2, Y_2 > d_2) \\ &= \int_{c_1-\Delta_{11}}^{\infty} \Phi\left(\frac{-d_1+\rho u}{\sqrt{1-\rho^2}}\right)\phi(u)du \\ &\quad + \int_{c_2-\Delta_{12}}^{\infty} \Phi\left(\frac{c_1-\Delta_{11}-\gamma u}{\sqrt{1-\gamma^2}}\right) \\ &\quad \times \Phi\left(\frac{-d_2+\rho u}{\sqrt{1-\rho^2}}\right)\phi(u)du,\end{aligned}$$

where $\phi(\cdot)$ denotes the pdf of the standard normal distribution; also $\Delta_{11} = \gamma\Delta_{12} = \Delta_1$. This expression is obtained by setting $\Delta_{21} = \Delta_{22} = 0$ in the secondary power expression (8).

Hung et al. (2007) investigated the choice $d_1 = d_2 = z_\alpha$ (where z_α is the upper α critical point of the $N(0,1)$ distribution) as Strategy 1 on the basis that the primary and secondary endpoints are tested in a fixed sequence, and so H_2 can be tested at level α following rejection of H_1 at level α. We show in Proposition 1 that this choice does not control the FWER for all ρ and Δ_1.

PROPOSITION 1: *If $d_1 = d_2 = z_\alpha$ then for $\rho = 0$,* FWER *is increasing in Δ_1 and* FWER $\to \alpha$ *as $\Delta_1 \to \infty$ (and thus* FWER *is controlled for all Δ_1). In fact,* FWER $\to \alpha$ *as $\Delta_1 \to \infty$ for all $\rho \geqslant 0$. However, for $\rho = 1$, $\max_{\Delta_1}$ FWER $> \alpha$ and thus* FWER *is not controlled for all ρ and Δ_1.*

To illustrate the results of Proposition 1, we considered the OF boundary for the primary endpoint with $\alpha = 0.05$ and $n_1 = n_2 = n$, which uses $(c_1, c_2) = (1.678\sqrt{2}, 1.678)$. Figure 1 shows the behavior of FWER as a function of ρ and Δ_1 when $d_1 = d_2 = z_{.05} = 1.645$. We see that the maximum FWER with respect to ρ and Δ_1 is attained when $\rho = 1$ and Δ_1 is close to 1, and this maximum is approximately 0.08. On the other hand, when $\rho = 0$, FWER $\leqslant \alpha$ for all $\Delta_1 \geqslant 0$. Furthermore, FWER $\to \alpha$ as $\Delta_1 \to \infty$ for all $\rho \geqslant 0$.

As seen from Figure 1, the choice $d_1 = d_2 = z_\alpha$ can be liberal for some $\Delta_1 > 0$ when $\rho > 0.4$. As an ad hoc solution to this problem, Hung et al. (2007) suggested Strategy 3 that uses $d_1 = d_2 = z_{\alpha/2}$; however, this choice can be shown to be too conservative. Instead of an ad hoc solution, we can numerically search to find the max FWER with respect to Δ_1 and ρ for different choices of $d_1 = d_2 = z_{\alpha^*}$ such that max FWER $=$ 0.05. Figure 2 shows that this is achieved when $\Delta_1 = 0, \rho = 1$, and $d_1 = d_2 = 1.876$, which corresponds to $\alpha^* = 0.0303$. Note that this is the PO boundary for $\alpha = 0.05$. We will show this numerical result analytically in Proposition 3.

In Propositions 2–4 that follow, we study three different combinations of the α-level boundaries: $c_1 = d_1, c_2 = d_2$ in Proposition 2, $c_1 > d_1, c_2 < d_2$ in Proposition 3, and $c_1 < d_1, c_2 > d_2$ in Proposition 4. From Tang and Geller (1999) it follows that all three boundaries control the FWER. In the first two cases we show that max FWER $= \alpha$, whereas in the third case we show that max FWER $< \alpha$. Hence it is possible to use a more liberal (d_1, d_2) boundary with level $\alpha' > \alpha$ in the third case.

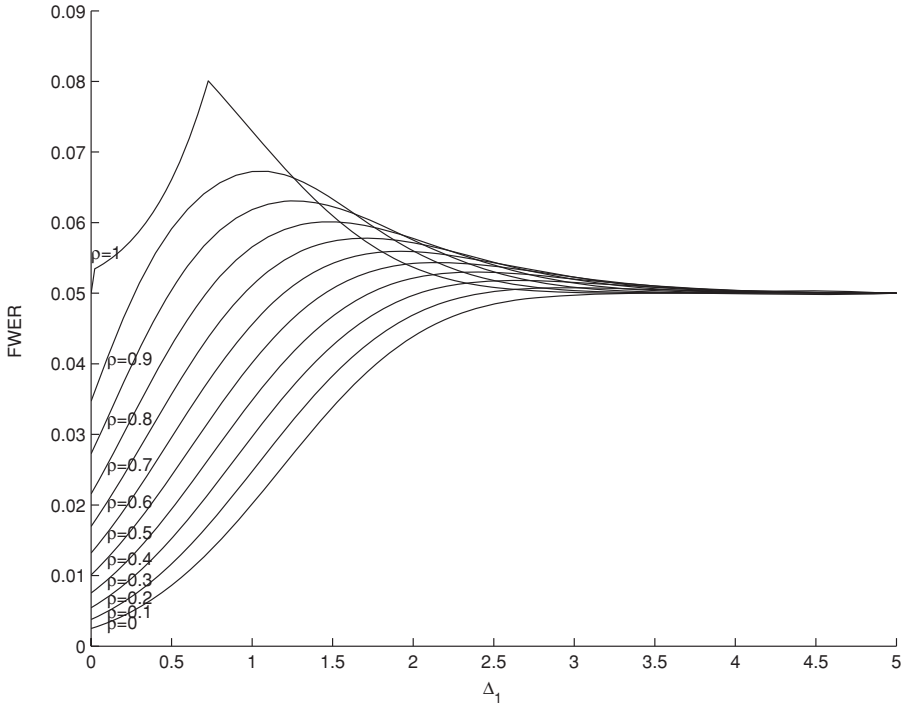

Figure 1. FWER as a function of ρ and Δ_1 when $(c_1, c_2) = (1.678\sqrt{2}, 1.678), n_1 = n_2$ and $d_1 = d_2 = z_{0.05} = 1.645$.

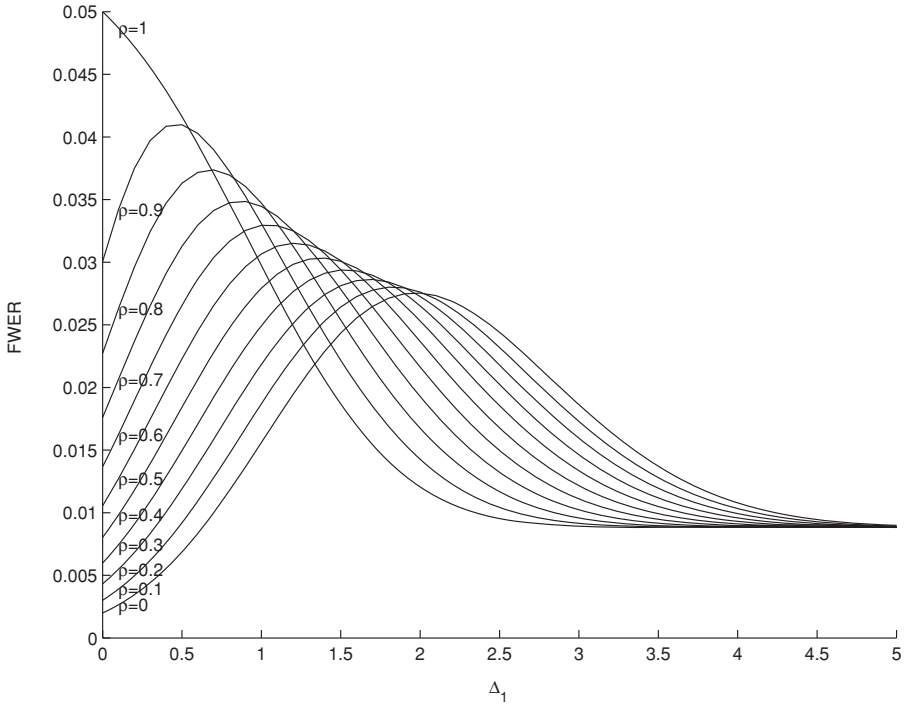

Figure 2. FWER as a function of ρ and Δ_1 when $(c_1, c_2) = (d_1, d_2) = (1.678\sqrt{2}, 1.678)$ and $n_1 = n_2$.

PROPOSITION 2. *If $(c_1, c_2) = (d_1, d_2)$ is an α-level boundary for the primary and secondary endpoints then for $\rho = 1, \max_{\Delta_1} \text{FWER} = \alpha$ is attained at $\Delta_1 = 0$ and for $\Delta_1 = 0, \max_{\rho} \text{FWER} = \alpha$ is attained at $\rho = 1$.*

Note that this proposition does not show that the global maximum of FWER is attained at $\Delta_1 = 0$ and $\rho = 1$, which we conjecture to be true but have been unable to prove. However, plots in Figure 3 of numerically evaluated FWER as functions

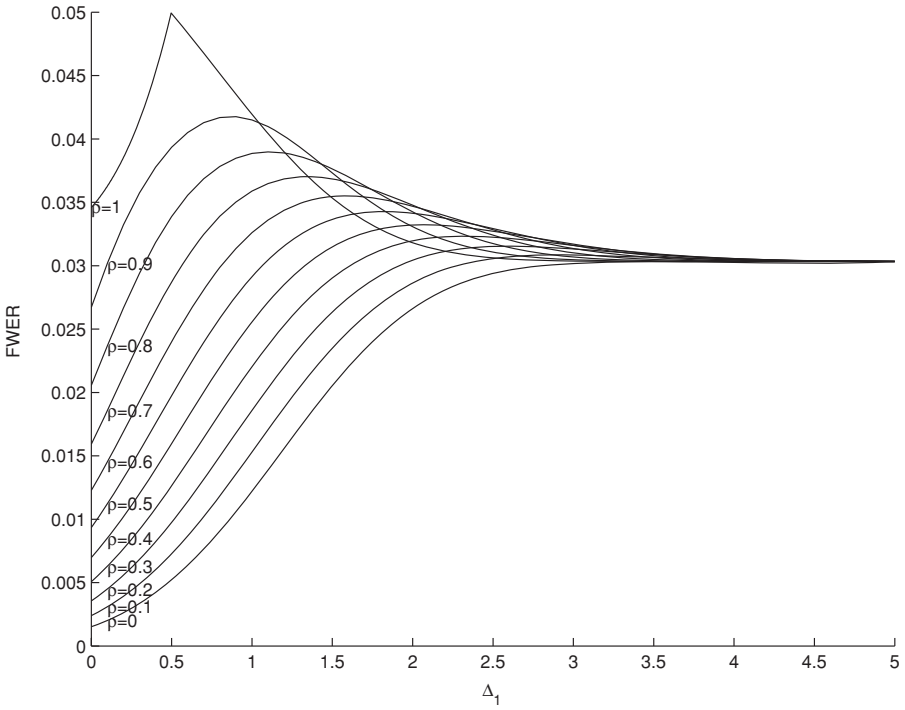

Figure 3. FWER as a function of ρ and Δ_1 when $(c_1, c_2) = (1.678\sqrt{2}, 1.678), n_1 = n_2$ and $d_1 = d_2 = z_{0.0303} = 1.876$.

of Δ_1 and ρ when $(c_1, c_2) = (d_1, d_2)$ is the OF boundary for $\alpha = 0.05$ show that the global maximum of the FWER with respect to ρ and Δ_1 is indeed attained at $\rho = 1$ and $\Delta_1 = 0$, and this maximum equals 0.05. In fact, as seen from Figures 1 and 2, even when $(c_1, c_2) \neq (d_1, d_2)$, the global maximum of FWER with respect to ρ is attained at $\rho = 1$ and $\Delta_1 > 0$.

PROPOSITION 3. *If (c_1, c_2) and (d_1, d_2) are α-level boundaries for the primary and secondary endpoints such that $c_1 > d_1$ and $c_2 < d_2$ (e.g., if (c_1, c_2) is the OF boundary and (d_1, d_2) is the PO boundary) then for $\rho = 1, \max_{\Delta_1} \text{FWER} = \alpha$ is attained when $\Delta_1 = \Delta_{11} = c_1 - d_1$.*

PROPOSITION 4. *If (c_1, c_2) and (d_1, d_2) are α-level boundaries for the primary and secondary endpoints such that $c_1 < d_1$ and $c_2 > d_2$ (e.g., if (c_1, c_2) is the PO boundary and (d_1, d_2) is the OF boundary) then for $\rho = 1, \max_{\Delta_1} \text{FWER} < \alpha$ is attained when $\Delta_1 = \gamma\Delta_{12} = \gamma(c_2 - d_2)$. Therefore, $\max \text{FWER}$ can be increased to α by decreasing (d_1, d_2) to (d_1', d_2') so that (d_1', d_2') forms an α'-level boundary with $\alpha' > \alpha$.*

Propositions 2–4 are illustrated by calculations in Table 1 of the secondary boundary (d_1, d_2) that controls the FWER at level α for the two choices of the α-level primary boundary (c_1, c_2) for $\alpha = 0.05, n_1 = n_2 = n$, and $\rho = 0.0(0.2)1.0$. For the OF boundary, $(c_1, c_2) = (1.678\sqrt{2}, 1.678)$ and for the PO boundary, $(c_1, c_2) = (1.876, 1.876)$. The same two choices are used for the secondary stopping boundary in the following way:

$$\text{OF Boundary: } d_1 = d\sqrt{2}, d_2 = d, \quad \text{PO Boundary: } d_1 = d_2 = d.$$

Table 1 lists the d-values for the four combinations of the primary and secondary boundaries.

In addition, we considered another boundary to address a referee's practical concern over having the secondary boundary with a level $\alpha' > \alpha$, as shown in Proposition 4. Note that, although the marginal error rate for the secondary endpoint exceeds α, the joint FWER for the primary and secondary endpoints is controlled at α. However, the referee thought that this subtlety may be difficult to explain to clinicians, and in fact suggested that both d_1, d_2 should be $\geqslant z_\alpha$.

For the PO secondary boundary, $(d_1 = d_2 = d)$, from Proposition 1 we have $d = z_\alpha$ for $\rho = 0$. From Figures 1 to 3 we see that $\max_{\Delta_1} \text{FWER}$ is an increasing function of ρ and so d is an increasing function of ρ; therefore, $d \geqslant z_\alpha$ for all $\rho \geqslant 0$, as can be checked from Table 1. Thus, the OF1–PO2 and PO1–PO2 combinations satisfy the referee's practical condition; as will be seen in the next section, the OF1–PO2 combination generally gives higher power for both the primary and the secondary endpoint, and is therefore preferred.

For the OF secondary boundary, $(d_1 = d\sqrt{2}, d_2 = d)$, it is possible to have $d_2 < z_\alpha$ and still control FWER. If this occurs then we set $d_2 = z_\alpha$ and then find the smallest $d_1 \geqslant z_\alpha$ such that $\max \text{FWER} = \alpha$ is maintained. We refer to this modified boundary as an AH boundary. We computed this secondary boundary for two additional combinations with the OF and PO as the primary boundaries. Note that we did not employ the AH boundary for the primary endpoint in our study. This gave a total of six combinations of the primary and secondary boundaries.

The following observations can be made from the numerical results in Table 1.

(1) The value of Δ_1 that maximizes FWER for fixed ρ is a decreasing function of ρ.

(2) If both primary and secondary boundaries are of the same type (e.g., both OF or both PO) then we get $(c_1, c_2) = (d_1, d_2)$ and $\Delta_1 = 0, \rho = 1$ gives the max FWER as shown in Proposition 2.

(3) If the primary boundary is OF and the secondary boundary is PO so that $c_1 > d_1$ and $c_2 < d_2$ then for $\rho = 1, \Delta_{11} = c_1 - d_1 = 1.678\sqrt{2} - 1.876 = 0.497$ gives max FWER, as shown in Proposition 3. Note that in Figure 3 we found $d_1 = d_2 = 1.876$ by numerical search, but it could have been found directly by applying Proposition 3.

(4) If the primary boundary is PO and the secondary boundary is OF so that $c_1 < d_1$ and $c_2 > d_2$ then for $\rho = 1$, the maximizing value of Δ_1 is given by $\gamma\Delta_{12} = \gamma(c_2 - d_2) = (1/\sqrt{2})(1.876 - 1.678) = 0.140$. However, using $(c_1, c_2) = (1.876, 1.876)$ and $(d_1, d_2) = (1.678\sqrt{2}, 1.678)$ gives FWER $= 0.0386 < 0.05$. So, as shown in Proposition 4, (d_1, d_2) can be decreased until max FWER increases to 0.05. The final boundary, $(d_1', d_2') = (1.570\sqrt{2}, 1.570)$, and the corresponding maximizing $\Delta_1 = 0.216$ were found numerically.

4. Power

There are two powers of interest: (i) primary power (Power_1), the probability of rejecting a false H_1, is just the power of GSP for a single endpoint that has been studied previously in the literature; and (ii) secondary power (Power_2), the probability of rejecting a false H_2. Note that $\text{Power}_1 \geqslant \text{Power}_2$ for the same Δ_{ij} $(i, j = 1, 2)$. The overall power for rejecting H_1 and H_2 when both are false or only H_2 is false is just the secondary power.

Consider an arbitrary configuration with $\Delta_i \geqslant 0$ $(i = 1, 2)$. The primary power is given by

$$\text{Power}_1 = P_{\bar{H}_1}(X_1 > c_1) + P_{\bar{H}_1}(X_1 \leqslant c_1, X_2 > c_2),$$

where $X_j \sim N(\Delta_{1j}, 1)$ $(j = 1, 2)$ with $\text{corr}(X_1, X_2) = \gamma$. Noting that the conditional distribution of X_2 given $U = X_1 - \Delta_{11} = u$ is $N(\Delta_{12} + \gamma u, 1 - \gamma^2)$, we get

$$\text{Power}_1 = \Phi(-c_1 + \Delta_{11}) + \int_{-\infty}^{c_1 - \Delta_{11}} \Phi\left(\frac{-c_2 + \Delta_{12} + \gamma u}{\sqrt{1 - \gamma^2}}\right) \phi(u) du. \quad (7)$$

The secondary power is given by

$$\text{Power}_2 = P_{\bar{H}_2}(X_1 > c_1, Y_1 > d_1) + P_{\bar{H}_2}(X_1 \leqslant c_1, X_2 > c_2, Y_2 > d_2),$$

where $X_j \sim N(\Delta_{1j}, 1), Y_j \sim N(\Delta_{2j}, 1)$ $(j = 1, 2)$ have the correlation structure shown in equation (4). In the first term, the conditional distribution of Y_1 given $U = X_1 - \Delta_{11} = u$ is $N(\Delta_{21} + \rho u, 1 - \rho^2)$. Similarly, in the second term, the conditional distribution of (X_1, Y_2) given $U = X_2 - \Delta_{12} = u$ is bivariate normal with mean vector $(\Delta_{11} + \gamma u, \Delta_{22} + \rho u)$ and covariance matrix

$$\begin{bmatrix} 1 & \rho\gamma \\ \rho\gamma & 1 \end{bmatrix} - \begin{bmatrix} \gamma \\ \rho \end{bmatrix} [\gamma, \rho] = \begin{bmatrix} 1 - \gamma^2 & 0 \\ 0 & 1 - \rho^2 \end{bmatrix},$$

i.e., conditional on $U = X_2 - \Delta_{12} = u, X_1$ and Y_2 are independent $N(\Delta_{11} + \gamma u, 1 - \gamma^2)$ and $N(\Delta_{22} + \rho u, 1 - \rho^2)$, respectively. Using these results we get

$$\text{Power}_2 = \int_{c_1 - \Delta_{11}}^{\infty} \Phi\left(\frac{-d_1 + \Delta_{21} + \rho u}{\sqrt{1 - \rho^2}}\right) \phi(u) du + \int_{c_2 - \Delta_{12}}^{\infty} \Phi\left(\frac{c_1 - \Delta_{11} - \gamma u}{\sqrt{1 - \gamma^2}}\right) \times \Phi\left(\frac{-d_2 + \Delta_{22} + \rho u}{\sqrt{1 - \rho^2}}\right) \phi(u) du. \quad (8)$$

We studied the primary and secondary powers for the six combinations of the boundaries for which Table 1 is given. The powers were numerically evaluated (not simulated) using equations (7) and (8). It can be shown that the OF boundary is uniformly (in Δ_1) more powerful in terms of primary power than the PO boundary. Jennison has proved this result more generally for any two α-level boundaries, (c_1, c_2) and (c_1', c_2'): if the test statistics have the monotone likelihood ratio distribution property then (c_1, c_2) is uniformly more powerful than (c_1', c_2') if $c_1 > c_1', c_2 < c_2'$.

The secondary power is more complicated to study because it depends on the particular combination of the boundaries used for the primary and secondary endpoints as well as the values of ρ and Δ_{ij}. The plots of secondary powers as functions of Δ_1 for the six combinations of the boundaries studied in Table 1 for $\alpha = 0.05, n_1 = n_2, \rho = 0.4$, and $\Delta_2 = 1, 2, 3$ are given in Figures 4–6; the patterns of the power functions are similar for other values of ρ and hence are not shown here. These figures show that generally the combination of the OF boundary for the primary endpoint and the PO boundary for the secondary endpoint (the OF1–PO2 combination) gives the highest secondary power. With the OF1–PO2 combination, we have $d_1, d_2 \geqslant 1.645$, so the practical concern raised earlier is not an issue. The OF1–AH2 has secondary power very similar to that of the OF1–PO2 combination, because their critical boundaries are similar. The reason that the secondary powers of the OF1–PO2 and OF1–AH2 combinations are close is because their critical boundaries are close. For $\rho = 0.4$, the OF1–PO2 boundaries are $(c_1 = 1.678\sqrt{2}, c_2 = 1.678)$ and $(d_1 = 1.686, d_2 = 1.686)$ whereas the OF1–AH2 boundaries are $(c_1 = 1.678\sqrt{2}, c_2 = 1.678)$ and $(d_1 = 1.714, d_2 = 1.645)$.

The power curves show that, for large Δ_1, any combination using OF2 is markedly worse than those using either PO2 or AH2. This is explained by the fact that when Δ_1 is large, there is a high chance of rejecting H_1 at the first look; so OF2 presents a higher threshold for the secondary endpoint to exceed at that look. Also note that for large Δ_1, the secondary power of OF1–OF2 is lower than that of PO1–OF2, which in turn is uniformly dominated by OF1–PO2 for all Δ_1 as well as the three values of Δ_2 that were studied. Thus, PO1–OF2 is not a contender. OF1–OF2 combination has a slight edge for small Δ_1. However, if OF1–OF2 is disqualified based on the referee's practical concern then OF1–AH2 may be preferred for small Δ_1 as well.

In looking at these plots, one may be intrigued by the fact that the secondary power of each combination of boundaries decreases with Δ_1 for large $\Delta_1 (\Delta_1 \geqslant 2)$ for fixed Δ_2. The reason is that the probability of rejecting H_1 at the first stage increases as Δ_1 increases. Therefore, for large Δ_1, H_2 gets

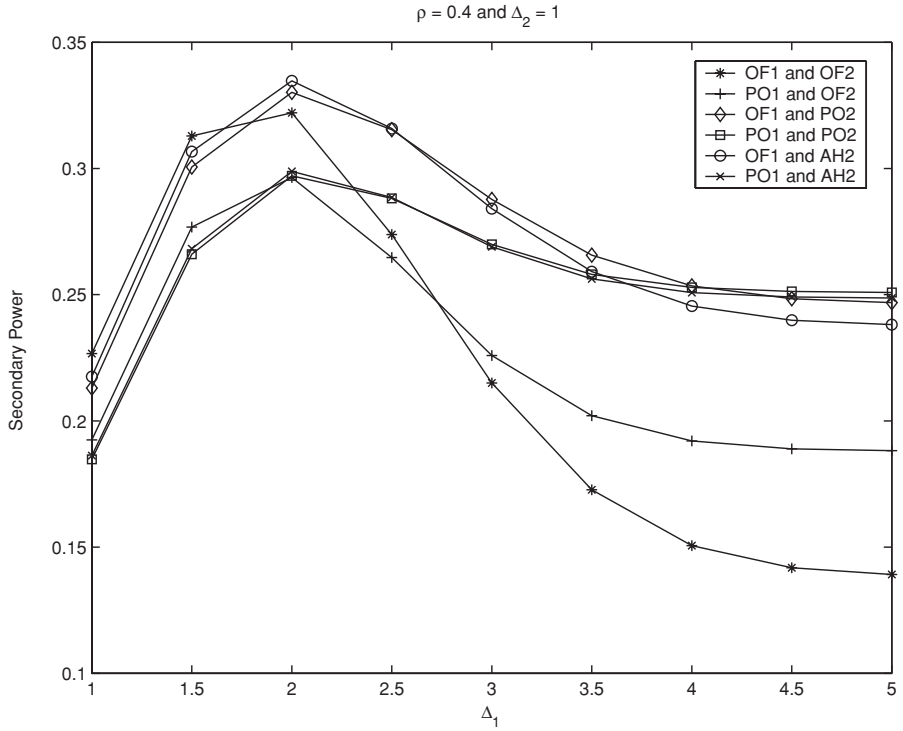

Figure 4. Secondary powers of the six boundary combinations (OF = O'Brien–Fleming, PO = Pocock, AH = ad hoc, 1 = primary, 2 = secondary) for the primary and secondary endpoints as functions of Δ_1 for $\rho = 0.4$ and $\Delta_2 = 1$ when $\alpha = 0.05$ and $n_1 = n_2$.

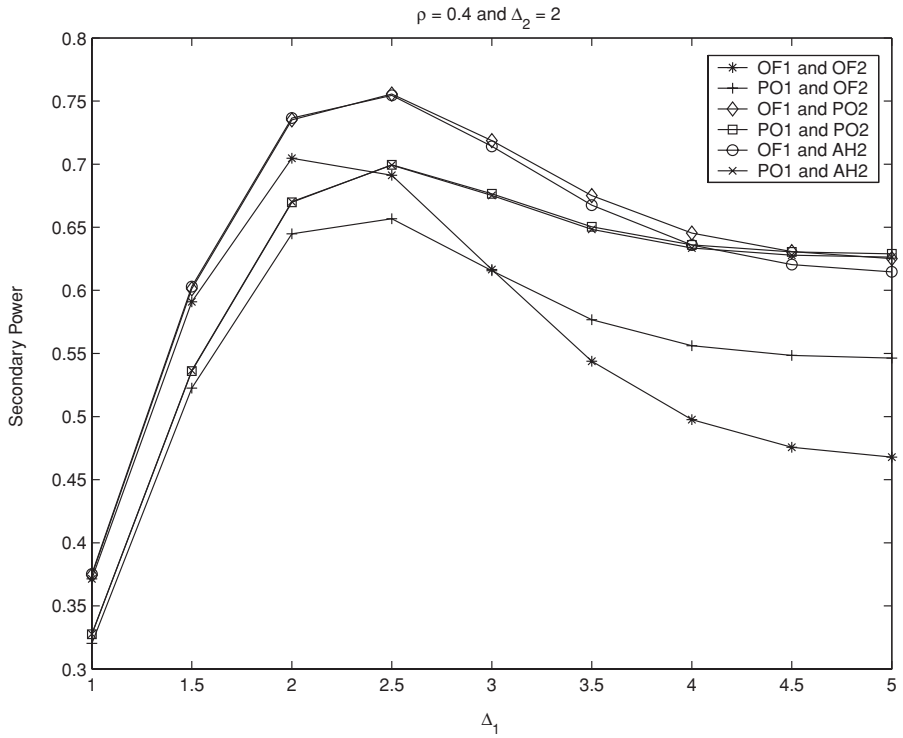

Figure 5. Secondary powers of the six boundary combinations (OF = O'Brien-Fleming, PO = Pocock, AH = ad hoc, 1 = primary, 2 = secondary) for the primary and secondary endpoints as functions of Δ_1 for $\rho = 0.4$ and $\Delta_2 = 2$ when $\alpha = 0.05$ and $n_1 = n_2$.

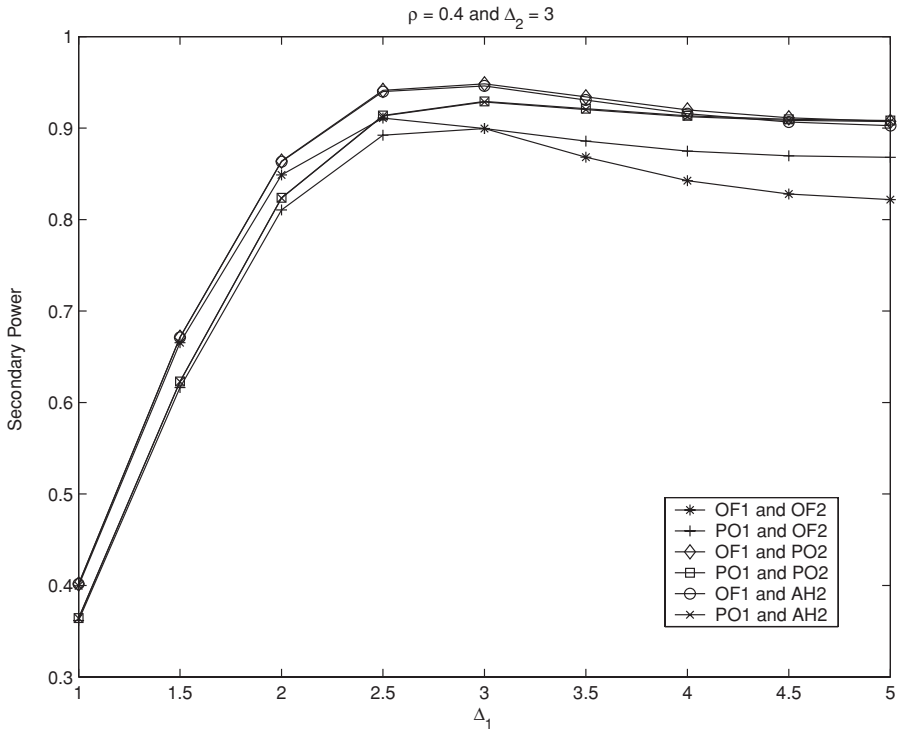

Figure 6. Secondary powers of the six boundary combinations (OF = O'Brien–Fleming, PO = Pocock, AH = ad hoc, 1 = primary, 2 = secondary) for the primary and secondary endpoints as functions of Δ_1 for $\rho = 0.4$ and $\Delta_2 = 3$ when $\alpha = 0.05$ and $n_1 = n_2$.

only one chance to be tested at stage 1 and so the secondary power decreases as Δ_1 increases, approaching $P(Y_1 > d_1)$ as $\Delta_1 \to \infty$. On the other hand, for small Δ_1, H_2 gets two chances to be tested; therefore, the secondary power increases as Δ_1 increases.

5. Example

To illustrate the methods discussed in this article, we consider the CAPTURE study (Simoons et al. 1997) in which a randomized trial was conducted to compare abciximab with placebo for coronary intervention in refractory unstable angina. It was planned to enroll 1400 patients to placebo and abciximab in equal proportions. The primary endpoint was a composite of death, myocardial infarction, or urgent intervention for treatment of recurrent ischemia, within 30 days of randomization. From previous studies in this patient population, the placebo event rate was estimated to be 15%. The investigators wished to provide good power to detect a 5% reduction in the event rate for the abciximab arm. With a sample size of 1400, a single-look design has 81% power to detect this treatment effect, using a one-sided test at the $\alpha = 0.025$ level of significance. The trial was, however, designed for a group sequential test with three equally spaced looks and OF-type stopping boundaries derived from the error spending function (5). Because of the possibility of early stopping, there is a power loss of about 1% with the group sequential design relative to the corresponding fixed sample design. For illustrative purposes we will ignore the possibility of early stopping at the first look and consider this to be a two-look design. This is a reasonable approximation because this error spending function yields an extremely conservative stopping boundary at the first look with a negligible chance of early stopping even under the alternative hypothesis of a 5% drop in the event rate.

The main secondary endpoint for this trial was death or myocardial infarction within 30 days of randomization. Although in the actual trial there was no formal testing strategy to protect the FWER of the two endpoints, we will consider two such strategies here. Under Strategy 1, we prespecify in the protocol that the secondary endpoint will be tested only if the null hypothesis for the primary endpoint is rejected; moreover, it will be tested at the same look and with the same spending function as the primary endpoint. This may be considered a conservative strategy because the secondary endpoint will face the same conservative OF-type stopping boundary as the primary endpoint. Suppose, however, that abciximab is expected to reduce the risk for the secondary endpoint to a lesser degree than that for the primary endpoint. It might then be advantageous to consider a more aggressive Strategy 2 in which the type I error for the secondary endpoint is spent at a faster rate, resulting in a less demanding early stopping boundary at the first look. For example, one might prespecify the use of PO-type stopping boundary derived from the error spending function (6) for the secondary endpoint.

Let π_{1e} and π_{1c} denote the true event rates for the experimental (abciximab) and control (placebo) arms, respectively,

Table 2
Interim results and test statistics

Endpoint	Event rates		Interim statistic
	Placebo	Abciximab	
Primary	84/532 (15.8%)	55/518 (10.6%)	$X_1 = 2.485$
Secondary	44/532 (8.3%)	26/518 (5.0%)	$Y_1 = 2.123$

with respect to the primary endpoint. Let π_{2e} and π_{2c} be the corresponding true event rates for the secondary endpoint. Define $\mu_i = \pi_{ic} - \pi_{ie}, i = 1, 2$. Let n_1 and n_2 be the sample sizes per arm at the two stages and $N = 2(n_1 + n_2)$ be the planned maximum total sample size. The Wald statistics for the primary and secondary endpoints at the interim look are given by

$$X_1 = \frac{\left(\hat{\pi}_{1c}^{(1)} - \hat{\pi}_{1e}^{(1)}\right)}{\hat{\sigma}_1^{(1)}}\sqrt{\frac{n_1}{2}} \quad \text{and} \quad Y_1 = \frac{\left(\hat{\pi}_{2c}^{(1)} - \hat{\pi}_{2e}^{(1)}\right)}{\hat{\sigma}_2^{(1)}}\sqrt{\frac{n_1}{2}},$$

where $(\hat{\pi}_{ic}^{(1)}, \hat{\pi}_{ie}^{(1)})$ denotes the maximum likelihood estimate of (π_{ic}, π_{ie}) at the interim look and

$$\hat{\sigma}_i^{(1)} = \sqrt{\hat{\pi}_{ic}^{(1)}\left(1 - \hat{\pi}_{ic}^{(1)}\right) + \hat{\pi}_{ie}^{(1)}\left(1 - \hat{\pi}_{ie}^{(1)}\right)} \; (i = 1, 2).$$

The Wald statistics for the primary and secondary endpoint at the final look are given by

$$X_2 = \frac{\left(\hat{\pi}_{1c}^{(2)} - \hat{\pi}_{1e}^{(2)}\right)}{\hat{\sigma}_1^{(2)}}\sqrt{\frac{n_1 + n_2}{2}} \quad \text{and}$$

$$Y_2 = \frac{\left(\hat{\pi}_{2c}^{(2)} - \hat{\pi}_{2e}^{(2)}\right)}{\hat{\sigma}_2^{(2)}}\sqrt{\frac{n_1 + n_2}{2}},$$

where $(\hat{\pi}_{ic}^{(2)}, \hat{\pi}_{ie}^{(2)})$ denotes the maximum likelihood estimate of (π_{ic}, π_{ie}) at the final look and

$$\hat{\sigma}_i^{(2)} = \sqrt{\hat{\pi}_{ic}^{(2)}\left(1 - \hat{\pi}_{ic}^{(2)}\right) + \hat{\pi}_{ie}^{(2)}\left(1 - \hat{\pi}_{ie}^{(2)}\right)} \; (i = 1, 2).$$

Applying the large sample results of Jennison and Turnbull (2000, Chapter 3), the above sequentially computed Wald statistics have independent increments, so that (X_1, X_2, Y_1, Y_2) have the same distributional properties as the corresponding random variables defined in Section 2. We may thus avail of the results presented in Section 3 for controlling the FWER. No data are available to estimate the correlation coefficient ρ, so we will assume the least favorable configuration $\rho = 1$.

In the CAPTURE study, an interim analysis was performed after data were available on 1050 subjects out of a planned 1400 subjects. Thus, the relevant information fractions are $t_1 = 1050/1400 = 0.75$ and $t_2 = 1400/1400 = 1$. We compute the stopping boundary (c_1, c_2) from the spending function (5) by solving

$$P_{H_1}(X_1 > c_1) = \alpha_1(t_1) \quad \text{and} \quad P_{H_1}(X_1 > c_1) + P_{H_1}(X_1 \leqslant c_1, X_2 > c_2) = \alpha.$$

For Strategy 1 we set $(c_1, c_2) = (d_1, d_2)$. For $\alpha = 0.025$, we obtain $(c_1, c_2) = (2.340, 2.012)$. (The classical OF boundary would be $(c_1, c_2) = (2.327, 2.015)$.)

For Strategy 2, we compute the stopping boundary (d_1, d_2) from the spending function (6) by solving

$$P_{H_2}(Y_1 \geqslant d_1) = \alpha_2(t_1) \quad \text{and} \quad P_{H_2}(Y_1 > d_1) + P_{H_2}(Y_1 \leqslant d_1, Y_2 > d_2) = \alpha.$$

For $\alpha = 0.025$, we obtain $(d_1, d_2) = (2.040, 2.258)$ with the same (c_1, c_2) as earlier. (The classical PO boundary would be $(d_1, d_2) = (2.126, 2.126)$.)

The following event rates (presented by K. Anderson at the 2002 Conference on Clinical Trial Data Monitoring Committees, Barnett International) were obtained for the primary endpoint at the time of the interim analysis: 84/532 (15.8%) for the placebo arm and 55/518 (10.6%) for the abciximab arm. Although no data were available for the secondary endpoint, we extrapolated from the published results of Simoons et al. (1997) to give 44/532 (8.3%) for the placebo arm and 26/518 (5%) for the abciximab arm. The test statistics are displayed in Table 2 and the stopping boundaries are displayed in Table 3.

The trial was terminated at the interim look because the test statistic X_1 for the primary endpoint crossed its early stopping boundary c_1. To formally claim statistical significance in the product label for the secondary endpoint, the test statistic Y_1 would have to cross its early stopping boundary d_1. It is seen that the secondary statistic does not cross its early stopping boundary under Strategy 1, but it does so under Strategy 2. Thus, depending on what was prespecified in the protocol, the investigators might or might not be allowed to claim statistical significance for the secondary endpoint.

6. Extensions

The scope of this work is somewhat limited. First, we have considered only one-sided tests on two endpoints with two stages. Second, the data on the primary and secondary endpoints are assumed to be bivariate normal with known standard deviations and a nonnegative correlation coefficient that is either assumed to be known or treated as a nuisance parameter with the FWER being maximized with respect to it to determine the critical boundaries for the endpoints. These limitations are in part dictated by the analytical and

Table 3
Critical boundaries for two strategies

Look	Strategy 1		Strategy 2	
	Primary: $\alpha_{\text{OF}}(t)$	Secondary: $\alpha_{\text{OF}}(t)$	Primary: $\alpha_{\text{OF}}(t)$	Secondary: $\alpha_{\text{PO}}(t)$
Interim	$c_1 = 2.340$	$d_1 = 2.340$	$c_1 = 2.340$	$d_1 = 2.040$
Final	$c_2 = 2.012$	$d_2 = 2.012$	$c_2 = 2.012$	$d_2 = 2.258$

computational difficulties involved. Given the practical importance of this problem, we expect that this work will serve as a springboard for further extensions, some of which are outlined later.

6.1 *Unknown Correlation*

The plots of FWER in Figures 4-6 and the critical constants tabulated in Table 1 show that $\max_{\Delta_1}$ FWER is an increasing function of ρ with $\rho = 1$ being the least favorable case. If an upper bound on ρ can be specified then it can be used to find a more powerful boundary for the secondary endpoint than the conservative boundary corresponding to $\rho = 1$. A practical approach to this problem would be to use the upper limit of a one-sided confidence interval on ρ estimated from the data after stage 1 to find the stage 2 boundary with a suitable adjustment to the required FWER to take into account the error probability of the true ρ exceeding the upper confidence limit. The ideas from Berger and Boos (1994) can be used in this context as follows.

Let ρ^* be a $(1-\varepsilon)$-level upper confidence limit on ρ (e.g., calculated using Fisher's large sample arctan hyperbolic transformation), i.e., $P(\rho \leqslant \rho^*) = 1-\varepsilon$. Then, using the property that $\max_{\Delta_1}$ FWER(Δ_1, ρ) is an increasing function of ρ, the overall maximum FWER can be written as follows. Let $\Delta_1^*(\rho)$ be the value of Δ_1 that maximizes FWER(Δ_1, ρ) for fixed ρ. Then

$$\begin{aligned}
&\max_{\{\Delta_1, \rho\}} \text{FWER}(\Delta_1, \rho) \\
&\quad = \max_{\{\rho \leqslant \rho^*\}} \text{FWER}(\Delta_1^*(\rho), \rho) \times P(\rho \leqslant \rho^*) \\
&\qquad + \max_{\{\rho > \rho^*\}} \text{FWER}(\Delta_1^*(\rho), \rho) \times P(\rho > \rho^*) \\
&\quad = \text{FWER}(\Delta_1^*(\rho^*), \rho^*) \times (1-\varepsilon) + \text{FWER}(\Delta_1^*(1), 1) \times \varepsilon.
\end{aligned}$$

We want to determine the smallest possible (d_1, d_2) (to maximize the power) subject to the earlier expression being $\leqslant \alpha$. This problem can be solved iteratively on a computer. For illustration, suppose we choose the PO boundary for the secondary endpoint, i.e., $d_1 = d_2 = d$. For fixed d we can evaluate FWER$(\Delta_1^*(\rho^*), \rho^*) = \alpha'(d)$ (say) and FWER$(\Delta_1^*(1), 1) = \alpha''(d)$ (say) where note that $\alpha'(d) < \alpha''(d)$. Then by numerical search we can find the smallest d such that $\alpha'(d)(1-\varepsilon) + \alpha''(d)\varepsilon = \alpha$. In this case also one may want to restrict $d_1, d_2 \geqslant z_\alpha$ for practical reasons.

6.2 *Multiple Stages*

In this article, we have restricted to a GSP with two stages. However, most practical applications involve more than two stages. Restricting still to two endpoints but $K \geqslant 2$ stages, denote the cumulative standardized test statistics for the primary and secondary endpoints at the kth stage by X_k and Y_k, respectively. The FWER control problem is to determine the critical boundaries $(c_1, c_2, \ldots, c_K)$ and $(d_1, d_2, \ldots, d_K)$ such that the FWER is strongly controlled at α for the decision rule that continues sampling as long as $X_k \leqslant c_k$ for $k \leqslant K$. If $X_k > c_k$ then H_1 is rejected and H_2 is tested. H_2 is rejected if $Y_k > d_k$ and accepted if $Y_k \leqslant d_k$; sampling terminates in either case or if $k = K$. If $(c_1, c_2, \ldots, c_K)$ and $(d_1, d_2, \ldots, d_K)$ are α-level boundaries then from Tang and Geller (1999) it follows that the FWER will be controlled at level α. However, a more detailed analysis can be carried out to sharpen the secondary boundary as a function of ρ as was done for $K = 2$ stages. It would be useful to develop a software that would compute these boundaries for any choice of the primary boundary, e.g., from the class of boundaries based on error spending functions of Lan and DeMets (1983) and Wang and Tsiatis (1987).

6.3 *Multiple Endpoints*

Multiple endpoints are of less practical interest than multiple stages, but it may be of interest to study three ordered endpoints (primary, secondary, and tertiary). Denote the test statistics for the three endpoints at the kth stage by (X_k, Y_k, Z_k) and let $(c_1, c_2, \ldots, c_K), (d_1, d_2, \ldots, d_K)$, and $(e_1, e_2, \ldots, e_K)$ be the critical boundaries for the three endpoints. Let H_1, H_2, and H_3 be the null hypotheses for the three endpoints. One may consider the following decision rule: Continue sampling as long as $X_k \leqslant c_k$ for $k \leqslant K$. If $X_k > c_k$ then reject H_1 and test H_2. If $Y_k > d_k$ then reject H_2; otherwise terminate sampling without testing H_3. If H_2 is rejected then test H_3; reject H_3 if $Z_k > e_k$; otherwise accept H_3. Terminate sampling in either case or if $k = K$. It is easy to show that if all three boundaries are α-level then the FWER will be controlled. A challenging problem is how the knowledge (or estimates) of the correlation coefficients between the three endpoints can be used to sharpen the boundaries for the secondary and tertiary endpoints. This problem is complicated by the fact that there are three correlation coefficients.

Finally, we mention another extension for $K = 2$ in which one may continue sampling to the second stage and test H_2 if H_1 is rejected but H_2 is not rejected at the first stage, effectively offering a second chance for the secondary endpoint to be tested.

7. Supplementary Materials

The Web Appendix referenced in Section 1, which gives the proofs of all propositions, is available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors are grateful to a co-editor and two referees whose comments helped to improve the article. Prof. Ajit Tamhane's research was supported by National Heart, Lung and Blood Institute Award 1 RO1 HLO 82725-01A1.

References

Berger, R. L. and Boos, D. D. (1994). P-value maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89,** 1012–1016.

Dmitrienko, A. and Tamhane, A. C. (2007). Gatekeeping procedures with clinical trial applications. *Journal of Pharmaceutical Statistics* **6,** 171–180.

Dmitrienko, A. and Tamhane, A. C. (2010). Gatekeeping procedures in clinical trials. In *Multiple Testing Problems in Pharmaceutical Statistics*, A. Dmitrienko, A. C. Tamhane, and F. Bretz (eds), Chapter 5. Boca Raton, Florida: Taylor & Francis.

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

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.

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

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

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

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.

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

O'Neill, R. T. (1997). Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials* **18,** 550–556.

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

Simoons, M. L. and the CAPTURE investigators. (1997). Randomized placebo-controlled trial for abciximab before and during coronary intervention in refractory unstable angina; the CAPTURE study. *Lancet* **349,** 1429–1435.

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

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43,** 193–200.

Received April 2009. *Revised December* 2009.
Accepted December 2009.